Performance Management and Analytics

Business Intelligence and the Digital Enterprise

Version 12.0 | August 2015

Sponsors:

arcplan

DATAWATCH
GET THE WHOLE STORY™

OPENTEXT™

pmOne

UNISERV

Viscovery

Author:

WOLFGANG MARTIN TEAM
powerful connections

www.wolfgang-martin-team.net

Copyright

Disclaimer

The use of names, trade names, trademarks etc. within this document without any special marking does not imply that such names are free according to trade mark laws and can be used arbitrarily by anybody. The reference to any specific commercial product, process or service through trade names, trademarks, manufacturer names etc. does not imply an endorsement of S.A.R.L. Martin.

Preface

The present White Paper "Performance Management and Analytics" is the twelfth edition on the evolution and progress of Business Intelligence. In January 2004, I started version 1.0 with my co-author Richard Nußdorfer. After the sudden and unexpected death of Richard in October 2008, I continued our joint work. The current version 12.0 describes business and technical architectures of operational, tactical, and strategic performance management and analytics under the aspects of the progressing digitalization of enterprises and markets.

Performance Management is defined as a business model enabling an organization to align continuously business goals and processes and keep them consistent. It works as a closed-loop model for managing the performance of business processes on all three levels with support for planning, monitoring, and controlling. From a business point of view, this is one logical model, but from a technological point of view, rather different technologies with completely different roots are clashing together: Traditional Business Intelligence meets Business Process Management (BPM). The convergence happens via the model of a service-oriented architecture (SOA).

In version 6.0, I ended up with the term "Performance Management". This was after some time of confusion in the market when different vendors used different terms for their product offering for performance management. We have seen terms like "Business Performance Management – BPM", "Corporate Performance Management – CPM", "Enterprise Performance Management – EPM", and others. I now use Performance Management as an umbrella term for BPM, CPM, EPM, etc. Consequently, in version 6.1, I replaced the former term "enterprise information management" by "information management".

Analytics *is "the science of analysis". A practical definition, however, would be that analytics is the process of obtaining an optimal or realistic decision based on existing data.*[1] The objective of analytics is to derive knowledge from internal and external data and information for controlling an enterprise. Digitalization makes analytics more and more important, since digitalization provokes the explosion of data volumes. It is a real data deluge. In 2013, the volume of the World Wide Web is supposed to be about 4 zettabytes (10^{21} B = 1 billion TB). This volume is expected to be doubled within 2 years. Welcome to **Big Data**. The most important drivers of this data explosion are social media, sensors, server logs, web clickstreams, the mobile internet and the internet of things.

Big data provides a huge potential. It is the source of deep knowledge. You only have to unlock it. But it is by far not as easy and straightforward as you may think. There is a huge variety of sources providing a mix of an enormous amount of data, of fragmented data, and of unmanageable data that makes it difficult to identify relevant data, and to extract, to store, to administrate and to analyze it. Indeed, this requires new approaches and new technologies for analytics.

Goal of this white paper on "Performance Management and Analytics". Organizations developing performance management and analytic solutions will have to decide which platform to choose for performance management and analytics and which additional best-of-breed-products will be required. They also have to decide how to deploy performance management and analytics,

[1] see http://en.wikipedia.org/wiki/Analytics

via on premise or by cloud computing. The focus of this White Paper is to assist any decisions in the described environment. It is aimed to help and to guide management to transform their business into a digital business through establishing a solid foundation by analytics and performance management: A digital enterprise is an information-driven enterprise!

The author. Since 1984, I have collected lots of experience in Information Technology in management functions, as an analyst and as a strategic consultant. From 1973 till 1984, I was a scientist. Given these backgrounds, I combine theory and practice. As an analyst, I have been dealing since 1996 with strategic deliberations and future developments in information technology and its impact to the business. I was with Meta Group till 2001, and I am working as an independent analyst since then.

The presented white paper on "Performance Management and Analytics" is divided into two parts. In this general part 1, benefits, concepts and facilities of performance management and analytics as well as its reference architecture are discussed. In part 2, platforms and solutions of selected vendors for performance management and analytics are presented. The following white papers are currently available[2]:

arcplan, BOARD, Clueda, Cortex, Cubeware, epoq, geoXtend, IBM, Informatica, Lixto, Kapow Software, Metasonic, Panoratio, PitneyBowes MapInfo, SAP, Stibo Systems, TIBCO/Spotfire, Tonbeller

Version 11.1 of this white paper was first published in March 2015. This version 12.0 from August 2015 is a revised and extended version. I added chapter 3.2 on "Operational Intelligence", and chapter 8 "Ethical Aspects of Analytics" has been rewritten. Furthermore, I have updated the themes "Planning in Digital Enterprises" (chapter 4.5), "Trends in Data Mining" (chapter 5.6), "Business Activity Monitoring und Complex Event Processing" (chapter 7.1), "Hadoop and Spark – Technical Answers to Big Data Challenges" (chapter 7.4), and "Hadoop, Spark, Data Lake and the Data Warehouse" (chapter 7.5). As always, chapters 10.3 to 10.5 (vendor list and classification), and 12 (glossary and abbreviations) have been updated.

The author will be delighted to receive reader feedback, commentary, criticism - and compliments, of course!

Annecy, August 2015

Dr. Wolfgang Martin
Wolfgang Martin Team

[2] Free download at http://www.wolfgang-martin-team.net in the sections „White Paper" and „Research Notes"

The author's biography:

Dr. Wolfgang Martin

Biography

Designated one of the top 10 most influential IT consultants in Europe (by Info Economist magazine in 2001), **Wolfgang Martin** is a leading authority on Customer Relationship Management (CRM), Business Process Management (BPM), Information Management, Information Governance, Business Intelligence (BI), Performance Management, Analytics, Big Data, and Cloud Computing (SaaS and PaaS). He is a founding partner of iBonD, Contributor of the ODBMS.ORG, Research Advisor at the Institute for Business Intelligence at the Steinbeis University, Berlin, and member of the Boulder BI Brain Trust.

After 5½ years with META Group, latterly as Senior Vice President International *Application Delivery Strategies,* Mr. Martin established the **Wolfgang Martin Team.** Here he continues to focus on technological innovations that drive business, examining their impact on organization, enterprise culture, business architecture and business processes.

Mr. Martin is a notable commentator on conference platforms and in TV appearances across Europe. His analytic skills are sought by many of Europe's leading companies in consulting engagements. A frequent contributor of articles for IT journals and trade papers, he is also an editor of technical literature, such as the Strategic Bulletins on BI, Big Data, BPM, CRM and SOA, as well as the text books "Data Warehousing – Data Mining – OLAP" (Bonn, 1998), "Jahresgutachten CRM" (Würzburg, 2002, 2003, 2004, 2005 & 2007), and "CRM Trendbook 2009" (Würzburg, 2009).

Dr. Martin has a doctoral degree in Applied Mathematics from the University of Bonn (Germany).

For more information see www.wolfgang-martin-team.net

Contents

1 Management Summary

“In the Age of **Analytics**, as products and services become ‘lighter’ (i.e., less physical and more digital), manufacturing and distribution costs—while still important—will be augmented with new metrics—the costs of know, the flow of know, and the costs of not knowing.” Thornton May[3], Futurist, Executive Director, and Dean of the IT Leadership Academy

Performance Management and Analytics are principles, methods, and tools for driving, controlling, and governing a business. They get in particular indispensable, when revenues start growing slower and are even becoming flat or are shrinking, when budgets become tighter and tighter, when market dynamics are increasing, and when competitiveness is skyrocketing. Then corporate management’s challenges increase significantly. Identifying potentials for profit, rigorously cutting cost, intensifying customer contacts as well as precisely calculating where to optimally spend the remaining resources are key issues not only for top management. Geopolitical uncertainties make planning much more difficult, but more important than ever. More and more regulations, for example in financial reporting and consolidation, do not simplify things at all! What really matters today is the ability to change strategies and tactics on the fly, the power to innovate, and to grasp upcoming opportunities immediately. This is the only way to survive in the **digital era**.

The digital world

The digitalization of the world is pushed by innovations in information technology: Cloud computing, social media, mobile, big data, and the internet of things revolutionize the world as fundamentally as the invention of the steam engine in the end of the 18th century, when the industrial revolution brought fundamental changes and a new order for society, industry, political systems, and even governments. At the center of both of these upheavals was, and is, is the human being – the person whose life and working world and its requirements undergo an equally fundamental change.

The digital era is also the era after the economic crisis of 2008. Some call it the “**new normal**”. It is different to the old normal we used to live in before the crisis. Nothing is and will be as before. “Strategy as we knew it, is dead”, said Walt Shill already 2010. At that time, he was heading the North American management consulting practice of Accenture[4]. "Corporate clients decided that increased flexibility and accelerated decision making are much more important than simply predicting the future." When strategies and forecasts have to be updated weekly or even daily, then we have to perform that way. The digitalization of corporations accelerates speed and dynamics even more. The mobile internet makes information ubiquitous. Social Media speed up tremendously the reach and speed of information. The Internet of Things is approaching: machine-to-machine and robot-to-robot communication start to produce huge amounts of new data – mostly in real-time. Welcome in the digital world!

[3] Thornton May: „The New Know“, Innovation Powered by Analytics, 2009

[4] see Wall Street Journal (Jan., 25th, 2010) http://online.wsj.com/article/SB10001424052748703822404575019283591121478.html#printMode

In other words, traditional corporate management hits the wall. The old principles of **"operational excellence"** are no longer sufficient. Operational excellence created **industrialization** of a business. Industrialization means automation and standardization. It speeds up and optimizes business processes, increases throughput as well as it improves quality. This is still needed, but now in the digital world, we need more. Managing a business now requires **"agility"** as a third imperative beyond traditional efficiency and effectiveness. Agility means the power to continuously change and adapt its business models and processes to a steadily changing market and customer dynamics. Acting fast and appropriately in times of change and uncertainty matters! Acting properly is getting a new name: "**smart conduct**", i.e. intelligently conducting. Life cycles of business strategies and processes get shorter and shorter. As a consequence, speed of changes must increase faster and faster. Corporate management today and tomorrow must be agile and smart not only for surviving, but also for prospering.

Indeed, corporate management must now go beyond the traditional instruments of planning, execution, monitoring and controlling. The new challenge is **governance, risk, and compliance (GRC)**. ***Governance*** means controlling of all activities and resources in the organization directed on respectful and durable value creation based on longevity. ***Risk*** management encompasses all activities of risk identification, minimization, and avoidance. ***Compliance*** means to act in accordance to all management policies as well as to all legal and regulatory requirements. Given the actual economic and financial crisis, more and more severe regulations are even to be expected. Management must follow these guide lines.

For today's organizations in the digital era, agility, industrialization, and compliance together with smartness are key differentiators making up winners or losers in the world's global and digitized markets.

An additional major advantage of industrialization, agility, compliance, and smartness is the capability to implement innovative and creative ideas on the spot. New products and services as well as new pricing models can be brought to market in shortest time. Standard processes like dunning processes can be made customer oriented for regaining revenues already believed lost. Cross and up selling is powered and new revenue potentials can be tapped by exploiting all information about customer profile and behavior. Just in time identification of risk and problems eliminates defective goods and returns. In other words: competiveness increases, competitors are stunned and customers are enlightened. Digital enterprises exhibit these capabilities and have the potential for **innovation.**

The digital enterprise

What differentiates a digital company from a traditional company? A digital company is formed by digital transformation taking place within the company itself, and it is the role of information technology that changes fundamentally. Whereas in traditional organizations, information technology has the function of auxiliary support processes, it now triggers innovation of business models as well as of business processes and services in digital companies. New digital business models, processes, and services mean new sources of revenues and disruptive competitive advantages. The product portfolio is complemented by digital products, and information is used as a strategic advantage. In other words: the organizations reinvents itself.

The second important property of a digital company is that it is competent in digital communication: Ever increasing quantities of new media and channels are continually integrated into all levels of company communication, and with all business partners. Only then can the digital customer be followed and the traces he leaves behind in the digital world be found and exploited. Big data can be filtered: Smart customer data is the result.

Besides the cultural change needed, the roadmap to a digital enterprise that is industrialized, agile, compliant, and smart is based on a comprehensive approach to a service oriented **business process management (BPM)**. Process and service orientation is definitively the right approach, but process and service orientation must go hand in hand with **information management**. This fundamental principle of "**no process without data**" is often neglected, even overlooked.

Indeed, the digital world is driven by data. In the old world, the customer paid for products and professional services using country-based currencies. But following digitalization, a new currency has emerged: Data! If you (i.e., the customer) give me your data, I (Facebook, Google, etc.) will give you a "free" service. Thus, data is increasingly becoming the currency of the virtual internet world. Digitalization has created the breakthrough: The real world has merged with the virtual world - and data has become a world currency.

Data in the virtual world can be enriched, e.g., with spatial coordinates from relevant sensor data. Localization and navigation data from smartphones and other mobile devices enable a customer to be localized and offered locally available services. It is upon this basis that the largest data collectors (Apple, Facebook, Google, etc.) have risen to become some of the highest rated companies in the world. But this is hardly surprising, because a company that uses data on an analytic basis, also has the power to make targeted campaigns in the customer's experience world. Smart customer data (customer related and appropriately filtered data created from big data) transforms a customer into the (much) longed for "transparent" customer.

Conversely, the power of data is available for the customer too. In the digital world, prices, product details and professional services are transparent. This means that the digital market is transparent too – far more transparent than the traditional market. The customer uses the digital market when hunting for the best offers. A flight from Berlin to Heathrow can be cheaper than a taxi journey from Heathrow to central London. The struggle for power will continue in 2015 and the years thereafter. The customer will pay increasingly with data, and the companies will become increasingly hungry for that data. Companies will provide the customer with even better experiences - and thereby induce the customer to provide even more data.

Information Management

In the past years, businesses have considered information management only in the restricted context of **business intelligence (BI).** Extensive BI projects and initiatives were performed to provide availability and accessibility of information to support corporate management by facts. The result is: Information is available in excess. But due to the scope of traditional BI, the provided information is not put into the context of business processes. In consequence, information is often of rather limited use and value for management. Now, we have to take a new approach for dramatically improving the value of BI and for achieving a break through. **Information must be put into the context of business processes and Services.** This is the basic idea of performance management, and consequently, we need an information management beyond the restricted view of traditional BI.

In a digital enterprise, business process management, performance management and information management belong together.

Business Process Management is all about managing the life cycle of business processes and business services: from analysis and design via implementation and operations to planning, monitoring, controlling, and enriching by analytics.

Performance Management is defined as a business model enabling a business to continuously align business goals and processes and keeping them consistent. The task of performance management within BPM is planning, monitoring, and controlling of processes and services and their performance.

Analytics denotes the process of gaining information and of deriving a model for utilization of information (e.g. a predictive model as a result of a data/text mining process) as well as enriching business processes and services by analytic models. The idea behind is to create "smart" processes or services.

The goal of ***Information Management*** is to create ***trusted data*** in the sense of the “single point of truth”. Information management includes data definition (the enterprise terminology), data modeling (the enterprise semantic), meta and master data management (transparency and traceability), data quality management (relevance and accuracy), and data security and protection.

Performance Management

is an important step towards optimal planning, monitoring and controlling business processes and services and their performance. It acts on all levels: operations, tactics, and strategies. Performance management is based on metrics associated with processes. Performance management starts when designing and engineering processes: metrics (“performance figures”) have to be derived simultaneously and in parallel with the operational process design. Goals have to be metricized. Achievement of goals has to be continuously monitored. Actions must be taken for continuously and real-time controlling the performance of processes. Performance management can do this, because it is a closed-loop model, and because it is part of the overall Business Process Management closed-loop model.

Under the current economic conditions, performance management is the next strategic step towards agile corporate management to make your business fit for mastering the challenges of transformation into a digital enterprise. The motto is “**You only can manage what you can measure”.**

Service-Orientation

Processes and services make up the competitiveness of a corporation. They decide on winning and losing in the global market depending on their power to innovate as well as on their quality, flexibility and compliance. In other words: The new focus of management is on processes and services. Service orientation of business processes provides the required flexibility. Processes are composed by business services.

For BPM, Analytics, Performance and Information Management, an appropriate IT support through the right infrastructure is essential. A service-oriented architecture (**SOA**) is required as an infrastructure for closed-loop management of business processes. BPM, Performance and Information Management on a SOA enable and empower automated, standardized, reliable, audit-proof, and flexible processes across business functions, departments and even across enterprises. This cuts cost, enlightens customers and employees, and boosts revenues. SOA based processes are independent of the underlying IT systems and applications. Hence, business can change processes with the speed of market dynamics and customer needs. You keep sailing close to the wind. This is why SOA means “software for change”.

A SOA is also the prerequisite for using cloud computing to provision performance management and analytics. Service-orientation enables the deployment of complementary functionality via SaaS or the deployment of a PaaS for analytics and performance management via a hybrid cloud model.

Analytics

Furthermore, a SOA makes it happen: **Analytics** can be embedded into processes. Processes are enriched by "intelligence" and get “smart”.

Analytics play a key role for planning, monitoring and controlling processes and their performance. The challenge is to identify problems in right time for taking preventive actions in right time. The results are early warning systems, a foundation for risk management.

> An ***example*** from day-to-day life explains how predictive models act as early warning and alert systems: In a department store the sales areas are stocked up at the right time, before products are out of stock. This avoids the situation when a customer wants to buy a product, and he finds himself standing in front of empty shelves.

In the digital era with its **Big Data**, analytics becomes more and more important. Big Data does not only mean the huge ocean of data, but a mix of structured and poly-structured data linked via complex relationships. Within an organization, there is already a large amount of structured and poly-structured[5] data (Typically, 20% of enterprise data are structured and 80% poly-structured). But the real big volume of data is in the web. The challenge of Big Data is not only the data deluge, but also the diversity of data sources: portals, web applications, social media, blogs, videos, photos and much more. It is all kind of web content.

Furthermore, additional data sources like localization and navigation data from the mobile internet enable new ways of location related interactions. Sensor data empowers the monitoring and controlling of machines and industrial processes. They generate data streams requiring analysis in real-time. In the end, these are the basic principles of driverless cars like the Google car. Finally, the analysis of machine data allows an extensive automation of controlling and optimization of the underlying processes.

The analysis of such new data sources in combination with enterprise data provides completely new insights you never had before: Analytics beat intuition. But the analysis of such data requires ethics about the usage of data. Indeed, customers give away their data

[5] The term “poly-structured” replaces the old term “unstructured”. This makes sense, because “unstructured“ data may contain some hidden patterns that provide certain structure information in the end.

voluntarily, for example their social profiles to Facebook or their search profiles to Google. But customers want a return for their data. From Google, they get information they hardly could get elsewhere. But all the Facebooks and Googles of the world as well as governments that eagerly collect data about their citizens must provide guarantees that the collected data is only used for the benefits of customers and citizens. This should be governed by ethics when dealing with data and data analyses. This is why topics like data privacy and security, transparency, commensurability and added value are more than important and decisive. Organizations and governments that apply and live these ethical principles will have satisfied and faithful customers and citizens.

Big Data drives technology innovation. This includes new technologies for real-time analytics, a new category of integration tools for agile web and cloud integration and extraction as well as innovative database technologies for analyzing the petabytes or even exabytes of information. The up-to-now dominant relational databases are no more in a position to analyze this volume of data in reasonable time. New innovative database technologies gain interest and traction: **analytic databases** and **NoSQL data management systems[6].** They combine innovative algorithms for data access and storage management with innovative methods like column orientation and innovative technologies like in memory processing.

Traditional BI tools are no more really appropriate and sufficient for Big Data analytics. **Data Discovery,** an interactive approach to analytics, has now gained importance and wide acceptance. It comes with filtering and visualizing of data, collaborative tools for team work, intuitive user interfaces, and a new generation of devices like tablets. This empowers the business, and improves productivity. **Location Intelligence**, the expansion of Business Intelligence by the dimension of „space“, is also regaining interest, because in the mobile internet, information, time and space are converging. Localization data from smartphones and navigation devices enable completely new types of analyses and insights. Furthermore, there are new analytic methods and approaches for analyzing poly-structured data like **text analytics**. Text analytics combines linguistic methods with search engines, text and data mining as well as statistical learning. That makes up a completely new arsenal for analytics.

All these new methods and technologies require new roles, for instance a **Data Scientist**. They act as a mediator between IT and the business driving the continuous improvement of collaboration, supporting management of big data, and enabling the boost of big data potentials. Indeed, this new role requires new skills and a repositioning of IT: In the age of big data, IT must put its main focus on data management.

Performance Management and Analytics versus traditional business intelligence

The difference of performance management and analytics to traditional BI is not only that we must now deal with Big Data, but also that we can now put decisions into the context of processes and services. Performance management and analytics can act operational and in real time. Traditional BI tools (reporting, ad-hoc querying, OLAP – online analytic processing, data mining etc.) failed to deliver the right information to the right location in the right time for the right purpose. Traditional business intelligence tools did not meet management expectations: results to be applied to processes and strategy for turning information into

[6] NoSQL = not only SQL, SQL = sequential query Language

value. Return on investment (ROI) in the old tools was typically rather low, if measurable at all.

Traditional business intelligence tools were difficult to master. Information remained a privilege in many enterprises. Only a handful of experts (the power users or business analysts) were in a position to exploit information via the traditional tools. Therefore, many management decisions and actions were based on guesses, much less on facts. But in the meantime, many of these problems have been resolved. "**Self-service BI**" is a principle of data discovery. It empowers the occasional user of performance management and analytics. They get access to facts and information they need in the context of BI governance. Social Media style concepts à la Facebook and Twitter have been carried over to BI tools' interfaces. Mobile solutions on smartphones and tablets have shown the way to go for more software usability.

Usability does not only deal with the ease of access and use, but also with the automation of work flows and analytic processes and services. Traditional BI was mainly based on manual tasks for analysis and provisioning of information. It was error-prone. Now, automation removes these obstacles and provides reliable and secure analytics. Enriching processes by analytics through a SOA is another way to improve usability: It enables a **Mashing Up**[7] of analytic services with operational and collaborative services. The SOA advantage is less integration efforts and very high flexibility. The result is business process and service innovation and a new quality far beyond traditional workflow systems. This is smart!

Take away: Performance management and analytics are essential foundations for driving, controlling and governing a business. It supports the implementation of the four management principles of a digital enterprise: industrialization, agility, compliance, and smartness. It is the coherent and consequent evolution of traditional business intelligence. Analytics unlocks the potential of Big Data. The result is valuable knowledge for corporate management. Via process and service orientation, performance management applies this knowledge to all actions and interactions from strategy to operations. The result is an **intelligent enterprise**. That makes up the benefits and value of performance management and analytics.

Provisioning performance management and analytics can be done via a traditional on premise model or via cloud computing. Both models are equally valid, and can be combined to a hybrid cloud model.

[7] **Mash Up** means the creation of new content by seamless (re-)combination of existing content.

2 From Business Intelligence to Performance Management and Analytics

Performance management and analytics can be understood as the successors of Business Intelligence. Therefore, we first discuss the roots and evolution of Business Intelligence, the goals of Business Intelligence, and what have been the problems to be solved and the challenges to be mastered.

The oldest known source mentioning the term Business Intelligence (BI) originates from the year 1958. In the October edition of the IBM Journal, Hans Peter Luhn wrote about „A Business Intelligence System“. In 1989, Howard Dresner, at that time analyst with Gartner Group, seized the term BI. Indeed, it was Gartner Group that shaped and promoted BI in the 90s. In parallel to the concepts of BI, the idea of a Data Warehouse (DW) was also born in the US in the 90s. Furthermore, the first technology change also happened then. Mainframe architectures for BI solutions were replaced by client server architectures.

In the years 2005/06, more and more people question the term of "BI", and new terms appeared in the markets: "Business Analytics" and "Performance Management". These new terms addressed the different tasks of traditional BI quite well. In parallel to these discussions, the next technology change happened: Client server architectures are succeeded by service oriented architectures.

About 2010, another new term appeared: "Big Data". It became rapidly wide-spread. Big Data not only provides new opportunities and use cases for analytics, but also introduces new architectural structures and technologies: Cloud and mobile computing. Today, this evolution leads to the digitalization of the world and to the digital enterprise. Analytics is now ubiquitous!

Despite of all this evolution and progress, there is still a different thinking about what BI is. Even within one and the same organization, there are different opinions about BI and what the advantages and benefits of BI are. So, let us start with a definition of BI that is rather widely accepted by the market:

Definition: **Business Intelligence** means the capacity to know and to understand as well as the readiness of comprehension in order to exercise this knowledge and understanding for mastering and improving the business. In somewhat more detail, we define: Business Intelligence is a model consisting of all strategies, processes, and technologies that create information out of data and derive knowledge out of information so that business decisions can be put on facts that launch activities for controlling business strategies and processes.

The idea of BI principles and concepts is all about to put decisions on facts and to make "better decisions". BI should give answers to questions like

- Do you know which of your suppliers is mission critical to your production? Will their failure bring down your production for hours or even days?

- Do you know what percentage of supplier revenue is due to your spending? Do you get good terms and conditions from suppliers, using this information?
- Do you know who your most profitable customers are? Are you providing superior services in order to retain them and are you able to service them, up-sell/cross-sell at appropriate points when interacting with them?
- Do you know in Q1 that you will miss your sales target in Q4, because your actual volume of leads is insufficient?
- Do you know what revenue you are actually loosing because customers cannot connect to your call center due to peak demand?
- Do you know how much business you miss by not fully exploiting cross-sell opportunities in face-to-face encounters, outlets, and web shops?
- Do you know how much money this means for your enterprise? Do you know how to find it, get it and keep it?

We practice IT-supported BI since 70s and 80s. Did we get all the answers we have been expecting to get from BI? Not always, because the problem has been to use and to apply BI concepts and principles in the daily operational business as well as on the strategic level for getting answers that matter. Especially about the year 2000, we got frustrated about BI triggering a new approach to BI: Performance management and analytics were born and should give life to a new BI.

2.1 Pitfalls of Traditional Business Intelligence

Up to now, business intelligence enabled decision support in the context of strategic planning and tactical analysis. The goal of traditional BI was to base decisions on facts. Unfortunately, in many cases this did not deliver the expected added value and enterprise wide acceptance. Reports, indicators, analytic applications and others: where is the real value? Indeed, BI tools more than often failed to deliver. It was always difficult to measure value achieved by business intelligence and the data warehouse. Reason is: information per se does not create any value. Value is created, when information is applied, used and turned into decisions and actions.

What was wrong with traditional Business Intelligence?

- Business Intelligence was bottom up and not process-oriented. Lines of business people were not sufficiently involved. Genuine, process-oriented business requirements had not been addressed at all. Traditional BI lacked business-oriented relevance.
- Business Intelligence was just an information access model for decision support (i.e. Bill Inmon’s “Information Factory“; Inmon, 1996). This means, information and the analytic processes for information exploitation were mashed together. The results are inflexibility and unnecessary complexity. Any innovation gets discarded from the very beginning**.** Acceptance decreases drastically.

- Business Intelligence did support decision making to a certain degree, but the feedback component for closing the loop was missing. Taking actions based on decisions was not part of the model. Indicators that are not in the context of a process bring only limited value. The real value of information is only achieved when information is deployed in the context of processes.

  Example: As soon as an indicator on the strategic level is in red, the owner of that indicator has to make decisions for launching tactical and operational actions. Information is deployed, decisions are based on facts, and a much higher value is achieved than with the traditional BI model where feed-back is not part of the model.

- Operational aspects of Business Intelligence were left out. Traditional BI was based on a data warehouse as the single point of truth. This architecture excluded BI from use in operational environments. BI was isolated and limited to tactical and strategic analysis. The potential of real-time analysis was completely neglected.

- Business Intelligence was retrospective. Focus was on analysis and diagnostics only. The potential of predictive models for identifying of problems and risks in the right time was ignored.

  Example: A midrange manufacturing plant analyses quality of production at the end of each shift. This guarantees that problems in production are identified as soon as possible so that actions can be immediately taken to ensure. This ensures that the identified problems will not occur in the subsequent shift. Pro-active BI creates significant value to be exploited.

- Business Intelligence tools did not supply the information consumer and manager sufficiently. Either information was not accessible (or even hidden and retained), or there was an information deluge. Again, this lowered acceptance of BI dramatically.

- Business Intelligence was a tools-centric approach based on proprietary technologies. Each analytic component played its own role in an isolated environment. Incompatibility and inconsistency were the consequences, and stove-piped information silos were the results. On board level, numbers did not match any more.

Taking into account theses pitfalls, it was time to **reinvent Business Intelligence.** The old and original idea of fact based decisions is not bad at all. It has been the view point of advocates of BI long since. But now, a new and even more important driver for evolution and reinvention of BI is the fundamental change that is happening in the global markets: We are now living in the digital world, that produces **Big Data** and lives in Big Data: Information has become one of the most important assets: Today’s businesses are data-driven.

2.2 The digitalization of the world

The economic and financial crisis in 2008 caused fundamental changes of markets and market mechanisms with heavy and worldwide impact to corporations: nothing will be as before. Complexity and dynamics of business has tremendously increased. Flexibility in operational business is more important than ever. The speed to grasp looming opportunities, to act and to react smartly[8], and to turn them into profit is decisive. Drivers are manifold. There is the new distribution of power between the old, ripe, and saturated markets on the one hand and the new, merging, and expanding markets on the other hand. There is the continuously increasing interconnectedness and globalization. Masterminds like Peter Hinssen create the term "the new normal" to describe this situation.

> "A number of new rules will apply in the New Normal. Consumers will have zero tolerance for digital failure. They will expect to get internet access anytime, anyplace. Internet and connectivity will be just as ubiquitous as electricity. Consumers will demand fulfillment of their information needs instantaneously. The effect on companies will be tremendous."[9]

In this citation, Peter Hinssen also talks about another important driver: The digitalization of the world. Information technology, communication, economy, and social areas converge and fuse. The mobile internet makes information ubiquitous: We are always "on". The internet transports information with the speed of light, and Social Media spread it to all corners of the world. The mobile internet fuses information, time, and space. That makes up the digital world. Consequently, corporations have to adapt and to transform into digital corporations.

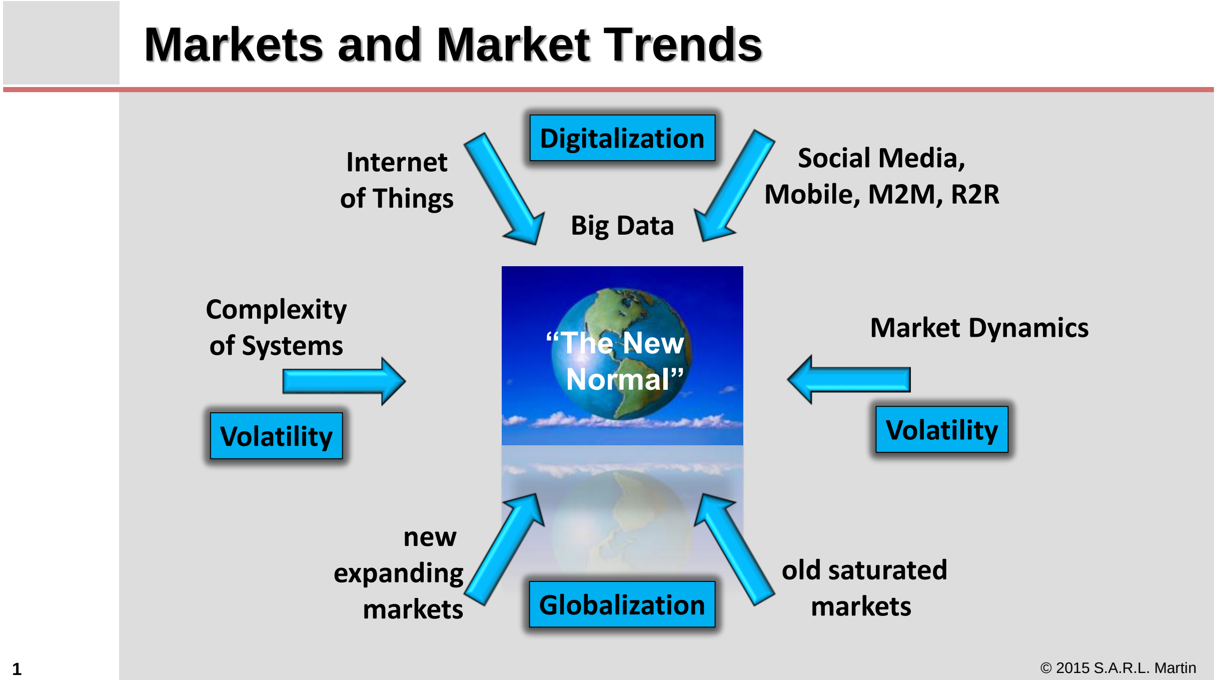
Mastering **volatility** and speed of the markets are challenges in the New Normal. There is nearly no more stability. Unpredictability and increasing change rates stress traditional management methods. Agility and smartness become key success factors. In today's markets, lack of agility and smartness simply means to lose business and market positions. Management now has to scope these new conditions: Traditional corporate management does not work anymore. Information becomes the decisive resource for mastering complexity and dynamics of the digital world. (Fig. 1)

Up to now, we have seen an even sometimes massive application of BI methods and tools for decision support. But traditional corporate management was mainly based on experience acquired in the past. Decisions were intuitively carried over from the past to actual and present situations and rather successfully applied to the future. This does not work anymore in the New Normal, since experience is only valuable and provides "correct" decisions, if the business model is the same in the past, presence, and future. In the New Normal, this is no more a given. Due to increasing complexity and dynamics, experience acquired in the past should not blindly be applied to the presence and – not at all – to the future. Experience based decisions can very well lead to wrong decisions in the New Normal. **Wrong decisions** are fatal in today's market dynamics, since a revision of wrong decisions is no more possible: There is no time anymore! Decisions taken too late are another source for wrong decisions. Early and real-time identification of problems and risks is necessary for taking smart counteractions in right time. This holds especially in the hyper-dynamics of the New Normal.

To summarize, **right decisions** are based on **information in real-time** and on a comprehension of the dynamics of the business model that maps the business to the market conditions.

[8] Please see glossary for a definition of the term "smart".

[9] „The New Normal", Peter Hinssen (2010) http://www.peterhinssen.com/books/the-new-normal/synopsis



Markets and Market Trends

Digitalization

Internet of Things

Big Data

Social Media, Mobile, M2M, R2R

Complexity of Systems

"The New Normal"

Market Dynamics

Volatility

Volatility

new expanding markets

Globalization

old saturated markets

1

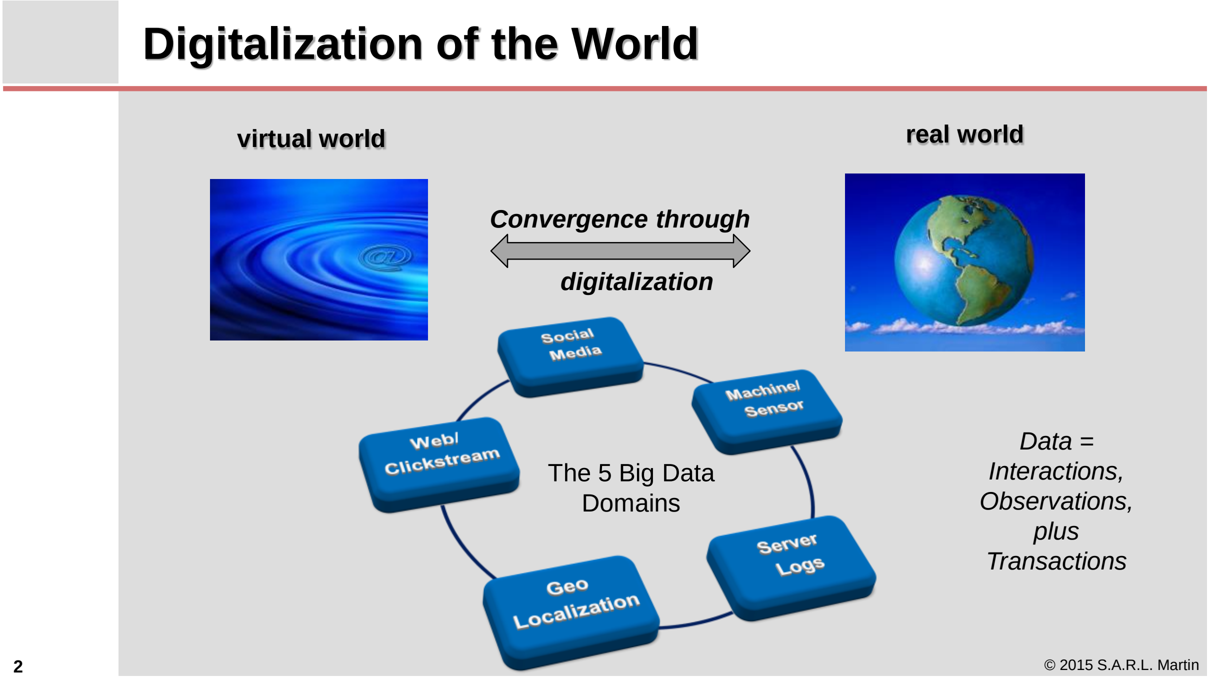
© 2015 S.A.R.L. Martin

Figure 1: Three forces drive the „New Normal", globalization of markets, digitalization of the world, and volatility of markets. The bottom line: A profound change happened, and the rules of the game have completely changed. Globalization creates tensions and distortions, volatility drives shakiness and uncertainness, and digitalization has brought an unbelievable acceleration of all processes. It sources Big Data, and alters the world disruptively. (M2M = machine to machine; R2R robot-to-robot)

In the pre-digital world, the business model used to be stable over long periods: In stable markets, and in particular in growing markets, an organization can be continuously and steadily managed with one and the same strategy and organization structure. In these circumstances experience was very useful and valuable: It was excellent for navigation in a calm sea. But the value of information for corporate management was rather low. Information had a certain importance, but was not always decisive. We have seen this in the deployment of reports that nobody used. We have seen this in the usage of spread sheets where numbers were smoothed and "massaged" until information and experience were in accord. We have seen this in the lack of acceptance of dashboards and analytic tools. We have seen this in the pitfalls of BI discussed in the previous chapter. In the "Old Normal", we could do it "without" information. Business was running without big needs for corporate governance when once correctly set up.

The New Normal changes everything. Due to the market dynamics, decisions must be made quickly and beyond experience. Now, it is essential to have the right information in right time with the right relevance for the decisions to be taken. But this is not as easy as it sounds. The challenge is to filter the right information out of the huge volume of irrelevant and hence superfluous and distracting information.

This filtering gets more and more difficult, because the digitalization of the world produces another challenge: It generates more and more data. Welcome to **"Big Data".** The essential drivers of this data deluge are social media, machines and sensors, server logs, web clickstreams, the mobile internet, and the internet of things. (Fig. 2)



Digitalization of the World

virtual world

real world

Convergence through

digitalization

The 5 Big Data Domains

Data =
Interactions,
Observations,
plus
Transactions

2

© 2015 S.A.R.L. Martin

Figure 2: Digitalization of the world triggers a convergence of real and virtual world. It creates five big data domains providing structured, poly-structured data and data streams. They present new data domains not available before. So now, data presents interactions, observations, or transactions. Transaction data was the first data that was managed by information technology. Interaction data came up with the usage of the internet since the early 90s, whereas observation data is rather new: It is mainly the result of the digitalization of the world. To summarize, more data and more detailed data is generated by digitizing the world, and in particular, the combination of various big data domains, and its integration with enterprise data creates new insights, the most important drivers of innovation. This makes up the value of big data.

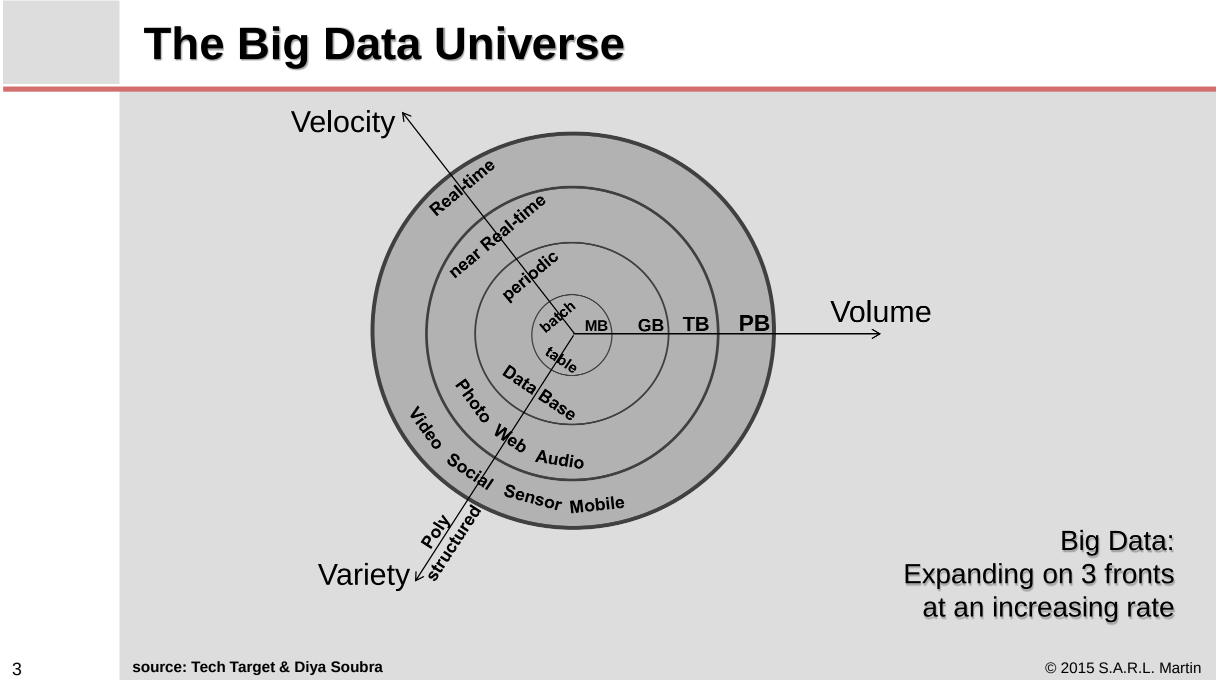
2.3 Big Data and Real Time

The „Big Data Challenge“: More and more users want to analyze data from the more and more increasing data flood and from more and more divers data sources in almost real-time.

This Big Data challenge is visualized by figure 3. It describes quite well what Big Data is and what it means:

- <u>Extreme and continuing growth of data volume.</u> In 2012, the world-wide production of new data already approached 3 ZB. (1 zettabyte = 1 billion terabytes). IDC estimates that the yearly data production will increase to 40 ZB in 2020. The majority of data will be poly-structured.
- <u>The number of data sources likewise increases massively.</u> This is not only due to social media, but also to machine generated data like localization and navigation data from the mobile Internet or measurement data from intelligent meters (telephony, electricity, gas, water, RFID etc.).

- More and more businesses and within a business more and more specialist departments want to benefit from the information and knowledge that is hidden in this extreme volume of data. Consequently, the number of members of staff that need information is rapidly growing.
- Information attains the maximal value, when it is new and actual. In the digital world, things are happening now, simultaneously and everywhere. This is why we need information in real-time, here and now.



The Big Data Universe

Big Data:
Expanding on 3 fronts
at an increasing rate

source: Tech Target & Diya Soubra

© 2015 S.A.R.L. Martin

Figure 3: The 3 “V” dimensions of big data depict the technological evolution over time. This is the foundation of the 4th “V” = value. In today’s infonomics, data drives value.

Each of these four trends per se is a big challenge for information technology. But now, these four challenges are to be mastered all together. If not, we will not get the required answers and insights. Unfortunately, traditional business intelligence technologies are not at all equipped or suitable for analyzing big data due to lack of performance, scalability and functionality. The big data challenge puts new requirements to the „New Intelligence“, and these new requirements are to be added to the requirements from the list of pitfalls of traditional BI. Consequently, we need more innovation in BI. Indeed, innovation in BI is right now happening on all levels of BI technology and tools. In the following chapters, we will describe these innovations that are ongoing in database systems, in information management, and in methods and tools.

To summarize, in the New Normal, corporate management needs **real-time information**, i.e. the right information in real-time for taking the right decisions in time. This is the prerequisite for monitoring and controlling business and business processes proactively.

> ***Example:*** Let us assume *“term of delivery”* is a goal of the shipment processes. First of all we have to make this goal measurable. As a metric, we could define that 90% of

all shipments should be within two days. This is a strategic metric. An operational business metric could be a predefined *"threshold for stock"* in a dealer warehouse. If stock falls below the threshold, an order is automatically executed. The outcome of this metric on stock level launches an action. It is a pro-active metric measured in real-time to avoid the risk of sold out. This real-time information on stock level allows identifying the risk in time and fixing the problem before any damage is caused.

Based on this example, we can now define "real-time":

Definition: Real-time in business means the availability of right information in right time at the right location for the right purpose.

So, "real-time" is relative, and the concepts and principles of "real-time" in business are not necessarily related to clock-time, but to the speed of business and the corresponding business processes and services. Monthly, weekly, or daily provision of information could be "real-time", if the underlying processes and services are running with the corresponding speed (Take timetable information when booking a travel versus information on delays when travelling.). This shows that "right time" is a more appropriated expression for information provisioning than "real-time".

The example on "*term of delivery*" also shows that metrics for monitoring and controlling processes and services do not only support diagnostics as they did in traditional business intelligence, but also forecasting. Forecasting enables proactive controlling. Problems and risks can be identified in right time, and counteractions in right time can prevent damage and loss. This is corporate management in **real-time**. It saves time, resources and cost.

2.4 Big Data Benefits and Potentials

Big Data promises a big potential, especially deep and profound insight. But exploiting Big Data is not easy at all, because such a mix of interwoven, huge, non-transparent and fragmented data sets makes it difficult to identify, to extract, to store, and to analyze the relevant data. But before we tackle the question of how to exploit Big Data, let us first better understand what the benefits really are and how we can profit from Big Data and Big Data insights. We start with two examples.

Example: Retail and Big Data. Retail was one the first vertical markets that met Big Data already some time ago: Sales slip data. It includes valuable customer knowledge, for instance „product profitability per customer". Indeed, this is an important metrics for controlling personalized campaigns and real-time recommendations, thus a frequently used metrics of analytical CRM for outbound and inbound customer communication. But traditional BI tools did not do a good job when calculating customer/product profitability: Due to the huge amount of data to be analyzed, it took many hours, sometimes even more than a day to come up with the results. In the end, due to the long lasting analysis duration, this important metrics could not be used for personalized recommendations in real-time customer interactions. Now, Big Data technologies address this problem, and calculation time goes down to some minutes or even seconds.

Example: Sentiment analysis based on social media data. Makers of fast moving consumer goods are especially interested in getting insight into the opinions of all market actors about their own products and brands as well as about products and brands of their competitors. Social media now provide a new and rich source for such knowledge that offers the opportunity to target customer segments with chirurgical precision. But social media also hold risks: Data extracted from blogs, forums and Twitter is not representative in the statistical sense, and it may contain complaisance and even real lies. Thus, analysis of social media communication must be applied with some caution when interpreting the results. Analysis starts with the identification and extraction of relevant Big Data sources. The extracted data is then analyzed by mathematical and statistical methods. The result is information about relevant traces in the web, how frequently certain topics are discussed, and more important, it also provides the tonality of all contributions by sentiment analysis. Based on such a **social media monitoring,** we can build **social media interaction.** It enables the business, to react immediately when necessary, and even to intervene in the same social media, where a critical contribution popped up. In particular, customer service and campaigns for introducing new products to markets can benefit from social media interaction, because a communication with communities in the web can be established and sustained. For example, in various service call centers, agents are becoming social media agents maintaining a multi-media communication with customers via traditional and social media channels. First experiences have shown improved time-to-market and improved customer loyalty. Indeed, this is the step from outbound and inbound customer communication to unbound customer communication. It provides an improved time to market as well as improved customer loyalty in relation to a relatively low investment in social media analytics.

Looking more closely to these examples, we can identify five different types of benefits through Big Data.

1. ***Transparency through Big Data.*** Tourism, for instance, is another vertical market besides the market of fast moving consumer goods that is interested in sentiment analysis. A hotel chain is keen on getting electronic feedback of their customers and/or their evaluations of competitors. Furthermore, publically available satellite images enable a completely new type of monitoring competitors. They could give hints about plant facilities, identify expansions or indicate topological constraints preventing a competitor from expanding. If such data is accessible and exploitable, completely new insights can be gained. In combination with enterprise data, businesses can not only establish the always targeted 360° view of customers, but even a 360° view of the total market: competitors, customers of competitors, press, market multipliers etc., because Big Data reflects the complete market with all market actors.

   But for benefitting from transparency through Big Data, the traditional “silo”-thinking within a business must be stopped definitely. Collection of departmentally related data is not sufficient for building customer and market knowledge through analyzing Big Data. Unfortunately, in finance, many enterprises still keep separate data about financial markets, monetary transactions, and lending. This prevents the composition of coherent customer views as well as the comprehension of relations and interactions between financial markets.

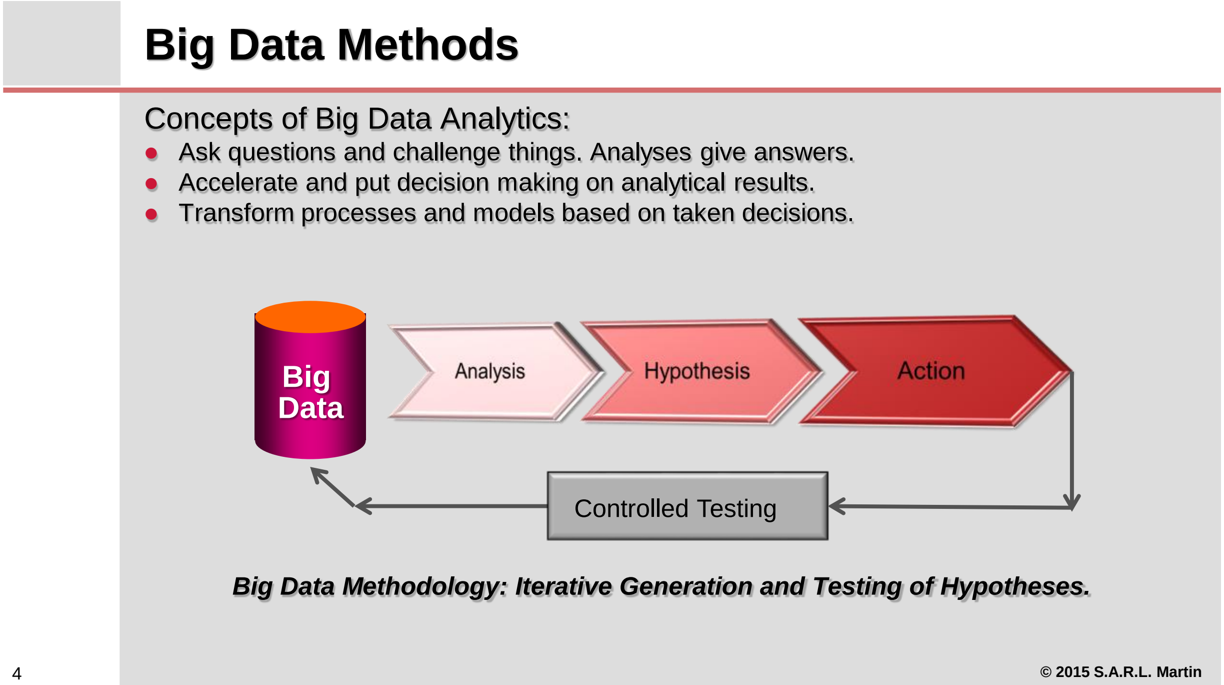
2. ***Generation of Hypotheses and testing of all decisions.*** In many projects, big data analytics attempts to avoid traditional statistical sampling and its impacts to analytics. It is assumed that N = all, i.e. the assumption is that the (very large) data set to be analyzed represents the total existing data and completely describes the phenomenon. Such a point of view has to be criticized. Let us just take a sentiment analysis of tweets as an example. Indeed, then at best, you analyze the sentiment of the Twitter channel, but not of the total market. The assumption that the analyzed tweets reflect the market's sentiment simply does not hold. Such an analysis may reflect the sentiment of Twitter users, but it is highly probable that the market's sentiment is different since the profile of Twitter users differs from the profile of the average market's consumers. People who tweet are just a special subgroup of consumers.

   There is another point that raises a lot of critique: Big data analytics delivers correlations that present not necessarily any cause and effect relations between correlated properties. Consequently, they do not necessarily represent relations in the real world. Hence, per se, they do not have any business value.

   In other words, big data analytics may deliver results to be criticized and to be interpreted with sufficient care. But if you consider such results just as a new hypothesis, then you can check its validity by traditional statistical methods of testing. In this sense, big data analytics is used to create new hypotheses out of data without human expertise. It complements human experts, and it widens traditional scientific approaches. This is particularly interesting when big data analytics creates hypotheses based on up-to-now not available data that are beyond human reasoning.

   So we can say, big data analytics changes the way of taking decisions fundamentally: Big Data analytics creates hypotheses purely based on data, then we test the hypotheses by controlled experiments, and finally, decisions and actions are based on tested facts (fig. 4). Big data analytics is then not in contradiction with traditional scientific methods, but it complements them. Furthermore, such procedures also permit to distinguish between cause and effect relations and mere correlations. Now, big analytics has a business value that can be exactly calculated.

   Big Data forerunners like Amazon and eBay were amongst the first that used such controlled experiments for improving conversion rates of web shop visitors. They changed structures, functionality and links between web pages in a specific way and measured the impact. In the end, they could identify factors influencing conversion rates. The mobile internet carries these concepts from the virtual world of the internet into the real world. For instance, for the first time measuring and optimizing the efficiency of outdoor advertising in relation to the place of location becomes possible by measuring click rates of QR codes placed on outdoor advertisements. This enables cross media marketing by analyzing Big Data. Another example of melting virtual and real world is videotaping of customer movements in combination with customer interactions and order patterns based on transaction data offers another approach. Through controlled experiments, product portfolios and placement as well as pricing can now be continuously improved. The result is cost savings by a perhaps possible reduction of the range of products without the risk of losing market share. Another result is an increase of margin by selling more premium products.



Big Data Methods

Concepts of Big Data Analytics:

- Ask questions and challenge things. Analyses give answers.
- Accelerate and put decision making on analytical results.
- Transform processes and models based on taken decisions.

Big Data Methodology: Iterative Generation and Testing of Hypotheses.

4

© 2015 S.A.R.L. Martin

Figure 4: Big data analytics comes with new methods complementing traditional scientific analytics: Big data analyses are used for generating hypotheses that can be controlled by traditional statistical testing. Since such hypotheses are created by using until now not available or not accessible data from various sources, they can provide new insights in observed phenomena and create hypotheses beyond reasoning and imagination of human experts.

3. ***Personalization in real-time.*** Customer and market segmentation has a long tradition. Big Data now offers new and additional ways of real-time personalization of customer interactions. In retail, strategies for personalization in real-time are well known from the Big Data forerunners like Amazon and eBay. Similar strategies are also known in social media, where friendships are proposed. But, many different vertical markets also profit from personalized customer interactions, for instance insurance. Insurance policies can be individually targeted to customer behavior and social and demographic attributes. Continuously adapted profiles of customer risks, changes in wealth, or localization provide the necessary data. Vehicles, for instance, can be equipped with special emitters so that they can be retrieved after theft.

4. ***Controlling and automating processes.*** Big Data expands the usage of analytics for controlling and automating processes. Sensor data of production plants, for instance, can be used for the auto-control of production processes. The result is cost saving by optimized material usage and eliminating manual interactions. Simultaneously, throughput is increased which influences the margin positively. Proactive maintenance provides another example. Engines can be continuously supervised by sensors so that irregularities can be immediately identified and removed in-time, before causing damages

or downtime. The advantage is not only avoiding risks, but also improving margins by continuous, uninterrupted functioning of engines.

Other examples come from the vertical market of consumer goods. Producers of beverage or ice-cream use the daily weather forecasts for fine-tuning their demand planning processes to the actual weather. Measurements of temperature, volume of rainfall and daily sunshine duration provide the necessary data to be analyzed. Such knowledge empowers process optimization by improving the forecasts by some percentages. Such a sometimes small improvement can cause clearly increased profits.

5. ***Innovative information driven business models.*** Big Data also drives new, innovative business models based on information. In the past, information about prices was mainly kept secret. Today, in the age of the internet and internet retail, prices are public. Hence, internet and other retailers can now monitor prices of competitors. So, they can react to price changes just in time. But this also empowers customers. They can also get price information and can use this advantage in achieving the best price for a desired product. Some vendors have grasped the need for price information, and made a business model out of this. They consolidate, aggregate and analyze information about prices and sell the results to consumers and others. This does not only work in retail, but also in health care, where such information-brokers are bringing transparency into the pricing policies of therapies.

But finally, some critique about Big Data should not be concealed, because more information does not necessarily mean better information. A big problem is the variety of data sources: It impedes the comparability of data, because unrelated sources generate data in dissimilar quality and consistency. For statisticians, there is even another problem: Information extracted from Big Data is typically not representative.

Despite all the critique about Big Data: The Big Data forerunners Amazon, eBay, Facebook and Google show, that Big Data potentials are real and can bring money and profit. Despite of all skepticism about the Big Data hype: IT vendors invest big money in Big Data and expect a lot from this fast growing market. Finally, we should not forget that all the mentioned data sources are springing. The digitized world produces an enormous amount of information, and mathematics, statistics, linguistics and artificial intelligence provide mighty tools and methods for detecting and identifying hypotheses that provide stupefying insights. As in traditional data mining, this makes up the appeal for finding „nuggets" in Big Data, but now even bigger and more precious.

2.5 Process-Orientation – a new context for Business Intelligence

Business Intelligence must be placed into the context of business processes and services for achieving business relevance.

What are the relevance and importance of business processes and services? What can be achieved by process-orientation? To get an answer, let us look back: In the 90s, it was common belief that enterprises could run exclusively on a single instance ERP application. Enterprises became application oriented. Ideally, all business-relevant data was meant to reside in a single database and all business functions were meant to have been supported by standard (ERP) functionality. Unfortunately this ideal world was never achieved. What lessons have been learned?

- ***"One size ERP application fits all" does not work.*** The majority of enterprises run several heterogeneous instances of ERP plus legacy and other systems. Enterprises have an average of 50 mission critical OLTP systems, large to very large enterprises even 100s of these systems.

- ***IT performance suffers.*** The huge number of point-to-point interfaces necessary to link applications drives up costs for implementing new applications. The budget for maintaining these interfaces killed IT innovation. IT became a legacy and a blocker of innovation.

- ***Process automation is minimal to non-existent.*** Data has to be manually re-entered from application to application. This makes process quality low and results in mistakes, failures and lost money.

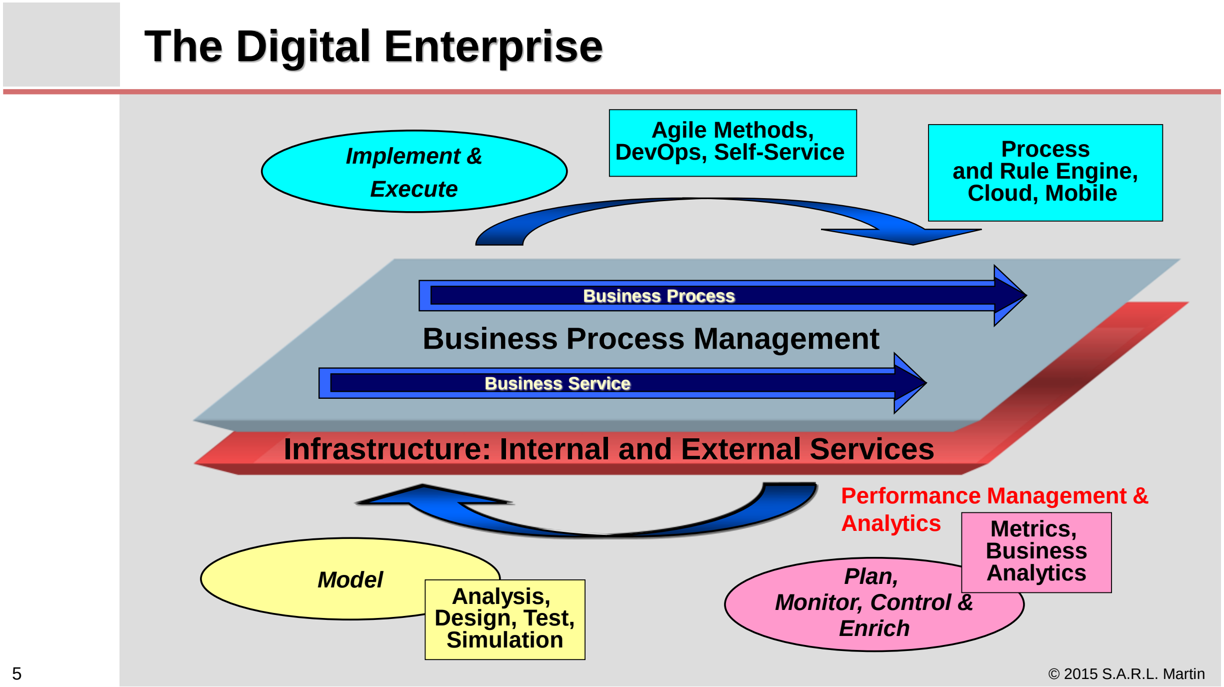
- ***Process integration is modest to non-existent.*** Processes end at the boundaries of applications making collaboration with suppliers, partners, and customers impossible. As a result, enterprises are sluggish and unable to react to changes in the market. Costs are driven sky-high.

- ***Changing your strategy and adapting your business processes to the speed and dynamics of the markets is impossible.*** Because business processes are hard-coded in the applications, if you need to change the business process, you need to change the application and every other application with which it interacts. In consequence, IT dictates the business, not strategy. This is not practical. Application-oriented enterprises are not agile and will ultimately lose to the competition.

- ***Master data is caught in applications.*** Each application has its own business vocabulary. Product or order numbers are defined completely differently from one application to the next. Collaboration across networks of suppliers requires master data translation. Each time you add a new supplier, customer, or product you must create new translation tables and/or add the new item to all the translation tables. This makes changes slow, error prone and costly.

- ***Information management is impossible.*** Timely access to business information across application islands becomes a luxury enterprises can't afford. The price of not having access to business information is even higher. Bypassing the problem by spreadsheet analysis à la Excel is not a good idea: It would just boost inconsistencies of data and information across the enterprise. Indeed, the risk is that you are neither audit-proof nor fail-safe.

How can the traditional enterprise be transformed from an application-centric focus to process-centric model? The answer is **Business Process Management (BPM).**



The Digital Enterprise

Implement & Execute

Agile Methods, DevOps, Self-Service

Process and Rule Engine, Cloud, Mobile

Business Process

Business Process Management

Business Service

Infrastructure: Internal and External Services

Performance Management & Analytics

Metrics, Business Analytics

Model

Analysis, Design, Test, Simulation

Plan, Monitor, Control & Enrich

5

© 2015 S.A.R.L. Martin

*Figure 5: **Business Process Management (BPM)** is a closed-loop model. Management of business processes and services together with innovation management becomes the center point of all entrepreneurial actions and activities. Processes are planned, modeled, implemented, executed, monitored, controlled, enriched and continuously improved. The infrastructure should consist of internal and external services**. A service-oriented architecture (SOA)** is to be recommended. **Performance Management** is a second closed-loop model for managing planning, monitoring and controlling of business processes and their performance within BPM. Analytics is used for deriving analytical models and services for enriching processes and making them "smart" ("intelligent"). Such a process-and service-orientation is the foundation for being industrialized, compliant, smart and agile, basic properties of digital enterprises.*

Definition: **BPM** is a closed-loop model consisting of four phases (fig. 5):

Phase 1: Analyzing, planning, modeling, testing, and simulating business processes. Sometimes, this phase is called "design".

Phase 2: Implementing business services by agile methods and DevOps concepts, and as far as possible through self-service of the special departments.

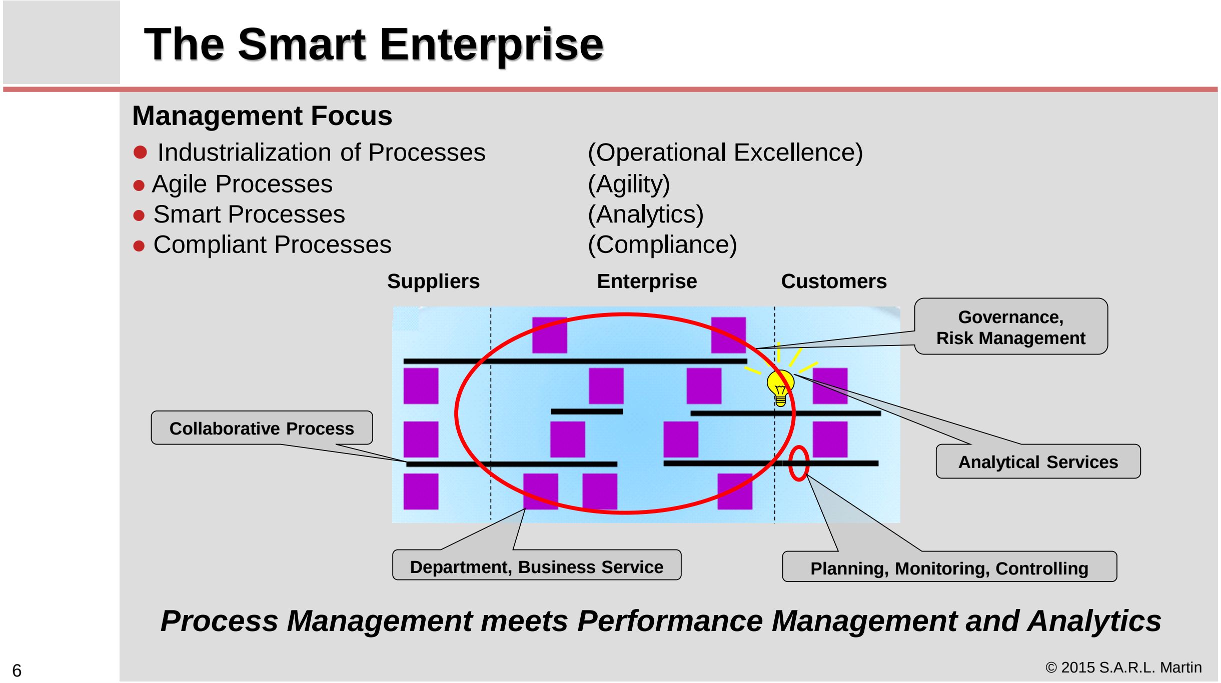
Phase 3: Mobile execution of business processes through cross-departmental work tasks ("process flows") via a process and a rule engine running on a SOA (service oriented architecture) infrastructure. The infrastructure is typically provided by a hybrid cloud.

Phase 4: Planning, monitoring, controlling, and enriching of processes and services, their performance, and the interaction of the ensemble of all business processes and services.

To summarize, BPM means closed-loop management of business processes. It enables synchronization of execution and exception management with continuous and comprehensive planning, monitoring, and controlling. This synchronization keeps business processes optimized in line with real time events and intelligent planning and forecasting.

Business processes are becoming the common communication platform between business and IT people. **For the first time we can create a genuine dialogue between business and IT**. The benefits of process-orientation are obvious (fig. 6):

- ***Processes become the common communication platform between business and IT.*** The specification of business requirements is now based on a common language jointly understood and spoken by the two parties, business and IT. Technical design of executable processes and back-end services providing application logic becomes straightforward when based on a common business design of processes.
- ***Processes become independent from applications.*** Collaboration makes enterprises shift to end-to-end processes across applications and platforms that are executed by rules-based process-engines running on an integration hub for application and data, the infrastructure for business process management and service management. An important point is that we are now dealing with cross-functional, cross-departmental, and even cross-enterprise processes that exploit the application logic of the existing application landscape.



The Smart Enterprise

Management Focus

● Industrialization of Processes (Operational Excellence)
● Agile Processes (Agility)
● Smart Processes (Analytics)
● Compliant Processes (Compliance)

Suppliers
Enterprise
Customers
Governance, Risk Management
Collaborative Process
Analytical Services
Department, Business Service
Planning, Monitoring, Controlling

Process Management meets Performance Management and Analytics

6

© 2015 S.A.R.L. Martin

Figure 6: A smart enterprise puts emphasis not only on „operational excellence" but also on the flexibility ("agility") of the business model, on compliance and on intelligent usage of information. Process and service orientation are the prerequisites. All together, these are properties of a digital enterprise that provides three domains for the "new" Business Intelligence: 1) Performance management for planning, monitoring and controlling of processes and services 2) Analytics for smart processes and services by embedding analytical services, 3) Governance and Risk Management. Here, the role of BI is to provide the right information in right time, for instance via early warning systems. Business Intelligence metamorphoses into "Performance Management and Analytics".

- ***Processes benefit from the advantages of service-orientation.*** A SOA is business-driven. The granularity of the process model determines the granularity of business services managed in a SOA. Furthermore, the SOA maps technical services from existing back-end applications to business services. This is 100% protection of investment in the existing IT architecture. With service-orientation we do the next step and build on top of the existing IT investments.

- ***Processes run across the underlying application data models.*** In order to automate event-driven processes across functions, departments, and enterprises, commonly-used application touch-points and data across the enterprise must not only be integrated and synchronized, but data models must be aggregated into a common information model to support collaboration processes. This common business vocabulary is the heart of master data management. Uniquely defined and centrally managed 'meta' data provides a common platform for all business terms and items across different applications and business constituents. This is essential when defining new products, gaining new customers, or adding suppliers to the business network. One simple update in the master database propagates changes safely and automatically to all related systems and services.
- ***Processes consume and publish services.*** The shift here is from application-oriented thinking to SOA-enabled processes (fig. 5). For a specific business process, operational, analytic, collaborative and information services are composed by a rules and a process engine. The result is that a business process either becomes a service or a group of services. Certain re-usability can be achieved by avoiding redundant implementation of functions and data. Redundancy was inherent in the old application-oriented model; service-orientation helps to overcome this problem.
- ***Processes drive the transformation to intelligent real-time enterprises.*** Business intelligence is gleaned from metrics associated with each business process. Business metrics are defined by goals and objectives to manage a process and its performance in a measurable and proactive way through information, metrics, key performance metrics (KPM)[10], rules, and predictive models. As an example, we have already discussed strategic and operational metrics for *"term of delivery*" in chapter2.3.

Real-time information empowers forecasting. Now, metrics can be anticipative, and we go beyond traditional diagnostics. Based on anticipative metrics and predictive models, processes get the power to act proactively: Problems and risks are identified in right time, and decisions and actions can be taken to prevent damages. In other words:

We have reinvented Business Intelligence. We have put Business Intelligence into the context of business processes and services. As a result, there are now three domains where BI can be applied to (cf. fig. 6):

1. **Performance Management.**

   Definition: In a process-oriented enterprise, performance management is the model enabling a business to continuously align business goals and processes and keeping them consistent. Performance management means planning, monitoring, and controlling of processes as well as to base analysis and forecasts on process content.

[10] Metrics are also called „Indicators“, respectively KPMs as KPIs. We prefer the terms metrics and KPM, because these terms better express the relationship to "measuring".

2. **Analytics.** Analytics refers to the process of extracting information and deriving a model for exploiting information (predictive models developed by a data mining process, for example) as well as to embedding the model into and applying it to a business process.
3. **Governance, Risk Management and Compliance (GRC).** Governance means an organization and a controlling of all activities and resources in the enterprise directed on respectful and durable value creation based on longevity[11]. Risk management encompasses all activities of risk identification, minimization, and avoidance. Compliance means to act in accordance to all management policies as well as to all legal and regulatory requirements. Here, the role of BI is to provide the right information in right time for decision making. Risk management gives the best example. Early warning systems for a proactive identification of risks for risk avoidance or risk minimization show successful adoption of BI.

2.6 Service Orientation – the new paradigm

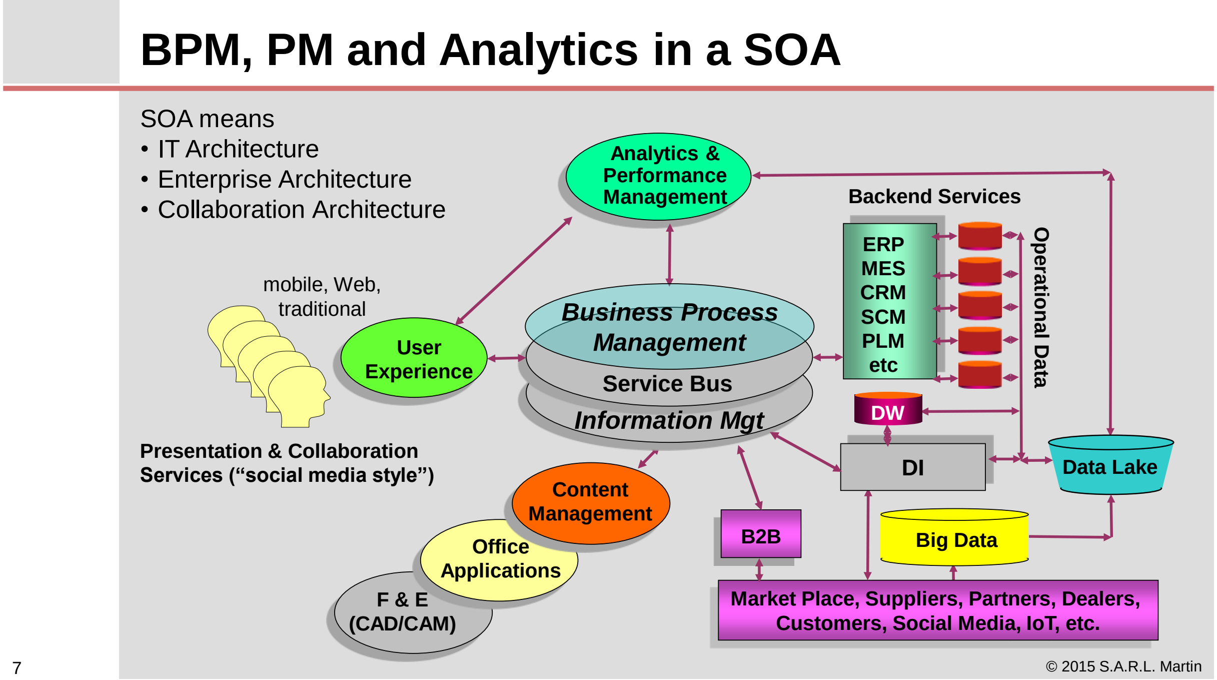
Finally, we have to define the infrastructure for BPM and performance management so that we can embed analytics into processes. As we have already seen (fig. 5), this is done through a SOA (fig. 7). From the IT point of view, agility **and** industrialization are two contradictory requirements, but if the infrastructure for managing processes is a SOA, this contradiction disappears. The reason is the nature of a SOA. It is a special architecture for providing **“Software for Change”** based on the principle of service orientation (SO). SO is a rather reasonable and most notably non-technical concept. It describes the collaboration between a consumer and a provider. The consumer is looking for a particular functionality (a “product” or a “service”) offered by the provider. Such collaboration works according to the following principles:

Service-Orientation

- Principle 1 – **Consistent Result Responsibility**. The service provider takes responsibility for the execution and result of the service. The service consumer takes responsibility for controlling service execution.
- Principle 2 – **Unambiguous Service Level**. The execution of each service is clearly agreed to in terms of time, costs and quality. Input and output of services are clearly defined and known to both parties by the Service Level Agreement (SLA).
- Principle 3 – **Proactive Event Sharing**. The service consumer is informed about every agreed change of status for his work order. The service provider is required to immediately inform the service consumer of any unforeseen events.
- Principle 4 – **Service Bundling.** For service provisioning, a service can invoke and consume one or more other services, and can be invoked and consumed by other services.

[11] Prof. Dr. Matthias Goeken, Frankfurt School of Finance & Management, on the occasion of the opening event of the “**Zukunftswerkstatt IT”**, Frankfurt/Main, April 19th, 2007

Such a service orientation provides a flexible framework: A service can be understood as a fulfillment as requested and matching with the conditions of an SLA. The SLA determines time, cost, and resources necessary for service provisioning. Furthermore, service input and output are well defined. Services can also be sourced externally from third parties via a "software as a service" (SaaS) model. Services provide business and decision logic that traditionally was included in applications. Processes now orchestrate and choreograph services – i.e. the business logic – according to the process logic.



BPM, PM and Analytics in a SOA

SOA means
- IT Architecture
- Enterprise Architecture
- Collaboration Architecture

Analytics & Performance Management
Backend Services
mobile, Web, traditional
User Experience
Business Process Management
Service Bus
Information Mgt
ERP
MES
CRM
SCM
PLM
etc
Operational Data
DW
DI
Data Lake
Presentation & Collaboration Services ("social media style")
Content Management
Office Applications
F & E (CAD/CAM)
B2B
Big Data
Market Place, Suppliers, Partners, Dealers, Customers, Social Media, IoT, etc.

7

© 2015 S.A.R.L. Martin

Figure 7: A SOA describes the design of the infrastructure for BPM and BI. The implementation is based on an service bus supporting management and life cycle management of processes and services including back-end services, information services (DI = data integration) and meta/master data services. It also provides the B2B interface. Other business domains like content and knowledge management, office, and research and development (F&E - CAD/CAM) can also be incorporated via the integration hub. Analytics and performance management act as the brains of the digital enterprise. They provide the "intelligence" for optimal monitoring and controlling all business processes and their performance. Analytics is embedded into the processes for anticipating problems and risks. The human interface defines and supports human interactions and experiences through collaboration and presentation services enriched by social media tools. Information management manages, monitors and controls data provisioning via data integration, the data warehouse, and the Data Lake. This will be discussed in more detail in chapters 6 and 7. (ERP – enterprise resource planning, MES – manufacturing execution system, CRM – customer relationship management, SCM – supply chain management, PLM – product life cycle management, DW – data warehouse, B2B – business to business, IoT – internet of things)

The sub service principle has an interesting drawback when compared with the corresponding sub process principle. A service acts like a process. In consequence, a process can be a service, and a service can be a process. In the sense of IT language, a service is then defined as follows.

A **Service** is a functionality typically triggered by a request-response mechanism via a standardized interface and consumed according to an SLA. In consequence, a service is a special instantiation of a software component.

Now, we are in a position to define “SOA”. It should be noted that the term SOA consists of two parts, the SO (service orientation) and the A (architecture). We have already defined SO and service, so we now have to define “architecture”. Unfortunately, this term does not have a commonly accepted definition, but with the support of several encyclopedias, we can state:

In IT, architecture specifies the interplay of components of a complex system. It describes the translation of business requirements into construction instructions. Hence, architecture has characteristics and consequences.

Finally, we can summarize:

- SOA is a design model for a special enterprise architecture and a special enterprise software architecture
- SO means loosely coupled. In an SOA, traditional application logic is decoupled into process and business logic. This is due to the roles of service consumer (process), and service provider (service).
- SOA is independent from technology. Technology for implementation can be independently picked.
- SOA is an evolution of component architectures (principles of “LEGO” programming)
- SOA services are business driven. The granularity of the process model determines the granularity of business services.

Standardization[12] is another property of an SOA. Service access is standardized (web services), as well as orchestration and composition of services (business process execution language – BPEL) and infrastructure services like authentication and identity management. Such collaboration across the IT industry is new and drives specifications and adoption of standards. Web services, for instance, are already generally accepted and frequently adopted.

As a prerequisite for SO, we need a **business vocabulary** so that all SOA based processes use the same notation and specifications. A repository is necessary for uniquely defining all meta and master data. The repository for meta and master data plays a similar role as the integration hub within a SOA. So, the architecture of the repository should be hub and spoke so that all meta and master data can be synchronized and versioned across all back-end systems and services. This is the role of **master data management (MDM)**. In chapter 6, we will discuss MDM in more detail.

For further reading about SOA (in German) we refer to Martin (2008).

Note: ROI is typically not achieved by a SOA per se, but by the implemented, SOA-based processes.

[12] See for instance http://www.cio.com/article/104007/How_to_Navigate_a_Sea_of_SOA_Standards

2.7 Mobile BI

The McKinsey Global Institute has identified 12 disruptive technologies that will have a massively economic impact from today till 2025[13]. The mobile internet is seen as the most important disruptive technology. The McKinsey researchers underpin their choice with impressive numbers: more than 1.1 billion people use smartphones and tablets. Last year, the number of smartphones has grown by 50% with an increase of app downloads by 150%. In 2013, smartphones and tablets have already outnumbered PCs. It is estimated that in 2025, 80% of all internet connections will be mobile. Finally, McKinsey estimates the economic potential of the mobile internet at $10.8 trillion.

Therefore, the mobile internet is no more an add-on to the internet enabling the field to stay connected with the organization and its customers. In the meantime, it morphed into the internet. Information is now ubiquitous and anytime available. This requires a new thinking. Mobile access to processes and applications is no more an add-on, mobile comes first. Later on, there will be mobile only. Mobile is the first model, because the world gets mobile.

This carries over to mobile BI. But if mobile BI is to be the first model for BI, then there are new requirements on traditional BI. They can be concluded by five bullet points:

1. A mobile BI solution should be part of a BI platform in order to be easily integrated into the process and application landscape.
2. A mobile BI solution should be independent from device types. It should be usable on all established smartphones, tablets and other mobile devices with any customization.
3. A mobile BI solution should fit into the existing IT infrastructure, standards and security policies.
4. A mobile BI solution must be able to access all relevant data sources of an organization and must support all relevant output formats.
5. A mobile BI solution combines the full functionality of BI solutions with the ergonomic advantages of mobile devices. This includes a read/write access so that data can be updated and new data can be captured.

Requirements 2 and 5 are extra tough. They provide rather strong blockers for a deployment of mobile BI. They particularly block BYOD[14] concepts, since the support of various device types causes high costs and needs many resources for porting and adapting mobile solutions to the specific device types. Indeed, one of the toughest challenges for BI app designers is designing for different device types. BI apps need to be ported to the device's operating system and to be matched to the different display sizes of devices – from desk tops, laptops, tablets to smart phones. Size is just one aspect to be considered, other factors play a role, too. Orientation, for instance, depends whether a device has been optimized for portrait or landscape, pixel density (pixel per inch – ppi), as well as navigation controls. It exist two alternative approaches for solving the problems, but both are not really satisfying:

[13] McKinsey Global Institute "Disruptive technologies: Advances that will transform life, business and the global economy." http://www.mckinsey.com/insights/business_technology/disruptive_technologies, May 2013.

[14] BYOD = bring your own device. This is a concept allowing employees to use their own device(s) within the enterprise for enterprise purposes. It requires that IT has to support all different device types.

- ***Native BI apps.*** Until recently, they have been considered as the most advanced app solutions, because they provide full interactivity. They also offer periodic caching improving the performance of the app plus an offline use. So, users of native BI apps get all advantages of mobile BI - ubiquitous access to information at any time with the full power and ergonomics of their device – and are absolutely happy and fully satisfied. But the drawback is, such apps have to be developed in a device specific way, i.e. rather expensively and resource intensively: All reports and dashboards need a device type specific app. If an organization goes BYOD, this is inevitable! It costs many resources, much time, and from the point of maintenance, it is a waste of time and resources.
- ***Web apps.*** Mobile devices come with mobile browsers so that web based HTML BI applications will work. HTML5 enables dynamic, interactive BI apps that run on both, PCs and mobile devices. In fact, HTML5 consumes more system resources than do native apps, but given the power of today's mobile devices, this does not matter anymore. Furthermore, HTML5 is a recognized web standard that is future-proof, i.e. a protection of investment in developed apps. That looks good, but there is still a problem. Web apps indeed run anywhere, but screens of PCs and various mobile devices have different dimensions so that a web page does not fit all devices. Furthermore, devices have different navigation controls, so that the ergonomic advantages of a given device cannot fully be exploited. This lowers the acceptance of such solutions by users, whereas development prefers this approach.

Today, "**responsive web design**" is a way out of the difficulty of device specific apps. This is a design and technical approach based on HTML5 for developing web pages that fit automatically to the device type and its properties. Thereto, design rules have to be developed for the various device types to be supported. They are filed into a style guide. If a web app is invoked, a media query for device identification is executed first. It allows pulling the corresponding design rules in the context of the app. The visualization of a web page is now done according to the device specific display requirements. It includes arranging and presentation of components as well as execution of navigation controls.

Goal of responsive web design for mobile BI is a visualization that fits any device. Hence, users get the impression of using a native BI app, whereas an app is just once developed for all device types. Some additional cost is to be added for the development of the style guide. But during usage of the style guide, it is good practice to continuously improve and extend the style guide. This improves the acceptance, and it considerably lowers cost in development and maintenance. First mobile BI solutions based on responsive web design came to market end of 2013. One of the forerunners is arcplan with its mobile version of arcplan 8 (fig. 8).

The deployment of mobile BI in an organization requires a program, because different IT groups are challenged by various requirements. It begins with **mobile device management (MDM[15]).** MDM can be considered as an extension of traditional client management to an additional type of hardware. This is why nearly all established vendors of software for PC management meanwhile offer solutions for mobile devices. They include inventory, software

[15] please note: MDM may also mean "master data management", see chapter 6.5.

distribution, and security policies complemented by specific functions for smartphones and tablets for preventing data leakage.

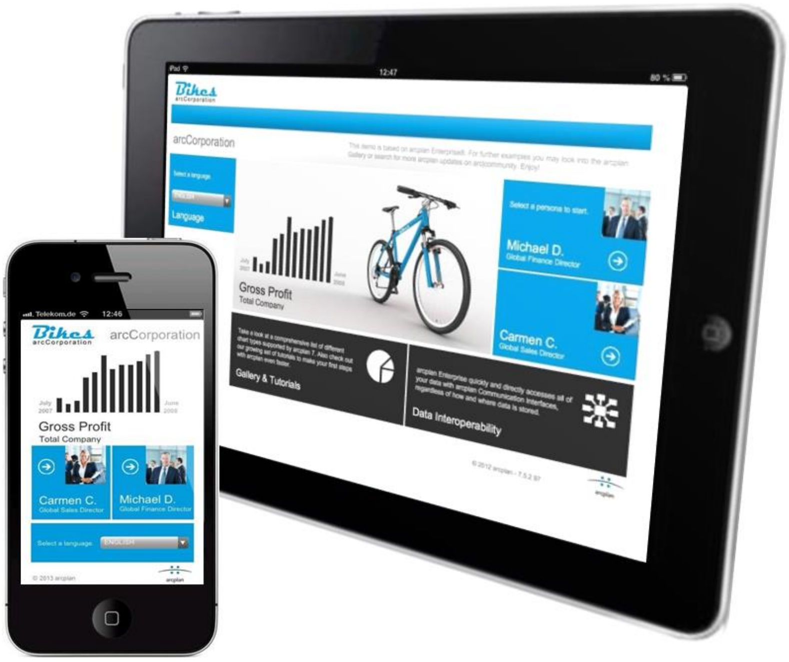


Figure 8: Example for responsive web design: the mobile version of arcplan 8. A dashboard automatically fits to the different screen dimensions of a smartphone and a tablet, and it also supports the different navigation controls.

- ***Data security.*** It is mandatory that business critical data can be always encrypted during transfer to and local storage in a mobile device.
- ***Theft/loss of a mobile device.*** For these cases, a clear and exact policy is needed that includes locating of devices and tele-deletion of critical data for preventing unauthorized use.
- ***Use of private devices in an organization (BYOD).*** BYOD requires considering additional organizational and legal aspects (for example usage rights or private usage patterns).

Meanwhile, **Mobile application management (MAM)** offers alternative solutions for BYOD. The idea is to put applications and data for non-managed devices into a specific area. Such an enterprise app store can be protected by encryption and by controlling the execution of these apps by policies. For instance, the administrator can dictate the conditions of starting an enterprise app and can chirurgically delete apps and its data with affecting private data of a user.

Organizations supporting both scenarios for mobile BI, i.e. an app store with enterprise apps plus a BYOD policy, need a solution that covers MDM and MAM. The combination of both approaches is called **enterprise mobility management**. A management concept like this for mobile BI should also include data synchronization. This is needed when mobile devices are not used as primary devices, but complement the enterprise PC. When transitioning from device to device, users should get their data and files always up-to-date.

3 Strategies, Processes, People, Metrics, and Governance

3.1 Process and Service Oriented Business Intelligence

As we have seen in the previous chapter, performance management and analytics put business intelligence into the context of processes. The obvious consequence is: performance management and analytics also put BI into the context of strategy and people. Today, processes are cross-functional, cross-departmental, and cross-enterprise. They link the suppliers of the suppliers with the customers of the customers. Let us recall the definition of a business process (cf. fig. 6).

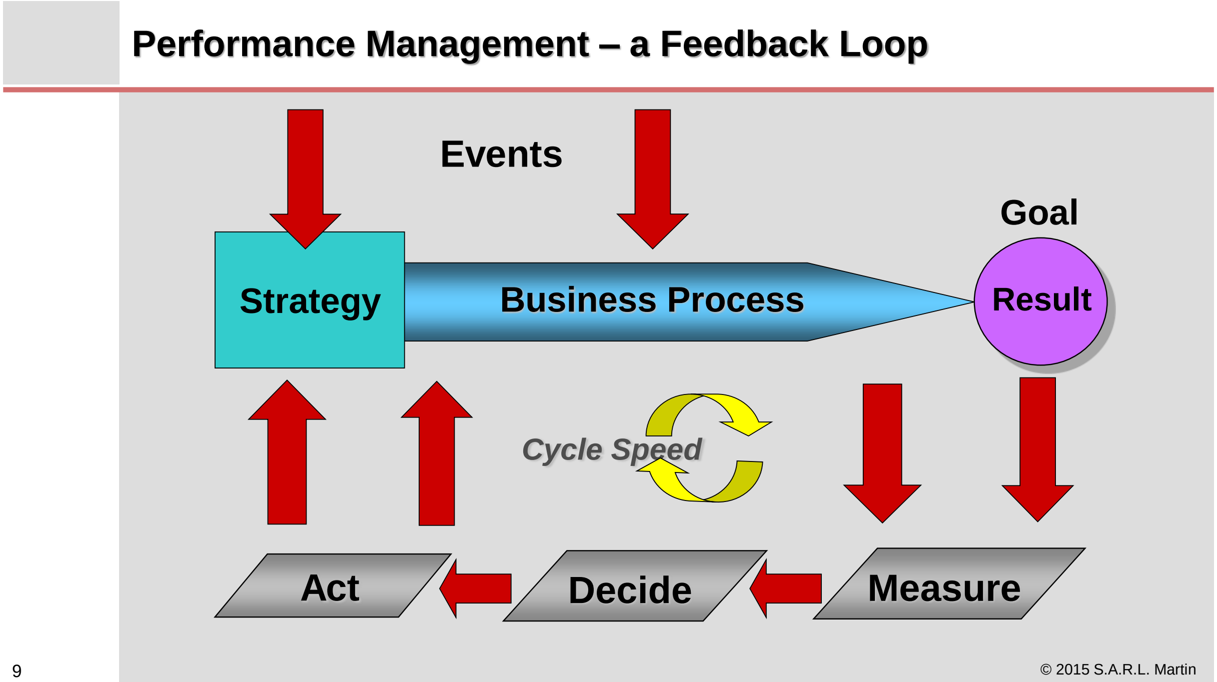
Definition: **A business process is…**

a set of activities and tasks carried out by resources
(services rendered by people and machines)
using different kinds of information
(structured & poly-structured)
by means of diverse interactions
(predictable & unpredictable)
governed by management policies and principles
(business rules & decision criteria)
with the goal of delivering agreed upon final results
(strategies & goals)

The benefits and advantages of integrated end-to-end processes are obvious:

- Faster and more reliable processes cut costs. Automation improves speed and quality of processes. The result is higher throughput with fewer resources.
- Integrating processes shortens time-to-market. The ability to respond quickly to new opportunities, customer needs, market dynamics and problems simply translates into increased revenue and profitability.
- Safe and reliable processes minimize risk. High process quality means less costly aftershocks to the bottom line. In addition to savings realized by reductions in post-sales service, enterprises can benefit from high customer satisfaction and, ultimately, market share. The ability to anticipate problems, customer needs, and market dynamics makes the intelligent real-time enterprise a reality.
- Through flexible process management - independent of applications - you maximize business flexibility and agility. By removing the constraints in hard-coded processes intrinsic ERP- and other standard application packages your processes will move in line with market dynamics.
- Process-orientation creates transparency and traceability. There is no alternative to compliance with the regulations of public authorities and the requirements imposed by auditors.

Therefore, “Business Process Management (BPM)” is one of the most important challenges for today’s digital enterprises. BPM and performance management are the process-oriented latest version of managing an enterprise: planning, execution, and performance management have always been the three basic categories of all management (“make a plan, execute it, and manage to keep the actual in line with the plan”).



Performance Management – a Feedback Loop

9

© 2015 S.A.R.L. Martin

Figure 9: Metrics-Oriented Management is a top down model for information-based business management. Measurable goals and objectives are derived from the strategy. Based on strategy, goals and objectives, business processes and business metrics for efficient process control and continuous optimization are modeled in parallel. Technical implementation of processes and metrics follows the principles of a SOA (service oriented architecture) by operational, collaborative and analytic services. Based on monitoring, decisions are taken either manually by people or automatically by decision engines. Decisions lead to actions for controlling the process and its performance (tactical and operational performance management) as well as updates strategy, goals and objectives (strategic performance management). The loop closes. Synchronizing monitoring, decision and action taking with the speed of the business process and business dynamics is key – indeed, this is a foundation of the digital enterprise that must be ready to work in real-time.

Performance management within the BPM model includes all processes that extend across all functions within a business, and beyond to all other relationships in business to business and business to consumer. Metrics-oriented management is the top down principal of performance management for optimal enterprise management by a closed-loop approach (fig. 9). Business strategy determines which business processes are to be executed and managed by the enterprise. Business metrics are associated with each business process. Business metrics are defined by goals and objectives to manage a process in a measurable way with information, performance indicators, rules, and predictive models. In the end, this means that the performance of white- and blue-collar workers becomes transparent and can be judged in an objective way. Consequently, salary and bonus payments can well depend on objectively measured performance. But note that in certain countries, this has be in compliance with labor law.

Embedding analytics in processes requires a new approach to process modeling as well as a new approach to business intelligence. Modeling process logic and flow only as in the past is insufficient. We now have to model simultaneously metrics, responsibilities and roles. We have to link strategy and goals to processes, metrics, and people and to build the closed-loop. This is all about **Governance**.

> ***Example:*** Monitoring and controlling of sales processes. Sales methodologies describe and structure the sales activities across the sales cycle. The sales cycle is typically defined as the time period between the identification of a lead and the payment of the bill according to a signed contract. The methodology describes the various levels of qualification of a lead and the actions to be taken to move a lead from one level to the next. These levels correspond to the different states of the sales process, where the desired end result and final state is the payment of the bill. The number of levels of qualification depends on the selected sales methodology (and does not matter in this example). Now let us apply performance management to this sample process. The metrics for monitoring and controlling of our sales process are the number of qualified leads per level, the estimated/achieved value of a deal, and the transition rate and transition time from each level to the subsequent level. Based on these metrics, we can now act proactively. Assume that the objective of our sales process is a revenue of x € in six months. We then can estimate the number of leads per level that is necessary to achieve this goal by evaluating the defined metrics and compare the result with the actual sales data. If the result shows that we will not achieve our goal, we still have time to preventively counteract by taking actions for controlling the process, e.g., an additional lead generation for filling up the sales funnel.

Metrics-oriented management is based on information management (see chapter 6). Information has to be available in "right-time" (see chapter 2.3) for triggering manual or automated decisions for process control. This corresponds to the "**information supply chain**" paradigm: supply the right information in the right time to the right location to the right information consumer to trigger the right decision. So "right time" means synchronization of information supply with information demand.

Business metrics represent **management policies** within metrics-oriented management. The idea behind is obvious:

You can only manage what you can measure.

So, flexibility of changing and updating any metrics is one of the top requirements of the model. Furthermore, business metrics must be consistent. Metrics specified to control the execution of a particular group of processes should not contradict other metrics. Indeed, metrics are cross-functional and cross-process: The performance of a business process may influence and interfere with the performance of other processes.

> For ***example***, *term of delivery*, a supply chain related metric (see chapter 2.3), may influence *customer satisfaction*, a customer relationship management metric.

These issues are addressed by **business scorecards**. A business scorecard aligns all management policies presented by all metrics across the enterprise. It presents the aggregated top management policy of the enterprise as well as all details for all employees. Examples of particular business scorecards are Norton/Kaplan's balanced score card or the six sigma model. The balanced scorecard, for instance, is a collection of metrics that is not only based on financial parameters, but uses also customer, employees and shareholders loyalties to provide a look to the corporate performance beyond the purely financial quarterly results. It presents indeed one particular style of management policies. Despite the wide variety of these metrics, the final goal remains the same: transform data into information and knowledge and maximize its value for the business by closing the loop: We now base planning, monitoring and controlling of processes on information, facts, and knowledge.

But caution, do not exaggerate the principles of management by metrics, because there is another principle to be regarded:

You only should measure, what can be used by managers and can launch actions.

Furthermore, besides the question of consistency of metrics, there are other factors to be considered. They are essential for the success of performance management[16]:

- Dysfunctional compensation plans: Compensation plans often do not keep up with where organizations need to go. This will skew managers' and employees' decisions.
- Poor metric selection, for example counting leads instead of actual resulting sales.
- Lack of leading indicators: Most managers are more comfortable with solid lagging indicators than they are with squishier leading indicators. While leading indicators require a great deal of thought to get right, you must include them in your key metrics, lest you create a company of backwards-looking managers.
- Poorly-defined metrics: These tend not to be taken seriously, since they may unfairly reward or penalize people and departments. Metrics must be based on clearly-defined variables.
- Self-fulfilling metrics: These are potential leading metrics where management losses sight of the point and accidentally makes their value a self-fulfilling prophecy.
- Blind benchmarking: The strategic mistake that managers make in benchmarking is that they try to converge blindly to the industry average. Benchmarks should be tools of understanding, not instruments of oppression.

Performance management is applied to all business domains like customer relationship management, supply chain management, human relations etc.

> **Example: *Financial Performance Management*** like any other analytic solution is a closed loop process characterizing the performance and information management of financial information and processes. The process stretches from planning, budgeting,

[16] Source: Dave Kellog's Kellblog http://kellblog.com/2014/11/11/dont-be-a-metrics-slave/, access January 15th, 2015

forecasting and strategic planning to audits via financial metrics including the statutory legal financial reporting and consolidation requirements. This corresponds to legal compliance. Financial performance management includes profitability analysis as well as simulations and what if analysis. Decisions are then made based on the financial metrics and analysis and fed back into the planning, budgeting and forecasting activities: The loop is closed.

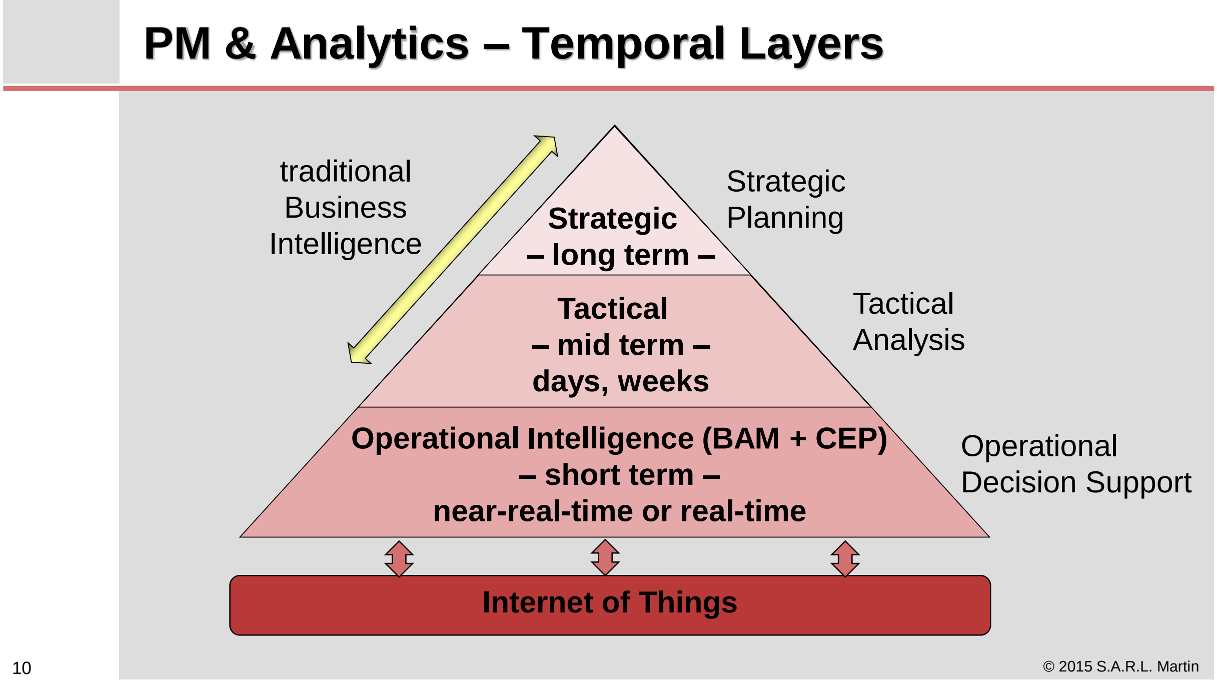
3.2 Operational Intelligence

As Figure 9 already implies, performance management spans across three levels, the operational, tactical, and strategic level (fig. 10). Traditionally, business intelligence mainly focused on enabling decision support in the context of strategic planning and tactical analysis. The idea was to use metrics designed for long term outlooks, and the basic concepts were to measure and to monitor the achievements of strategic goals, for example customer satisfaction, customer value, term of delivery, supplier value, staff fluctuation etc. Long term here relates to the dynamics of the process. Question is how fast can actions influence the process and significantly change its behavior and their indicators. Here, the tactical level comes into play. Achievement of tactical goals can be considered as mile stones towards strategic goals. Actions targeting the achievement of tactical goals typically address a time frame between some few days to several months.

Today, process orientation operationalizes business intelligence, i.e. operational processes can be monitored and controlled in near-real-time (“right time”) or even in real-time via intelligence. Such an “**Operational Intelligence”** – indeed, it might be better to call it “Operational Performance Management” or “Process Performance Management (PPM)” – should enable decision support of running processes and services so that interventions and actions based on a continuous and real-time monitoring can alter the performance or can even change the processes and services on the fly.

The infrastructure for operational intelligence is information management, because operational intelligence typically requires parallel and simultaneous access to analytic, historic and operational data for contextually relating these three data sources. As an example let us consider fraud detection in credit card transactions. Significant events pinpointing to fraudulent use of credit cards must be put into the context of the corresponding customer profile, for instance, the typical credit card use by each individual customer. If not, there is the risk that regular and by its customer intended transactions are considered as a fraud and are blocked causing a bad surprise to say the least.

Operational Intelligence includes **“Business Activity Monitoring (BAM)”** and **“Complex Event Processing (CEP)”**. In the digital world, BAM and CEP are important concepts for monitoring and controlling events happening in the internet of things. As examples, we may think of real-time analysis of data streams like machine or sensor data as well as server log data. Concepts of BAM and CEP are discussed in chapter 7.1 in more detail.



PM & Analytics – Temporal Layers

traditional Business Intelligence

Strategic – long term –

Strategic Planning

Tactical – mid term – days, weeks

Tactical Analysis

Operational Intelligence (BAM + CEP) – short term – near-real-time or real-time

Operational Decision Support

Internet of Things

10

© 2015 S.A.R.L. Martin

Figure 10: Performance Management (PM) is the process of managing the performance of business processes by applying metrics, deciding on the outcome of the metrics, and launching actions for controlling the performance and/or the process, a closed loop model for corporate management. Performance management spans from operations to strategies. It should be noted that operational performance management is a key concept for monitoring and controlling of objects in the internet of things. A key issue for all performance management approaches is to put the metrics into a monetary context. This requires process-oriented accounting principles like activity based management/costing. But in practice, this approach is not very well adopted and used.

These ideas of performance management originate from control theory. As room temperature is monitored and controlled by a closed loop feedback model in near-real-time, business processes should also be monitored and controlled not only on the strategic and tactical level, but also on the operational level, i.e. in real-time. Let us come back to our model of an information supply chain. Its real-time principles already provide a monitoring and controlling of operational systems, and let us recall, the key principle of an information supply chain is the availability of the right information in the right time at the right location for the right purpose. Consequently, information is treated as the duty of the information provider. This is in contrast to the data warehouse model, where information is treated as the duty of the information consumer. In performance management in contrast to the data warehouse model, the provider of information is now responsible for propagating information. The implementation of our new model, for example, can be done through publish and subscribe communication methods.

> ***Example.*** In a web shop, *product availability* is a valuable metric when controlling the order process. *Product availability* is an operational metric. It measures stock via sales and supply transactions. Hence, *product availability* is synchronized with transactions. When *product availability* gets below a certain pre-defined threshold, an alert can be launched. Such an alert could automatically trigger an additional shipment. If shipment is not an option, then the product could be blocked in the product catalogue so that customers cannot place any orders for this product. This is a pro-active action that avoids cancelling customer orders. In the end, delivery

expenses and the frustration of customers due to the unavailability of a product are minimized. Furthermore, the blocked product could be tagged by a note stating when the product will be available again.

An automated teller or a vending machine runs in a similar way. If such a machine is interconnected with a logistics system, the machine can automatically order when the reserve gets lower than a predefined threshold. This is a typical example from the internet of things.

This example shows how information can be used to monitor and control operational business processes in a preventive way. This is especially interesting in the internet of things. Processes are automated, and manual interactions of product managers are minimized. By the way, what is the meaning of "real-time" in this example? Here, *product availability* in web shops is typically measured twice a day. This is an empirical experience balancing cost of measuring with cost of risk ignoring *product availability* for controlling the process. In the internet of things, this question is answered differently, because machines with local intelligence measure and log events without noteworthy additional cost.

Operational performance management has been first addressed by vendors coming from process engineering and business integration. They added reporting and graphical features for visualizing operational performance metrics. Via activity based management and costing these metrics can be also put into a monetary context. But technically, this requires having access to financial data in a data warehouse.

Vendors of traditional business intelligence tools have been the first to address tactical and strategic performance management. They moved from the data warehouse model and business intelligence tools to analytic applications and closed loop processing. These two different approaches to operational performance management are converging already since 2005. Today, the concepts of operational performance management carry over to the internet of things.

3.3 Governance: Roles, Responsibilities, Rights and Security

An important success factor for performance management and analytics is governance. Governance generally speaking means a framework for organization as well as monitoring and controlling of a business unit according to agreed principles. These principles can be put from the outside, for example by the legislation, or from the inside by internal rules and management policies. Therefore, governance can be applied to various domains. Governance applied to a business as a whole is called corporate governance, and applied to IT, it is called IT governance. So, let us first define the overall term "governance":

Definition: **Governance** means management and behavior compliant to rules and policies. All activities of all enterprise resources – people, machines, systems – must be compliant to all management policies and guidelines. Everybody acts as he should act.

Governance links processes, metrics, people, roles and organization. Now, we put the principles of governance into the context of performance management and analytics. This addresses the following question: Who needs what information, where, when and why? This is governance of information provisioning that we will now call for the sake of simplicity "BI governance". We define:

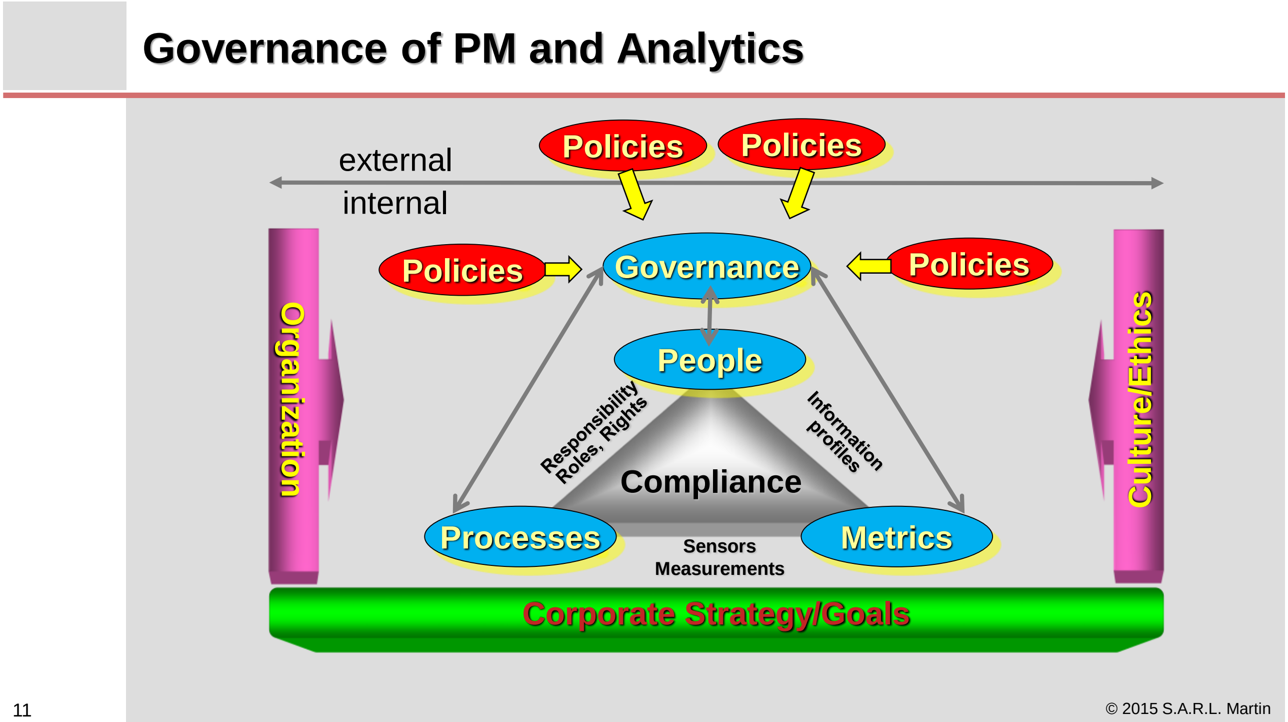
Definition: BI governance means the set of all processes, metrics and structures for managing and protecting all enterprise information so that the right information and the right tools are enterprise-wide available for analysis and all tasks of monitoring and controlling.

This is reflected by a **process ownership model.** It associates roles, responsibilities and organizational units to processes. The traditional process model considered mainly process logic („workflows"). In the context of performance management, it is now extended by modeling in parallel process logic plus metrics, and in the context of governance, it is extended by process ownerships.

The process ownership model describes who of the constituents (employees, partners, suppliers, customers etc.) participates in and is responsible for what processes and activities and what is his/her role. This enriches the process model and links processes to people and organizational structures. In metrics-driven management, the process ownership model also includes **information profiles**. They describe the links between the process ownership model and the metrics for monitoring and controlling the process and its performance. This can be understood as information sharing and filtering. The constituents share data, information and knowledge in the context of their process-oriented communication and collaboration. Data and information that does not belong to this context is filtered out. Consequently, a by-result is a top down security model based on the process ownership model. The information profiles describe and filter exactly the information that is needed by all constituents based on the context of collaboration. This specifies the relationship between roles and the related metrics. Finally, as a result, the structure of a corresponding business scorecard is defined (fig. 11).

A **business scorecard** visualizes information specified by an information profile. Technically, it is embedded as a portlet into a portal or, in the mobile internet, as an app (fig. 12). A portlet as well as an app is a container for a certain amount of information and/or collaborative tools. Portals as web based man-machine interfaces have evolved from intranet and extranet solutions to the central point of control for collaboration. A portal is defined as a system that enables sharing and filtering of data/information, functions/functionality, content/knowledge, and processes. The same role of a portal plays a corresponding app in the mobile internet. By the information profile, this sharing and filtering is related to the functional role of a collaborative team within the process ownership model. A collaborative team is a group of people representing the various constituents that work together according to the collaborative goals and objectives of the team. In this way, portals and/or apps support cross-functional, cross-departmental, and cross-enterprise virtual teams. As a special case, a team could also consist of an individual portal user. In this sense, portals support governance within performance management and analytics.

A man-machine interface like this (see fig. 7) can be understood as an abstraction layer linking and aggregating content and services as well as reducing the complexity of their access. In this sense, the team-context defines the collaboration bandwidth by the relationships between roles and information profiles, i.e. which data/information, functions/functionality, contents/knowledge, and processes are exposed to the collaborative team. This includes the corresponding business scorecard and the appropriate collaborative tools. Each user gets a personalized environment that can be further individualized.



Governance of PM and Analytics

external
internal
Policies
Policies
Policies
Governance
Policies
Organization
Culture/Ethics
People
Responsibility
Roles, Rights
Information
profiles
Compliance
Processes
Sensors
Measurements
Metrics
Corporate Strategy/Goals

11

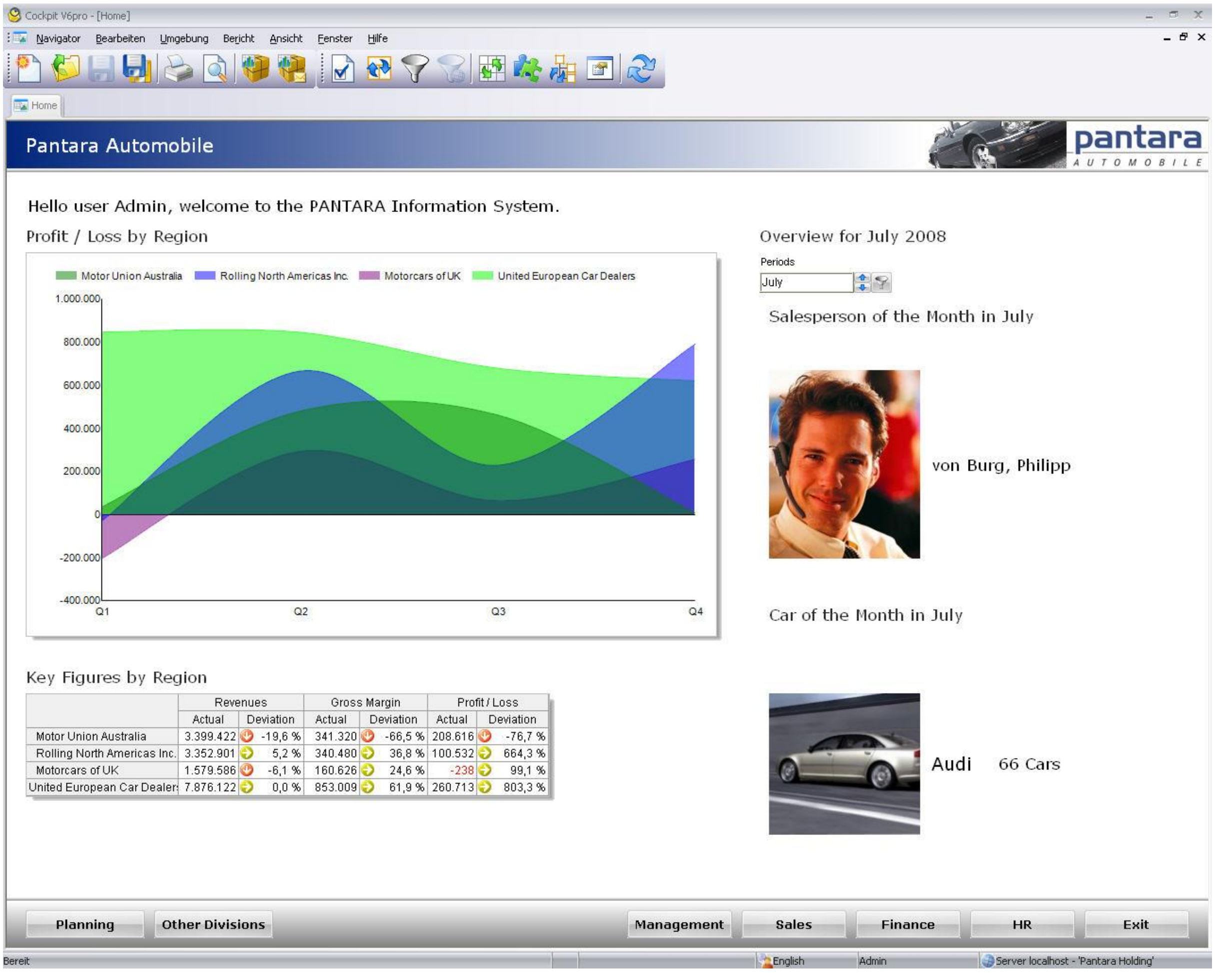
© 2015 S.A.R.L. Martin

Figure 11. For implementing corporate strategies, people, processes and metrics must be linked to each other to ensure that the organization acts as it should. "Compliance" is all about this. It means proactive management and behavior in conformance with all the rules. The rules are either specified externally in the form of policies (for instance by the legislature) or are imposed internally by either the organization or by the culture and ethics of the organization. This is controlled by governance. Governance regulates the relations between people, processes and metrics. Policies, responsibilities and authorities, roles and rights are described by the relations between people and processes. The relations between processes and metrics are described by sensors and measurements, which are intended to monitor and control the performance of processes in the scope of performance management (PM). Finally, the relations between people and metrics are described by information profiles, which precisely describe the set and structure of the metrics needed by a team or a person within process responsibility in the scope of governance. An information profile finally depicts the structure of a business scorecard so that the corresponding metrics can be visualized.

Man-machine interfaces can also be understood as an integration technology. The ultimate integration is done via a human interaction, i.e. within the team-context. A user can execute a message transfer between content and services within his collaborative context. Collaborative tools also provide synchronous and asynchronous tools, e.g., e-mail, blogs, co-browsing, chat, forums, instant messaging, web-conferencing, wikis etc. This is where **social media** technologies come into play. Since social media technologies are service oriented, they optimally fit to the concepts of a SOA and support collaboration of people and teams. In the mobile internet, this is again provided by apps, where typically, an app corresponds to a mashing up of several portal services.

Today, nearly everybody uses social media or mobile devices in private life. So, people are used to social media tools and mobile devices, and like their ease of use. Hence, people expect that enterprise tools work in a similar intuitive way. This makes up a critical success factor for enlightening users by technology and creating acceptance for governance.

Take away: Goal of **BI governance** is to implement the BI strategy, to reinvent BI as performance management and analytics, and to deploy BI into the day to day operations.



Hello user Admin, welcome to the PANTARA Information System.

Profit / Loss by Region

Overview for July 2008

Salesperson of the Month in July

von Burg, Philipp

Car of the Month in July

Audi 66 Cars

Key Figures by Region

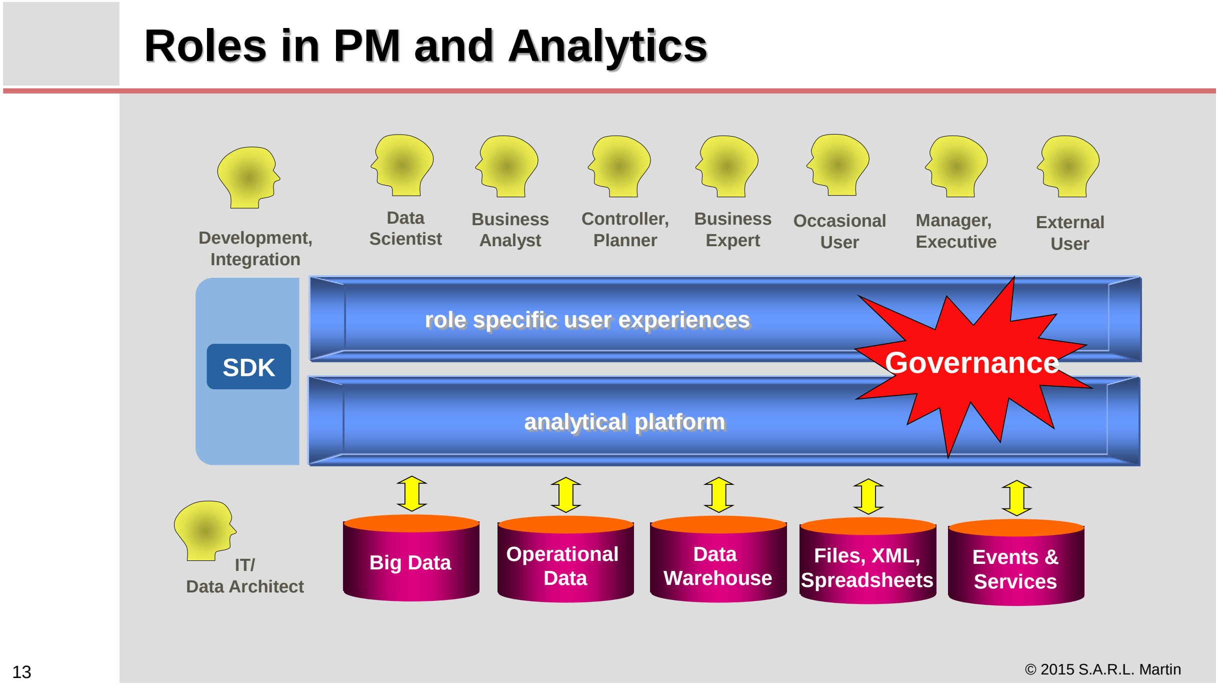
	Revenues		Gross Margin		Profit / Loss	
	Actual	Deviation	Actual	Deviation	Actual	Deviation
Motor Union Australia	3.399.422	⬇ -19,6 %	341.320	⬇ -66,5 %	208.616	⬇ -76,7 %
Rolling North Americas Inc.	3.352.901	➡ 5,2 %	340.480	➡ 36,8 %	100.532	➡ 664,3 %
Motorcars of UK	1.579.586	⬇ -6,1 %	160.626	➡ 24,6 %	-238	➡ 99,1 %
United European Car Dealer:	7.876.122	➡ 0,0 %	853.009	➡ 61,9 %	260.713	➡ 803,3 %

Figure 12: Example of a BI portlet as entry point into „his“ business scorecard (fictitious enterprise, screenshot provided by Cubeware: Various components like chart, table with traffic lights indicating trends, top-1-tables with traffic lights using pictures, titles, logos and background can be placed anywhere. Navigation is done via push buttons.) Today, as in this example, the concepts of visualization of portlets are inspired the design of online publications. For such an approach, the mail order company Quelle GmbH has won the BARC BI Award 2008 at the official Forum BI at CeBIT 2008 (cf. chapter 4.4).

3.4 Working with Performance Management and Analytics

At this point, we have to address the question about the skills needed for benefitting from performance management and analytics. Which are capabilities and support needed to successfully, effectively, and efficiently turn information into knowledge and to take decisions?

We have just learned that a combined approach of **organizational procedures and technological facilities** is best for succeeding with performance management and analytics. It is the combination that matters. In the past and still today, technological considerations were leading the discussions. The question of tools had highest priority. As a rule, BI tools were selected by IT. Technological aspects and concepts were rated highest. People, the users of BI technology, were neglected. Common believe was that usage of tools could be learned by training. But this assumption turned out as a major mistake. Many BI projects did fail, because the potential users did not accept BI tools that were too complex and too complicated to be used. As a consequence, business analysts and power users evolved as a new class of people that were empowered by these tools. Specialist departments became dependent on this new type of information empowered employees. So, information became a kind of luxury product that was not available for everybody.



Roles in PM and Analytics

Development, Integration

Data Scientist

Business Analyst

Controller, Planner

Business Expert

Occasional User

Manager, Executive

External User

SDK

role specific user experiences

Governance

analytical platform

IT/ Data Architect

Big Data

Operational Data

Data Warehouse

Files, XML, Spreadsheets

Events & Services

13

© 2015 S.A.R.L. Martin

Figure 13: A combination of organizational procedures and technological facilities makes the difference: It is the basis for success with performance management and analytics. Technology must come with an intuitive handing of the BI tools, automation of information provisioning and analysis, role specific tools supporting BI governance, and finally with a mature data integration and connectivity. The architecture should be service oriented. This facilitates flexible data integration and is a prerequisite for agility. (SDK = software development kit; an SDK is necessary for customization and extensibility of the performance management and analytics platform.)

Therefore, many BI consultants tried to stop this approach. They turned the tables and emphasized the organizational approach. Technology had to step back. But this approach lacked the necessary balance between the two poles, organization and technology. It is correct that organizational procedures are absolutely necessary, but organizational procedures do fail when not appropriately supported by tools and technology. Technology is necessary for creating acceptance, and technology could even trigger excitement. Therefore, technology is one of the two critical success factors for performance management and analytics. It helps to experience the true value and benefit of information. Technological aspects and organizational procedures must be balanced, both matter.

Today, social media, smartphones, and tablets rule the way of how to handle BI tools. This is the technology that really triggers excitement. It supports organizational procedures in an optimal way. Now, BI governance comes into play. It models the information needs and links it to the organizational and functional roles. Organization and technology play together. It is not only the intuitive way of using tools that is important, but also the ease of how to implement the BI tools in relation to the roles (fig. 13). If this is done well, BI governance will not be considered as a restricting set of rules, but as an empowering framework. If BI tools are well tuned to the information needs of a role, reservation fades away, barriers disappear, and the necessary excitement is triggered: Motivation is created. The goal of BI governance can be attained, a compliant management and behavior of each employee in the context of performance management and analytics.

Besides an easy, intuitive handling of role specific BI tools, the acceptance of these tools also depends on the capabilities of automation of analytic tasks and processes. The flaws of traditional BI like manual information provisioning and manual analysis have to be replaced by reliable and secure automation. Hence, the quality of BI tools with respect to architecture and process and service orientation is another decisive element when selecting tools and technology.

An appropriate technology does not only provide intuitive handling and automation of tasks, but autonomy of the specialist departments from IT. Due to the software ergonomics of BI tools, many analytic tasks can be completely addressed within a specialist department without the help of IT. Today, this is called **self-service BI** (cf. chapter 5.1 and 5.2), an approach that gets together technology and organization.

Definition: Self-Service BI. In 2011, Claudia Imhoff and Colin White defined self-service BI as “the facilities within the BI environment that enable BI users to become more self-reliant and less dependent on the IT organization”[17].

According to Imhoff/White self-service BI should have four objectives: easy access to data for reporting and analysis, user friendly BI and analysis tools, simple and customizable user interfaces of tools, and data warehouse technologies that can quickly be deployed like appliances or cloud based systems.

But self-service BI is not necessarily a fast-selling item. First, one notices that cost of training is typically rather underestimated. Good tools and easy access to data are important, but there is more: Self-service BI starts in the heads of people. If this is not the case, one ends

[17] cf. InformationAge „Self-Service Business Intelligence”, June 2013 http://www.information-age.com/technology/information-management/123457131/self-service-business-intelligence

up in a report chaos, achieves total confusion of users, and gets higher support costs than with traditional BI, where reports are built and analysis is done by IT.

Therefore, self-service BI is to be set up as a program. It starts with asserting the different requirements for self-service BI within the organization: Who are the users? Which degree of self-service is desired? For instance, technology adepts will understand and live the concepts of self-service BI rather quickly, whereas an occasional user will be already happy, when self-service allows altering some parameters in a report for an alternative view on data.

Governance remains a critical success factor, even if IT partially transfers control of the data analysis processes. But in the end, this requires a close collaboration between IT and special departments. Data definitions for the most important performance metrics have to be communicated for consistent reports and analyses. IT and specialist department managers should continuously monitoring the usage of self-service software for detecting possible conflicts with compliance or stopping queries in-time that could block the whole BI system.

Reusable building blocks in self-service BI platform are another critical success factor. BI developers should code these predefined and fixed performance metrics. BI managers should create start libraries with report patterns and standard analysis procedures: Self-service BI users should just select, pick, and combine. IT should not hesitate to invest during the introduction of self-service BI, because it builds the necessary standardization and eases use.

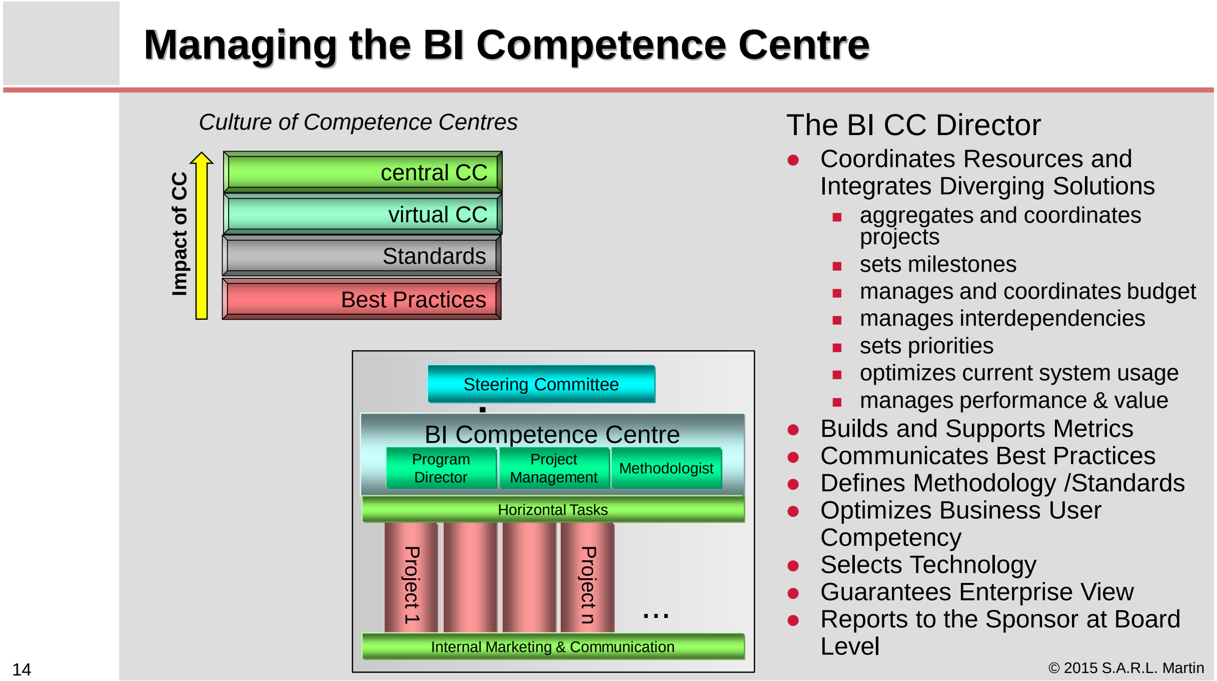
Finally, collaboration within BI teams is critical, too. Collaborative tools support should be part of self-service BI. This includes annotations and comments to reports and analytical results, evaluations of reports and dashboards, recommendations of reports and other results via chat functionality for sharing etc. We will discuss collaborative tools later on in this text.

In the end, this “new” BI becomes a commodity in day to day operations. It improves the productiveness of all people. They now can again fully focus on their functional tasks. Technology really supports them and is no more an end in itself. On the other hand, the higher autonomy of specialist departments discharges IT. BI governance now defines and controls the new and clearly defined division of work between IT and specialist departments. The often stressed relationship between IT and specialist departments is unbent.

3.5 The Business Intelligence Competence Center

As any governance, BI governance consists of four elements, the BI governance processes, the corresponding management policies, an organizational structure, and a technology platform. The **BI governance processes** and policies monitor and control the enterprise business intelligence program. These processes and policies are the processes of the BI strategy, design, implementation and execution. Furthermore, there is a continuous improvement process for continuously updating the BI program by gained experience and lessons learned. This BI governance process model corresponds to the ITIL V3 and COBIT model and uses best practices from management of IT and services.

A **BI competence center (BI CC)** is an organizational construct for BI governance used by many enterprises. It is a cross-functional unit within the organization that acts as an interdisciplinary team and has the responsibility to advocate the deployment of BI in the business. It consists of a steering committee chaired by the BI sponsor, the proper BI CC, and the business analysts and data stewards. The BI sponsor should be an executive or board member for guaranteeing that the BI strategy and the policies of BI governance are respected, applied and executed. The business analysts and data stewards are located in the specialist departments and act according to information governance: BI governance meets information governance. Key assignments to information governance, its relationship to BI governance and the role of data stewards will be discussed in chapters 6.8 and 6.9.



Managing the BI Competence Centre

Culture of Competence Centres

Impact of CC

central CC

virtual CC

Standards

Best Practices

Steering Committee

BI Competence Centre

Program Director

Project Management

Methodologist

Horizontal Tasks

Project 1

Project n

...

Internal Marketing & Communication

The BI CC Director

- Coordinates Resources and Integrates Diverging Solutions
  - aggregates and coordinates projects
  - sets milestones
  - manages and coordinates budget
  - manages interdependencies
  - sets priorities
  - optimizes current system usage
  - manages performance & value
- Builds and Supports Metrics
- Communicates Best Practices
- Defines Methodology /Standards
- Optimizes Business User Competency
- Selects Technology
- Guarantees Enterprise View
- Reports to the Sponsor at Board Level

14

© 2015 S.A.R.L. Martin

Figure 14: A BI competence center (BI CC) can be organized in different ways. The chosen type of organizational structure determines the impact and authority of the BI CC. In the lower right corner, the architecture of a BI CC is depicted, and in the upper right corner, the culture. One can start with a simple collecting and marketing of BI best practices. In the next step, standards can be developed and published as recommendations, and finally, proactive use of standards can be supported by a virtual or central team and shared services can be offered to the BI projects.

The **BI CC** centralizes management of the BI strategy, methods, standards, rules, policies, and technologies (fig. 14). The guiding principle is: **The BI CC plans, supports, and coordinates BI projects and makes sure efficient resource allocation of people and technology.** Its tasks are:

- Control of the application landscape for performance management and analytics,
- Standardization of methods, tools and technology,
- Coordination of specialist departments and IT in all issues related to performance management and analytics,
- Identification and communication of best practices in BI scenarios,

- Methodological and functional support in all questions about and in all projects related to performance management and analytics,
- Cross-sectional tasks spanning from internal marketing, communication, and training to change management,
- Identification and preparation of data via information management assuring data quality. This could be also done by an information management competency center. We discuss the interferences between the competency centers for BI and information management in chapter 6.8.

Traditional BI roles like business analysts and power users are now allocated to the BI CC. But these roles are changing. Due to better ease of use of performance management and analytics tools, business analysts will be less engaged in providing standard information upon request. So, they can spend more time for interactive analytics (“data discovery”). This creates more value for the enterprise. Plus, a new task is attached to them: management of the performance management and analytics methods and technologies (cf. chapter 4.1 and 4.2). Identification and communication of best practices of analytic scenarios are also part of their responsibilities. This requires a close cooperation and collaboration with the information consumers. If an information consumer will be confronted with a new, not yet encountered problem in analytics or in identifying relevant information, a new analytic scenario is jointly developed with a business analyst of the BI CC. Once solved, the new scenario will be added to the analytic scenario portfolio for re-use. In this sense, the BI CC has a continuous learning curve. Its solution portfolio is continuously extended and improved. This corresponds to continuous improvement process.

The value of a BI CC has been proven by many BI programs across all industries. BI initiatives get faster starts and blocking interferences can be avoided. This definitely lowers costs. Finally, BI CCs can be implemented either as part of the IT organization of as part of an operational specialist department, for instance as part of the financial or marketing department. Finally, there is a fundamental rule for BI CCs: They should be enterprise specific and should be adapted to the business ethics. Furthermore, there are several success factors that should be noted.

- All parties involved and all stakeholders should be convicted that a BI CC creates value. This is definitely decisive. Here, the role of the sponsor comes into play. He/she is necessary for a sustained deployment of a BI CC and for overcoming any objections from various enterprise levels. A well-communicated position of the board is also crucial for success.
- A BI CC should have a mission statement clearly defining goals, objectives, and competencies. This includes a link to the enterprise process map. As a result, in many cases, potential resistance and scrap for competencies can be proactively administered and avoided.
- The introduction of a BI CC should not be done in a big bang, but step by step. The first steps should address the scope of duties that have the highest recognizable added value and where success can be achieved easily and fast and can be well communicated.

- A BI CC is a service provider. The portfolio of services is to be properly defined and to be documented. The goal is to point out to all involved parties how they can be supported by the BI CC and how their own work and their own responsibility are changed and what the interfaces are.
- A continuous improvement process (that is part of any governance) is to be established. This notably includes performance management of the BI CC via metrics like average implementation of requirements, number of resolved cases, improvement of BI maturity etc. and of course, compliance with service level agreements.
- The BI CC team should be interdisciplinary. Competences should encompass IT, business specific and enterprise specific know how. This is essential for a successful communication between the BI CC and the specialist departments. This also means that in a BI CC, a certain number of experienced business people should be merged with innovate persons and with process know how.

3.6 Agile Methods in BI-Projects

Special departments always complain: Application development takes too long. This is especially true for BI projects. The reason among others is the usage of traditional application development methodologies like the waterfall model for project management. Such methodologies have also been applied to BI projects, and consequently, all disadvantages have been inherited. Since then, new methodologies came up for application development like agile methodologies that have been proven to be superior to the waterfall model. Projects become faster and are finished in time and in budget. In the meantime, agile methodologies have been also successfully introduced in BI project management. So, what is new, and why is it better?

Agile methods provide a big advantage: Functional and usable results can be immediately implemented and executed. This is in contrast to the waterfall model that produces results not before a long, tedious, and for the special departments non-manageable development process. The development process is broken up in small and manageable packages. Packages are treated within given time windows. Planned functionality that cannot be implemented within such a time window is omitted. The development process is now guided by the so-called “agile manifesto”.

The study group “Agile BI“ of TDWI Germany e.V. has worked a memorandum for agile BI[18]. Here, we summarize the “agile BI values“ and “agile BI principles“:

Agile BI values

1. Business benefit is more important than sticking to methods and architectural concepts.
2. Continuous collaboration and interaction between requestor and supplier is more important than processes and tools.

[18] see Trahasch, S. et al. (Hrsg.) : Memorandum für Agile Business Intelligence – Entstehungsgeschichte, Werte, Prinzipien und Fallbeispiele; dpunkt, Heidelberg, 2014.

3. Comprehending change is more important than sticking to a plan.
4. Functional BI solutions are more important than detailed specifications.

Agile BI principles

1. The overruling goal of agile BI consists of providing business benefits as soon as possible.
2. Agile BI requires structure, discipline and understanding.
3. Depending on the level of BI agility, various process models, methods, and tools can be used and combined.
4. Various business requirements to a BI system that need a different degree of BI agility are to be embedded into a holistic concept, but to be separated in architecture and organization.
5. BI is a continuous process. Therefore, for the sake of sustaining BI agility, all three domains of BI, architecture, processes and organization have to be regularly examined.
6. The introduction of a new or the exchange of an existing architectural component requires an initial phase for setting the big picture and reconciliation with the master plan.
7. Team members need the capability and willingness to consciously use process models, methods, and tools.
8. Project teams are interdisciplinary.
9. A project team consists of developers and of requestors that mange the project jointly.
10. BI agility can only be sustained, if agile BI principles are part of BI governance.

Agile BI projects take care of the well-known fact that not all requirements of the intended solution are known in advance and that a development process cannot be completely planned. Therefore, the development process is treated as a learning process. Some of the resulting constructs of agile project development have been proven to be very successful and valuable for agile projects:

- **User stories.** Requirements on a software solution should be described in a generally comprehensible natural language. (IT shaped gobbledygook is definitely to be avoided.)
- **Product backlog.** There is a responsible for the product backlog. He/she prioritizes the demands. Demand can always be added or discarded. The Product Backlog is a „living system“.
- **Sprints.** Sprints belong to the basic ideas of agile methods. A sprint is used for working a predefined set of requirements from the product backlog. The goal of sprints is to deliver ready to be used software components in a given time window.
- **Test-driven development.** Testing is performed in small steps just like development. Test-driven development means that first, the test is defined, and then program code is written and not vice versa as in traditional testing.

Agile methods mean to abandon well-known and proven process models. This makes up the main difficulty in adopting agile methods. Organization of projects and control of agile projects are not easily established and cause a lot of resistance because they are in contradiction with traditional methods and corporate culture. Agile methods are indeed new and difficult to be integrated into corporate culture because they require a cultural change.

Agile BI projects also suffer from their many interfaces to other projects and systems, in particular to operational transaction systems causing a high degree of harmonization and coordination. This is especially difficult, if projects running in parallel are managed through different methodologies. A lot of friction will arise when project leaders of traditionally managed projects require fixed definitions of deliverables at the start of a project that cannot and will not be promised by an agile BI project, because this is against the agile manifesto. Here, management decisions are the only way out of such dilemmas.

Furthermore, agile BI projects cannot be controlled by traditional budgeting standards. The reason is simple: Agile projects allow changes of requirements, but changes of requirements also mean changing budgets. Consequently, applying agile projects methodologies also requires new budgeting models that have been developed in the meantime, but require again a change of corporate practices.

Despite all these challenges, agile methods for BI projects are to be recommended because they fit very well with the nature of BI. Success of agile BI has also been proven by a couple of studies highlighting well the advantages and benefits of agile BI[19].

3.7 Roles in Big Data Analytics

Big Data analytics requires new skills and roles that have been not required in traditional analytics. From an organizational point of view, these new roles should be attached to the BI competency center. In some organizations like Amazon, eBay, Facebook, Google, Twitter etc. that have already gained a lot of experience in managing Big Data, such new roles making up a Big Data team have been developed. McKinsey & Company[20] described the five most important roles. Originally, that contribution focused on Big Data marketing projects, but it is sufficiently general for being transferred to other Big Data projects, too. In particular, the McKinsey approach identifies roles instead of job descriptions. Roles have the advantage that the necessary expertise and job description can be derived directly from a role description. Let us have a look to these five roles:

- Data Hygienists make sure that data coming into the system is clean and accurate, and stays that way over the entire data lifecycle. This data cleaning (see chapter 6.6) starts at the very beginning when data is first captured and involves all team members who touch the data at any point.

[19] see Krawatzeck, R., Zimmer, M., Trahasch, S. : Agile Business Intelligence – Definition, Maßnahmen und Herausforderungen. In: HMD – Praxis der Wirtschaftsinformatik, 50 (2), 2013, S. 56-63.

[20] see the Blog „Five Roles You Need on Your Big Data Team" in *Harvard Business Review* (July 2013) http://blogs.hbr.org/cs/2013/07/five_roles_you_need_on_your_bi.html?utm_source=Socialflow&utm_medium=Tweet&utm_campaign=Socialflow

- Data Explorers scan through tons of data to discover the data you actually need. That can be a significant task because so much data out there was never intended for analytic use and, therefore, is not stored or organized in a way that is easy to access.
- Business Solution Architects put the discovered data together and organize it so that it is ready to analyze. They structure the data to ensure it can be usefully queried in appropriate timeframes by all users. Some data needs to be accessed by the minute or hour, for example, so that data needs to be updated every minute or hour.
- Data Scientists take this organized data and create sophisticated analytics models that, for example, help predict customer behavior and allow advanced customer segmentation and pricing optimization. They ensure each model is updated frequently so it remains relevant for longer.
- Campaign Experts turn the models into results. They have a thorough knowledge of the technical systems that deliver specific marketing campaigns, such as which customer should get what message when. They use what they learn from the models to prioritize channels and sequence the campaigns.

Roles like Data Explorers and Campaign Experts need skills like cognitive science and behavioral economics. Such an expertise is important for identifying which data is needed for the project, and which not. It is also necessary when interpreting results and deciding on actions. This makes up the value of the McKinsey role model: When tasks and roles are understood, the choice for needed skills and experts in the project team is well defined.

Finally, let us have a closer look to **Data Scientists**. They have the following profile:

- Technical expertise: Profound knowledge in science or engineering is required. This is the foundation for being successful as a data scientist. Thus, data scientists should be selected from that group of people, and be further tested on their expertise in the other required skills.
- Consciousness: the capability, to break up problems into testable hypotheses.
- Communication: the capability, to transport complex issues via anecdotes, and to present them through easily comprehensible and communicable facts, especially via anecdotes.
- Creativity: the capability, to look at problems differently, and to tackle them alternatively. This is the famous „thinking out of the box“.

The role of a data scientist could be split into some “sub-“roles. This is especially interesting for larger organizations, where typically one will find two types of data scientists. On the one hand, a new generation of business analysts will evolve. They work directly for the decision makers, are responsible for the analytic results, and communicate results in the language of decision makers. The second role puts a focus on statistics and mathematics, designs models, and is responsible for the correctness and completeness of data.

“**Data scientists** turn big data into big value, delivering products that delight users, and insight that informs business decisions. Strong analytical skills are given: above all a data scientist needs to be able to derive robust conclusions from data.” *Daniel Tunkelang, Principal Data Scientist, LinkedIn*

In the end, data management is becoming the proper and main task of IT[21], whereas the main focus of the business is mastering of processes and analytics. This implies that data scientists will typically be recruited from the special departments, and not from IT.

[21] This is underpinned by several market studies, see InformationAge http://www.information-age.com/channels/information-management/features/1687078/its-focus-shifts-to-data-management.thtml

4 Methods and Technologies

As we have already seen, performance management is fundamentally different from traditional BI. Focus of BI was tools, e.g., OLAP, spreadsheets, reporting, ad-hoc querying, statistical and data mining tools, etc. Performance management comes with new methods and technologies. Goal is to empower everybody collaborating in the context of a business process by analytics without the need to become a specialist in analytics. This principle is not only applied to employees, but also for suppliers, partners, dealers, and even customers. Analytics must become consumable by everybody.

4.1 Business Components of Performance Management

Metrics and Key Performance Metrics – Metrics are used to measure and to manage the performance of a process and/or to monitor and to control a process. They are derived top down from metricized goals out of strategy and process analysis. Metrics work like sensors along the reach of a process flow. The final goal is the proactive identification of risks and problems. Early warnings become possible so that preventive actions can be taken to bring a process instantiation back on track. (See the example of monitoring and controlling sales processes on p. 40).

Metrics consist of indicators and scales. Scales define how to interpret instantiations of indicators, how to interpret them, and what decisions to take. As we already pointed out, metrics should be modeled in parallel with the process model. This includes the definition of the scale. A scale is typically derived from planning, since planning defines goals and objectives of processes. Alternatively, a scale could also be derived from service level agreements or from management policies.

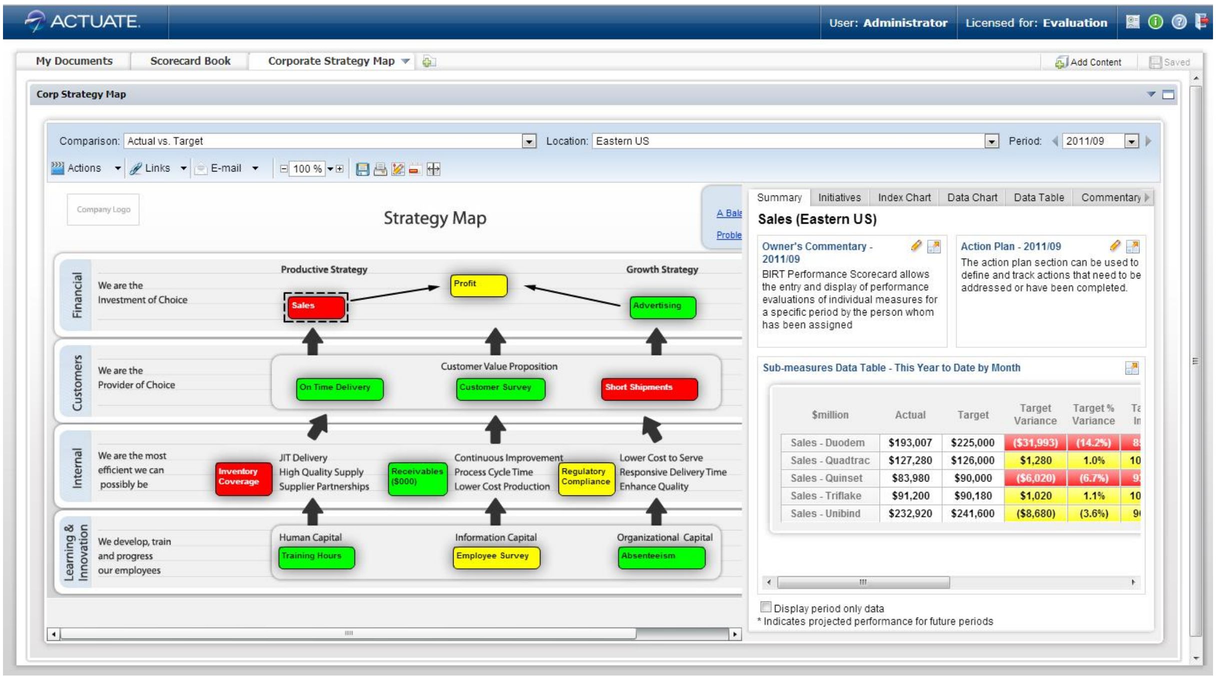
A key performance metric (KPM) is a composite, aggregated metric. *Term of delivery* is an example for a KPM. It is aggregated from detailed metrics like time of delivery across all customers within a certain time period. Typically, an employee will have a lot of detailed metrics according to his roles and responsibilities, but just some selected KPMs. KPMs should be related to the personal goals and match the model of management by objectives. So, KPMs typically have an impact to certain components of the salary.

The information profiles of BI governance define the responsibilities of enterprise resources for the interpretation of instantiations of metrics, for making the right decisions and for taking the right actions. Resources can be persons or machines. In the example about *term of delivery* used as a KPM, a decision maker, i.e. a person, has this responsibility. In case of such human interactions, scales are typically visualized by traffic lights, speedometers and/or other suitable symbols. Green, yellow, and red lights ease and speed up the identification of deviations and exceptions, and improve the interpretation of instantiations of KPMs and metrics. Additional visualization of trends behind metrics and KPMs is a good practice. For example, a red traffic light could have a green trend indicating that we are still in a critical situation, but doing better than last time. In the web shop example about managing the order process by *product availability* (see p. 43), the interpretation is automated by a decision engine, i.e. the resource is a system and visualization is not necessary.

***Business Scorecard* –** We already introduced the concept of a business scorecard in chapter 3.3. It visualizes information profiles. So, it can be understood as a consistent and comprehensive group of metrics together with a management policy for monitoring and controlling the performance of a group of processes, a business division or even the total enterprise. Consistency of metrics is very important, because metrics should not be contradictory and cause conflicts between roles and collaborative teams working in different contexts. The term business scorecard was first developed for strategic performance management, but is now used for all levels of performance management. Known models of business scorecards are the already mentioned balanced scorecard of Kaplan and Norton (www.bscol.com), Baldridge's scorecard model (http://www.nist.gov/baldrige/), and the Six Sigma model (www.isixsigma.com). It should be noted that he majority of enterprises does not exactly apply one of these models, but uses its own customized scorecard model that is a derivative of one of these models. Furthermore, a fundamental difference between a traditional business scorecard and a business scorecard in the context of performance management is to be noted: In our performance management model, all the metrics presented by a business scorecard is linked to corresponding business processes in accordance to BI governance (cf. chapter 3.3).

***Strategy Maps* –** Strategy Maps (fig. 15) are a visual presentation of a strategy based on the cause-and-effect relationship of input and process metrics to their respective output metrics. The well-known and typically used indicators of traditional BI were too much biased by financial data and did not sufficiently consider investments in people, IT, customer relationships as well as in supplier and partner networks. This is why the standard planning and reporting systems like profit and loss and cash flow based on traditional BI indicators are not applicable for monitoring and controlling of resources beyond finance. In genuine performance management, we overcome this problem by the concepts of process-orientation: We use metrics as sensors and cause-effect relationships between the various goals and objectives within a strategy. Strategy determines the goals and objectives of value creation by processes. This is depicted by strategy maps, and the business scorecards provide the translation into decisions and actions for monitoring and controlling processes in a proactive way. Strategy Maps as well as Business Scorecards are not static. Market dynamics, customer needs as well as changes in the organization drive and change strategy. Hence, strategy maps as well as business scorecards must be easily and quickly adaptable to new situations.

***Business Rules* –** They represent cross process decision logic in the context of business expertise and management policies (refer to the definition of a business process on p. 38). Modeling of rules is either top down by an expert system type approach or bottom up by generating predictive models (e.g., a customer behavior model developed through a data mining process). Ultimately, business rules can be modeled by a combined top down, bottom up approach aligning predictive models with expert rules. Business rules must be managed centrally in a rules repository. They must also be managed independently of business processes. The reason is the n : m relationship between rules and processes: a rule can belong to several processes, and a process can have several rules. When business rules are hard coded into the processes, then a chaos for maintenance of rules is inevitable after some short time, since the consistency of rules will be in danger. Furthermore, re-usability of rules would not be possible.



Comparison: Actual vs. Target
Location: Eastern US
Period: 2011/09

Sales (Eastern US)

Owner's Commentary - 2011/09

BIRT Performance Scorecard allows the entry and display of performance evaluations of individual measures for a specific period by the person whom has been assigned

Action Plan - 2011/09

The action plan section can be used to define and track actions that need to be addressed or have been completed.

Sub-measures Data Table - This Year to Date by Month

$million	Actual	Target	Target Variance	Target % Variance	Ta In
Sales - Duodem	$193,007	$225,000	($31,993)	(14.2%)	8
Sales - Quadtrac	$127,280	$126,000	$1,280	1.0%	10
Sales - Quinset	$83,980	$90,000	($6,020)	(6.7%)	9
Sales - Triflake	$91,200	$90,180	$1,020	1.1%	10
Sales - Unibind	$232,920	$241,600	($8,680)	(3.6%)	9

□ Display period only data

* Indicates projected performance for future periods

Figure 15: Example of a Strategy Map of a Balanced Scorecard Model built with OpenText/BIRT 360+. In the Strategy Map illustrated here, input and process metrics are shown in a cause-and-effect relationship to their respective outcome metrics. This pictorial representation of the strategy allows the organization to evaluate its effectiveness by tracking key measures relating to each corporate objective.

Alerts – Event-orientation enables alerting services. An event is indicated by the arrival of information that is external to the process. It indicates an exception to be managed.

> ***Example:*** Let us take the marketing process „campaign management“. A competitor’s campaign influencing our campaign is an event that is to be detected and to be counteracted. An event like this is operational. But there are also events acting on the strategic level, for instance the market entry of a new competitor. In such a situation, it might be necessary to examine all actual sales and marketing processes and perhaps even to remodel some or all of them. Now, speed of change matters. If it takes weeks or even months to implement the new processes taking into account the new competitor’s market penetration strategy, we could easily lose market share. This is one of the reasons why IT departments are moving from application orientation to SOA based business processes.

If an event occurs, the first step is its identification. In a second step, an alert has to be activated by an automated notification for making a decision what to do. The decision could be made by a responsible person or by a system. Here, the already mentioned **BAM tools** come into play. If the event is operational, time matters, and a reaction in real-time is required. All information that is necessary to process the event/alert should be available to all recipients in right time for making the right decision and taking the right action. Again, right time means to synchronize the speed of the process with the delivery of information via the propagation. If speed is high, and the delta between event/alert and decision/action becomes small, then a human interaction may be to slow: The decision / action taking must be automated. Examples for automated decision/action taking can be found on various web sites where recommendation engines are working. Rules engines or rule services are state-of-the-art technology for automated decision taking (for more information see chapter 4.3 and chapter 7.1).

Information access (BI widgets). BI widgets ease the access to information and analytical content: They deliver in-time, personalized and intelligent information especially for the occasional user. Via a rather simple drag and drop interface, users can access directly all relevant BI content from a desktop or a mobile device. They also can organize and adapt the BI content by themselves. Thus, BI widgets allow a simplified access to analytics, improve the productivity of users, and lower IT cost by this kind of self-service.

***Broadcasting* –** These are services for delivering personalized messages to millions of recipients via SMS, e-mail, fax, Twitter, pager, mobile services etc. RSS („real simple syndication“) feeds are becoming the leading technology for data syndication pushed by the more and more widespread usage of Web 2.0 concepts. Using exception conditions and recurring schedules as triggers, events can be automatically created and propagated to processes and people within the enterprise or to any external community. Content can be personalized to the individual subscriber, preventing information overload and ensuring that security requirements are strictly enforced.

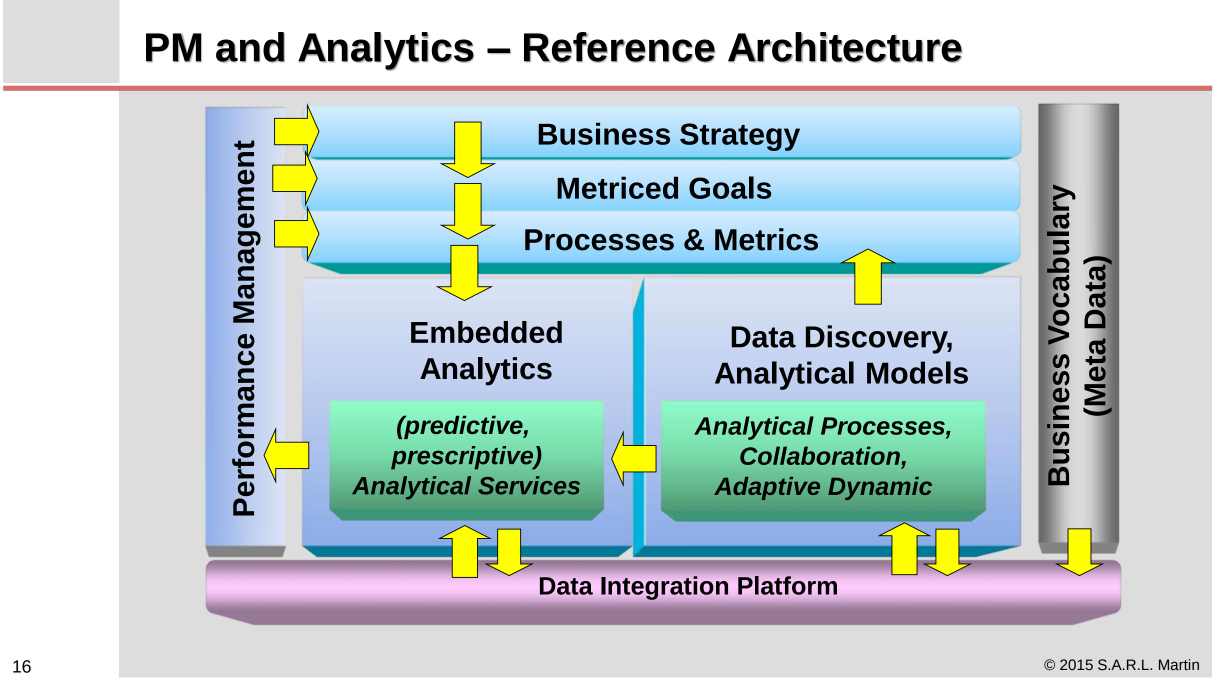
4.2 From Business Intelligence to Performance Management and Analytics

Process-orientation drives the evolution from BI to performance management and analytics. The business components for this require new BI tools and services, a new architecture for positioning the tools into the context of performance management and analytics (fig. 16) as well as a new fresh thinking in terms of processes and services. We are making the next step from Business Intelligence to **Performance Management and Analytics**. Here, we list the fundamental differences to traditional BI:

Performance Management and Analytics are now process-driven, no more data-driven. Performance management and analytics link business strategy to processes, metrics and people according to their roles in collaborative teams: the use and value of information now goes beyond power users and business analysts that in the past were the only people benefiting from information provided by traditional BI tools. Performance management and analytics now empower all participants of the enterprise value network, suppliers, partners, dealers as well as customers. It fully targets the business, no more IT.

Analytics is predictive. It is aimed at responding to unforeseen events and revealing new insights and unexpected discoveries. This makes analytics an indispensable prerequisite of risk management. Therefore, analytics is not limited to the analysis and diagnostics of historical data in a data warehouse or in data marts.

Embedded Analytics – from strategy to operations. A SOA makes it happen. Enriching operational processes by embedded analytics allows the synchronization of information delivery with process speed and the interaction of information at the speed of business so that decisions and actions can be taken in right-time. Through embedded analytics, processes become intelligent and event-driven.



PM and Analytics – Reference Architecture

Performance Management

Business Strategy

Metriced Goals

Processes & Metrics

Embedded Analytics

(predictive, prescriptive) Analytical Services

Data Discovery, Analytical Models

Analytical Processes, Collaboration, Adaptive Dynamic

Business Vocabulary (Meta Data)

Data Integration Platform

16

© 2015 S.A.R.L. Martin

Figure 16: Reference architecture for performance management and analytics. Joint modeling of processes, metrics, and governance as well as the top down implementation of metrics by analytic services, bottom up by "data discovery", and deriving analytical models is the most critical success factor. A data integration platform is the foundation for performance management and analytics. It provides parallel and simultaneous access to internal and external operational and analytic data via services within the framework of the SOA. In this model, the traditional Data Warehouse becomes a component of the data integration platform, and the business vocabulary provides the "single point of truth" which is no more the Data Warehouse (cf. chapter 6).

Performance Management and Analytics need information management – Traditional BI tools worked exclusively on the data warehouse. The data warehouse provided the "single point of truth", i.e. reliable and high quality information. This prohibited the application of BI to operational environments. BI was restricted to strategic and tactical analysis. The potential value of real-time analytics was discarded. But, performance management and analytics should also have access to operational data for operational, "just in time" (real-time) monitoring and controlling. A data integration platform (in the context of SOA also called "enterprise service data bus") now delivers the single point of truth by information management. In SOA, it links performance management and analytics to both, operational databases and the data warehouse. The data integration platform provides information services (fig. 17) that are composed of any operational and data warehouse data. In SOA, the data warehouse now acts as a backend service (fig. 7) providing in particular historical data. Analytical applications running on certain NoSQL data bases that simultaneously manage operational and analytical data, offer alternative solutions. This will be followed up in chapters 7.3 to 7.5. For more details on information management, we refer to chapter 6.

Data Discovery – analytic processes and collaboration. Data Discovery (formerly called data exploration) is an ad-hoc, dynamical, easy to handle, analytic, collaborative process. The goal is to provide new analytics, e.g., profiles, predictive and prescriptive models, rules, scores, and segmentation for a better insight into markets, customers, risks etc. In this sense, data discovery is a bottom up development environment for metrics as well as predictive and prescriptive models. For more details on data discovery, we refer to chapter 5.2. A typical example for data discovery is the development of predictive models by data or text mining, text analytics, or traditional statistical methods. The final predictive model is then implemented in a rules engine controlling an operational process.

> ***Example.*** Let us consider the process of credit approval in banking. Standard rules for checking a customer situation for solvency and credit approval can be rather easily modeled by a financial consultant. This top down model can be complemented by a bottom up model describing the risk of credit failure. This can be identified by data mining customer data and providing a risk based customer segmentation. A combination of the expert rules and the generated predictive model provides the final rules. The process of credit approval can now be automated, its workflow is controlled by a rules engine, and customers can now run credit approval as a self-service on a web site, for instance. And if a customer does not agree with automatically generated credit approvals, an embedded alert function could invoke the intervention of a human consultant within such a credit approval process.

Other examples can be found in the context of cross/up-selling and customer attrition. It is important to note that special knowledge how this intelligence works is not necessary when working in the context of intelligent processes. Analytics embeds intelligence into the process, and works as a black box. So, Analytics including even sophisticated approaches like data mining, text mining and web mining is made consumable for everybody, not only for some thousands of specialists, but for millions and more information consumers.

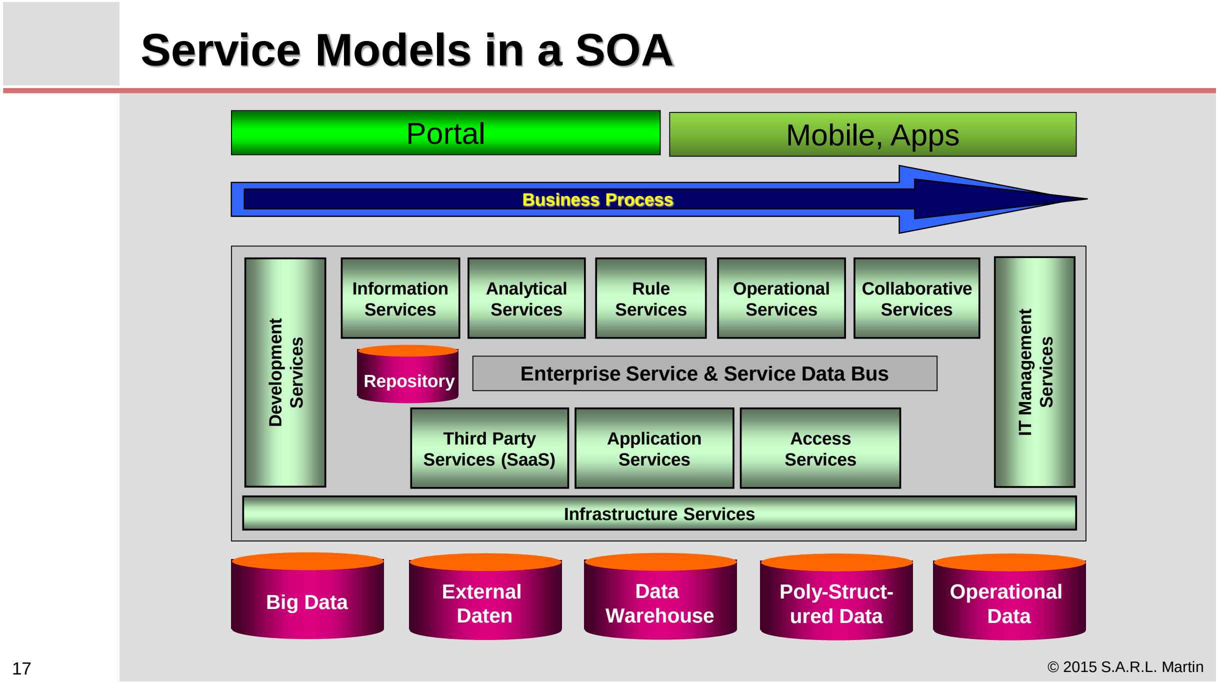
Analytics enables intelligent (“smart”) processes. Operational processes can now be enriched by embedded analytics and can be monitored and controlled “in real-time”. Service-orientation eases the embedding of analytics. Traditional BI tools metamorphose into analytic services. In a SOA, performance management is implemented through analytic services (cf. fig. 17).

4.3 Performance Management and Analytics in a SOA

We have already introduced the concept of a SOA. Cross departmental and cross enterprise processes are based on a SOA and are implemented as **Composite Applications** (or: **Business Mash Ups**) – see also Martin (2008) – orchestrating and composing business services according to the process logic. Business services present and publish the business logic from existing back-end systems (see also fig. 7), or have to be developed and/or acquired, if the necessary business logic has not yet been implemented.

There are five categories of services providing business logic (fig. 17):

- ***Operational Services.*** They provide transactional business logic like creating new customer, new account, placing an order etc.
- ***Collaborative Services.*** They provide services supporting human interactions and person to person communication like setting up a meeting, search services, communication services like embedded e-mail, chats, SMS, voice etc. Web 2.0 tools can be as well embedded, since they are typically service oriented and provide Web services.
- ***Rule-Services.*** Rules define the decision logic. Now, let us recall: A process typically uses several rules, whereas a rule can be used in several processes. This is why we have to strictly separate process and decision logic. In a SOA, rules are considered as rule services that are orchestrated by the process engine as any other service. A rule service can be also understood as an encapsulation of complex rules. Indeed, a rule service could use another rule service as a sub-rule.



Service Models in a SOA

Figure 17: Business processes orchestrate and choreograph services within a SOA. The main idea of service orientation is to split process and business logic. There are five SOA service models providing business logic, information, analytic, rule, operational, and collaborative services. These service models present the "business services". They are composed out of "technical services" provided by 3rd parties (for example out of the cloud via SaaS – software as a service), backend applications, and the various types of data sources. Furthermore, we need development services for both, process logic and business logic, and IT Management Services for administration, execution and security of services. The Enterprise Service Bus together with the Enterprise Service Data Bus is a kind of intelligent middleware enabling service and data brokerage. It also includes the repository as a service registry listing and publishing all available services via service catalogues. The user interface to processes or services consists of either a traditional portal solution that should today include social media functionality, or an interface to mobile devices that is typically implemented via apps.

- ***Analytic Services.*** They provide analytic business logic like a threshold for *product availability*, a predictive model for customer behavior or customer risk, a forecasting service for sales, etc.
- ***Information and Data Services.*** They provide composite information based on structured and poly-structured, operational and analytic, internal and external data sources like customer address, customer value, term of delivery etc. Information and data services also include meta data and master data services.

The term poly-structured data is just replacing the old term “unstructured“, because the term “poly-structured“ already indicates that “unstructured“ data may contain some hidden patterns that provide certain structure information in the end. “Poly-structured” also allows to summarize all data types, structured and unstructured. The objective is to use facts and information provided by poly-structured data, in order to enrich and to expand the traditional analysis of structured data.

.

Definition: **Poly-structured data** denominates data with unknown, insufficiently defined or multiple schemas, for instance machine-generated event data, sensor data, system log data, internal/external Web Content inclusive social media data, text and documents, multi-media data like audio, video etc.

In this white paper, we now focus on analytic services (chapter 5.1) and information and data services (chapter 6). Before doing so, let us make another note to rule services. They can also be used to automate human decision making by using rules engine as a decision engine. A decision engine should have scheduling features for follow up of events by intervening actions. For instance, if a customer has visited a web site, given a positive response, but did not come back within a certain amount of time, then the decision engine should be able to detect this “non-event”, and send a trigger, for example to a call center agent for follow up. Decision engines enable intelligent interactions with all business constituents. For example, they can enable intelligent real-time interactions with customers in the web or call center channel. In cross/up-selling, decision engines execute predictive models reflecting customer behavior. The right customer gets the right offer in right time. This boosts revenues as various business cases have shown.

4.4 Performance Management and Analytics meet Enterprise 2.0 (Social Business)

The term Enterprise 2.0 was created by Andrew McAffee[22] in 2006. He also gave a short, but comprehensive definition.

[22] See „Enterprise 2.0 – The Dawn of Emergent Collaboration” http://adamkcarson.files.wordpress.com/2006/12/enterprise_20_-_the_dawn_of_emergent_collaboration_by_andrew_mcafee.pdf

Definition Andrew McAffee: Enterprise 2.0 means the use of social media concepts by social software platforms in an enterprise or between enterprises and its customers and partners.

Today, the term "enterprise 2.0" has been replaced by "social business", but in this text, we shall still stick to enterprise 2.0. Enterprise 2.0 is all about use and use patterns of social media technology in an enterprise, this is more than just using social media technologies per se. To some extent, enterprise 2.0 requires considerable changes in enterprise culture, especially in enterprise communications. Utilization of social software platforms implies collaborative working, maintaining, and applying social computing while safeguarding the existing organizational and technological structures.

Before going further, let us first recall how social media technologies evolved. It all started with the term Web 2.0.

Web 2.0 concepts. Web 2.0 began as a social initiative for using the www. Everybody participates and contributes. A consumer turns into a producer. But the Web 2.0 concepts reach beyond the communication people to people practiced in YouTube, Twitter or Facebook. Let us have another look at Web 2.0.

Tim O'Reilly is recognized as one of the fathers of the term Web 2.0. He wrote his fundamental article "What is Web 2.0" [23] in 2005. Let us take his article as a point of start and let us look beyond. This will give us some insight into common points of BI and Enterprise 2.0.

- ***From Applications to Services.*** It simply means to abandon monolithic applications and to move to a service orientation. The Web 2.0 idea is, to make up services for mash ups for all consumers in the www. Via mash ups, the information consumer turns into an information producer, and he/she is integrated into the life cycle management of analytic processes and services in a targeted and controlled way. In the domain of BI, there are the so called CPM Toolkits that support mash ups, for instance (cf. chapter 10.2).

  > ***Example.*** The result of a data analysis may be a certain, interesting cluster, but mapping the data would provide a much better comprehension and interpretation of the detected clusters. If mapping requires an application to be programmed by IT, then mapping is either not possible, because IT does not have the resources for such a development, or it is too expensive, or it takes too much time, or even all of that! But if the information consumer can act as a producer and brings the data himself into a mapping system, then visualization is quick, simple, and even not too expensive. Self-service eases the analysis.

  The principle means full flexibility. This is exactly one of the drivers of SOA as we have shown in chapter 2.6. In this sense, this web 2.0 concept is also a SOA concept: collaborative services are linked to analytic services.

[23] See http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html?page=1

- ***Architecture for collaboration.*** This Web 2.0 concept can also be mapped to SOA. The basic concept of SOA is a service provisioning and consuming model controlled by service level agreements. This is a collaboration model par excellence.
- ***Using collective intelligence.*** In the original text, Tim O'Reilly thinks of open source concepts. But let us think a bit further and let us apply the basic principle of "everybody contributes and produces" to the collaboration of specialist departments and IT, then users become developers. Do you believe that? Indeed, in selected areas and domains this is already practiced. Here, Business Intelligence even has a certain pioneering task. Users should create their own reports without IT that was already a dream in the 80s/90s. The tools at that time were insufficient, but today, the Web 2.0 principle of "trust in users as co-developers" looks interesting. Again, we end up with mash ups, but here, we can do even more. It is about team working. BI is used for better decision making, but decisions are set up in teams. The idea is a "board room style of decision making". So, the chance of a new level of collaboration between business people and IT is opened: **Self-Service BI**, we already discussed in chapter 3.4.
- ***Simple interfaces and models.*** Tim O'Reilly speaks of light weight programming. Such a "keep it simple" is a good approach for designing and implementing services in a SOA. Simple and standardized interfaces are provided by Web Services, a today widely accepted standard. This makes it easy to embed analytics into business processes and fosters the enrichment of processes by intelligence.
- ***Multimedia software.*** That fits perfectly a SOA. SOA enables the convergence of various heterogeneous IT disciplines like business intelligence, business process management, document management, office etc. This is in particular beneficial for BI because BI should not only deal with structured data, but also with structured data. Indeed, the combined analysis of structured and poly-structured data provides new and not yet known insights. In chapter 5.7, we will discuss text analytics and give more details into these new opportunities.
- ***The long tail impact.*** One of the most interesting Web 2.0 features is reaching the "long tail" via communities. This principle can be also found in the concepts of SOA, when the concept of "SaaS" (Software as a Service) or apps is linked to SOA. The long tail impact can now be understood as the business opportunity to identify and use SaaS services or apps as complements to internal services: It is simply easier to detect such services in the web or in an app shop.
- To summarize, Web 2.0 concepts empower a new style of collaboration and communication by providing far reaching collaborative services in a SOA (cf. Fig. 17).

Enterprise 2.0 meets Performance Management and Analytics. As discussed in chapter 2.6 (cf. fig. 6), a successful organization is based on the principles of industrialization, agility, compliance, and smartness. We also identified the contradictory objectives of industrialization and agility. Industrialization means automaton and standardization, while agility stands for creativity and innovation. We have already seen that these conflicting goals can be harmonized by SOA. We had a discussion on a technical level, but now applying Web 2.0 concepts as collaborative services, we add human and social aspects.

Example: Deployment of an online newspaper as web based intelligent reporting (Quelle GmbH, a German mail order house, winner of the BI Award 2008, CeBIT 2008). The jury especially appreciated the new and innovative approach for knowledge management based on Web 2.0 elements. Quelle created a user friendly and innovative solution by integrating pictures, text elements and classical performance metrics. The use of a wiki for the definitions and documentation of performance metrics and for best practices in interpreting reports and analysis contributed well to a high user acceptance. The Quelle solution provided not only numbers, but transparency of content and usage of reporting for all information consumers (see: *is report* 2008).

The Quelle example shows nicely the way to go with performance management and analytics in the context of Enterprise 2.0. In chapter 2.1, we already discussed the pitfalls of traditional BI. In Enterprise 2.0, we can tackle and solve these challenges in a different way by applying Enterprise 2.0 principles:

- All depends on people! Everybody in the enterprise should contribute. Consumers should be turned into producers. The Quelle example also showed us that a moderated communication could be very helpful.
- Active participation and contribution to problem solutions should be honored. Intensive collaboration could be honored by a bonus or by the rating of content by the community. Here, we also could apply concepts of **gamification**. Gamification means the application of game theory concepts and techniques to non-game activities, i.e. to real life situations. The goal of gamification is to engage the participant with an activity he finds fun in order to influence his behavior.
- A critical mass of information and consumers/producers is a critical success factor. It eliminates initial blockers. So, traditional applications should be transferred into the Enterprise 2.0 environment. Quelle succeeded in doing so by moving its reporting system to the online newspaper.
- Personal reputation of an employee is a strong motivation for an active engagement. Anonymous contributions should not be admitted.
- Management should trust in its employees. Control taken by the community should not be underestimated. This is another critical success factor for Web 2.0. Enterprise culture must invite for frankness and openness.
- A simple and easy to be used user interface empowers all users whatever their expertise is. Quelle succeeded in doing so by its intuitive visualization.

Take Away. Enterprises should consider social media concepts and technologies and how they could be used for building the enterprise 2.0. Enterprise 2.0 is a way of marrying industrialization with creativity and innovation. In particular, enterprise 2.0 can stimulate a better usage of performance management and analytics by linking them to knowledge management. The result will be a dramatical improvement of its acceptance.

4.5 Planning in the digital enterprise

In addition to monitoring and controlling of processes, planning is the third pillar of performance management. Indeed, planning and budgeting are core processes of corporate management and controlling. They consistently serve business critical goals and objectives: Strategies have to be implemented via business models, processes, and actions, finance and cash flows to be optimized, and decisive enterprise variables like revenues, cost, contribution ratios, and gains to be controlled across enterprise divisions. Professional planning is one of the critical success factors of enterprises.

But in many corporations, planning is not a best practice – to say the least. Since years, concepts like driver-based planning or zero budgeting are discussed in various papers, but actual studies show that such methods are not at all common practice or even not adopted. Very often, the planning process is fragmented, and consequently long term goals cannot be aligned with short term decisions and actions. The closed-loop model of performance management (cf. fig. 9) is broken. Furthermore, there is a big lack of integrating various operational budgets. As a result, planning silos pop up. In many planning circles, for example, sales is planned in a completely isolated manor causing that profit planning is without any solid baseline.

Imperatively, corporations have to do much better. An important step for improving planning processes and towards professional planning is the use of dedicated, special planning tools. Such tools remove many of the known problems and provide an end-to-end support of planning processes. They accelerate planning (too slow planning processes is indeed another source of trouble), and they also secure planning in contrast to manual, Excel-based planning. The use of such planning tools is a right step into the right direction, but not yet sufficient. There are additional challenges.

In the digital world, planning is much more difficult and more complex than ever, because predictable and long lasting trends do not exist anymore. Even if today business is running well, it could abruptly stop and break tomorrow. In other words, how should we and how can we plan in such a volatile world? Traditional yearly planning nearly completely fails in the digital world. Who knows how markets and customers will act and what their needs will be in a year’s time? Traditional planning habits and methods do not help anymore. We must reinvent planning, and move to agile planning. Agility as a key success factor for businesses in the digital world is about all: flexible and quick reactions to changing market and customer demand, quick and flexible changes to processes and services, if strategy has to change, as well as flexible, all-embracing and farseeing planning.

Agile planning imposes new thinking. In the past, in most enterprises, planning was restricted to financial planning. In some cases, financial planning was complemented by strategic planning. But only leading enterprises had an integrated strategic and financial planning. The rule was to have at best two isolated planning silos, and no integrated operational planning at all. Now, agile planning means an integrated, end-to-end planning process. The gap between isolated planning silos for resources, programs, products, processes, services etc. is to be closed, since strategic, operational, and financial planning are mutually dependent.

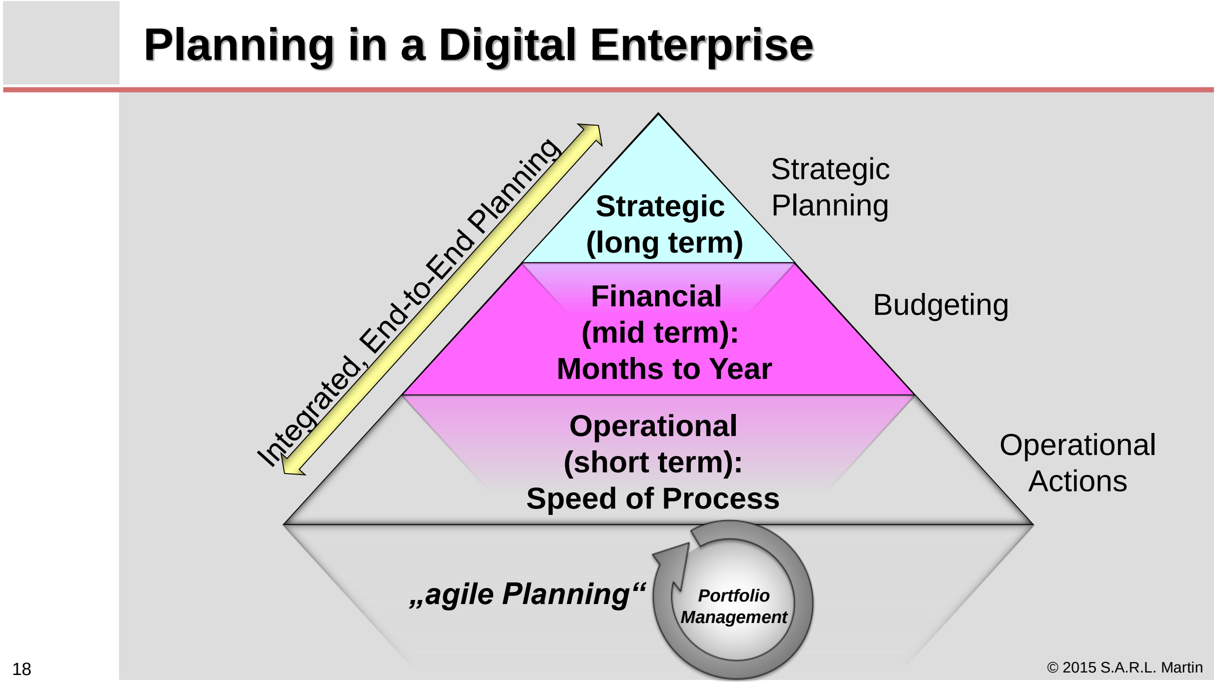
Agile planning means even more. Planning periods have to be shortened. Planning methods like rolling, driver-based, or top down/bottom up planning now come into play. As we have already seen, these methods are not really new, but in the digital world, they finally come into focus of many enterprises. Such a planning is completed by forecast scenarios. One single plan is no longer sufficient to encounter the volatility of the markets. Due to volatility, changes to market behavior and direction can abruptly happen. As a result, even the best plan becomes obsolete. Therefore, various scenarios are required assuming different market evolutions under different influencing parameters, and a plan can be derived for each of the scenarios. Forward thinking planners go even one step further. They use simulations based on mathematical models for evaluating the various scenarios and to balance decisions. A planning horizon is established for driving the enterprise in a more comprehensive way and enabling fast reactions when changes occur.

Agile planning also means a more intense interaction and communication between all participants in the planning processes. By definition, planning is a highly collaborative process. It is driven by people: Everybody has to be involved and integrated. As a consequence, leading planning tools come with social media style concepts and tools. The use of wikis acting as knowledge base and help functions is wide spread. It helps to create a common language between controllers and other constituents of the planning processes. For instance, the gap between controllers and finance and accounting can be closed. This is important for the agility of financial planning: Peaks in financing have to be identified and to be bypassed. This is part of risk management.

Documentation and annotation of planning assumptions and decisions are also very important for making planning agile. This is a top requirement for today's planning tools. It enforces the participants to concentrate on the numbers and to comment deviations. A co-requirement is autonomy for all participants in planning. All basic functions must be accessible and usable by everybody. There is no agility without autonomy. All participants in planning need creation of tables and graphs, aggregation of data, application of even complex mathematics etc. In the past, this was provided by spreadsheets, but given the needs for integrated planning and collaboration, an isolated spreadsheet environment is insufficient, unfeasible, and even useless for agile planning. Furthermore, a new requirement is coming up that cannot be addressed by isolated spreadsheets: mashing up of various planning services for building new planning processes on the fly.

To conclude, agile planning consists of new planning methods, social media style communication and collaboration in planning, and the integration of strategic, operational and financial planning. The real challenge is to close the gap between strategic and operational planning. This requires the dissolution of existing planning silos and the migration into an integrated planning platform, where methods and tools of strategic and financial planning are consolidated and merged with those of operational panning. This also requires a new generation of planning tools, in the end a planning platform. Such platforms do exist in the meantime, but unfortunately, they are not yet mainstream.

Portfolio Management now becomes a key element of planning. It is a well-qualified approach to optimize business by balancing strategy against scarce resources – people, assets and money. To do so, we combine strategic and financial planning with investment analysis, and capacity, demand, program, resource, and change management. In this way, we create integrated planning. Operational planning of business processes now is a direct consequence of strategic and financial planning. We close the planning loop in the same way as we have closed the control loop in performance management (fig. 18).



Planning in a Digital Enterprise

Integrated, End-to-End Planning

Strategic (long term)

Financial (mid term): Months to Year

Operational (short term): Speed of Process

Strategic Planning

Budgeting

Operational Actions

„agile Planning“

Portfolio Management

18

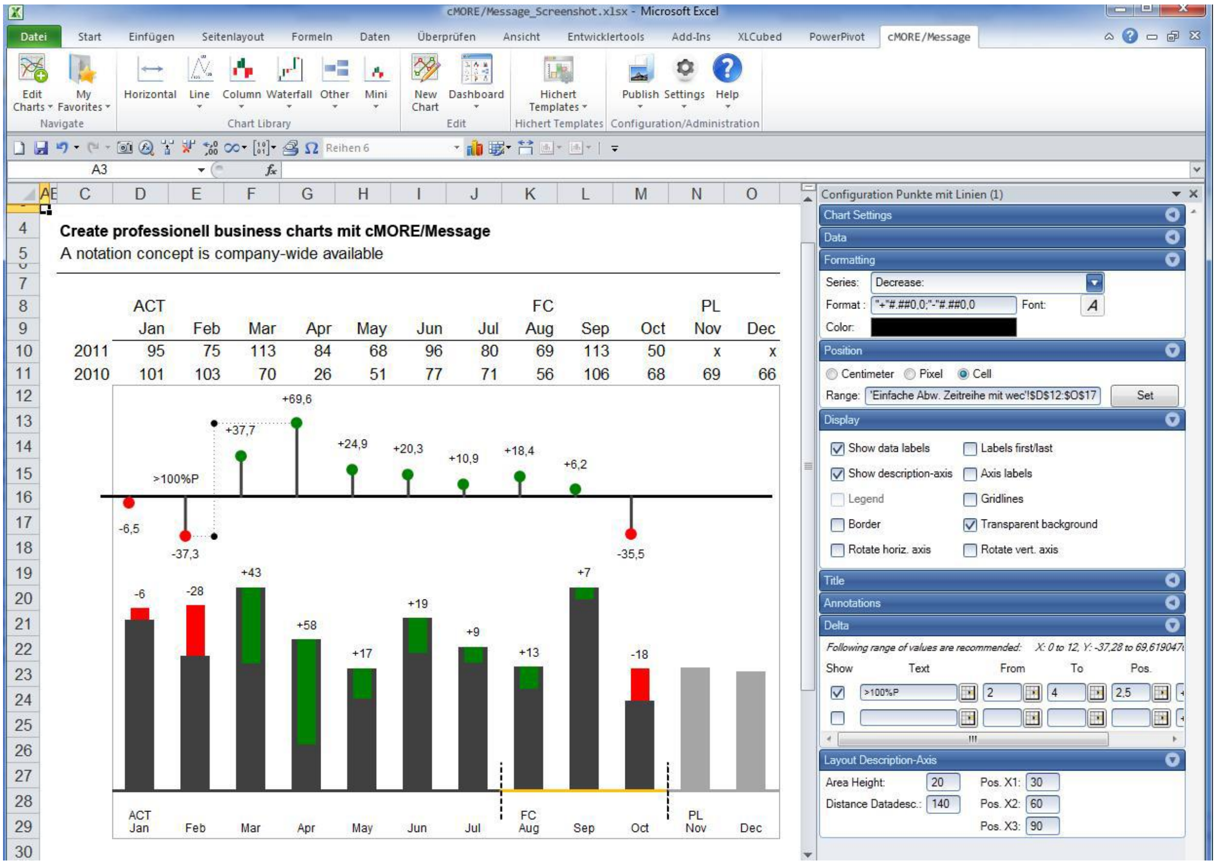
© 2015 S.A.R.L. Martin

Figure 18: Market conditions of the digital world require new, alternative approaches to planning (Yearly planning cycles are no more valid!). Collaborative, social media style planning tools are now a must. Integrated planning is absolutely necessary: The gap between strategic and financial planning on the one hand and operational planning on the other hand must be closed. Portfolio management becomes a key element in integrated planning. We are moving towards agile planning.

5 Analytics – the Foundation of Performance Management

Analytics can be classified into three succeeding types of tools and methods:

- Basic analytics. The question is: *What happened?* This corresponds mainly to traditional BI: Canned batch reports and interactive drill down/up reports (fig. 19), fixed analytic dashboards (may or may not be event driven), and BI automation (alerts and recommendations).



Create professionell business charts mit cMORE/Message

A notation concept is company-wide available

	ACT							FC			PL	
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2011	95	75	113	84	68	96	80	69	113	50	x	x
2010	101	103	70	26	51	77	71	56	106	68	69	66

Configuration Punkte mit Linien (1)

Chart Settings

Data

Formatting

Series: Decrease:

Format : "+"#.##0,0;"-"#.##0,0 Font: A

Color: ■

Position

○ Centimeter ○ Pixel ● Cell

Range: 'Einfache Abw. Zeitreihe mit wec'!$D$12:$O$17 Set

Display

☑ Show data labels ☐ Labels first/last

☑ Show description-axis ☐ Axis labels

☐ Legend ☐ Gridlines

☐ Border ☑ Transparent background

☐ Rotate horiz. axis ☐ Rotate vert. axis

Title

Annotations

Delta

Following range of values are recommended: X: 0 to 12, Y: -37,28 to 69,619047

Show	Text	From	To	Pos.
☑	>100%P	2	4	2,5
☐				

Layout Description-Axis

Area Height: 20 Pos. X1: 30

Distance Datadesc.: 140 Pos. X2: 60

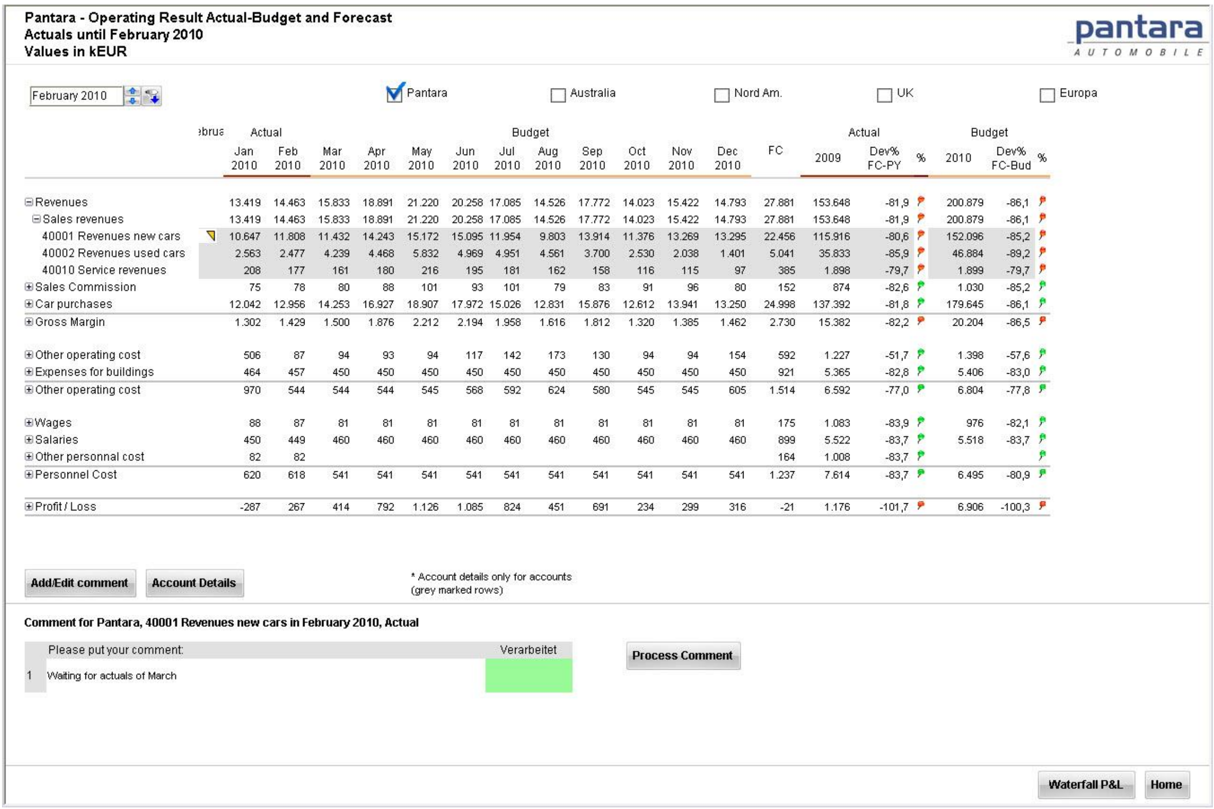
Pos. X3: 90

Figure 19: Basic analytics: The most common traditional BI tool is Excel. But, if Excel is used for an individual data management, then the risk is to end-up in an inconsistent BI landscape. But when Excel is used as a front-end only, it can be applied as a service, and it excels by remarkable user friendliness, as well as by a considerable number of add-in offerings that enhance the functionality rather slickly. Here, as an example, the Excel add-in cMORE/Message from pmOne. It creates with a few mouse clicks meaningful business charts from a data table (at the top) integrating several types of data (actual, budget and forecast data). In addition the deviation tree displays easy-to-understand the percentaged difference between actual data and forecast. The principles of notation comply with the SUCCESS rules stated by Dr. Hichert[24].

[24] For Dr. Hichert's success rules, please see http://www.hichert.com/en/success

- Standard analytics. Here, we want to explore: *Why did it happen?* This corresponds still to traditional BI like interactive analytic dashboards with drill up/down, slice/dice, etc., OLAP, BI automation (decision analysis workflows), and customizable data mash ups and BI widgets
- Advanced analytics. Now we want to *predict what may happen/investigate new opportunities.* Here we use data and text mining, text analytics, location intelligence, analytic modeling and statistical/predictive analytic functions, as well as advanced visualization. These tools are mainly used in data discovery as well predictive and prescriptive analytics.

5.1 Analytic Services



Pantara - Operating Result Actual-Budget and Forecast
Actuals until February 2010
Values in kEUR

February 2010

☑ Pantara □ Australia □ Nord Am. □ UK □ Europa

	ebrua	Actual		Budget											Actual			Budget		
		Jan 2010	Feb 2010	Mar 2010	Apr 2010	May 2010	Jun 2010	Jul 2010	Aug 2010	Sep 2010	Oct 2010	Nov 2010	Dec 2010	FC	2009	Dev% FC-PY	%	2010	Dev% FC-Bud	%
⊟ Revenues		13.419	14.463	15.833	18.891	21.220	20.258	17.085	14.526	17.772	14.023	15.422	14.793	27.881	153.648	-81,9	🟥	200.879	-86,1	🟥
⊟ Sales revenues		13.419	14.463	15.833	18.891	21.220	20.258	17.085	14.526	17.772	14.023	15.422	14.793	27.881	153.648	-81,9	🟥	200.879	-86,1	🟥
40001 Revenues new cars	◥	10.647	11.808	11.432	14.243	15.172	15.095	11.954	9.803	13.914	11.376	13.269	13.295	22.456	115.916	-80,6	🟥	152.096	-85,2	🟥
40002 Revenues used cars		2.563	2.477	4.239	4.468	5.832	4.969	4.951	4.561	3.700	2.530	2.038	1.401	5.041	35.833	-85,9	🟥	46.884	-89,2	🟥
40010 Service revenues		208	177	161	180	216	195	181	162	158	116	115	97	385	1.898	-79,7	🟥	1.899	-79,7	🟥
⊞ Sales Commission		75	78	80	88	101	93	101	79	83	91	96	80	152	874	-82,6	🟩	1.030	-85,2	🟩
⊞ Car purchases		12.042	12.956	14.253	16.927	18.907	17.972	15.026	12.831	15.876	12.612	13.941	13.250	24.998	137.392	-81,8	🟩	179.645	-86,1	🟩
⊞ Gross Margin		1.302	1.429	1.500	1.876	2.212	2.194	1.958	1.616	1.812	1.320	1.385	1.462	2.730	15.382	-82,2	🟥	20.204	-86,5	🟥
⊞ Other operating cost		506	87	94	93	94	117	142	173	130	94	94	154	592	1.227	-51,7	🟩	1.398	-57,6	🟩
⊞ Expenses for buildings		464	457	450	450	450	450	450	450	450	450	450	450	921	5.365	-82,8	🟩	5.406	-83,0	🟩
⊞ Other operating cost		970	544	544	544	545	568	592	624	580	545	545	605	1.514	6.592	-77,0	🟩	6.804	-77,8	🟩
⊞ Wages		88	87	81	81	81	81	81	81	81	81	81	81	175	1.083	-83,9	🟩	976	-82,1	🟩
⊞ Salaries		450	449	460	460	460	460	460	460	460	460	460	460	899	5.522	-83,7	🟩	5.518	-83,7	🟩
⊞ Other personnal cost		82	82											164	1.008	-83,7	🟩			🟩
⊞ Personnel Cost		620	618	541	541	541	541	541	541	541	541	541	541	1.237	7.614	-83,7	🟩	6.495	-80,9	🟩
⊞ Profit / Loss		-287	267	414	792	1.126	1.085	824	451	691	234	299	316	-21	1.176	-101,7	🟥	6.906	-100,3	🟥

* Account details only for accounts (grey marked rows)

Comment for Pantara, 40001 Revenues new cars in February 2010, Actual

	Please put your comment:	Verarbeitet
1	Waiting for actuals of March	🟩

Figure 20: Planning and simulation: Cubeware Cockpit V6pro supports forecasting, what-if simulations and multiple planning options including top down budgeting, a bottom up approach or a mixture of the two. Users can easily incorporate the latest actual data and assign budget values using different allocation options such as percent/absolute increases or reference values. A wide range of visualization options – such as the columns representing the difference between top down and bottom up planning results as shown in this screenshot – offer added support.

In a SOA as infrastructure for BPM, embedded analytics is implemented via analytic services. Analytic services are encapsulated, component based modules that can, but need not communicate via services (see fig. 17). They publish business logic. This includes all kinds of analytic content and functionality, e.g., customizable and extensible templates for metrics in business scorecards and analytic tools. It also includes all necessary development services for managing the life cycle of analytic services (implementing, customizing, maintaining). This is why analytics extends the traditional data warehousing centric BI. It puts intelligence into the context of strategy, goals and objectives, governance, processes, and people via metrics and predictive models, and it implements intelligence through analytic services.

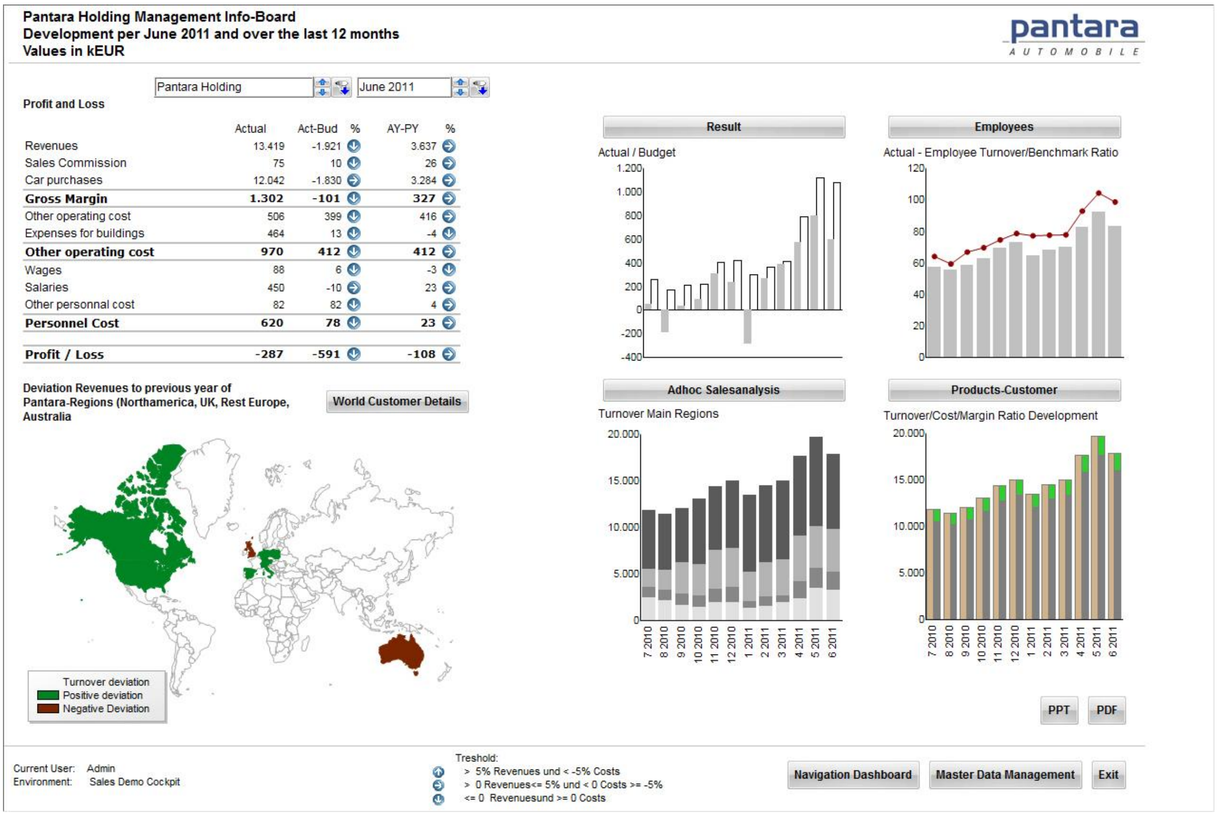
***Reporting and Analysis Services* –** Reports can be classified into three groups: management reports (strategic level) for visualizing the performance of an enterprise, business reports for periodically reporting the performance of sales, marketing, production, logistics etc., and operational reports for ad-hoc and periodically status reports in warehousing, spending, outstanding items etc. Typically, management reports have the smallest number of users, whereas operational reports have the largest number. This puts technical constraints to the performance and scalability of reporting tools as well as business requirements for the functionality of the tool like visualization, design, distribution and collaboration features.

Furthermore, the different groups of reports also have different producers. Management reports and to a certain extend also business reports, are produced by business users. Therefore, they need a self-service approach to reporting as part of a **self-service BI** (cf. chapter 3.4), an important requirement when selecting reporting tools: It should be easy and intuitive to use the tools. The tools should provide a good visualization and include lots of pre-manufactured components that can easily be mashed up. In contrast, IT or the BI competency center is typically in charge of provisioning operational reports. Finally, the BI competency center is also responsible for the support of all users of the self-service reporting tools.

In a SOA, the functionality of basic and standard analytics is implemented through components providing analytic services that can be embedded in any process. Thus, business processes can be enriched by analytic services. These services can use any information service for data supply so that these services can now act on composite data out of analytic and operational data sources. Analytics now goes real-time whenever relevant for the business. Furthermore, composite reporting and analytic services can be created by mashing up. Indeed, this is one of the technical requirements for self-service BI. The pre-manufactured components of self-service reporting tools should be analytical services.

***Planning and Simulation Services* –** Planning is a typical cross-departmental process that is best implemented as a SOA based process. So, planning functionality is implemented as planning and simulation services providing full flexibility and adaptability of this process to changing business scenarios. The advantage of implementing planning through a SOA is obvious, the planning process can be composed out of any analytic and other services avoiding the redundancy in analytic functionality by implementing a planning application in a traditional data warehouse/ business intelligence architecture and by fostering a rigorous and audit-proof planning by a controlled process instead of spreadsheet based manually driven planning processes (fig. 20). SOA based planning is therefore the best technological prerequisite for planning in a digital enterprise (cf. chapter 4.5).

***Dash-Board Services* –** A dashboard (fig. 21) visualizes large volumes of information and data from various data sources in an aggregated way. Degree of aggregation as well as type and style of visualization correspond to goals and profile of a user. For instance, different metrics from different reports can be jointly depicted and be compared. Dashboards can also be used to visualize business scorecards (cf. chapter 3.3). It is either embedded in a portal as a portlet or displayed on mobile devices. The information profile of the business scorecard user describes the personalization of dashboards so that each user gets exactly the right information according to the process owner model according to the BI governance. Traditional deployment of dashboards is passive: The information consumer uses search and navigation services to access its metrics and is guided by an analytic workflow. Today, state of the art deployment of dashboards is active: In case of escalation, events or alerts important information is automatically sent to the information consumer by special channels, e.g. RSS feeds, instant message, e-mail etc. for triggering decisions and actions. This enables **management by exception**.



Pantara Holding Management Info-Board
Development per June 2011 and over the last 12 months
Values in kEUR

Pantara Holding | June 2011

Profit and Loss

	Actual	Act-Bud	%	AY-PY	%
Revenues	13.419	-1.921	↓	3.637	→
Sales Commission	75	10	↓	26	→
Car purchases	12.042	-1.830	→	3.284	→
Gross Margin	**1.302**	**-101**	↓	**327**	→
Other operating cost	506	399	↓	416	→
Expenses for buildings	464	13	↓	-4	↓
Other operating cost	**970**	**412**	↓	**412**	→
Wages	88	6	↓	-3	↓
Salaries	450	-10	→	23	→
Other personnal cost	82	82	↓	4	→
Personnel Cost	**620**	**78**	↓	**23**	→
Profit / Loss	**-287**	**-591**	↓	**-108**	→

Deviation Revenues to previous year of Pantara-Regions (Northamerica, UK, Rest Europe, Australia

Turnover deviation
🟩 Positive deviation
🟥 Negative Deviation

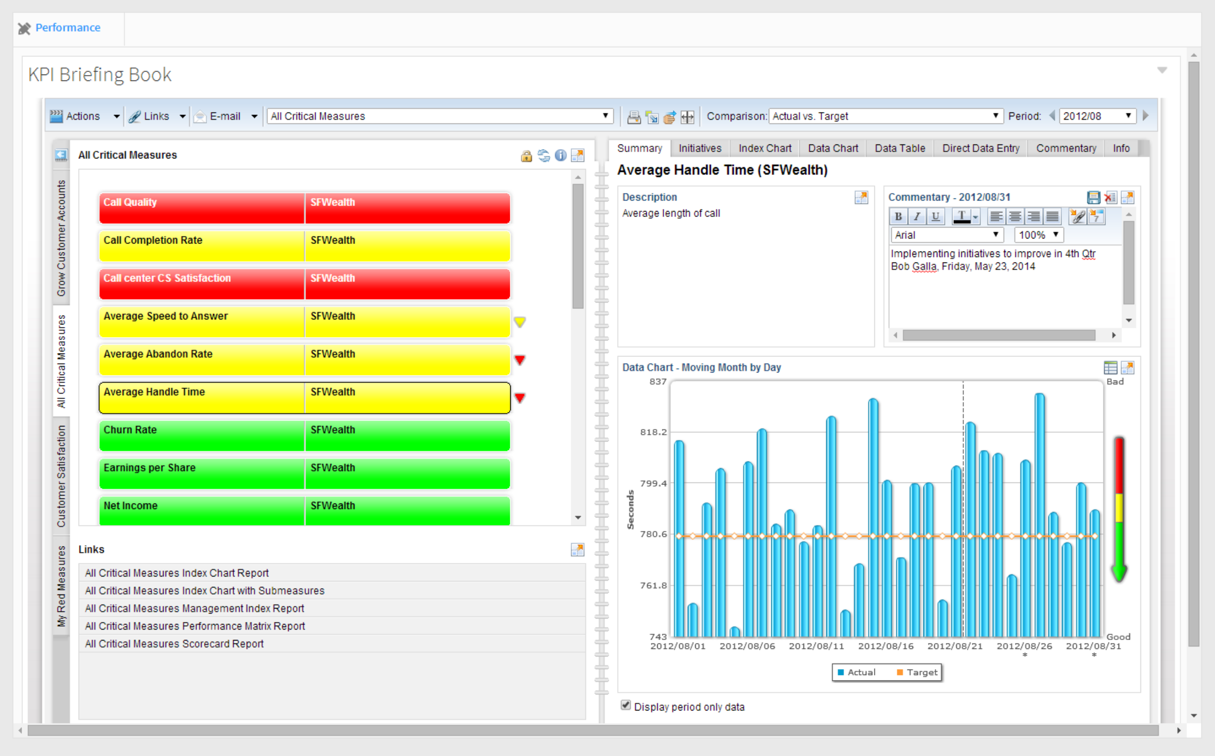
Current User: Admin
Environment: Sales Demo Cockpit

Treshold:
↑ > 5% Revenues und < -5% Costs
→ > 0 Revenues<= 5% und < 0 Costs >= -5%
↓ <= 0 Revenuesund >= 0 Costs

Figure 21: Sample dashboard created with Cubeware Cockpit V6pro: This dashboard provides an intuitive overview of the main corporate KPIs and the current development of various strategic areas. To create a similar report, you simply drag and drop various layout components onto the interface. In the screenshot above, each dashboard component is based on Data Views from different cubes. Users can create ad hoc queries or drill through to further details in the underlying data at any time. Creating and modifying dashboards is so easy that most users do all the work themselves.

"**Briefing books**" are a flexible enhancement of the reporting capabilities of dash boards. A briefing book consists of chapters. These chapters allow structuring the profile of an information consumer by assigning its analytic services to the defined chapters. A briefing book is a complete structured set of analytics that provides all information according the governance model in an intuitive, interactive and visual manner (fig. 22).

Information Management – As already mentioned, traditional business intelligence tools were restricted to the data warehouse, whereas analytic services work on a data integration platform. Such a data integration platform is the foundation for performance management and analytics. On a technical level, it links performance management and analytics via an enterprise service data bus to operational databases and the data warehouse. The data warehouse now provides backend data services for analytic services that can be mashed up with operational and real-time data (fig. 7 and 17). So, data integration provides information services for analyzing data, master and meta data, develop data models, prepare and profile any type of data, as well as ETL (extraction, transformation, load) services. For more on data integration and real-time concepts we refer to chapters 6 and 7.



All Critical Measures

Comparison: Actual vs. Target

Period: 2012/08

All Critical Measures

Call Quality	SFWealth
Call Completion Rate	SFWealth
Call center CS Satisfaction	SFWealth
Average Speed to Answer	SFWealth
Average Abandon Rate	SFWealth
Average Handle Time	SFWealth
Churn Rate	SFWealth
Earnings per Share	SFWealth
Net Income	SFWealth

Links

- All Critical Measures Index Chart Report
- All Critical Measures Index Chart with Submeasures
- All Critical Measures Management Index Report
- All Critical Measures Performance Matrix Report
- All Critical Measures Scorecard Report

Average Handle Time (SFWealth)

Description

Average length of call

Commentary - 2012/08/31

Implementing initiatives to improve in 4th Qtr
Bob Galla, Friday, May 23, 2014

Data Chart - Moving Month by Day

☑ Display period only data

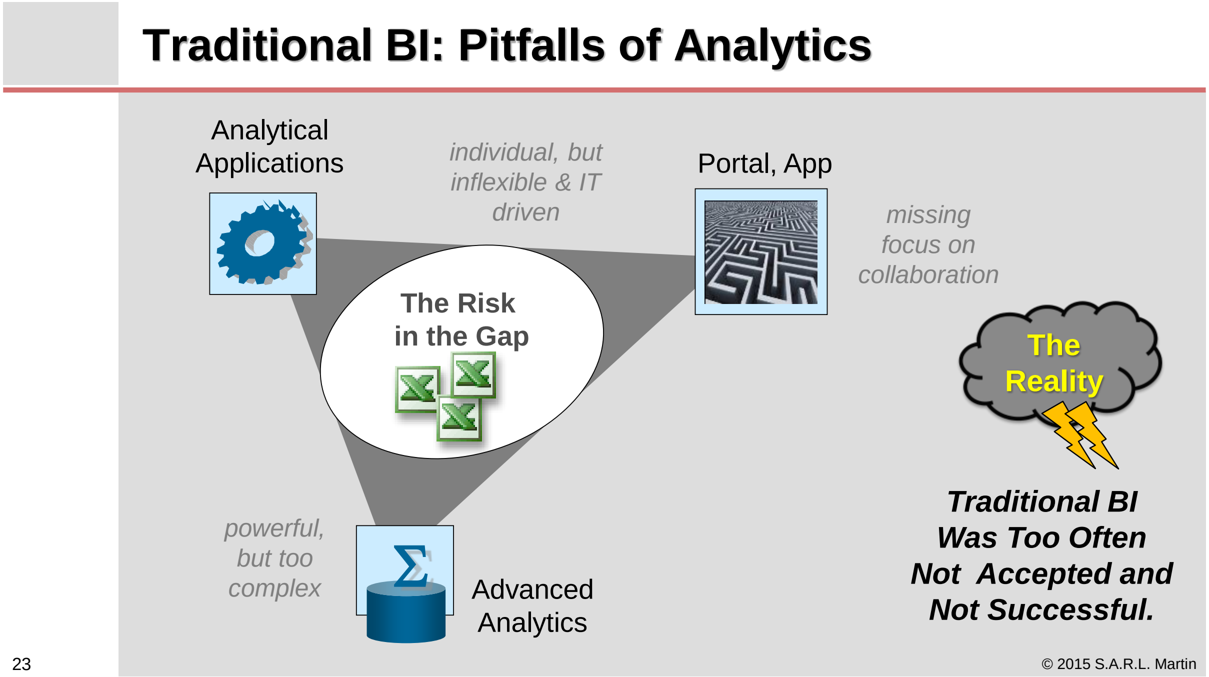
Figure 22: Example of a Briefing Book made by BIRT from Actuate. An intuitive usage is supported by the presentation as a "book". On the left hand, you see the user defined chapters that can be opened by clicking on the flags. Here, each chapter consists of subchapters depicted by the flags top right. In addition, a page flip function simulates the turning of pages.

5.2 Data Discovery

From Interactive Analytics to Data Discovery. Enriching business processes through embedded analytics is enabled by interactive analytics (data exploration). Metrics are not only derived top down from strategy, goals, and processes, but could also be derived bottom up from data. This is the purpose of interactive analytics on top of the data integration platform. We can now combine analytic and operational data if necessary. Up to now, interactive analytics used mainly traditional business intelligence tools like ad-hoc querying, OLAP statistical tools, data and text mining, text analytics etc. Especially OLAP has found a broad acceptance und proliferation.

OLAP (online analytic processing) is an analytic method enabling fast and interactive access to relevant information. It provides complex analysis functions and features based on a multi-dimensional data model. In such a data model, metrics are aligned by various dimensions, e.g. revenue in respect to customer, product, region, time period etc.

For further details, we refer to www.OLAP.com. Data and text mining as well as text analytics are discussed in chapters 5.6 and 5.7.



Traditional BI: Pitfalls of Analytics

Analytical Applications

individual, but inflexible & IT driven

Portal, App

missing focus on collaboration

The Risk in the Gap

The Reality

powerful, but too complex

Advanced Analytics

Traditional BI Was Too Often Not Accepted and Not Successful.

23

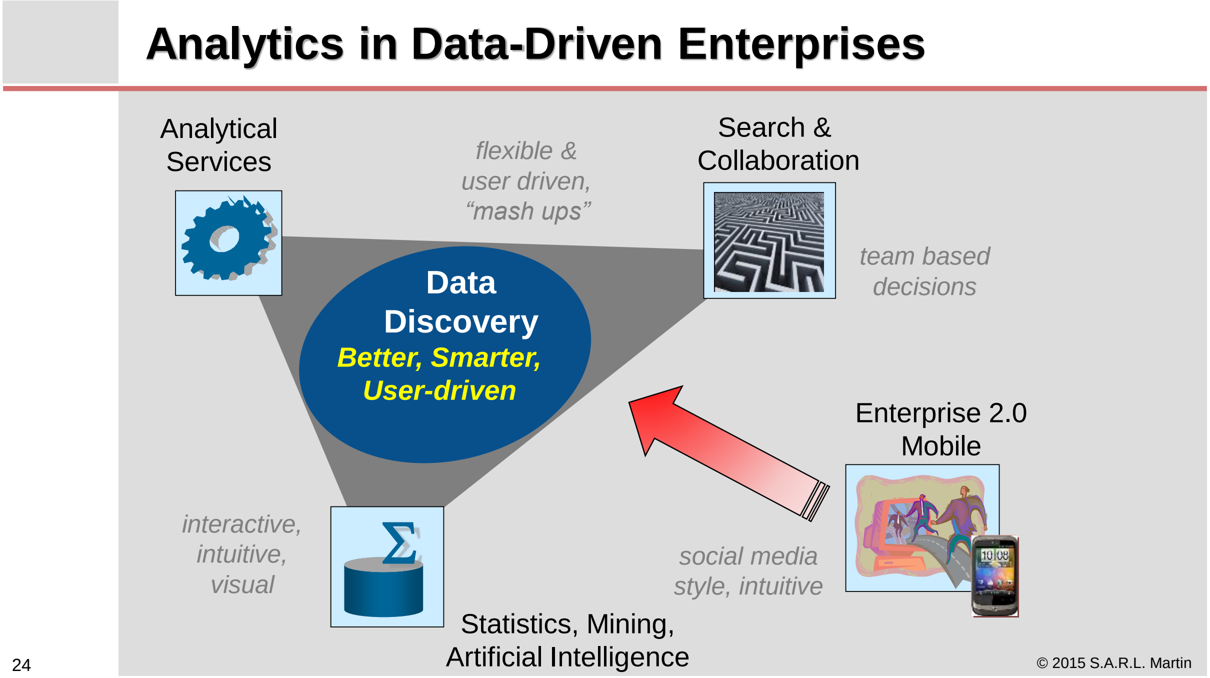
© 2015 S.A.R.L. Martin

Figure 23: Analytics with traditional BI tools did not get sufficient acceptance in business. Traditional portals lacked collaboration features and team based decision making. Traditional analytic applications were IT applications preventing specialist departments from agile actions and reactions and cemented the IT dependency. Advanced mathematical and statistical analytics was too complex for business users. Spreadsheets became the substitutes. But adoption of spreadsheets including data management in business is a high risk, because data consistency will be rapidly lost.

But the traditional BI tools for interactive analytics did not get sufficient acceptance in business. The sometimes high to very high expectations were not met at all. The tools required experts that had to be appropriately trained, because they showed all the deficiencies listed in chapter 2.1 on BI pitfalls. Consequently, despite all training efforts, they were not used. However, since certain analytics are indispensable, BI tools were substituted by spreadsheets. But, when spreadsheets with individual data management are used in business, data consistency is rapidly lost. The result is data chaos, and hence, the foundation for compliance is corrupted (fig. 23).

In the midst of the years 2000, new analytic tools came to market and have changed the game. They are in particular based on enhanced data visualization techniques and analytic workflows. They empower data discovery.

Data Discovery means a new generation of analytic tools for combining and investigating various data sources. Consciously, they do not provide predefined pathways with drill functionality, because a user should have full flexibility by interactivity. Furthermore, they excel by extraordinary user friendliness and flexibility. Furthermore, they use in-memory technologies for internal storing and processing. The big advantage of in-memory technology is performance: Therefore, Data Discovery tools are especially equipped and suited for Big Data analytics.



Analytics in Data-Driven Enterprises

Analytical Services

flexible & user driven, "mash ups"

Search & Collaboration

team based decisions

Data Discovery

Better, Smarter, User-driven

Enterprise 2.0 Mobile

interactive, intuitive, visual

social media style, intuitive

Statistics, Mining, Artificial Intelligence

24

© 2015 S.A.R.L. Martin

Figure 24: Data Discovery combines social media style collaboration tools, visualization technology and statistical methods, and is based on service oriented architecture. It follows the concepts of self-service BI empowering specialist departments. Supported by a BI competency center, business user can now perform their analytic tasks largely independent from IT. Data Discovery is also well suited for the mobile BI: Information is ubiquitous, and tablets simplify access and usage of information. The "data driven" business becomes real.

Data Discovery tools come with visualization of data, interactive intuitive analysis, as well as collaboration and autonomy of users. Interactive data visualization supports the human eye by intuition. It combines the capabilities of the human eye for pattern recognition with:

- visualization services providing different views on the data simultaneously,
- instant drill downs and dynamic queries,
- statistical methods and techniques as well as data mining,
- configurable and dynamic access to all relevant data sources beyond data marts (mashing up).

We will discuss data visualization in more detail in the following chapter 5.3.

Users of Data Discovery tools can access central data via client-server, Web browser or via apps through mobile devices like tablets. Tablets play an important role since they have been designed for intuitive usage. In the mobile internet, each given task can be addressed by a special app that is tuned to the business process it supports (see also chapter 2.7 “Mobile BI”). Since visual tools now run on tablets, they present definitely a break-through: Data discovery tools on tablets are not only accepted by the users, but even inspire users.

Data Discovery tools follow the concepts of **„self-service BI“,** we have already met in chapter 3.4 and during the discussion of reporting services in chapter 5.1. Users benefit from high autonomy, and the role of the IT evolves towards a service provider. IT deploys and operates the platform for self-service BI, as well as provisions the necessary consultancy for using the platform.

Another important feature of Data Discovery tools is a much improved team support in comparison to traditional BI tools. Applications can be exchanged with business partners via Web, e-mail or Social Media tools, and can be integrated into office or other applications. These collaborative aspects go even further: Annotations to data and results can be attached to certain views and can be shared with co-workers (fig. 24). We have already discussed this evolution in the context of BI competence centers in chapter 3.5.

***Analytics in real-time* –** This means high performance, largely scalable analytics that delivers results up to the second or even faster when analyzing huge and very huge amounts of data (Big Data). So, the issue is to accelerate analytics. There are three different principles to accelerate analytics that can also be combined.

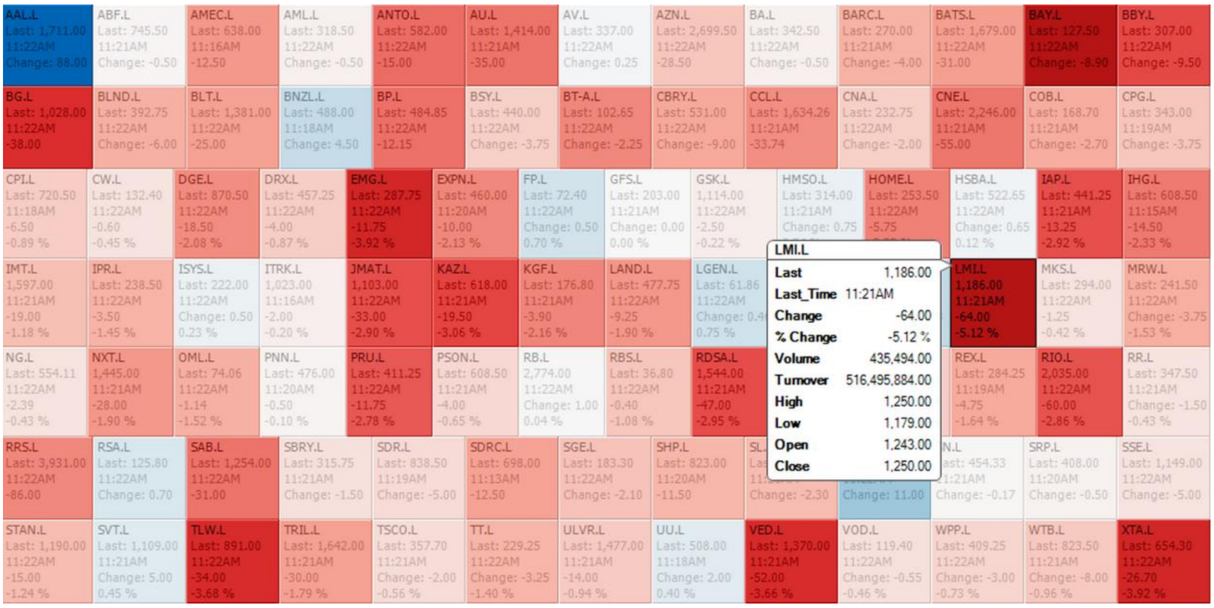
- ***Special Database Technologies* –** Special SQL technologies like compression, indexing, vector processing, memory-based caching, massive parallelization etc. or NoSQL technologies can dramatically improve the performance of data discovery tools. This speeds up the data discovery process by faster responses (from hours to minutes and seconds). Today, technologies in this category are rather mature.
- ***In-Memory Databases* –** This is one of the rather recent developments in database technology. Here, the total database is processed in memory providing even more performance than other specialized database technologies that still store data physically. In-memory databases especially benefit from a 64 bit address space. In addition, in-memory databases also use the above mentioned special database technologies
- ***Special access algorithms* –** They are used for reading and processing large amounts of data. An example is Google’s MapReduce, a programming framework and model for distributed processing. It is part of Hadoop. The goal is to overcome the limitations of traditional SQL and OLAP technologies. Many vendors combine various features with special database technologies as described above.

We will come back to these approaches, methods and technologies in chapter 7 (“Latency Matters”), and vendors are listed in chapters 10.3 to 10.5.

5.3 Data Visualization

Data visualization is a more and more used tool of data discovery. It is an ad hoc, interactive, solution oriented process designed by human interactions. Analytics via data visualization is a dynamic, user-centric approach based on the strong pattern recognition capabilities of the human eye. It can be further supported by analytical algorithms. Data visualization also enables team decision making via built-in collaborative services.

Data visualization is particularly equipped for the analysis of large volumes of data, of complex data structures, or of real-time data. Large volumes of data and complex data structures get clearly represented by appropriate visualization. Visualization is an excellent tool for detecting structural changes: You just see them! Data should not be aggregated, in fact, data visualization processes raw data. Data visualization supports self-service usage very well. It typically comes with an intuitive and visual user interface as well as a mighty library of various display formats like charts, heat maps and tree maps. It also combines data from internal and external data sources – big data sources included. The result is transparency (fig. 25).



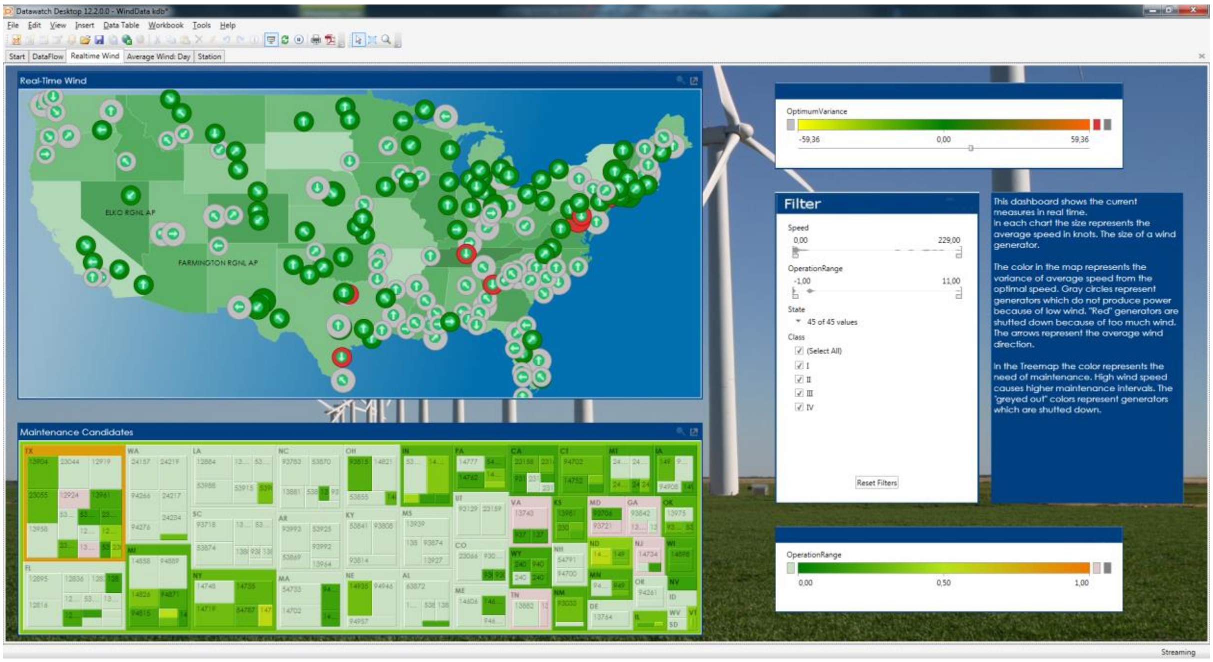
LMI.L

Last	1,186.00
Last_Time	11:21AM
Change	-64.00
% Change	-5.12 %
Volume	435,494.00
Turnover	516,495,884.00
High	1,250.00
Low	1,179.00
Open	1,243.00
Close	1,250.00

Figure 25. Examples of a heat map produced by Datawatch Desktop. It visualizes the percentage of changes of stock price in the FTSE 100 for a defined time period. Dark red indicates up to minus 5 percent, dark blue up to plus 5 percent changes. Just at one glance, important stock price evolutions are easily detected.

Data visualization is more than simply visualizing structured and static data. Indeed, it has much more potential and use cases, for instance, visualization of data streams generated by sensors or engines. Real-time data streams are either visualized by a time series representation or can be video-taped and made available as animations. Typically, such a visualization is coupled with event processing: Outliners can be immediately detected and trends can be identified and extrapolated. Sensors are used for monitoring and controlling functioning of engines. Organizations not only benefit from automation of engine control which means cost savings, but also from pro-active maintenance. Detection and forecasting of trends enable identification of risks like future problems caused by engines standing idle or taking damage. Such problems can be detected and solved in advance by data visualization of data streams. It saves time and cost. Furthermore, loss of revenues by engines standing idle are avoided. The results are smart processes that identify and solve problems before they arise.

> ***Example:*** Data visualization in real-time can also be applied for controlling the production of renewable energy. For instance, data visualization could identify appropriate locations for wind engines by exploiting weather data. Operations of wind engines then need network control. Again, weather data allows the prediction of the expected amount of produced energy. Data visualization can now be applied to visually monitor and record the actually produced energy. If a predefined maximal amount of produced energy is exceeded, a notification engine can directly switch off one or several windmills or influence the control of conventional power plants (fig. 26).



Speed
0,00 229,00

OperationRange
-1,00 11,00

State
45 of 45 values

Class
☑ (Select All)
☑ I
☑ II
☑ III
☑ IV

This dashboard shows the current measures in real time.
In each chart the size represents the average speed in knots. The size of a wind generator.

The color in the map represents the variance of average speed from the optimal speed. Gray circles represent generators which do not produce power because of low wind. "Red" generators are shutted down because of too much wind. The arrows represent the average wind direction.

In the Treemap the color represents the need of maintenance. High wind speed causes higher maintenance intervals. The "greyed out" colors represent generators which are shutted down.

Figure 26. Predictive maintenance with Datawatch Desktop: real-time visualization of data from US wind engines. The tree map shows all wind engines marked by colored spots that had been exposed to higher than expected wind velocities. Therefore, these wind engines should be preventively maintained.

Another example is given by visualizing semi and non-structured data, for instance data included in SAP reports, CSV file, log data, Web pages etc. or in Web click streams. Again, Web click streams are data streams providing optimal added value when they are analyzed in real-time. Then, they can help marketing to improve customer experiences across all channels. In the end, the RoI can be calculated: Transparency and proactive process control can be monetarily evaluated so that the added value of data visualization can be identified.

Data visualization can be applied by two different types of users. On the one hand, data scientists use data visualization for preparing fact based decisions corresponding with market dynamics and with the speed of business. Data visualization provides the better insight into markets, customer behavior and risks.

> ***Example.*** Many car manufacturers do not know which model variants generate profit or loss. This is due to traditional BI systems that only allow analysis of aggregated data. Data discovery and data visualization go the next step: They enable insights into non-aggregated data. So, a data scientist can easily and quickly give answers to questions about profitability of single models by combining the necessary data and by using tree maps, for instance, that visualize model variants causing loss by red spots

When such a solution has been delivered by data scientists, then a solution framework can be composed. It consists of the corresponding data acquisition, preparation and integration, as well as pre-selected display formats. The solution framework is then associated to the users that are responsible for monitoring and controlling of the solution scenario. Users can now work according to the principles of self-service data visualization.

In many cases, relational databases turn out to be show stoppers when organizations try to reap all these benefits of data visualization: Their technology hits the wall and is not sufficient to manage huge volumes of data, in particular poly-structured data. Therefore, data visualization tools should support big data technologies. NoSQL (not only SQL) technologies are especially important, because they are best equipped to manage semi and non-structured data as well as data streams. We will discuss NoSQL technologies in chapters 7.3 and 7.4 in more detail.

Take away: Goals of data visualization are fact-based decisions corresponding to the dynamics and speed of business, and better insights into markets, customer behavior and risks. Data visualization supports development of predictive models that enrich business processes by embedding analytical services into business processes and applications. Smart processes are the final result. They can identify and solve problems before such problems cause damage. Data visualization empowers users in special departments by better insights in risks and challenges as well as by enabling faster and better decision making.

5.4 Web Analytics

Web Analytics means the application of performance management and analytics to web data produced by (human) visitors when surfing on web pages. These traces can provide valuable information about the usage of the web page and the behavior of the visitors. Analysis of such web data can provide answers to questions like “how many visitors do I have when?”, “where do the visitors come from and where do they go to?” etc. The goal is obvious. The web page should be optimized. It should be made sure that goals to be achieved by the web page can be measured, monitored, and controlled like increase of visits and time of residence, increase of downloads, newsletter subscriptions and orders. This is exactly what performance management is all about, and web analytics can be understood as an instrument for protection of investment into a web page. Controlling of visitors’ behavior together with the continuous enhancement of its own web strategy can considerably improve internet presence and its efficiency.

Web analytics consist of two different methods. On the one hand, it uses performance management for continuously measuring and monitoring the effectiveness of a home page. On the other hand, it applies various analytic methods for the identification of point of failures for counteracting in the sense of web page optimization. In other words, web analytics is based on performance management and analytics.

Performance management relies on defining the right metrics. Typical metrics in web analytics are developing of revenue over time and number of visitors over time. But visitors are not equal. We have to distinguish between visitors who simply watch, visitors who put products into the basket, and visitors who really buy. Other metrics include the average value of an order, cost per campaign, and effectiveness of advertising by banners, newsletters etc. As in standard performance management, metrics are presented by reports or scorecards.

Analytics uses several scenarios. Click stream analysis helps to detect preferred and non-preferred parts of a home page. Segmentation supports the identification and classification of different groups of visitors, for example visitors coming from search engine X versus those from search engine Y. Analysis of visitor communication is used to monitor and optimize defined and important page sequences. Start and landing page optimization is achieved by testing the impact of changes to click and conversion behavior. Search engine optimization applies similar principles. Higher ranking of its own home page at the main search engines should be measured and monitored.

Web analytics is based on either the web servers' log data or on data generated by certain tags on the web page. Additional data for web analytics can be gained by web server plug-ins or net sniffers. Cookies are used to link a page call to a session and a session to a returning visitor. But many visitors do not like cookies at all because cookies break the anonymity of a visitor. By applying cookies, we enter into a conflict between the protection of the personality of a visitor and the interest of the operator of the home page to know its customers in the sense of customer orientation and relationship management. But note, in certain situations an IP address could be considered as personal data. Therefore, it is good advice to involve the company's data protection commissioner into all questions about web analytics.

A selection of main web analytics tools is given in chapter 10.4.

Take away: Web analytics is an important element of customer orientation and relationship management. It permits a better understanding of customers and to optimize marketing, increase revenues, and avoid fraud (click cheating, affiliate hopping etc.). ***But note:*** usage of web data for web analytics is subject to data protection laws.

5.5 Predictive Analytics

No one can foresee the future, no one can know data from the future, and certainly, no one can analyze it. But there are methods for predicting future trends and developments from historic data. This is task and target of **predictive analytics.**

Do you already use a forecasting system *(e.g., funnel management)* in your company's sales department to monetarily evaluate your leads according to their potential volumes, the likely period until completion of contract, and the probability of a successful conclusion? If so, you already implement predictive analytics. Do you have a suggestion machine in your web shop to give your customers purchase suggestions? If so, this is just one of the many uses of predictive analytics. Perhaps you use marketing models to help to decide which of your advertisements will appear where? All of these are different forms of predictive analytics *(the technique of estimating the probability of an event occurring in the future)*, such as:

- the completion of a contract,
- acceptance of purchase recommendations,
- the chances and the risks of measures about to be implemented,
- etc.

Together with data discovery *(discovering relationships and patterns in data sets)*, predictive analytics belongs to the family of analytical concepts; and these, together with the different aspects of performance management *(planning, monitoring and control using dashboards, reporting, other methods, etc.),* build up business intelligence.

Predictive analytics is already quite frequently used in CRM/CEM as well as in sales and marketing. Predictive analytics is based mainly on data mining. Traditional data mining methods include regression analyses, classification *(clustering)*, neuronal nets, as well as association analyses. Beyond this identification of patterns in (large) data volumes, predictive analytics also uses statistical calculations, machine learning and game theory elements; as well as other methods of operations research, such as optimization calculation and simulation processes. The background is made up of considerable amounts of mathematics and statistics, and also linguistics when text mining or text analytics is applied to unstructured data, such as text, blogs, tweets, etc.

Predictive Analytics is the most common term used today, but it appears together with descriptive analytics and prescriptive analytics. What is the difference?

- **Descriptive analytics deal with the past.** Descriptive analytics make the relationships between customers and products understandable. The idea is to learn from the past, and to apply this knowledge for making better decisions in the future. Typical examples are OLAP analyses. The problem of such analyses is that although correlations are found, they can be purely coincidental, and aren't enough for identifying cause-and-effect relationships. Nevertheless, descriptive analytics is an important first step towards achieving valuable, previously unknown insights into data.

- **Predictive analytics deal with the future.** Predictive analytics enable an assessment of the probability of a future event occurring. This might sound complicated at first, so here is an example for credit scoring. In this case, an assessment is made of the probability of a customer not being able to pay back future credit instalments. This assessment then enables risk evaluation of a credit application and provides solid support when deciding whether a given credit sum should be approved. Both historic and transaction data is used to determine possible patterns in the data, and statistical models and algorithms identify any relationships found in different data sets.
- **Prescriptive analytics provide suggestions based on predictive analytics.** Prescriptive analytics is based upon predictive analytics, but goes one step further. It provides explanations as to why a future event will occur, and gives recommendations for suitable reactions when it does occur. In the case of credit scoring, additional information is available to weigh up whether the customer is likely to default on repayment, and what decision is best - to grant a credit, or not. Therefore, prescriptive analytics attempts to assess the effects of potential future decisions, enabling them to be evaluated before they are made. Prescriptive analysis is still in its infancy, but it is one of the foundations of robots and self-driving cars.

But what, and how much, should the manager or expert in the specialist field know and understand about analytics? To keep things simple, we don’t want to differentiate between descriptive/predictive/ prescriptive analytics here, but simply to speak of “predictive analytics”.

Predictive analytics serves as an aspect of business intelligence, by providing support for making decisions. It establishes clearly understandable facts from within the framework of a given model, and enables decisions to be made based on these facts. Such decisions are made by managers, or experts in specialist departments, therefore it’s important to be able to understand the model, to correctly interpret the facts, and to draw correct conclusions. But it’s not necessary to understand how mathematics, statistics and linguistics are used within the framework of the model – that’s the job of special business analysts and/or data scientists.

Here are a couple of tips as to how the results of predictive analytics can be used to improve customer experience management, without any special mathematics/statistics/linguistics training being necessary:

- ***The Data:*** Despite big data, the biggest problem is that insufficient quantities of *suitable* data are available. A forecast of the future purchase behavior of customers requires information about their purchase behavior up until now. This information can be obtained from a customer loyalty program *(loyalty cards, etc.)*, or from analyses of purchases paid for by credit card. When different sales channels or customer contact points exist *(and this is usually the case!)*, data from these sources must be consolidated too. In other words: Professional information management is needed to create a customer data warehouse, with unique customer identification, and where data is conditioned as necessary. This is the prerequisite for successful predictive analytics, and for transforming big data into smart customer data. Therefore, predictive analytics should start with a state-of-the-art information management solution.

- ***Statistics:*** The most common data mining method used in predictive analytics is “regression analysis”. Its advantage is that it provides a model that can be used immediately; the regression equation. Let’s take another look at our example of credit scoring. Using regression analysis, we determine the regression equation and estimate the parameters. This model can now be used immediately for every new, unknown customer. We set the parameters given by the regression equation and then calculate the score. To ensure that everything functions correctly, both the quality of the data used for regression analyses and the efficiency of the analysts performing them are extremely important. A wide range of additional or alternative methods are used today and will be dealt within the subsequent chapters.
- ***The Model:*** The fundamental assumption in predictive analytics is that the behavior of the model in the past will remain unchanged in the future. This is referred to as the “stationary model”. In our example of credit scoring, this would mean that a customer has the same credit score from birth to death, irrespective as to what happens to him during his lifetime, or what changes occur in his environment. But this doesn’t sound very realistic: Personal circumstances change; markets change; customer behavior changes. Therefore, any acceptance of the stationary behavior of a model once derived in the past, should always be viewed critically. A model that is no longer realistic, is also of no particular help, and usually gives incorrect, outdated forecasts. Therefore, managers, and experts in specialist departments, should regularly question their analysts about the fundamental assumptions of the model, their effects, and when these are no longer applicable.

Based on this fundamental understanding, the results of predictive analytics can be discussed with analysts and data scientists. Answers can be found for the following questions, which are decisive for the models’ evaluation and its interpretation:

- Which/what data sources are used / not used?
- Is the data sufficiently representative for the given situation?
- How good is the quality of the base data?
- Are there any stray entries or aberrations? Is any data missing? Does this influence the analysis?
- What assumptions have been made?
- Under what conditions were the given assumptions no longer applicable?

This means, therefore, that both correct data, and the correct mathematic/statistic/linguistic model, must be available. It also means that care must be exercised when dealing with fundamental assumptions. In practical terms this is not always easy; but with this approach, accurate insights are possible, based on reliable facts. It allows better decisions to be made at specialist and management level, based upon improved knowledge and better understanding of the customer. It enables customer experience management that fulfils the expectations of the customer. And then it’s just a question of measuring the effects of these decisions in terms of performance management, and thus to develop monitoring mechanisms for ensuring that all decisions have the desired effect.

Take Away: Predictive analytics give companies a powerful analytic tool for extracting smart customer data from big data. Efficiently implemented predictive analytics allow better decisions to be made, and actions made more effective by forecasting probable developments in customer and market trends. The result is customer experience management that answers the customer's wishes in all respects. For a company, this gives a clear advantage over the competition.

The greatest challenge of predictive analytics is to successfully integrate the results into regular business. Smart customer data *(perhaps revealing which customers are likely to* cancel), is only useful when the company also draws the correct conclusions. False interpretation of the results can have exactly the opposite effects of those desired, such as an analysis of potential cancellation intentions of customers with an increased cancellation rate. Equally critical for success is the importance of making actions and measures resulting from predictive analytics, transparent for managers and experts in specialist departments.

5.6 Trends in Data Mining

Definition: **Data Mining** is a process that identifies and/or extracts information from a large or a very large amount of (structured) data; information that has not been known before, that is non-trivial, unexpected as well as it is important.

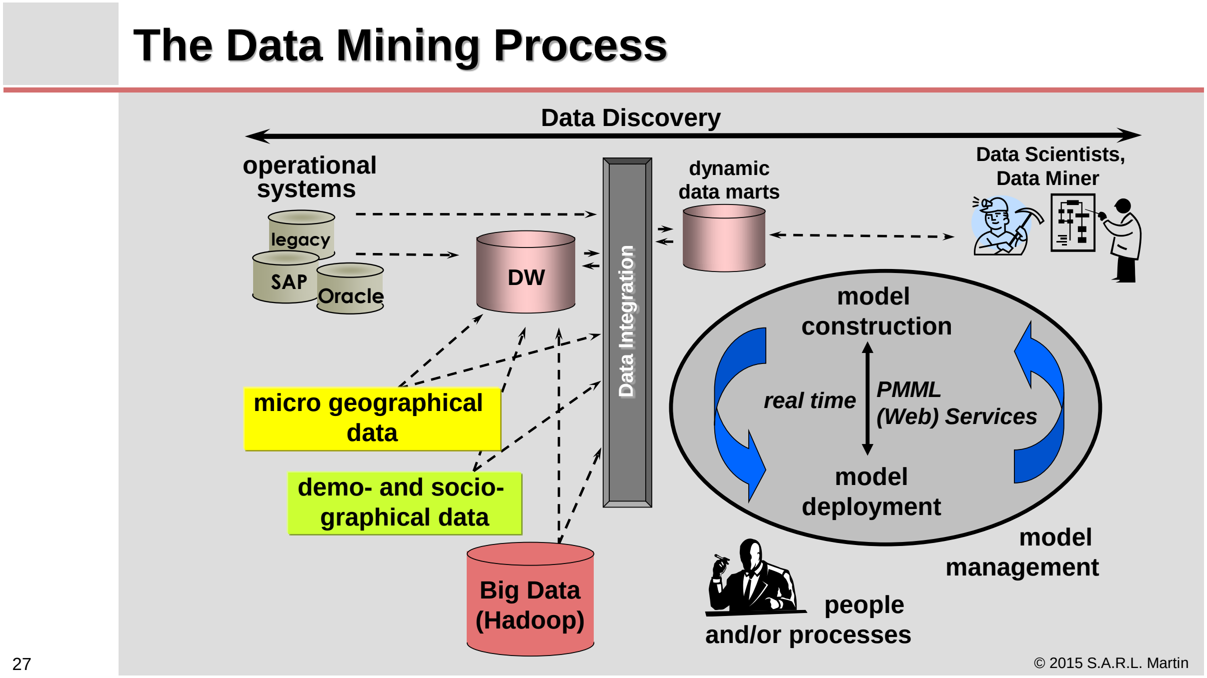
Data mining is the most important and main method of predictive analytics as we have seen in the previous chapter. It the truly “intelligent“ method and technology of analytics. It is a “bottom up” approach to discover patterns, structures and correlations to create hypotheses. Since the second half of the 90s, data mining has found its position within the instrumentation of sales processes. Thus data mining has always been a central part of CRM, and nowadays it's even more than that: Data mining has found its position in many enterprise-divisions, as in product-control, risk management, detection of fraud, money-laundering etc. Statistic tools have worked as forerunners; they are still used today to prepare and to amend data mining applications. Data mining on poly-structured data is known as **Text Mining**. Therefore, the following discussion about data mining can be transferred one to one to text mining.

Data mining comes with a methodology describing the iterative process of data mining lead by the business, as well as the tools supporting the processes (fig. 27).

Step one in a data mining process is data supply. No data – no data mining; sounds trivial, but: 60 to 80 % of all resources in data mining are plugged in availability, accessibility and quality of data for data mining (see chapter 6). Automation of data supply is still one of the biggest challenges in data mining. This is also a result of the explosion of data volume. For example, daily, 500 million tweets are sent[25]. Twitter data is mainly poly-structured, and

[25] see http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/2/#.U4R75suKDIU, Zugriff am 27.05.2014.

important information can only be gained by text mining and text analytics which will be discussed in chapter 5.7. In supply chain management, data volumes explode because of RFID-technologies for registration and monitoring of products and parts, and sensor data enable automated monitoring of machines and proactive maintenance.



The Data Mining Process

Figure 27: Data mining processes reveal challenges of data and text mining. On the left: A data integration platform should be used for data supply. The corresponding data architecture will be discussed in chapters 7.3 to 7.5. As data volume is high to very high, high performance and scalability of algorithms are needed, as well as handling of up to a thousand variables. On the right: Building and deploying of data mining models have been separated processes, but real-time challenges make them unite to a Closed Loop managed by performance management. As a general rule for deployment, models should be embedded in operational SOA-based processes. Furthermore, data mining models work more effectively when they are tuned to the granularity of the problem. As a consequence, up to hundreds of models need to be created in just a couple of days. More than ever before the amount of models requires model management.

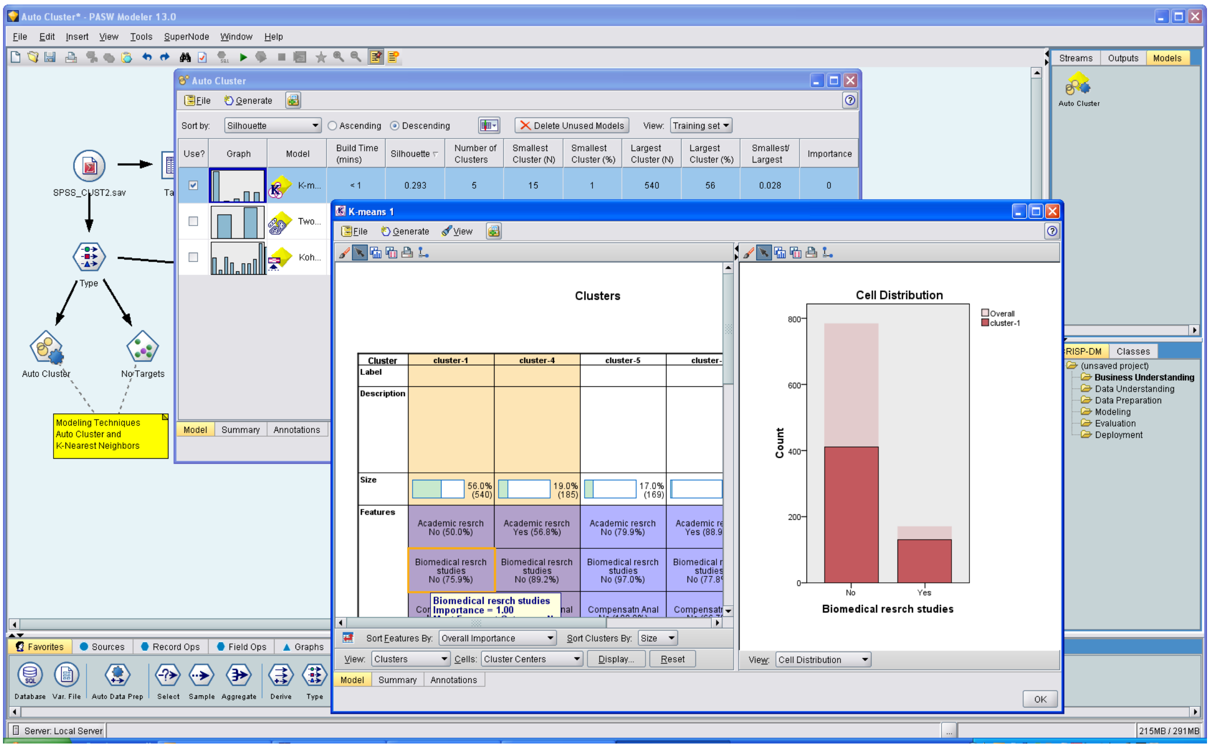
High performance is thus a challenge for data mining algorithms, as large amounts of data have to be handled; thousands and more of variables have to be managed due to data diversity.

Step Two describes the choice of methods, technologies and tools. In data mining different methods are used to solve business tasks:

- ***Classification.*** By means of this method new objects are to be sorted in given classes. Class-affiliation is given as a new attribute to a data record, as e.g. classification of request for loans into risk groups or sorting of clients into customer segments.
- ***Estimation.*** While classification involves discreet results, simply “yes” or “no”, estimation deals with continuous results. Based on input parameters the output quantity has to be

estimated. To cite an example: a scoring that quantifies the probability that a customer might either purchase an item or cancel his subscription.

- ***Prediction.*** Here models are deduced which describe data structures sequentially in order to make chronological prognoses. Possible examples are time series models in banking and financial services like prediction of stock exchange quotation. Also it is a matter of detecting chronologies as basic patterns of habitual buying behavior. In retail, for example, a typical application of prediction is the detection of sequential buying patterns.



Auto Cluster* - PASW Modeler 13.0

Auto Cluster

Sort by: Silhouette ○ Ascending ● Descending

View: Training set

Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/ Largest	Importance
☑		K-m...	< 1	0.293	5	15	1	540	56	0.028	0
□		Two...									
□		Koh...									

K-means 1

Clusters

Cluster	cluster-1	cluster-4	cluster-5	cluster-
Label				
Description				
Size	56.0% (540)	19.0% (185)	17.0% (169)	
Features	Academic resrch No (50.0%)	Academic resrch Yes (56.8%)	Academic resrch No (79.9%)	Academic re Yes (88.9
	Biomedical resrch studies No (75.9%)	Biomedical resrch studies No (89.2%)	Biomedical resrch studies No (97.0%)	Biomedical r studies No (77.8
	Cor	nal	Compensatn Anal	Compensat

Biomedical resrch studies
Importance = 1.00

Sort Features By: Overall Importance Sort Clusters By: Size

View: Clusters Cells: Cluster Centers

View: Cell Distribution

Server: Local Server

215MB / 291MB

Figure 28: Today's tools for data analysis and data mining like the Auto Cluster functionality of the IBM-SPSS PASW Modeler 13 support the calculation of data based decision models with different statistical procedures and provide a comparison of their predictive capabilities. A user can now pick the model of his preference, and the tools apply automatically the selected model.

- ***Affinity grouping or association rules.*** This method is able to detect correlations between different elements within a class. The question is to determine which things go together. A typical example is the analysis of shopping carts at the supermarket for detecting statistic correlations between products. In this way co-products can be identified, which are suited for cross-selling.

- ***Clustering.*** This is a method to segment objects into disjunctive classes based on self-similarity. As an example there is the deduction of a pricing model by analyzing the customer specific use patterns on telephone network, or the analysis of handling damages in automobile assurances.

- ***Descriptive models.*** They detect and describe correlations between variables, which enables to find explanations for special attitudes and behaviors.

Classification, estimation and prediction belong to *directed* data mining. Goal is to build a model that describes a particular variable of interest in terms of the rest of the available data. On the other hand, association rules, clustering and depicting models belong to *non-directed* data mining. It's all about deduction of correlations and relationships between variables.

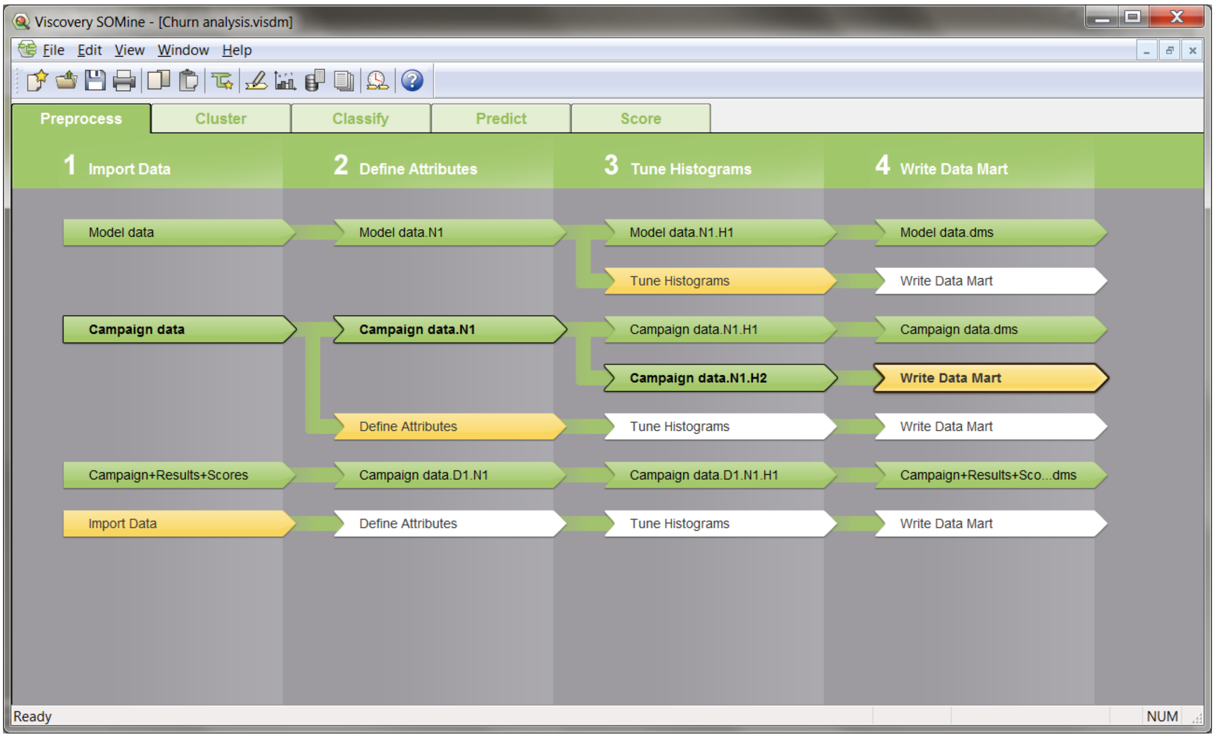


Figure 29: Modeling of data mining processes: as an example, Viscovery's visualization of the pre-process workflow of a churn analysis with its administration of user-defined versions is shown.

Data mining tools support these methods by different techniques. There are mathematic-statistical operations, such as general linear analysis, discriminant analysis, statistical regression, correlation analysis etc.; on the other hand there are methods based on cybernetics as for example neuronal networks, decision trees, rules of induction, self-organizing maps (SOM), support-vector processes, fuzzy-logics and knowledge-based systems. Finally, visualization provides important support. The human eye is an equivalent detector to mathematics, statistics and information theory when it comes to recognizing structures, if is appropriately supported by visualization.

> ***Example of a data mining platform.*** **R** is a free programming language already developed in the early 90s. It is published under a General Public License (GPL) of the Free Software Foundation (see http://www.r-project.org). R stands for **The R Project for Statistical Computing**. It is a software for statistical computing and graphics. A large number of additional packages complements the R functionality by methods from various statistical application areas. R can easily be extended by customer written functions. It can be combined with other programming languages like GRASS, Perl, Python, C or Java. In comparison to other frequently used program packages for mathematical statistical analysis like „SAS“ or „IBM/SPSS“, R offers the advantage of being license-free (under the free GPL). For some time now, R is used

> by many universities, especially in education. As a result, nearly all graduates have R skills that introduce knowledge into corporations without causing any additional cost. Consequently, R is the world's mostly used statistical computing environment. In chapter 10.4, we will give a reference to the top 20 most liked R software packages.

The problem is: which data mining technique will support which method. There is no one-to-one allocation. There are, of course, a certain amount of experience-based rules which help to choose a data mining technique as a method for solution, but these are heuristic rules. Typically you choose various techniques for trouble solving, so you are able to check and compare the results via assessment (fig. 28).

Step Three describes the development of the data mining model and the interpretation of its results. Data mining tools have been clearly improved; yet consultancy and know-how of experts will stay required in certain tasks and in certain circumstances. In this respect this step three is iterative: The data mining expert deduces the model and refines it together with the business division in charge with the project. In the meantime, there are some data mining tools enabling a kind of self-service data mining (in the sense of self-service BI). They support data mining power users in the specialist departments. A broader use of data mining in an organization comes true.

The division is responsible for the interpretation of results and checking of relevant usability and applicability. The data mining approach should be documented as a methodology, should be supported visually (fig. 29), and the resulting data mining process should be ideally a SOA based process.

Up till now creating a model took several days if not weeks. Due to this fact the amount of possible models was limited. As a consequence tasks were pooled in groups and all elements of a group were processed with the same data mining model. Of course it is preferable to create models faster than that, so each model can be refined.

> ***Example:*** Creating a clustering model for the whole of Germany will show good results; but creating a model for each federal state or even for each district will increase the results. Regional differences in demography are estimated more exactly and results will be more adequate. This means we have to be able to create a large amount of models simultaneously. Regarding the time dimension of the model, the amount of models will naturally increase. Models vary with time. There are considerable differences whether the model is created on Christmas, Easter, holiday time or when a staff members driving a campaign. The challenge is to create hundreds of models within a given time frame (1 or 2 days).

We need to manage the enormous amount of models. Data mining model management has always been a necessity, but due to the large amount of models used today, model management is more important than ever.

Step Four describes how a predictive model can be used within an operational process. Up till now, the data mining modeling process and operational use of the deduced predictive model have been separated rigorously. Embedding the predictive model is nowadays easier than ever. The predictive model is consumed as a service by the corresponding SOA-based process. But if modeling is still done off-line, we can run into a problem: after some time we are no longer able to know whether the model that has been deployed in the operational process is still valid and significant. Thus it is not sufficient to deduce the predictive models once or periodically, but we have to ensure that the predictive model describes the context actually: It has to be à jour.

So we need a model management that is able to link per closed-loop the process of model-building to the operational process. This leads to robust or adaptive algorithms.

- ***Robust algorithms*** are able to measure the precision of prediction in every interaction. Thus we are able to discover when an update or an adjustment is needed. In this way we can build a semi-automated process triggering the remodeling of the predictive model, as for instance sending a notification to the data mining specialist.
- ***Adaptive predictive models*** are self-learning and they deliver up-to-date prognoses in the context with the process data models. Using these automated and dynamic data mining solutions we are able to compile any prognoses on your clients, i.e. risk of churn, customer value and sales forecasts. Also, these solutions are applied for cross- and up selling prognoses in fast-reacting processes such as telephone-marketing and e-commerce.

Last but not least like any other process, data mining processes need performance management. It is imperative to define sensors within the process, to develop metrics and to monitor and to control the data mining process exactly like any other process via a closed loop model.

Finally there is an important guideline leading all data mining tasks: precise solutions as well as mathematic elegance of certain data mining techniques should be subordinated to the velocity of achieving solutions.

5.7 Text Analytics

Text analytics is a new type of analytics[26]. It combines linguistic methods with search engines, text mining, data mining, and machine learning algorithms. Social media are the big driver for text analytics.

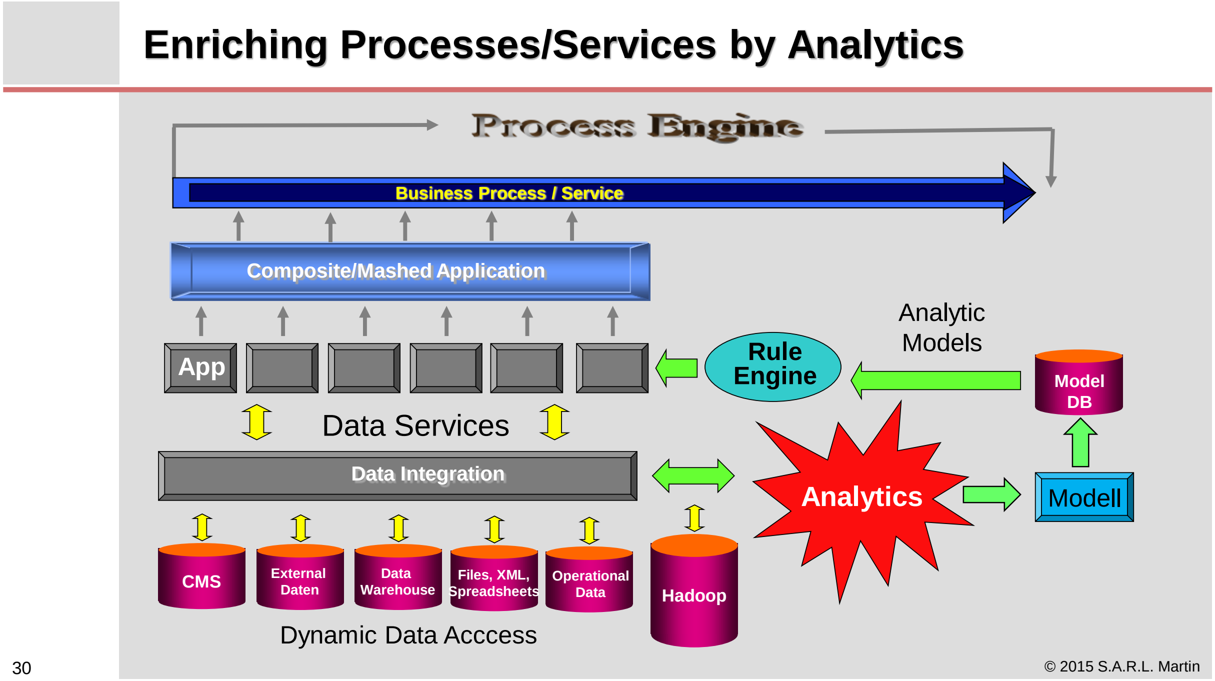
Social Media. There are hundreds of millions and more users of Facebook, LinkedIn, Xing and other social communities. This number is steadily growing. Furthermore, there are 10th of thousands mainly specialized blogs and forums. Not to forget platforms like Twitter for mini blogging. This shows the attractiveness of the initial ideas of Web 2.0, the experience web where everybody can join in and utter his opinions, sentiments and preferences. But market researchers, product managers, and anybody in marketing have a completely different look to social media. They

[26] See also http://www.intelligententerprise.com/blog/archives/2007/02/defining_text_a.html

recognize the new dimension of rich data about current and future customers, about market potentials, sentiments and trends in the market.

Social media invite everybody to share his/her personal data with anybody. In social media, sharing and participating is not only “in”, but the “big deal”. Some of them do exaggerate, and you may think people think that in social media we are dealing with “digital exhibitionists”: Personal and even very private data is exhibited publically. For marketers in the business, this makes up a real nugget of valuable information just waiting to be explored.

Several years ago, telecommunication industry has already started to explore web data systematically. In the meantime, we have seen banks and insurances follow this roadmap, as well as retailers and manufacturers of consumer goods, but nobody really likes to talk about. The crystal clear transparency of customers that has been created in the marketing and market research departments is still a top secret not to be shared with its clients! The advantages and benefits of the crystal clear transparency are obvious: A manufacturer of consumer goods wants to know how its brands are recognized by surfers, bloggers, and social media communities, and how they are rated and compared to its competitor’s brands. A global hotel chain is interested not only in ratings by its guests, but also how its guests rate the competitors. There are no limits any more, when all this data in social media is accessible and analyzable (fig. 30).



Enriching Processes/Services by Analytics

Process Engine

Business Process / Service

Composite/Mashed Application

App

Data Services

Data Integration

CMS

External Daten

Data Warehouse

Files, XML, Spreadsheets

Operational Data

Dynamic Data Acccess

Hadoop

Rule Engine

Analytic Models

Model DB

Analytics

Modell

30

© 2015 S.A.R.L. Martin

Figure 30. Data integration is the foundation of text analytics processes. It provides a dynamic, service oriented data access to all relevant internal and external data sources. The (very) large data volume requires high performance and scalability of the algorithms as well as managing thousands of variables. Ideally, the predictive model is deployed by a rule engine, so it can be embedded into a process as a service (cf. chapter 4.3). This advantage stems from the underlying SOA infrastructure as depicted in the figure. In the end, text analytics makes intelligent and smart processes. For instance, in customer experience management, a social profile can be associated to a customer so that precise and targeted buying recommendations by a call center or by a web shop can be derived and put into action.

First thing you need is a kind of “vacuum cleaner” for extracting relevant data from the web. This is done by **semantic web crawlers**. This technology for web extraction allow to access and to extract all public data in the web even if there are no published interfaces (APIs – application programming interface). These technologies are described in chapter 6.4.

Let us now discuss analysis. Traditional statistical and data mining procedures are insufficient, because web data typically is poly-structured or at its best semi-structured. These new requirements have helped to develop a class of analytic tools and methods: text analytics. Web 2.0 requirements to analytic tools are quite demanding. In social media, there is a lot of cynicism, sarcasm and polemics. Furthermore, tweets for example are semantically poor.

Since 80% of enterprise data is not stored in relational databases, but is stored as email and other poly-structured content, text analytics is not only a good approach to web 2.0 analytics, but also for new and advanced enterprise analytics. In many emails and in corresponding documents, you will typically find the context for correctly interpreting structured information. We can say that traditional BI with its OLAP, statistics, and data mining provide the “what” in the enterprise, whereas text data analysis enables the “how”.

Definition: **Text Analytics** is the extension of analytics, especially of data and text mining, reaching out to Content Management and into the World Wide Web.

Text analytics is both, technology as well as process for knowledge discovery and extraction in poly-structured data. The first goal of text analytics is to identify selectively entities (like names, data, locations, conditions) and their attributes as well as relationships, concepts, and sentiments between entities. It starts with adding structures to the text to be analyzed. As many words as possible are to be identified and to be matched to business domains. Content-oriented classification is the result. Examples for business domains are brand names or country names. The identification principally follows three procedures:

- Dictionaries or reference sets. This is typically used for identifying brand or country names. A dictionary is set up that includes all relevant terms of this domain. The identification is then simply a comparison where additional information can be linked to each term.
- Rules based on known patterns. This approach is quite appropriate for semi structured text elements like telephone numbers, credit card numbers, dates etc.
- Rules based on grammar. Such rules are derived from corresponding grammatical constructs of a language like declination, conjugation etc.

Results of such first identification are called annotations. They can be presented as tables per domain. These tables set the foundation for all subsequent steps of text analytics.

The second goal is to create and to visualize classifications based on the identified structures. As an example, an outcome of text analytics could be the identification of opinion leaders in social networks.

> ***Example.*** Let us assume a fictional phone company. Suppose its competitors are offering an aggressively priced family plan and customer service is getting calls about

> this. How, as a marketing team, are you to know this? It could constitute 10% of calls and be a part of a mountain of notes from customer service operations spread out geographically all over the world. If the phone company had the technology to parse through the sentences, pull out key words and phrases and see that people are asking about this calling plan, they could start to see pattern emerging and could react. What's more, it can be used to analyze information both inside and outside the firewall, so companies could use it to find patterns on social networks, for example.

This example shows an interesting application of text analytics: **Sentiment Analysis and Opinion Mining.** Leading European market research enterprises scan web blogs, discussion forums, and product ratings by automated sentiment analysis for online market research. Goal is the assessment of general opinion and attitudes of consumer groups about products, brands, and/or enterprises. Examples include opinion mining about hotels or consumer goods like washing agents or technical products like mobile phones. Beyond these assessments of attitudes, this analysis can also provide the general opinion and attitudes towards competitors and competing products, controlling of effectiveness and efficiency of marketing campaigns as well as provide recommendations for certain marketing activities. The capabilities of multi-lingual analysis today enable global analysis, for example the way a brand is recognized and rated in different countries.

Automated sentiment analysis is also applied in pharmaceutics for opinion analysis of new drugs, for competitive analysis, and for image analysis of brand and enterprise. Automated sentiment analysis has also found its way into financial services. Opinions and moods uttered in articles and other text documents about certain stock trigger recommendations about buy and sell. Sentiment analysis has already been applied in politics, for instance in the US presidential campaign 2008.

Text analytics is also successful as soon as the identification and classification of critical customers is concerned. Critical customers could be very helpful in removing product flaws, but could also be notorious grouches and wiseacres.

> ***Example:*** Automakers like BMW actively uses blogs. Experience with bloggers has shown that customers communicate sometimes more positively about BMW products than BMW's own slogans would dare to. (Attention: Sony has once tried to influence bloggers and blogs. When this was brought to light, damage to Sony's image was serious.) BMW created "M Power World", a social network about sportive driving for the special customer segment of buyers of M models. Here, customers are invited to exchange ideas with BMW developers and designers. Customer becomes product developer – this is the application of the fundamental Web 2.0 principle that consumers metamorphose to producers.

Automakers typically apply social media forward strategies: Social media principles become part of their CRM strategy. An alternative would be a passive strategy by automated observing of selected blogs and forums by text analytics for identifying critical situations and mood changes as quickly as possible. This is very well doable by text analytics, but it turns out that it is extremely difficult to launch the right actions in case of. You may legally enforce the deletion of blog entries, but in reality, they will pop up elsewhere. In the world of social media, the principle of "semper aliquid haeret" is inexorable. But recently, European laws enforced Web organizations like Google to respect "the right to forget", i.e. under certain circumstances, people and corporations have the right to claim that search requests do not reveal certain Web sources, i.e. are to be forgotten.

Text analytics as any analytics should be always related to a performance management according to our leitmotiv “You can only manage what you can measure.” In text analytics, we need metrics for calculating the relevance of data sources and of interconnectedness of data sources, business scorecards for visualizing and consolidating of monitoring and finally a reporting, especially an exception reporting for automatically alerting when social media peculiarities like increases of tags, of authors, of threads occur. Text analytics vendors are listed in chapter 10.4.

Many vendors offer text analytics solutions that look like out-of-the-box solutions. But this is overselling, because given today’s state of the art in text analytics, projects need a high level of consultancy and should be run by data scientists. But since data scientists are not easy to form and to hire, text analytics projects are still lucrative business for consultancy practices: The interpretation of results requires deep understanding of business context. The reason is obvious: Mathematical methods provide always correct structures and facts, but this does not imply that there is a relevance of these models to the real world. So, caution is mandatory, when facts derived from social media data are to be put into business context. Ratings could be related to friendship and not to real experience, opinions in blogs could be manipulated, and profiles in social media could be faked. Plausibility tests of outcomes are a must (see also chapter 2.4), but requires data scientists, business experts and/or consultants.

5.8 Location Intelligence – the Importance of “Where”

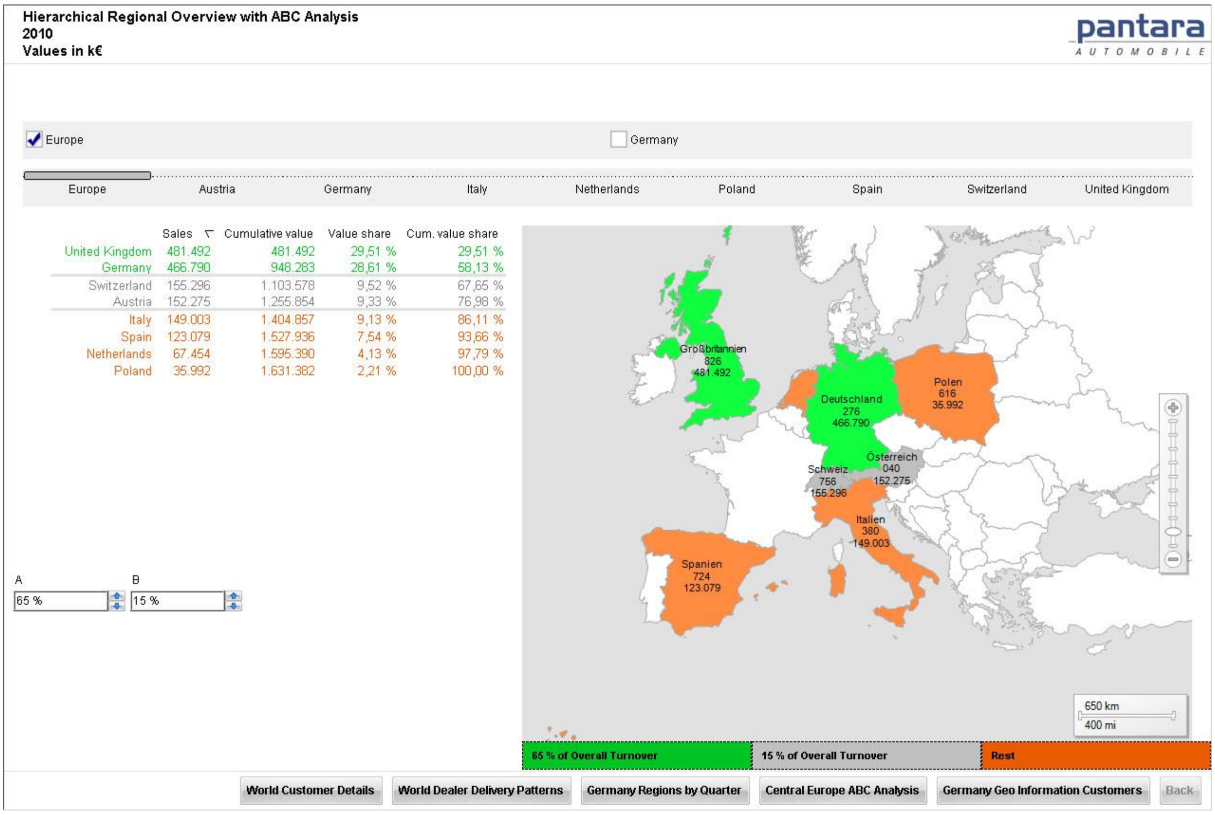
Location Intelligence means the geographic dimension of Business Intelligence. Location Intelligence uses geographic data describing **“where”** a customer, a supplier, a partner, an enterprise or products are located, or where a service is performed. It is standing for the capability to understand and to organize complex phenomena via geographic relationships being found in nearly all information about addresses and routes. 80% of an enterprise’s information is about location, about “where”[27]. Combining geographic and space-oriented data with other business data means optimizing intelligence, with the result that better decisions can be taken and business processes can be optimized.

Forerunners of Location Intelligence are Geographic Information Systems (GIS). Generally GIS have been a generic, department-oriented niche product, appreciated and used by few experts. Location Intelligence is applied to the entire enterprise and it is applied to processes and markets. Location Intelligence addresses every single staff member.

[27] See http://www.intelligententerprise.com/print_article.jhtml;jsessionid=SFANAIHXNPWYMQSNDLOSKH0CJUNN2JVN?articleID=181503114

Location Intelligence combines technology, data and services with domain knowledge. It enables enterprises to measure, to compare, to visualize and to analyze their business data in a geographical context.

Location Intelligence pursues the same goal as Business Intelligence: to extract information from data, knowledge from information, and knowledge is used for decisions and actions to control an enterprise (fig. 31). Consequently Location Intelligence is used in analytic scenarios as well as it can be embedded into operational, even real-time business processes. That is why Location Intelligence as well as Analytics is provided through services, to cover both functional areas: **analytics and operations**, and to provide appropriate IT support.



Hierarchical Regional Overview with ABC Analysis
2010
Values in k€

☑ Europe □ Germany

	Sales	Cumulative value	Value share	Cum. value share
United Kingdom	481.492	481.492	29,51 %	29,51 %
Germany	466.790	948.283	28,61 %	58,13 %
Switzerland	155.296	1.103.578	9,52 %	67,65 %
Austria	152.275	1.255.854	9,33 %	76,98 %
Italy	149.003	1.404.857	9,13 %	86,11 %
Spain	123.079	1.527.936	7,54 %	93,66 %
Netherlands	67.454	1.595.390	4,13 %	97,79 %
Poland	35.992	1.631.382	2,21 %	100,00 %

A: 65 %
B: 15 %

Figure 31: Data analysis in context with maps in Cubeware Cockpit V6pro. Especially sales data benefits from adding geographical and area information to the reports. Revenue comparisons, time line analysis, and changes in responsibility for sales regions are visually enhanced by maps. The screenshot shows an interactive Pareto analysis (ABC analysis). One look at the map tells you that underperforming regions are right next to each other.

Examples for analytic Location Intelligence are: Network-planning and –design in supply industries (water supply and electricity), telecommunication and IT, urban management and location planning as well as location analysis (public administration, Health Service, banking, retail and tourism), risk management (insurance), market- and customer-analysis (all B2C verticals). Examples for operational Location Intelligence are: CRM (acquisition of new customers, Cross-/Up-Selling, and customer loyalty) and other business processes in various branches as insurance (claim and risk management) and transport (breakdown services, emergency services as well as tracking and (real-time) routing management). In particular, operational Location Intelligence has a very high potential for mobile services, which before all need to know about “where”.

Data sources for Location Intelligence are various. To start with one’s proper customer data it continues with external data sources like client-demographic data, aerial views, satellite images and geographic data. This once more emphasizes the importance of service orientation for Location Intelligence. **Mashing up** of information from various sources and combination with further business data provides enormous added value. Of course this even works better if geographic data and information are available as web services or other standardized services, which nowadays is guaranteed in general: Using of Location Intelligence as a new dimension of Business Intelligence will be easier and faster than ever. An advantage of Location Intelligence that is easily and quickly achievable is to simply link customer data to spatial coordinates and visualize the enriched data. This already provides a big added value.

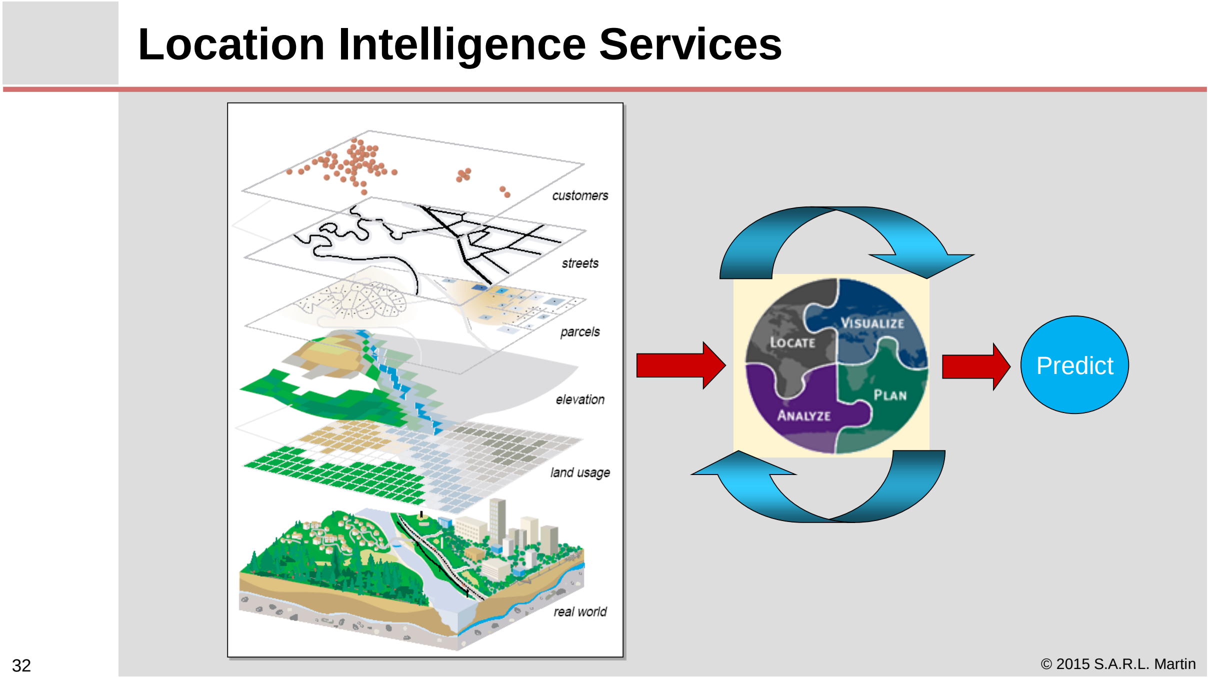

Location Intelligence Services

Figure 32: Location Intelligence includes geographical information services providing various views on the “real” world as for example information on streets, parcels, land usage, and elevation that can be combined with customer data. Location Intelligence processes describe the mashing up of geographical information services within the phases of localization, visualization, analysis, planning, and forecasting.

Tools of Location Intelligence are geo coding of data, descriptive mapping, visualization and analytic mapping up to predictive models (fig. 32). Geo coding includes two tasks. One task is coding of address data, the other one is coding of IP addresses. Coding of IP addresses is the prerequisite for locating users in the mobile internet. Tools should be interactive and intuitive. That is crucial for user acceptance. Speaking of a map, we are still used to imagine a static document. That is not comparable with interactive field mapping of today’s Location

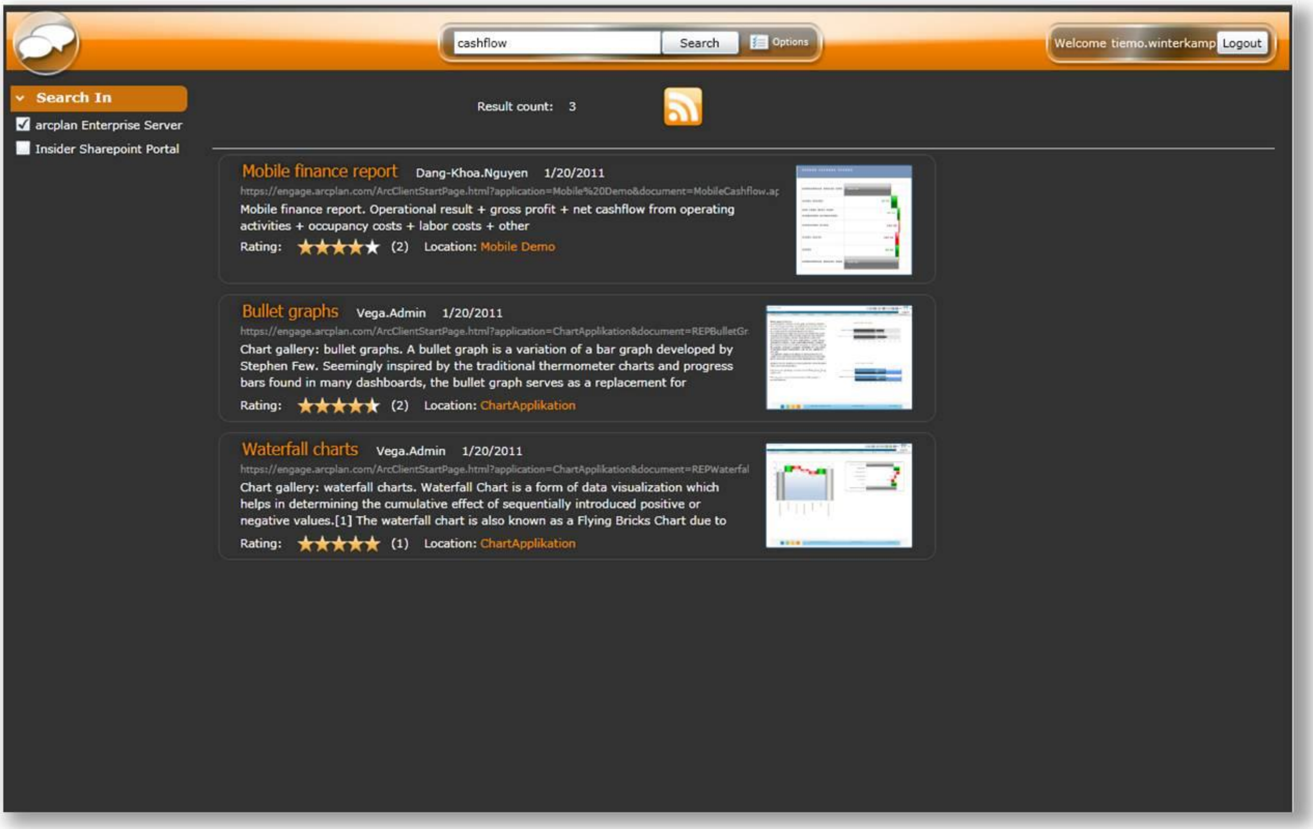
Intelligence. Interactivity enables human eyes to act as a detector for unknown patterns, trends and structures. This is the impact and potential of combining geographic and business data. That is why Location Intelligence tools offer a mighty complement to data mining and statistical analysis. "Where" is essential in Business Intelligence: Location matters!

5.9 Big Data meets Performance Management and Analytics

In the era of Big Data, performance management and analytics is complemented by social media functionality, by knowledge management, by new technologies (web and cloud integration tools, analytic databases, text analytics), and by new application areas (social media performance management, social media analytics). We now go into more details and discuss needs and benefits.

Social Media Functionality and Collaboration. Social media are not restricted to private use, but enter organizations. People used to live social media in his/her private life, want to see similar approaches and concepts in the organization. Hence, social media bring new working conditions and environments into performance management and analytics. One of the first things to change is the traditional portal. Facebook and other social media offerings have set the standards: They created new types of intuitive user interfaces that can be understood by everybody without any training. This is complemented by new ways of communication in the internet. Communication in social nets is a communication following the social net structures. It is a "many to many" communication. Facebook, LinkedIn, Twitter etc. are the examples. They consequently translate to performance management and analytics tools. Now for the first time, reports and metrics can be associated to authors. They can be annotated and discussed. Questions to numbers now get answers. Knowledge that has been "hidden" in the heads of experts now can be shared with everybody who is concerned. This creates transparency: Numbers are put into their semantic contexts, and are no more isolated in tables and graphs. Hence, numbers now can be better understood and interpreted, because it is no more a single person in charge, but the total team that is engaged. We challenge the "wisdom of the crowd".

Social media also come with a completely new approach to "search and find". This has been an old problem in BI. Where can I find useful information, which reports do I need, which metrics should be added to my dashboard, did anybody already had a similar question, a problem to solve? Theoretically, such questions can be solved by rigorous BI governance, but as practice has shown, acceptance of rigorous BI governance is rather low. Social media offer alternative approaches following the wisdom of the crowd principles. Reports and metrics are evaluated by the subscribers. This creates top lists of best rated numbers (fig. 33). Subscriptions of similar job profiles (if necessary made anonymous) can be displayed. For instance, in an Amazon style, a controller could get a recommendation that controller X in country Y is benefitting from a report that he/she is not yet using. So why not subscribing this report, at least as a trial? To conclude, social media tools and concepts can create bottom up BI governance that by definition is highly accepted by everybody. In the end, this creates und stimulates motivation. Social media concepts for finding the right information are

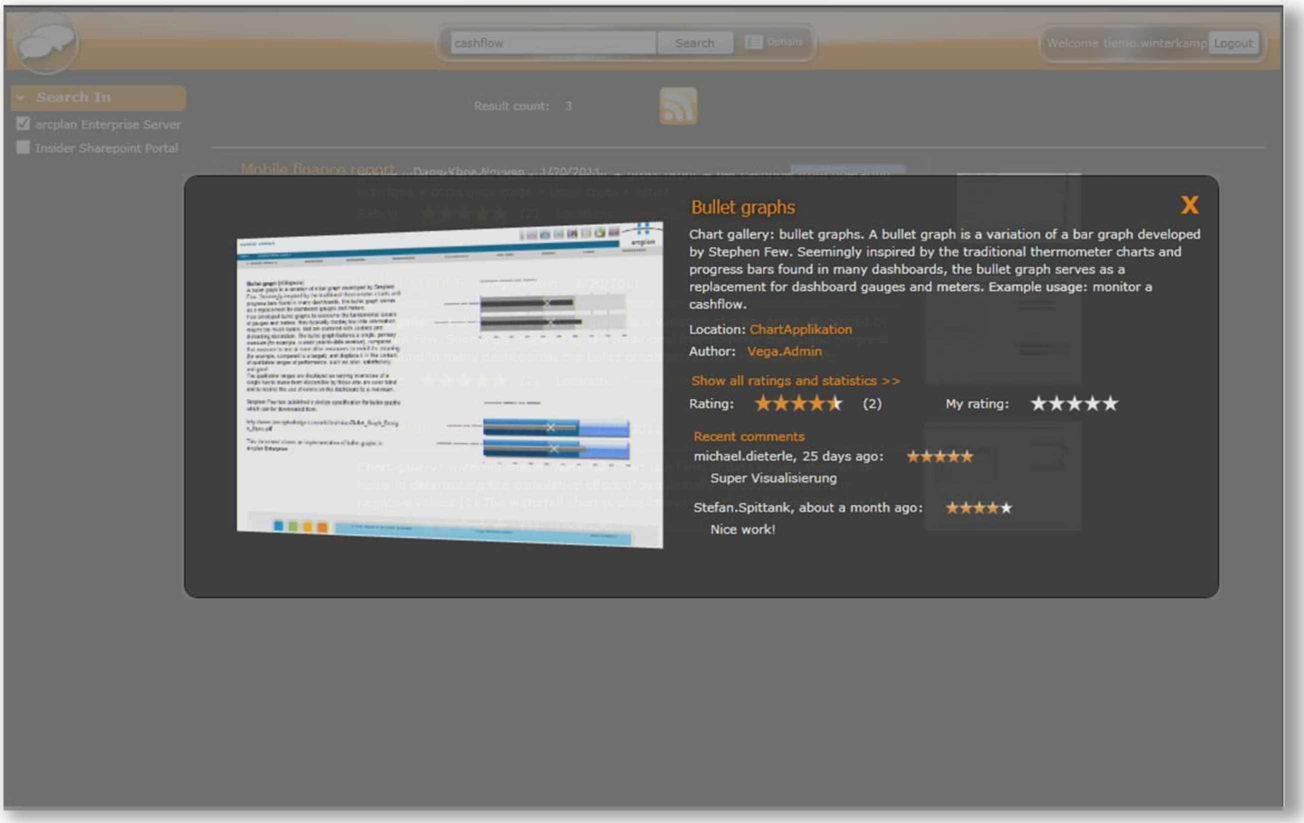


Result count: 3

Mobile finance report Dang-Khoa.Nguyen 1/20/2011
https://engage.arcplan.com/ArcClientStartPage.html?application=Mobile%20Demo&document=MobileCashflow.ap
Mobile finance report. Operational result + gross profit + net cashflow from operating activities + occupancy costs + labor costs + other
Rating: ★★★★★ (2) Location: Mobile Demo

Bullet graphs Vega.Admin 1/20/2011
https://engage.arcplan.com/ArcClientStartPage.html?application=ChartApplikation&document=REPBulletGr
Chart gallery: bullet graphs. A bullet graph is a variation of a bar graph developed by Stephen Few. Seemingly inspired by the traditional thermometer charts and progress bars found in many dashboards, the bullet graph serves as a replacement for
Rating: ★★★★★ (2) Location: ChartApplikation

Waterfall charts Vega.Admin 1/20/2011
https://engage.arcplan.com/ArcClientStartPage.html?application=ChartApplikation&document=REPWaterfal
Chart gallery: waterfall charts. Waterfall Chart is a form of data visualization which helps in determining the cumulative effect of sequentially introduced positive or negative values.[1] The waterfall chart is also known as a Flying Bricks Chart due to
Rating: ★★★★★ (1) Location: ChartApplikation



Bullet graphs

Chart gallery: bullet graphs. A bullet graph is a variation of a bar graph developed by Stephen Few. Seemingly inspired by the traditional thermometer charts and progress bars found in many dashboards, the bullet graph serves as a replacement for dashboard gauges and meters. Example usage: monitor a cashflow.

Location: ChartApplikation
Author: Vega.Admin

Show all ratings and statistics >>
Rating: ★★★★★ (2) My rating: ★★★★★

Recent comments
michael.dieterle, 25 days ago: ★★★★★
Super Visualisierung
Stefan.Spittank, about a month ago: ★★★★★
Nice work!

Figure 33. arcplan Engage as an example of social media style search and evaluation of reports, at the top: result of a search, at the bottom: selection of the second displayed report „Bullet Graphs“.

finally complemented by state of the art search methods and technologies. This enables search on both, content and its meta data.

“Mobile BI” (cf. chapter 2.7) provides another aspect. Trusted and device specific visualization of BI results on mobile devices is useful per se, but mobile BI means more. Indeed, social media are orientated to the mobile internet. Hence, mobile BI also comes with all collaborative functionality. Mobile BI users are always on. Wherever they are, they can participate in the wisdom of the crowd. New ideas are born in a faster and smarter way. Creativity of all participants is stimulated, but also requested. The result is a “common decision making”, true cooperation and collaboration in a team.

Knowledge Management. It has two tasks, the person to person transfer of knowledge as well as the documentation of knowledge. In a first approach, it is about bringing knowledge into the heads of all employees, but more important, it is about extracting knowledge from the experts so that it can be shared by everybody. In the past, knowledge management quite often failed due to lack of acceptance of tools. Knowledge management was considered as cumbersome duty. Initiatives for knowledge management were fizzled out. The first upcoming web 2.0 tools changed the situation, and today, knowledge management can find the necessary acceptance, if managed properly. It works best, if the employees do not at all notice that there are engaged in knowledge management. We have seen first successes with Wikis. These are good and appropriate tools for widely accepted knowledge management as experience shows in many enterprise that are moving towards a social business culture. The idea is: everybody actively participates, everybody gets involved turning the organization into a community. Then, communication flows à la Facebook. Today’s social media tools are Big Data provides a huge potential. But the exploitation of Big Data and to turn Big Data into “Big Knowledge” is by far not easy. The mix of the enormous amount of fragmented data makes it difficult to identify relevant data, to extract, to store, to administrate and to analyze it. This requires new approaches and new technologies. Traditional IT tools for data extraction and integration do not really help. Innovation in information management is needed. This means new tools for an agile integration of enterprise, web, and cloud data. The requirement is fast and flexible unlocking and benefitting from relevant Big Data sources. Here, we encounter another challenge: Not all Big Data sources to have sufficient APIs, many do not have APIs at all. Data extraction tools that can even extract data without APIs are needed! Those who are first in benefitting from “Big Knowledge” will be the winners. Knowledge is power, and time-to-market is crucial. Our opinion is shared by Gartner Group that believes, “organizations with an information management infrastructure for Big Data will be the future market outperformers[28].” We describe and discuss agile integration tools in chapter 6.4.

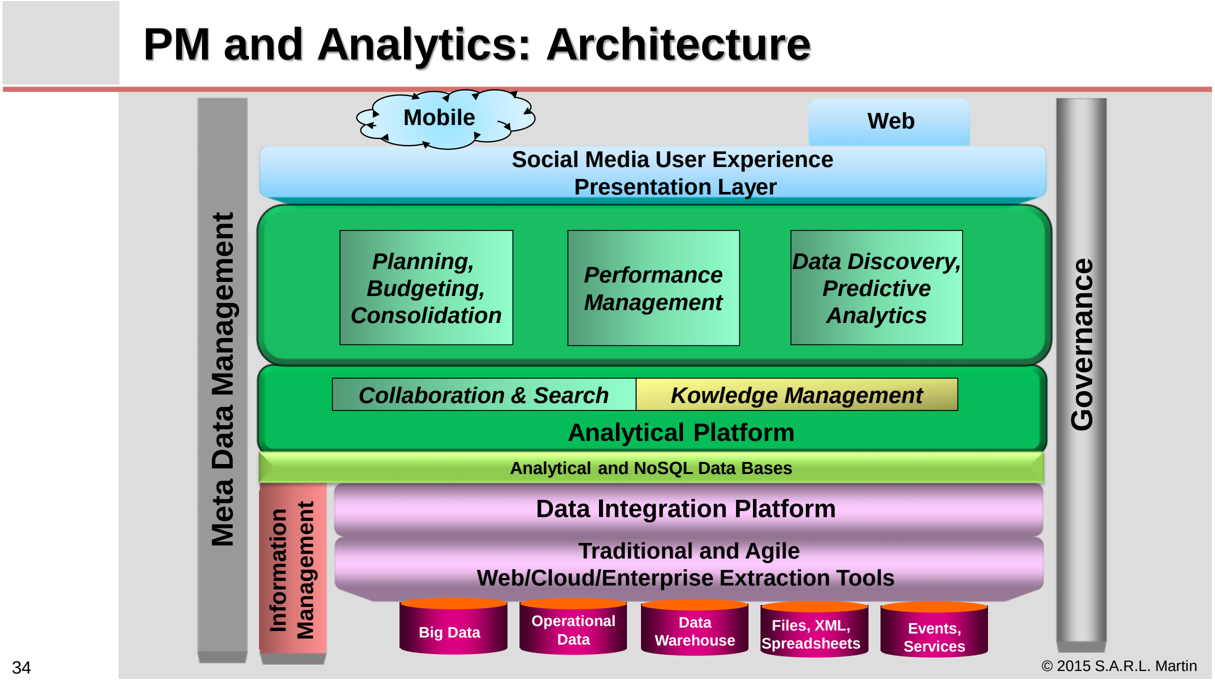
Corporate management is put on the facts revealed by analytics, and on the wisdom of the crowd presenting the knowledge of all. As a result, Social BI acts and works as traditional BI in the past should have acted and worked.

New Technologies. Performance management and analytics architecture consists of three layers (fig. 34). Up to now, the traditional BI portal layer presented the top layer. It is now redefined by a social media platform. In the layer below, we find the analytics platform with

[28] Gartner presentation “The Grand Challenges of Information: Innovation to Make Your Infrastructure and Users Smarter,” Bill Hostman and Mark Beyer, Oct. 2010.

the tools. Information management with its extraction tools and databases presents the bottom layer. Let us start the discussion with the information management layer.

Big Data does not only denote the huge volume of data, but also a mix of structured and poly-structured data linked via complex relationships. Data sources in the web are manifold: portals, web applications, social media, videos, photos and much more, all kind of web content. Data from the mobile internet and the Internet of Things is to be added. Besides the mere deluge, the other problem about Big Data is the variety and multiplicity of sources: The access to a Big Data source has to be redefined each time. This ends up in a time and resource problem.



PM and Analytics: Architecture

Mobile

Web

Social Media User Experience
Presentation Layer

Planning, Budgeting, Consolidation

Performance Management

Data Discovery, Predictive Analytics

Collaboration & Search

Kowledge Management

Analytical Platform

Analytical and NoSQL Data Bases

Data Integration Platform

Traditional and Agile
Web/Cloud/Enterprise Extraction Tools

Big Data

Operational Data

Data Warehouse

Files, XML, Spreadsheets

Events, Services

Meta Data Management

Information Management

Governance

34

© 2015 S.A.R.L. Martin

Figure 34. Big Data drives the consequent continuation and extension of performance management and analytics. The architecture consists of three layers. Information Management makes up the inferior layer. Here, we have new tools. Traditional ETL tools have been complemented by agile extraction tools allowing the extraction of data even if there is no API. This is important, because not all sources do have APIs or do have sufficient APIs. The analytical layer is in the middle. Today, analytics is no more restricted to relational data bases, but uses also analytical and NoSQL data bases. This layer includes all central analytical services like analytical workflow and rules management, collaboration and search services, knowledge management and the central administration of all resources. On top of this analytical infrastructure, we have the tools for planning, budgeting and consolidation, performance management as well as data discovery and predictive analytics. The presentation layer is the superior layer. It can be accessed either by web browsers or mobile apps. It provides a social media user experience. Furthermore, there is end-to-end governance and a central integrated meta data management.

Big Data not only drives agile web and cloud integration, but also the adoption of innovative database technologies for analyzing the peta and exabytes. Such **“analytic and NoSQL databases”** are discussed in chapters 7.2 and 7.3. We are seeing that more and more analytical platforms are running on these database technologies. It goes hand by hand with another innovation: Analytical tools move more and more functionality and algorithms directly into the databases because of performance advantages. The old paradigm “data to compute” turns into “compute to data”.

Tools for data discovery and predictive analytics now use new methods and procedures. Data visualization and location intelligence have made big progress. Furthermore, an evolution has taken place from data mining via text mining to **text analytics** (cf. chapter 5.7). Again, social media is the main driver for text analytics. Text analytics provides exactly the type of analytics that is required for analyzing large volumes of poly-structured data. But many organization successfully applying text analytics, keep as a top secret, because text analytics can create the ultimate customer transparency. In combination with enterprise data, we do not only get the famous 360° degree customer view, but even a 360° degree view on the total market: Social media mirror the total market with all market players.

New application areas. Just like performance management and analytics worked in the old world before the New Normal, we now get performance management and analytics applied in the social media context**: Social Media Performance Management** uses the same closed loop model as traditional performance management. Leading B2C organizations have already deployed **social media monitoring**. Its task is to sniff where, when and how in social media, an enterprise, a person, a product or a brand is talked about and discussed. Based on such a social media monitoring, the loop can be closed by **social media analytics** and **social media interaction.** An organization now can immediately react to relevant contributions and opinions in social media, and can intervene. This creates advantages in customer service or when introducing new products to market, because a communication can be built and sustained within social media communities. We already presented an example in chapter 2.4.

Take away. Big Data drives the evolution of performance management and analytics. Social media style user interfaces considerably ease the use of performance management and analytics and improve collaborative concepts. Social media style concepts also enable performance management and analytics to be complemented by knowledge management. Progress in data visualization and new analytic methods and tools like text analytics, agile web and cloud integration, as well as analytic and NoSQL databases extend the scope of analytic solutions. Organizations, for instance, now can benefit from a 360° view on the total market, because Big Data can be turned into Big Knowledge. New insight is provided into customers, competitors, and brands: New potentials can be reached.

6 Information Management

Information Management organizes the handling of information in an organization. It takes care of both, the utilization of internal and of external information that is necessary for achieving the organization's goals.[29]

Information management became a topic in traditional BI, when it was about to fill and to refresh a data warehouse. Solutions have been and will be extraction, transformation and load (ETL) processes. The goal was to create **trusted data**, and to make the data warehouse the "single point of truth".

In the course of process orientation, information management became much more important in the context of performance management and analytics. It also broadened its scope, and today, it is the foundation to supply processes with data and to deliver data for performance management and analytics, even in real-time. This changes the role of a data warehouse as we have already sketched out (Fig. 7). Information management now creates trusted data beyond the data warehouse. The single point of truth now shifts to meta and master data as we will point out later.

6.1 Scope of Information Management

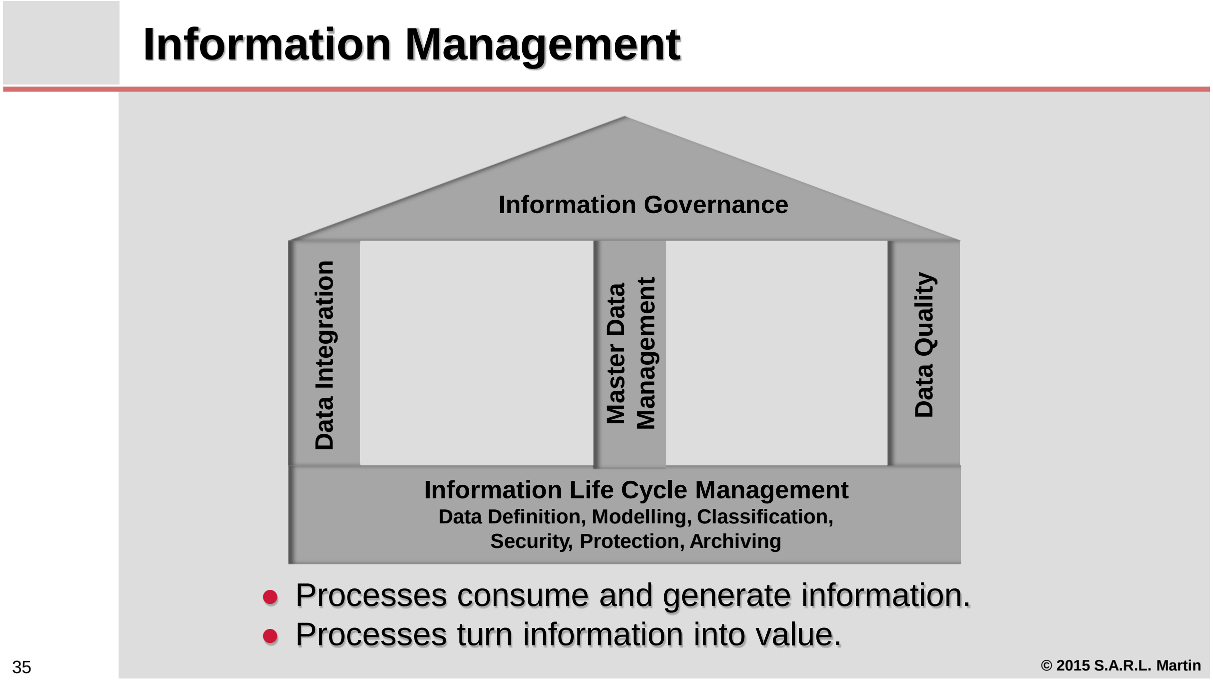
As already discussed in chapter 3.1, information management is the foundation of process orientation in organizations and in networks of organizations. Indeed, it is especially important when organizations are collaborating with suppliers, partners, and customers and are running common processes across border lines. Let us have a closer look to the scope of information management (fig. 35).

- ***Data definition per business vocabulary.*** The business vocabulary (also called "business glossary") plays a central role in a process oriented enterprise. For, processes need a unique and common terminology for modeling and communicating with all participants: employees in business and IT, suppliers, partners, and even customers. The business vocabulary represents the terminology of all business items and business knowledge within an organization and within a network of organizations. This is why the creation of a business vocabulary cannot be delegated to IT. If a business vocabulary is missing, then an organization will suffer from a "Babylonian" chaos. The impact is even well visible at board level. Terms like revenue, contribution ratio or returns have different meanings in different specialist departments, and often do not fit and match at board level. Other example: a question "who is our customer?" cannot be answered without

[29] translated from: Universität des Saarlandes, Fachbereich 5.6 Informationswissenschaften, Aufgaben des Informations-managements, accessed 22nd of January 2014. http://www.uni-saarland.de/campus/fakultaeten/fachrichtungen/philosophische-fakultaet-iii/fachrichtungen/informationswissenschaft/infowissthemen/wissensinformationsmanagement/aufgabendesinformationsmanagements.html

doubts. A good starting point to build a business vocabulary is given by predefined business vocabularies that offer standardized terminologies for different verticals.

- ***Data modeling.*** Relations between technical terms defined in the business vocabulary are captured in a data model. So, data modeling builds the semantics of a business. Critical success factor for data modeling is a business/IT alignment: Data modeling is a common and joint task of specialist departments and IT.



Information Management

Information Governance

Data Integration

Master Data Management

Data Quality

Information Life Cycle Management
Data Definition, Modelling, Classification,
Security, Protection, Archiving

- Processes consume and generate information.
- Processes turn information into value.

35

© 2015 S.A.R.L. Martin

Figure 35: Technical components of information management. The foundation of information management is information life cycle management. Based on this foundation, the three pillars are data integration, master data management and data quality management. Information governance provides the all-encompassing catenation connecting the processes of information management with the organization and its roles.

- ***Meta and master data management.*** Before the age of process orientation, organizations were application oriented. This has created a data and process fragmentation problem, since each application has its own processes, and its own meta and master data. Today, information management solves this problem by using information services to map isolated application meta and master data to central meta and master data managed by a **repository**. This is critical when defining new products, gaining new customers, and/or adding new suppliers to the enterprise's network. A simple update via a meta or master data service synchronizes securely and reliably all affected applications. We will discuss this issue in more detail in chapter 6.5.

- ***Data quality management.*** Data quality is another critical success factor of information management. Data quality means the completeness, relevance and accuracy of data. A best practice for data quality management is **total quality management (TQM)**: Build data quality into the processes from the very beginning. We will discuss data quality management in chapter 6.6.

- ***Data integration.*** Process orientation is based on end-to-end processes that are application independent. Processes go beyond application silos and even reach across the organization. Consequently, interfaces with applications and internal and external data have to be integrated and to be synchronized. In chapter 6.2 we will discuss data integration in the context of performance management and analytics.
- ***Data classification.*** A data classification means a partition of data into categories according to a criterion so that data within one category have the same properties. Data classifications can be derived bottom up by data / text mining, for instance. But they can also be defined in a top down approach by matching categories to enterprise goals. Top down data classifications are especially useful to support storage strategies and life cycle management.

  Example. In a top down data classification, a category could be "confidentiality" that consists of classes like "strongly confident", "confident", "internal", and "public". Other categories in this example could be accessibility, retention period, auditing acceptability etc.
- ***Data archiving.*** Data archiving is the process of rolling no longer actively used data out to a data storage device for long-term retention. Data archives consist of older data that is still important and necessary for future reference, as well as data that must be archived for compliance requirements. Data archives have search functions so that files and parts of files can be easily located and retrieved. Data archives should not be confused with data backups for restoring data in case it is corrupted or destroyed.
- ***Information life cycle management.*** Information management encompasses management of information over its whole life cycle. It starts with the creation, i.e. generation and capture of information. It continues with the phases of information preparation and deployment. Furthermore, information must be persistent. Thus, information has to be stored in data storage devices. Next phase is about linking of information to context and processing for creating knowledge. Then, phases of information distribution and usage follow up to the end of life time which is either archiving or deleting of information.
- ***Data security and protection.*** Finally, legal issues have to be respected when we manage data and information. The issue is to protect all enterprise data and to prevent unauthorized usage, malpractice and transmission. Similar to data quality management, a total quality management approach is best practice: data security and protection should be built into the processes from the very beginning.
- ***Content Management (Web, rich media).*** The much larger part of enterprise data is poly-structured (documents, contracts, letters, memos, e-mail, images, graphics, maps, videos, audios etc.). All this type of information has to be managed. So, information management is about management of structured and poly-structured data.

Take away: Information Management creates trusted data. It solves the problem of data fragmentation. Therefore, it is the prerequisite of process orientation. It is a joint and common task of business and IT, and because it is a critical success factor for industrialization, agility, and compliance, it should have top priority and should be sponsored by a board member.

6.2 From the Data Warehouse to Data Integration

Definition: **Data integration** is defined as merging and fusing data from various data sources of an organization and from Big Data into one common data model, where sources typically have different data models and structures. The goals of data integration are the creation of a trusted data foundation for analytics and performance management as well as empowering data discovery across data that does not necessarily comply with enterprise standards.

A data warehouse is a known concept from the 90s for a subject-oriented, integrated, time-related and sustainable storage of information for tactical and strategic decision support. It is a warehouse of data separate from all data in the operational IT systems. In the past, data warehouse data was mainly derived from transactional data, and often enriched by external data. For example, combining enterprise data with additional information provides a better foundation for data mining programs (see fig. 29). A data warehouse used to be supplied by ETL or ELT processes, mostly in batch, but sometimes also in real-time by a data integration platform. In either case, it served as the “single point of truth” for performance management and analytics.

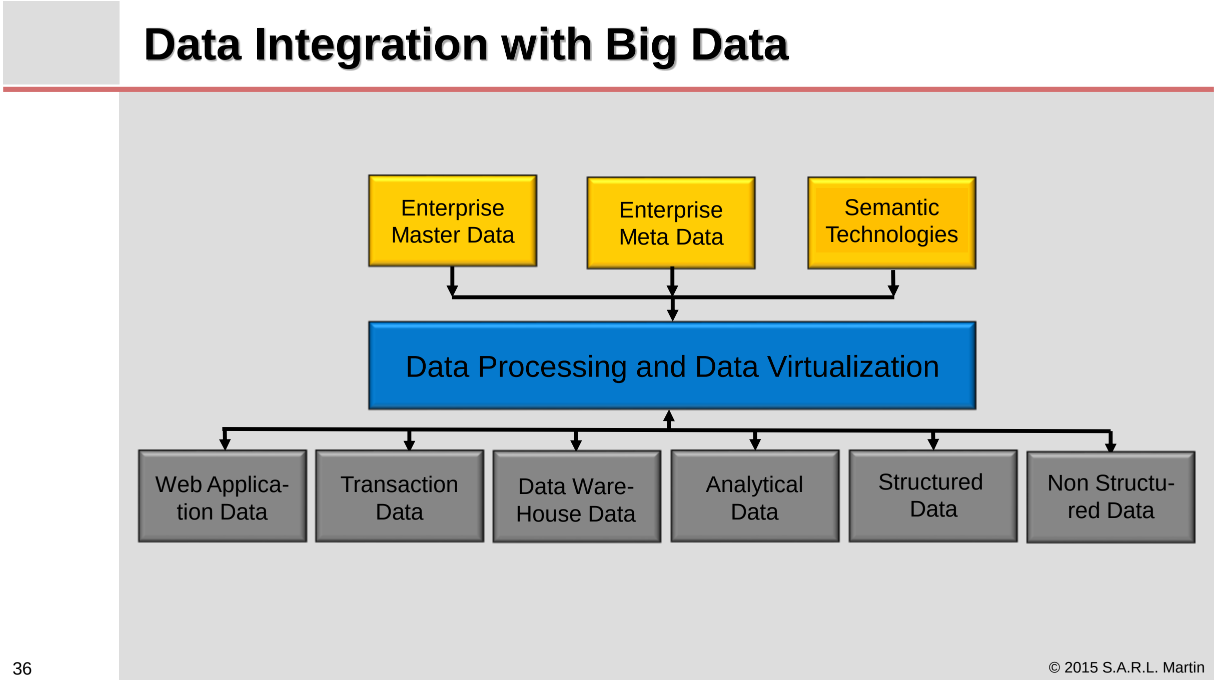
Big data now changes this situation considerably. Enterprise data and traditional external data are no more sufficient for answering the fundamental questions of business: Who buys or uses or products and services when, how, where and why? These questions can be put across various dimensions: geography, channels, campaigns, interactions etc.

> ***Example:*** Optimization and personalization of promotions and improving up- and cross-selling by better knowing customers and markets. In retail, big data fore-runners like Amazon and eBay have adopted these strategies since long. In the meantime, these best practices have been deployed in many vertical markets. Latest examples can be experienced in social networks where friendships or relations are proposed. In insurance, for instance, policies are fine-tuned to customer profiles based on data sources like risks, wealth, or location data. Vehicles can be equipped with special emitters so that in case of theft, they can be detected: In consequence, risk can be considerably reduced, and that means money for the insurer and the customer.

Big Data provides not only infinitely many additional and even completely new data sources, but also new data types like XML, clickstream data and poly-structured data from social media, server logs, machine data and many others. Now, the “old” data warehouse with its relational database technology gets challenged by new technologies and progress in technology: virtualization, cloud computing, Hadoop and NoSQL database technologies.

Hence, we need a new generation of Data Warehouses capable to integrate all the various Big Data sources and to master the corresponding data volumes. In fact, data integration is becoming the new data warehouse platform. The “old” Data Warehouse is transforming to just one component of the data integration platform. Data integration now becomes the central platform providing the “single point of truth” and it replaces this old role of a Data Warehouse. It gets also the platform for Data Discovery across data from various sources that do not necessarily correspond to such a single point of truth.

For developing such architecture, it is recommended to start with a categorization of all data types according to the characteristics of data and its requirements for processing. Afterwards, the processes of integration are to be modelled and to be implemented. As always, this needs a process and a rule engine to support the logic and flow of processing. Such an integration process finally combines Big Data with enterprise meta and master data as well as semantic technologies and taxonomies (fig. 36).



Data Integration with Big Data

Enterprise Master Data

Enterprise Meta Data

Semantic Technologies

Data Processing and Data Virtualization

Web Application Data

Transaction Data

Data WareHouse Data

Analytical Data

Structured Data

Non Structured Data

36

© 2015 S.A.R.L. Martin

Figure 36: Architecture of a data integration platform. Data integration works bi-directional. Incoming data are fused and integrated with meta and master data as well as semantic technologies via ETL/ELT/CDC processes or text processing, and finally loaded. This data presents the trusted source for analytics and performance management and is deployed via information services (cf. fig. 17).

.

Let us have a closer look at the data categorization:

- *Transactional data.* The traditional OLTP data belongs to this segment. In contrast to traditional data warehouse architecture, this data is now directly available for analytics and performance management.
- *Web application data.* This data includes clickstream data, e-/m-commerce data, and data about customer relationships as well as call center chat data etc.
- *Data Warehouse data.* The "old" Data Warehouse and its data marts now play the role of a component of the new architecture provided by the data integration platform. It should include all the different data warehouses and data marts in the organization where data is processed and stored for use by business users.
- *Analytical data.* This is data from analytical systems that are deployed currently in the organization.
- *Poly-structured data.* Under this broad category, we can include:
    - Text: documents, notes, memos, contracts.

- Images: photos, diagrams, graphical presentations.
- Videos: corporate and consumer videos associated with the organization.
- Social media: Facebook, Twitter, Instagram, LinkedIn, Forums, YouTube, community websites, blogs etc.
- Audio: call center conversations, broadcasts etc.
- Machine data: includes data from sensors on any or all devices that are related to the organization's line of business.
- Weather data: has become a vital component of predictive analytics and is used by both B2B and B2C businesses today to analyze the impact of weather on customer behavior and sentiment.
- Scientific data: applies mainly to medical, pharmaceutical, insurance, healthcare and financial services segments.
- Financial market data: used for processing financial data in many organizations to predict market trends, financial risk and actuarial computations.

- *Semi-structured data.* This includes emails, presentations, mathematical models, geospatial data etc.

With the different data types clearly identified and laid out, the data characteristics can be defined clearly. This includes the data type, the associated metadata, the master data elements, and the use of data. The later also describes the business users of the data from an ownership and stewardship perspective.

> ***Example.*** Enrichment of enterprise data by **social media data**. Social media data is a rich source for better knowing customers. The problem is that social media data either lacks of meta data or uses different meta data than the organization. For example, customers are hidden by pseudo or wrong identities. Thus, the question is how to match enterprise data with social media data. Such a problem is tackled by identity resolution. In the past, we already deployed identity resolution for removal of address data duplicates. But in Big Data, the problem is a bit more complex, because there are various social media sources, various languages with various characters and various transcriptions. That makes it difficult to derive a social profile of customers and to match it with the enterprise profile. We will discuss the various methods and technologies of identity resolution in more detail in chapter 6.7 on “Entity Identity Resolution”.

The challenges of an implementation of the new generation data integration platform consist of loading and availability of data, managing the data volume, performance of storage, scalability, load of analytics and performance management, and finally of operational cost. Today, loading and storing of data follows Hadoop concepts (Hadoop is discussed in chapter 7.3): Load and store data one-to-one without any transformations. Other challenges are addressed by **data virtualization**. Here, the integration is performed when accessing data. The foundation of data virtualization is a logical data model (canonical schema). It provides both, the interface to the data sources and their data models and an integrated central read-write interface to the federated data for accessing services via information services.

Definition: **Data Virtualization** means a virtual (logical) access to data via a data abstraction layer. Access to data is centralized avoiding data replication and duplication.

Data virtualization uses abstraction of location, storage, interface and access. All data operations are executed by logical views. Result sets are provided as views or information services at user request. The information services may include other services for data preparation and enrichment, for instance data quality checks for validations. Data virtualization is a further development of data federation, formerly called Enterprise Information Integration (EII).

Data virtualization is well designed for real-time analytics and enables zero-latency data integration, i.e. analytics is executed synchronously with transaction data. But up to now, such a solution was rather expensive due to performance requirements on the network and hardware infrastructure. Today, in-memory processing offers alternative and less expensive solutions. But even if in-memory processing is attractive, zero-latency is not needed in many analytic scenarios.

> ***Example:*** In chapter 2.3, we already discussed the business metric *product availability = <stock - threshold of stock>.* This is an operational metric that can be used to re-order automatically if this metric is negative or zero. It is a pro-active metric measured in real-time to avoid the risk of being sold out. But what is the meaning of “real-time” in this example? In today’s practice, *product availability* is typically measured twice a day (cf. chapter 3.1). This is an empirical experience balancing cost of measuring with cost of risk ignoring *product availability* for controlling the order process. With in-memory processing, we can certainly improve this practice and calculate *product availability* hourly or even faster, but doing so in each order process instance each time when a customer puts an item into his basket will still be overkill.

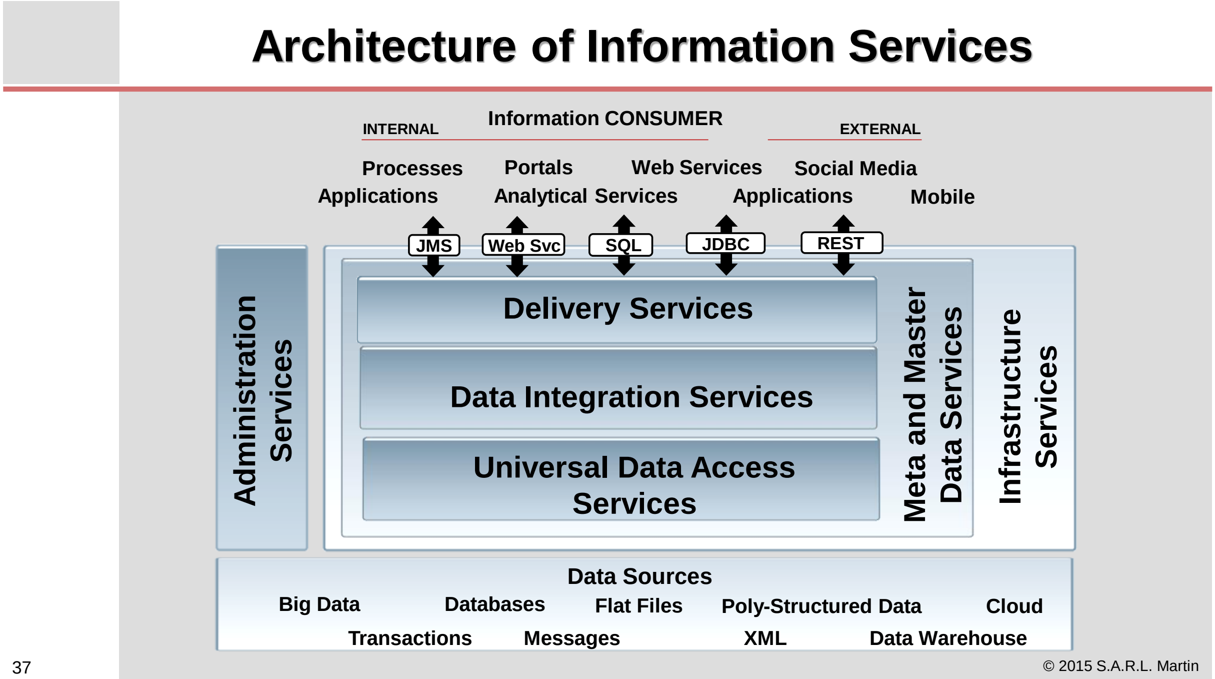
A low latency solution is much less expensive and mostly sufficient. Therefore, it is important to find out what is the tolerated latency in a given business process, because latency is related to costs: the lower the tolerated latency the higher the cost.

In the low latency model, all relevant transactional and analytic data is collected and stored in a so-called low-latency data mart (LLDM). This requires integration of the data integration platform with the enterprise service bus (ESB) where the processes and services across all backend applications are managed. The LLDM is refreshed either by message queuing or by batch, where the batch is executed in short periodicities according to the tolerated latency (e.g., hourly etc.). The LLDM can be used for low latency data propagation, a feedback loop for triggering events in operational systems via cross-process metrics. This coupling with operational systems requires managing the data integration platform like the ESB: The data integration platform is an operational system.

6.3 Information Services

We have already met information services as a SOA service model (fig. 17), and in the previous chapter we have discussed information services as an access method for data virtualization. Let us start with a more detailed definition of an information service.

Definition: An **information service** is a modular, reusable, well-defined, business-relevant service that enables the access, integration and right-time delivery of structured and poly-structured, internal or external data throughout the enterprise and across corporate firewalls. An information service can be a meta data service, a master data service, or a data service.



Architecture of Information Services

Information CONSUMER

INTERNAL

EXTERNAL

Processes

Portals

Web Services

Social Media

Applications

Analytical Services

Applications

Mobile

JMS

Web Svc

SQL

JDBC

REST

Administration Services

Delivery Services

Data Integration Services

Universal Data Access Services

Meta and Master Data Services

Infrastructure Services

Data Sources

Big Data

Databases

Flat Files

Poly-Structured Data

Cloud

Transactions

Messages

XML

Data Warehouse

37

© 2015 S.A.R.L. Martin

Figure 37. Information and data should be deployed by services. Information services include six different categories. Universal data access services provide service-oriented access to enterprise data or big data sources. Data integration services provide any type of mapping, matching, and transformation. Delivery services publish information to any information consumer – internal or external. Meta and master data services provide the common business vocabulary. Infrastructure services look to authentication and security. Administration services provide the functionality for administrators, business analysts and developers for managing the life cycle of all services.

Given the definition of an information service, the next step is now to look at the needs of information service consumers to identify the different categories of information services and their architecture (fig. 37).

- ***Universal Data Access Services.*** Data access services consist of the basic CRUD services for creating, reading, updating and deleting data from any backend systems, structured or poly-structured, internal or external. They also provide zero and/or low latency access to federated data, i.e. data virtualization.

- ***Infrastructure Services.*** Infrastructure services include basic functionality around authentication, access control, logging, etc.

- ***Data Integration Services.*** Integration services move data from source data models to target data models like synchronization, transformation, matching, cleansing, profiling, enrichment, federation, etc.

- ***Meta and Master Data Services.*** Their purpose is to manage and use the technical and business metadata and master data for audit, lineage, and impact analysis purposes***.***

- ***Data Delivery Services.*** They automate and standardize the publication of information to all consumers according to a request/reply model or a publish/subscribe model (data syndication). Delivery mechanisms are bulk and/or single records by batch, real-time messaging or delta mechanisms bases on change of data.
- ***Administration Services.*** These are services for the life cycle management of the other services, i.e. development, management, and monitoring and controlling.

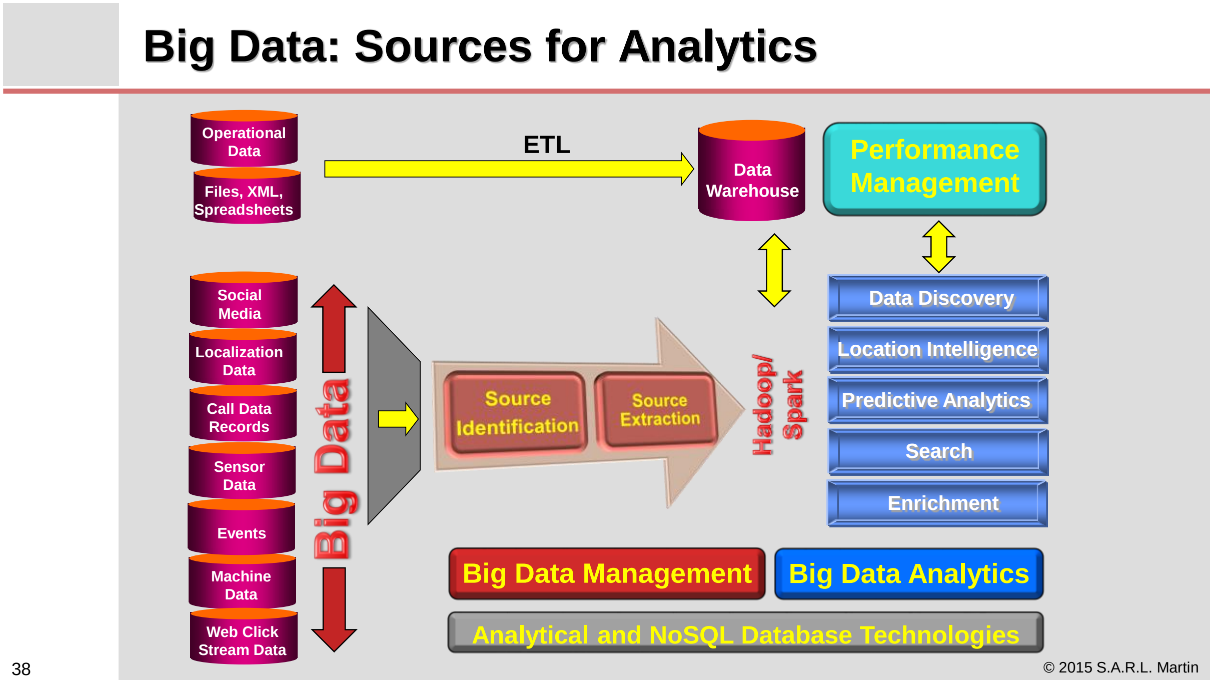
The model of service-orientation provides another advantage. Due to the sub-service principle, composite information services can be built for any purposes by mashing up. Typical examples include data warehousing, data migration, and data consolidation processes. We will discuss other examples like master data management and data quality management in the following chapters.

6.4 Agile Integration and Extraction Tools for Big Data

Big Data has quite an impact to enterprises: data is becoming more and more the driver for running the enterprise. We need data to get information and to finally turn it into facts, knowledge and competitive advantages. Time-to-market is vital. More than ever, access to and extraction of data also in real-time is getting the critical success factor in the New Normal. Now, we have data ad abundance, in the web, in the cloud, in the enterprise. But how can we move data rapidly, flexibly, reliably, and if necessary in real-time from the web and from the cloud into the information management infrastructure of an enterprise. How to move it into the cloud or from one cloud into the cloud of another provider?

An answer and a solution are provided by **agile integration tools**. They can extract and integrate enterprise, web, and cloud information in parallel and simultaneously from Big Data. They come with a new approach for data access and extraction. Agile integration tools work in a browser style and always use one and the same visual interface for all data sources (fig. 38).

That is a new approach and beyond what traditional IT tools for data extraction can do. Agile integration tools can exploit visually all data sources in the web, in clouds, and in the enterprise without a predefined interface and without programming. This is another advantage, because interfaces often are not available, and have to be specified and programmed. If interfaces are available, there are sometimes not reliable, difficult to use, or they do not provide what is needed. Furthermore, interfaces quite frequently restrict data access: Not all data is accessible. But a browser based integration tool provides unrestricted access to all visible data even in real-time. So, complete data sets can be accessed and extracted without the additional time and cost of programming interfaces. The interface is created on the fly by specifying visually and in a browser style what is to be extracted. Indeed, this extraction is not restricted to data, but any information from any web application can be extracted. Agile integration tools can access all layers of web applications and extract information. Another advantage, this visual interface is always the same for all sources.



Big Data: Sources for Analytics

38

© 2015 S.A.R.L. Martin

Figure 38. Big Data means both, the data deluge, and the variety of various sources in the web that typically do not have interfaces, or if interfaces are available, these interfaces only provide a restricted access to data and functionality. Hence, Big Data Management needs new approaches to web extraction like browser-based agile web extraction and integration tools that can extract all data that is visualized by a browser. Furthermore, semantic search engines support the identification of relevant Big Data sources. Extracts from Big Data are loaded into Hadoop and are integrated and analyzed through Spark. Hadoop and Spark are more and more becoming the platforms for Big Data, resp. Big Data Analytics. We will discuss the relationship between Hadoop, Spark and the Data Warehouse in chapters 7.4 and 7.5 in more detail.

Due to the visual specification, agile integration tools are well suited for joint IT/business cooperation. A joint team of an IT and a business expert can do the job of specification in a fast and in the end agile way. First task is the specification of the web sites where information should be extracted. This can already be partially automated by semantic search engines. They can identify the relevant sources based on semantic search patterns. After identification of the relevant sources, the extraction is visually specified. The tools then work like micro workflow controlled robots. The workflow can contain rules and loops.

Comprehensive workflow logic can be built so that even the most complex extraction scenarios can be set up without programming. Robots can be planned and controlled by a management console. They can operate in any speed, even as slow as a human visitor of the web page. This prevents detection of extraction robots by web masters. The final result is a reliable and trusted extraction of any information.

Today’s agile integration and extraction tools also come with built-in intelligence for extracting data from dynamic web pages. When certain positions of data to be extracted have changed, the robots can even detect these changes within certain limits. Even if such a change is out of scope of the robot’s intelligence, it can at least send an alert for fixing the problem by a human interaction. For vendors offering agile integration and extraction tools please see chapter 10.5. These tools are also very valuable in B2B collaboration. Examples are price comparisons and automated information exchange between portals of different players in a supply chain. Such solutions are offered by Lixto, for instance, Brainware offers a special solution for invoice processing, and Kapow Software has developed six solution scenarios for application areas of agile integration tools[30].

From a technical point of view, data extraction from Big Data is not the problem. It is rather automated, robust and not too difficult to be implemented. But there are legal issues. For publically available data on web sites, there could be a prohibition for automated extraction stated in the general terms and condition. Hacking protected data in social media is definitely a crime, and for extracting public data from social media, a customer opt-in is strongly recommended. But here, we do not like to enter these legal questions. They are out of the scope of this white paper.

6.5 Meta and Master Data Management

Meta Data. The foundation of process-orientation is meta data management. Meta data spans across all layers of a SOA, and it is crucial for a consistent data model including life cycle management. The objectives are an all-embracing comprehension and communication of data model, data quality, data protection and security.

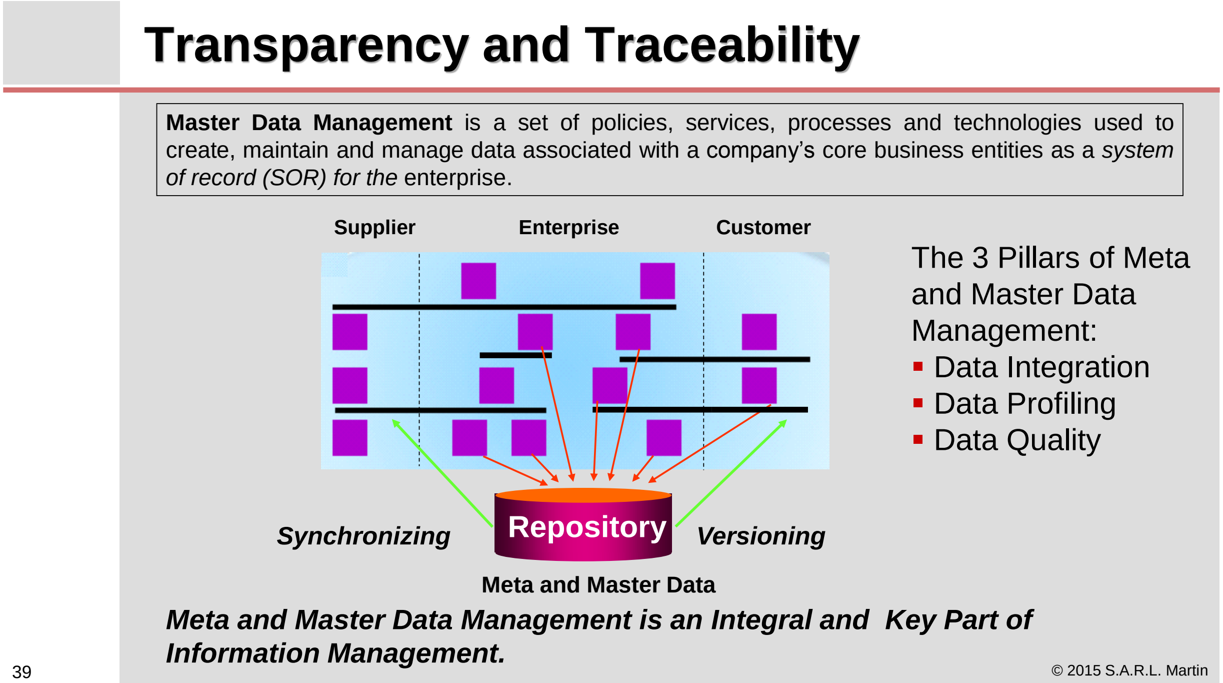
Meta data is presented by three layers:

- **Layer 1 – Master Data:** This is business oriented meta data providing the foundation of the business vocabulary. Master data is meta data describing business structures like assets, products and services, and the business constituents (e.g., suppliers, customers, employees, partners etc.) The goal is to provide a single view on all enterprise structures.
- **Layer 2 – Navigational Meta Data:** This is also business oriented meta data describing the information flows (e.g., sources and targets of data, cross references, time stamps)
- **Layer 3 – Administrational Meta Data:** This is organizationally oriented meta data defining the roles and responsibilities of all involved users (information profiles including responsibility, security, monitoring and controlling usage etc.)

Meta data provides the ***single point of truth*** that was traditionally claimed by the data warehouse. Today, this single point of truth is established through data integration. The **business vocabulary** plays the central role, because business processes need a common and uniquely defined terminology for modeling and for enabling a common understanding across all constituents, i.e. staff in the specialist departments, suppliers, partners and even customers. They all have to communicate using one and the same terminology. This implies that the business vocabulary controls both, BPM and performance management, processes and metrics. No processes, rules, and metrics without data.

Master Data describes all data objects related to processes, rules, and metrics. It consists of the business-oriented meta data of layer 1, and it represents the structures of an organization (fig. 39): customers, suppliers, dealers, partners, products, services, staff, and assets, in short, everything that composes an organization.

[30] see Research Note on Kapow Software http://www.wolfgang-martin-team.net/research-notes_dt.php



Transparency and Traceability

Master Data Management is a set of policies, services, processes and technologies used to create, maintain and manage data associated with a company's core business entities as a *system of record (SOR) for the* enterprise.

The 3 Pillars of Meta and Master Data Management:

- Data Integration
- Data Profiling
- Data Quality

Meta and Master Data Management is an Integral and Key Part of Information Management.

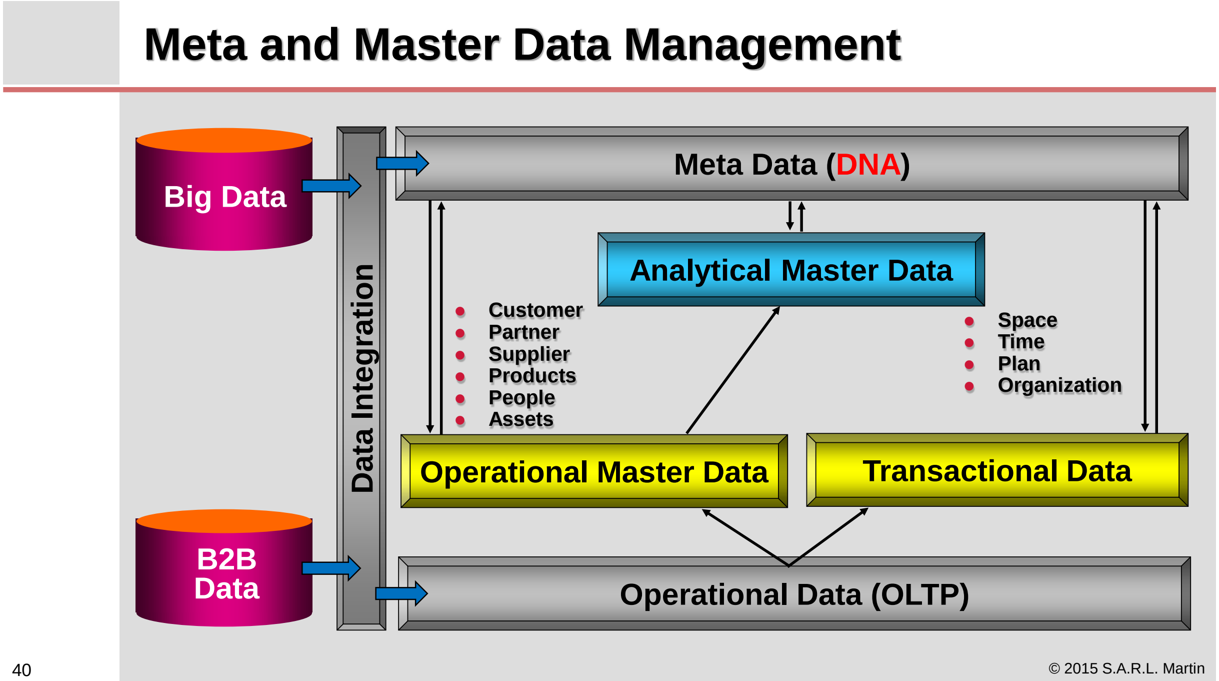
39

© 2015 S.A.R.L. Martin

Figure 39: Meta and master data management means to establish information services for synchronizing and versioning all meta and master data across all backend applications. The center of master data management is a repository that manages the common business vocabulary so that all processes can use unique and common structures and terms. Best practice architecture for such a repository is a hub & spoke architecture corresponding to the architecture of an ESB: meta and master data management processes should be also SOA based. The three pillars of MDM are discussed in chapters 6.2 (data integration), 6.6 (data quality management), and 6.8 (data governance).

Operational and Analytic Master Data. (Fig. 40) Operational master data is part of transaction data, where transaction data consists of (operational) master data and inventory data. The various types of operational master data can be derived from the basic structure of an enterprise. These are all resources (objects, systems and people) that are involved in executing and managing the business, i.e. processes, products and business constituents (customer, supplier, dealer, employees). Analytic master data is derived from performance management and the process ownership model. It represents the principles of measuring and responsibilities: time, space, plan and organizational units (e.g., cost center, cost objective).

Repository and Life Cycle Management. The repository provides a container of all meta and master data. It plays the role of an integration hub for meta data of all back end systems in the BPM model (fig. 7). When services of back end systems are invoked by a business process, then they can only communicate to each other, when they all use the common language based on the business vocabulary. A point-to-point communication would lead into the chaos of isolated silos and fragmentation. Therefore, the meta data model of each back end system must be mapped to the central business vocabulary (cf. fig. 17). Then all back end systems can speak to each other via the central business vocabulary, and adding additional back end systems becomes straight forward, easy, and fast. The organization gains agility.



Meta and Master Data Management

40

© 2015 S.A.R.L. Martin

Figure 40: We derive master data from operational and external data (OLTP – online transaction processing – systems and B2B – business-to-businss – data, i.e. partner and customer data) and we classify master data into operational and analytic master data. Master data is part of the business layer of meta data (D = Definition, N = Navigation, A = Administration).

Meta data and master data are not static. New structures in an organization, i.e. all organizational changes cause master data changes. These can be merger and acquisitions or carve-in and carve-outs. But also market changes, as well as process and policy changes cause changes to the master data definitions and structures. Indeed, any update of a business definition creates new meta data and master data. To summarize: master data has a life cycle. Consequently, it is absolutely insufficient to update meta data and master data and store the most recent and actual version in the repository. **For enterprise planning and auditing as well as for any comparisons between past, now, and future, the availability of the total life cycle of all meta data and master data is a must.** Hence, meta data management and master data management are to be based on life cycle management. Up to now, this is still a weak point, sometimes even a gap in vendor offerings and enterprise architectures. But without a professional meta and master data management, BPM, performance management and Big Data initiatives will fail.

Consequently, master data should be consistent, complete, actual, and correct" for attaining the very "natural" goal of master data management: *"delivering the right product to the right customer with the right quantity for the right price to the right location with the right invoice".* This sounds trivial, but in practice, it is a rather heavy task to be accomplished. Let us have a look into the situation with a typical organization.

Even today, business processes are still supported by business applications. There is an ERP, CRM, SCM, PLM system plus a data warehouse as well as various vertical legacy applications. Even in a midsize organization, we typically find 50+ applications that support the organization. The problem is fragmentation, because each application manages its own master data. As a result, master data is redundant across applications, and it is dispersed. Each application has its own life and life cycle, and develops its own terminology. Master data management gets a nightmare. Customer, product or order numbers in one application do not correspond to those in other applications. The IT department could help here with translation tables for transforming the master data terminology from one application into the terminology of another application. But this is no real solution, because now, collaboration with suppliers or customers becomes a cost driver. Each time, a new customer, a new supplier, a new product is to be created, a new transformation table has to be defined or the new term has to be added to all translation tables. The impact is changes are slow, error-prone, and in particular (very) expensive. Overall, the organization loses agility.

The impact of such a non-professional or even lacking master data management to the specialist departments is even worse.

- The specialist departments suffer from the extra cost for capture and maintenance of master data, when master data is managed redundantly across all applications. Additional extra cost is caused by non-availability of master data. Search for master data becomes a waste of time. This extra cost increases exponentially with the increasing number of deployed applications. Staff is kept busy by mere operational tasks. Time for business critical tasks is lacking. Consequently, new additional staff is necessary: Cost explode.
- Keeping master data redundantly increases data volumes, makes data quality doubtful, and impedes actuality. The operational business suffers from inconsistencies and errors.
    - Processes get error-prone. Substantial extra cost is caused by wrong lots, wrong prices, and wrong invoices. Even worse, error-prone processes cause return shipments, cancelations, and customer dissatisfaction und frustration.
    - Processes run in a deadlock or are aborted. This means a loss of productivity, and dramatically increased extra cost by manual interventions, by loss of time, and potential contractual penalties.
- Deficient master data management also causes inconsistencies and errors on the tactical and strategic level.
    - Inconsistencies in reporting drive wrong management decisions. The board discusses numbers and not the business. Opportunities will be missed.
    - Erroneous analyses can cause strategic mistakes with incalculable consequences.
    - Mistakes lower customer satisfaction and loyalty resulting in declining sales in the mid and long run.

To summarize, a professional master data management is absolutely necessary and the investment in master data management pays off. It avoids process cost that can be monetarily calculated per process. It increases time-to-market that can also be monetarily calculated. Finally, it increases revenues. Hence, monetary advantages of professional master data management can be summarized and turned into a budget.

The Golden Record in Master Data Management. It is the core of professional master data management. It is a master data record that unifies all relevant attributes of all relevant data sources. Therefore, it represents a super set of all attributes of all sources. The golden records are managed in the central repository. Data profiling and cleansing ensure data quality. Identity resolution (see chapter 6.7) matches similar records from different sources to the golden record. Duplicates are avoided, and data quality is further increased.

Due to the necessary versioning of master data, golden records must provide and manage all life cycle information in the repository. Golden records also include all links to all master data records in the various sources used to create the golden record. It guarantees that in case of changes of any attribute in any source this change is also performed in all other affected sources. Consequently, master data remains consistent, and need neither be moved physically nor kept redundantly.

Finally, our goal is reached: All data silos in an organization are synchronized. Fragmented master data belongs to the past. Furthermore, via rules, the management of golden records can be automated, and only if necessary, manual interactions become necessary.

Components of successful master data management. A master data management program is not restricted to technical requirements. Business requirements are the even more important second part of it. Indeed, it is a business-driven, technology-enabled program, and not vice versa! The following components have been proven to be critical success factors:

- ***Goals and vision***. An MDM program must be aligned with the overall business strategy; otherwise it will fail to provide value to the organization. It must be clearly formulated and communicated. It needs a business sponsor representing the goals at board level and anchoring the program within the overall business initiatives. An MDM vision provides the "what" and the "why" of the program.
- **Strategy**. After identifying a clear vision for the master data program, organizations need to create a strategy. This means considering the available resources and understanding the amount of time and money involved in executing an MDM plan. Identify the specific goals the company is trying to accomplish, the level of maturity of your current master data program, and what amount of resources it will take to get to the next level.
- **Metrics**. Before an MDM program is implemented, teams should identify what metrics are most important and most indicative of success. Those metrics should align with the goals and the strategy of the MDM project and, ultimately, of the business. If the goal of the program is to improve customer data, one metric to keep an eye on would be the percentage increase of the accuracy of customer data over time. Master data teams should also try to identify the specific business outcomes that metrics translate into. For a program where the metric is customer data accuracy, then pitches to executives should focus on the way improved accuracy could lend better insight into sales patterns and other business parameters that could influence business strategies.
- **Governance**. An effective governance structure at a large organization usually consists of a sponsor – someone in a position of authority who carries the necessary weight and cross-departmental authority – and of people working at multiple levels to guide work in the right direction. This is absolutely critical to master data success. An effective

governance program also requires a steering committee -- for making MDM governance a reality.

- **Organization.** One of the most important roles in the world of master data is that of the data steward (cf. chapter 6.8 and 6.9). Data stewards play the role of influencers in order to help build a culture that values the high level of data quality necessary in master data. They typically will guarantee the sustainable success of the program.
- **Master data and business processes.** It is key to understand which processes are affected by the changes being made with master data. Unless an MDM program is conceived with an understanding of the organization's business processes, it will not affect the overall operation of the business. There is little business value in a master data program that does not actually influence and improve the business process. The key is to break down the workflow and understand which specific application systems are used in each business process. Then adjust the master data plan to include how the business processes are currently supported and how the MDM program is expected to change and improve over time.
- **Infrastructure.** Getting the master data infrastructure right is absolutely critical. When selecting a master data management platform, teams should first agree on what they need from an infrastructure, weight and prioritize those criteria, and then evaluate the available technology products that can help get the job done.

6.6 Data Quality Management

Which day once in a year is most people's birthday according to all birthday data stored in all databases in the world? Nonsense question? Not at all. The result is striking. It is the 11th of November. Why? Well, if a new customer is to be entered into the customer database, then there are mandatory fields to be filled in and additional fields. Input into mandatory fields is checked (in many situations at least), but input into additional fields is typically not checked. Birthday data unfortunately is stored in additional fields. So what happens? Humans are lazy, and the easiest and fastest way to input a birthday date is "1-1-1-1"….

Enterprises have introduced ERP systems from SAP and others for many millions of euros. One of the drivers was to be more competitive based on all stored data about market and customers. CRM per self-service, coupons, pay-cards, communities, and weblogs are best practices for chasing the budgets of customers. Customer-orientation is the rule. Marketing, sales and service are working together supported by collaborative end-to-end processes. Inbound and outbound campaigns in customer interaction centers and/or in the web shops are enriched through analytics. Today, the demand and customer driven supply chain is reality. But as we have already noticed: Data quality is the prerequisite.

Example. In the apparel industry, people since long collect all sales transaction data in a data warehouse. Customer profiles are calculated, and a demand profile per boutique can be derived. According to these demand profiles, merchandise is individually attributed to all boutiques. As a consequence, a customer will typically find "his/her" products he/she is looking for in "his/her" boutique. Customer satisfaction and loyalty increases: In the end, customer profitability increases. There is another consequence: We also cut costs. If a boutique offers the right products to the right customers, then stock is lowered, and lower stock means less costs. An economy of 30% to 40% of cost of stock is achievable.

Data quality is the key element for being more successful with information. The principle "garbage in – garbage out" is without mercy. It is too late, when organizations notice that the quality of data stored in SAP and other backend systems is insufficient for their business processes after they have built performance management solutions.

Example. A leading European mail-order-house had a problem with its birthday data. Birthday data allows the calculation of the age of customer, an important parameter in customer relationship management. So, what to do, if your birthday data is not reliable? There is a solution that provides a good estimate of age of customer. You look to the customer's first name. First names follow trends. Customer age can be estimated based on patterns of attributing first names to children. But this is an expensive approach, and it will never achieve full reliability. Much better approach is to build data quality from the beginning into the operational processes.

A well-known concept is building quality from the very beginning into processes. Indeed, this is the idea of "**total quality management (TQM)**" that has been applied successfully in manufacturing 20/25 years ago.

Total Quality Management (TQM) means to declare quality as an overall, sustainable goal. It is an integrated, continuous, comprehensive, monitoring and controlling, as well as organizing set of activities across all departments of an organization. TQM was developed by the Japanese automotive industry, and it proved to be a successful model. For TQM, full support of all employees is a critical success factor.[31]

TQM for IT is not only an issue for today. When implementing ERP systems, the principles of TQM for assuring data quality should have been already applied. But still today, data quality is an issue. The Data Quality Check 2007 (Lehmann, Martin, Mielke, 2007), a market research study on data quality and data quality management in the German speaking markets, revealed:

- Just 10% of all surveyed enterprises use a „real" data quality management as described below, but nearly everybody is convinced of the highest importance of data quality.
- 61% of the surveyed enterprises do not at all use tools for data quality management.
- In many enterprises, a data quality director and sponsor on the executive level is the exception. Data quality needs management attention.

This survey is seven years old now, but have we seen real improvements in data quality management in the meantime?

[31] after Wikipedia http://de.wikipedia.org/wiki/Total-Quality-Management (access march 12th, 2013)

> ***Example.*** Let us assume we want to build a 360° view on customer. Goal is to know customers in order to optimally serve customers according to his/her customer value. 60% to 80% of cost for customer data integration is caused by infrastructure. Customer data integration means to synchronize and to harmonize customer data from various sources into one single customer data model. Data originates from various application islands, historical and archived databases, external market data, demographical data, social media data, web click stream data, and others. When building the customer data model, you may notice at once that in a backend system, there is a data table with data about customer that could be linked to a table in another system providing a new and not yet available customer insight. Great, but what if the data field that is to be used as key is not a mandatory field, but just an additional field? Typically, this is the end of the good idea: Will the owner of this application be ready to change the additional data field into a mandatory data field just because you tell him that would ease your job? An IT question turns into a business issue. Indeed, only business can decide on these questions that seem to be IT questions at first glance, but have to be tackled and solved in a collaborative approach jointly by business and IT.

Data quality needs management attention as this example proofs. Therefore, leading enterprises have a data quality director. He/she coordinates the roles of data stewards who are located in the lines of business. The data stewards have responsibility for data quality from the business point of view, i.e. content of master and meta data as well as validity rules for certain transaction data. In process-oriented enterprises, the role of a data steward can be played by the process owner. In other words: Data quality management needs information governance. This is why a DQ initiative should be started by the competence center for master data management or even better by the competence center for information management (cf. chapter 6.8).

Data quality should be implemented into operational processes by a TQM program. The four cornerstones of data quality are[32]:

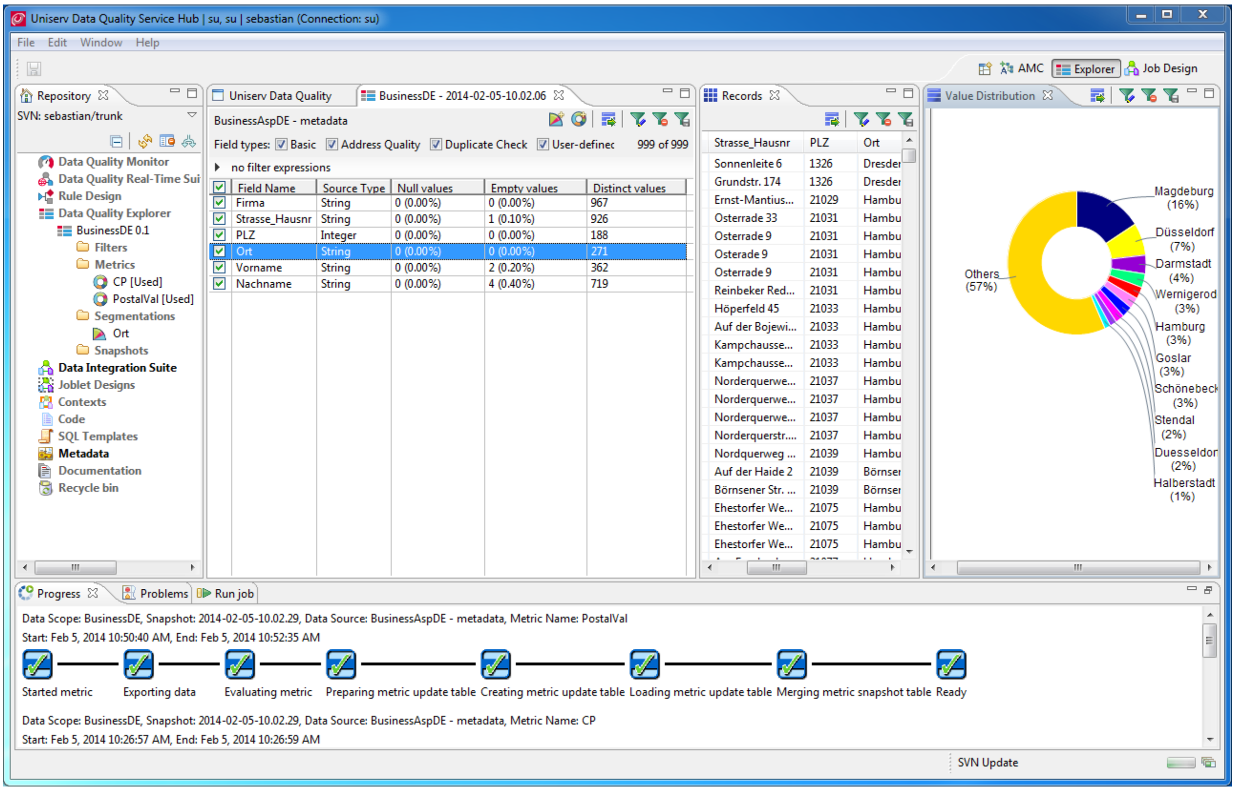
- Quality is defined as the degree of *compliance* with the requirements.
- The principle is: *Preventing* is better than healing.
- *The zero-defect principle* must become the standard.
- Cost of quality is the cost of non-compliance with the requirements.

But the actual situation in most organizations is different. Too many organizations believe: Our data is fine! But in reality, numbers in various reports and dashboards differ. Making decisions based on facts? No way. In the meantime, the number of interrupted transactions is increasing, because important basic data is erroneous. The number of cancellations is increasing, because customers get products they did not order. The number of returns is increasing, because address data is getting out of date. Then, there is point where management wakes up, because cost is increasing. In a slap-bang action, a data quality initiative is decided: A mass data cleansing is necessary. But such an action will not cure the problem, but only mitigate the symptoms. Anyway, before cleansing, a profiling is needed.

[32] after Philip Bernd Crosby http://de.wikipedia.org/wiki/Qualit%C3%A4t#Die_4_Eckpfeiler_der_Qualit.C3.A4t_nach_Crosby

Data Profiling. Data Profiling is used to analyze the properties of a given data set and to create a profile. There are three types of analysis:

- Column profiles. Analysis of content and structure of data attributes helps to identify data quality problems related to data types, values, distributions and variances.
- Dependency profiles for identifying intra-table dependencies. Dependency profiling is related to the normalization of a data source. It provides expected, unexpected and weak functional dependencies and potential key attributes.
- Redundancy Profiling. It identifies overlapping in-between attributes of different relationships. This is typically used to identify candidate foreign keys within tables, and to identify areas of data redundancy.



BusinessAspDE - metadata

Field types: ☑ Basic ☑ Address Quality ☑ Duplicate Check ☑ User-definec 999 of 999

no filter expressions

☑	Field Name	Source Type	Null values	Empty values	Distinct values
☑	Firma	String	0 (0.00%)	0 (0.00%)	967
☑	Strasse_Hausnr	String	0 (0.00%)	1 (0.10%)	926
☑	PLZ	Integer	0 (0.00%)	0 (0.00%)	188
☑	Ort	String	0 (0.00%)	0 (0.00%)	271
☑	Vorname	String	0 (0.00%)	2 (0.20%)	362
☑	Nachname	String	0 (0.00%)	4 (0.40%)	719

Strasse_Hausnr	PLZ	Ort
Sonnenleite 6	1326	Dresder
Grundstr. 174	1326	Dresder
Ernst-Mantius...	21029	Hambu
Osterrade 33	21031	Hambu
Osterrade 9	21031	Hambu
Osterade 9	21031	Hambu
Osterrade 9	21031	Hambu
Reinbeker Red...	21031	Hambu
Höperfeld 45	21033	Hambu
Auf der Bojewi...	21033	Hambu
Kampchausse...	21033	Hambu
Kampchausse...	21033	Hambu
Norderquerwe...	21037	Hambu
Norderquerwe...	21037	Hambu
Norderquerwe...	21037	Hambu
Norderquerstr....	21037	Hambu
Nordquerweg ...	21039	Hambu
Auf der Haide 2	21039	Börnser
Börnsener Str. ...	21039	Börnser
Ehestorfer We...	21075	Hambu
Ehestorfer We...	21075	Hambu
Ehestorfer We...	21075	Hambu

Data Scope: BusinessDE, Snapshot: 2014-02-05-10.02.29, Data Source: BusinessAspDE - metadata, Metric Name: PostalVal
Start: Feb 5, 2014 10:50:40 AM, End: Feb 5, 2014 10:52:35 AM

Started metric — Exporting data — Evaluating metric — Preparing metric update table — Creating metric update table — Loading metric update table — Merging metric snapshot table — Ready

Data Scope: BusinessDE, Snapshot: 2014-02-05-10.02.29, Data Source: BusinessAspDE - metadata, Metric Name: CP
Start: Feb 5, 2014 10:26:57 AM, End: Feb 5, 2014 10:26:59 AM

Figure 41: Uniserv screenshot (cut-out) as an example of data profiling. The various windows provide different views of the profiling process. Bottom: Progress (progress display of ongoing loading or metrics processing operations). Top right: Value distribution (graphical view of value distribution at field level), to the left, Records (display of the currently selected data records - filtering, sorting), Main Window: Accumulated basic information at field level.

Data profiling tools use rules and methods of descriptive statistics (analysis of distributions, tests for outliners) as well as of data mining (cluster analysis and decision trees). Data Profiling provides an analysis “as is”, and is a valuable tool for estimating and directing further investments in data quality. Data profiling tools (fig. 41) identify data quality problems much faster than any manual analysis.

To summarize, data profiling identifies weaknesses within data so that it can be cleansed for reestablishing the wanted data quality level.

Data Cleansing is based on different methods:

- Parsing. Compound data is decomposed.
- Semantic approach. Data is transformed into standard values and formats according to rules.
- Benchmarking. Internal data sources are compared with external sources for verification.
- Matching. Data of similar content in different fields is identified (match customer information that is stored in different applications to one and the same customer).
- Removing duplicates. (e.g., address data)
- Consolidating. Create a complete data record out of dispersed information (e.g., create one customer address record)
- House holding. Detect relationships between data (e.g., identify all persons belonging to one and the same household).
- Enriching. External data may enrich the value of cleansed enterprise data.

Data cleansing tools are based on probabilistic, deterministic and knowledge procedures. Probabilistic and deterministic procedures use appropriate algorithms, whereas the knowledge based approach uses country/language specific knowledge databases for composing addresses, names or legal entities.

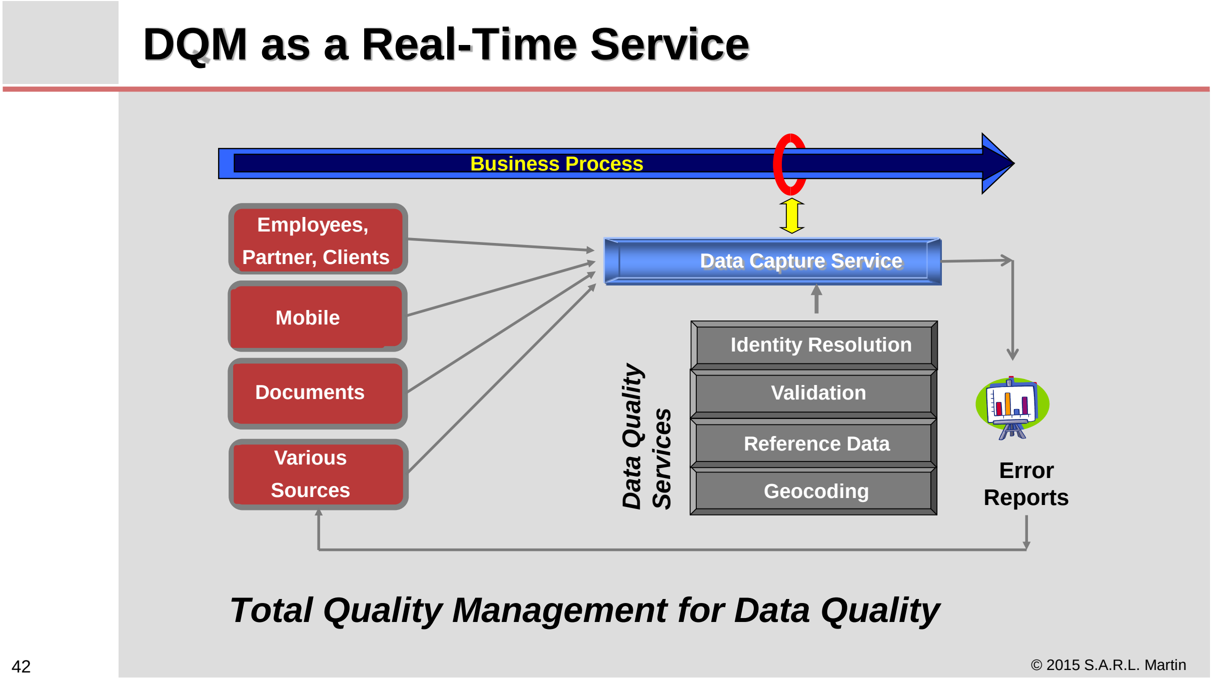
In many cases after a data profiling and cleansing, organizations wait until new data quality problems occur, and then repeat the procedure in an ad hoc manner. A better practice is a regular periodic profiling and cleansing. The result is a data quality level over time that looks like a saw tooth diagram. After cleansing, data quality is best, and then it degrades until a new cleansing is applied and data quality is brought back to its highest level – up, down, up, down.

Obviously, this is not a best practice. A good principle to be applied would be: Prevention is better than heal, or: avoid damage. This is a good principle for each business: Better prevent risks than repair damage, since preventing risks means less cost plus business processes that continue to run instead of standing idle or even being aborted. This also means a considerable gain of time. Finally, we are talking about risk management.

So, the idea is: Data quality management should be performed according to the principles of risk management. Data quality as a risk can be valued real money, since data quality determines process quality. Decisions based on wrong data end up in wrong decisions. Cost and loss of time by wrong decisions can be rather precisely evaluated case by case. Wrong data in operational processes mean higher process cost and slower flows. Wrong data stop processes, block automation, require escalation management or cancellation, cause returns – sometimes even with recovery claim. Therefore, cost and loss of time per process caused by wrong data can be precisely calculated. This is due to the well-known principle: no process without data. Data drive and control processes.

In fact, data quality management in the sense of risk management is a total quality management concept. Data quality is assured from the very beginning across the whole life cycle of data. Data quality management starts with data capture, and it does not end before deletion of data (fig. 42).

Data come from various sources and flow to business. It is captured by different channels, either manually by employees, partners, customers etc. or automated via document exchange (scan and fax technologies), electronic data exchange (EDIFACT, SEPA etc.), machine-to-machine (M2M) communication or mobile devices. Data capture is either triggered by a process or vice versa: An event creates data and triggers a process. This again shows the interconnection between data and processes.



DQM as a Real-Time Service

Business Process

Employees, Partner, Clients

Mobile

Documents

Various Sources

Data Capture Service

Data Quality Services

Identity Resolution

Validation

Reference Data

Geocoding

Error Reports

Total Quality Management for Data Quality

42

© 2015 S.A.R.L. Martin

Figure 42: TQM of data quality means a closed loop approach. When data is captured, it is simultaneously checked via data quality services. Erroneous data records are stored in a staging area, and an error report is produced and sent to the source so that corrections can be performed. If the corrections are successful, the data record in the staging area is marked and processed. Periodic reports enable a performance management for continuous improvement of this closed loop approach.

Now, the idea is to include real-time checking completeness, accuracy, and redundancy of data into data capture. This is done via data quality services. (Customer) identity resolution services are the first group of such services. They assure that a new data record is matched to the correct master data record. This avoids duplicates, for instance, since identity attributes of a customer, for example, can deviate due to transmission and transcription errors as well as to orthographic variants. The task is to identify similarities and to match correspondently (cf. the following chapter 6.7).

Another group of data quality services are validation services. These are rules describing the structure of certain data records. They check the filling degree of obligatory and additional attributes, data types, range of values, orthography and grammar as well as relations between attributes and records. Other data quality services check knowledge bases for checking plausibility and accuracy. This is important for specific country properties like address standards, for instance.

Finally, especially in the era of big data, geocoding services are getting more and more important. Geocoding allows an address evaluation in local markets, and it helps to locate customers and to unlock new potentials (cf. chapter 5.8):

1. Geocoding of data: each address is matched to a spatial coordinate (x-y coordinate).
2. Erroneous addresses or positions are selected and validated by a data cleansing service.
3. Each address is matched with a raster ID, so that additional attributes about socio demographics, buying potential, life style information etc. can be added. This is called data enrichment.

Geocoding complements data quality management in the sense of risk management. It acts like a profiling, identifies errors in address data, and does the cleansing. In parallel, data is enriched so that data can also be used in various other applications, for example for clustering in analysis of potentials.

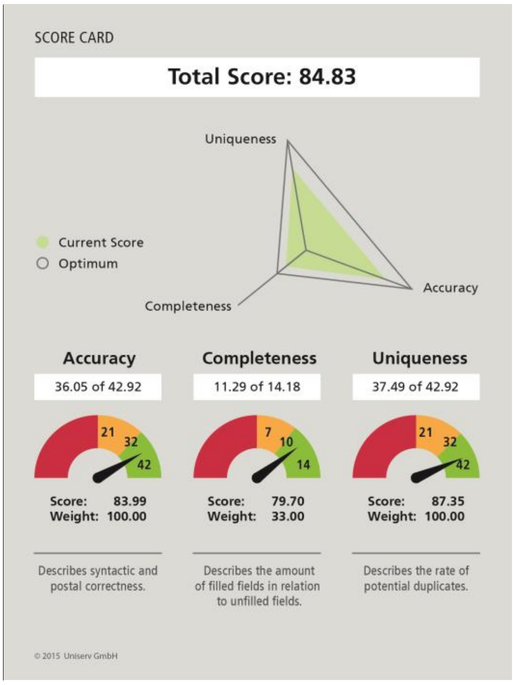


Figure 43: Data quality at a glance, a dashboard (Uniserv's Data Quality Score Card as an example) monitors key performance metrics, accuracy, completeness, and uniqueness. This lays the foundation for continuously improving data quality.

Data quality services can act in real-time. So, data quality services can be provided and consumed either as on premise services or as SaaS via cloud computing, i.e. Data Quality as a Service (DQaaS). Alternatively, data quality could also be provided as a hybrid solution, a combination of on premise and cloud.

Data records that have been identified as erroneous and that cannot be corrected automatically are written to a staging area, and an error message is sent to the source of data capture. Via an escalation management, the error is to be corrected. Typically, this is a manual task performed by a process manager with the necessary expertise. When the data record is cleansed, it can be transferred into the enterprise pool of data and can be used within the corresponding processes. It is also marked in the staging area. A data quality dashboard (or a simple report) monitors the performance of this closed loop approach for data quality management (fig. 43). Finally, it provides the performance management to continuously improve the closed loop as required by TQM. The result is data quality on a nearly constant level, a considerable progress towards the traditional saw tooth approach.

Big Data Quality. Data quality also plays an important role in Big Data. This is especially true, when enterprise data should be enriched by Big Data information, for instance customer data by social media data or in health care, patient data by therapeutic data. The basic tasks of data quality management remain the same: profiling, cleansing, and enriching by and checking with reference data. But Big Data Quality requires more: Since master and meta data from various data sources will rarely be identical, but just “similar”. So, we have to match meta and master data on similarity metrics for aggregating social data from various sources. This will be discussed in the subsequent chapter.

Finally, Big Data raises another question about data quality: How to treat outliners? For enterprise data, we had a hard rule: Outliners are to be eliminated. But does this make sense in Big Data analytics? Do not outliners in Big Data hide important information? Big Data analysts say: May be – and therefore, Big Data analysis is performed without a hypothesis, i.e. without eliminating outliners. This breaks the rules of traditional data management. Two aspects are to be considered:

- Decisions in Big Data analytics are based on milliards of analyzed data records. Trends are not derived by simple “X=Y” comparisons, but by milliards of these comparisons. Then, decisions will be taken via aggregates: They depict trends and reveal the probability of events. Big Data cluster do not aim to generate exact and precise numbers like finance, but to detect potentials based on the analyzed data. Big Data analytics is not about deterministic facts, but about fuzzy logic.
- Velocity of data production forces a new thinking in data quality. What is the value of data quality, when several terabytes of data are produced per day? When there is some “bad” data due to hard- or software problems, we will certainly somewhat loose precision, but next day’s data will wipe out this short term incidence.

Big Data volume and velocity reduce the impact of erroneous data within some data records. Indeed, this is an essential property of Big Data analytics: It does not focus on specific data points, but on meta trends. Here, we are entering virgin soil, and still have to learn a lot.

6.7 Entity Identity Resolution

Entity Identity Resolution is all about supporting organizations when managing entity identity data. This is data from various sources for specifically and correctly identifying entities like products, services, customers, suppliers, prospects, opinion leaders; patients, tax payers, criminals etc.

In different data sources, entities are typically not stored the same way, but differently. This puts up a problem, when we want to match data from different sources for creating enriched data records per entity. The simple solution of entity identity resolution by simply comparing character strings or applying match-codes does not work anymore.

> ***Example:*** Let us assume we are dealing with customers as entities. A customer registered as John Smith in our enterprise customer data base, may call himself Johnny Smith in a social network or could use the name John W. Smith when complaining about one of our products. Are these now three different persons or is it one person with three identity denominations?

Problems like this typically are not the exception, but nearly the rule. Causes are a natural variability as in our example about John Smith, but also unexpected errors through mistakes in writing or mistakes in transcription like nicknames, abbreviations, and notations in different letterings (like Arabic, Chinese, Hellenic, Cyrillic, Latin etc.). Furthermore, there are professionally created lies pretending wrong identities and customers may use pseudo names and anonymity. Indeed, entity identity resolution is one of the toughest challenges in information management.

The problem of entity identity resolution is not new. When the entity is a customer, then **Customer Identity Resolution** is a well-known task within data quality management since the times of establishing and running data warehouses or when creating a single view on customer in analytical CRM. Roots go back to direct marketing: Suppression of duplicates is an important task within address data management. Effectively managing customer identity is mission critical in managing and cleansing address files, external cleansing, list mix, cluster comparisons, negative and positive comparisons as well as in cross-country comparisons with different lettering. Indeed, Customer Identity Resolution is in the heart of data quality management.

Today, in the age of social media like Facebook, LinkedIn, Xing, Foursquare, Twitter, Pinterest etc., Customer Identity Resolution becomes even more important. Businesses now want to know what customers are saying and thinking when communicating in social media. Marketing, for instance, can gain not only additional attributes about customers and customers' behavior, but also completely new types of information that was not available before: Even information about customers of competitors is now accessible.[33] Information like this enables better addressing customer preferences and needs: Competitiveness can be well improved. Customer Identity Resolution makes it possible: It supports the matching of enterprise customer data with social media data. More and more information can now be put into the context of customers and can be cumulated. This provides better knowledge about and better insights into customer behavior. This is also known as **"social master data management (social MDM)"**. All the information can be put together to finally provide a complete picture of customers. It is a multi-channel image of customers so that better customer models for predictive analytics can be built for improving the bottom line.

If entity is product, we have an equivalent situation **("Product Identity Resolution")**. Price comparisons make up a typical case for Product Identity Solution. Again, this is the task of identifying and matching "similar" products, because products in one web shops will typically

[33] We assume that the access of such kind of customer data is compliant to the laws and rules of data protection.

not be identical to products in another web shop, for instance. So, we have to define similarity, and then to identify similar products by a fault tolerant search.

Methods of Entity Identity Resolution are based on country specific rules and terminologies, language related phonetic and entity specific fuzzy logic. (For more details, please look up the grey box at the end of this chapter.) Entity Identity Resolution should be provisioned by services for enabling the deployment of Entity Identity Resolution in batch for mass production, for example for cleansing of mass data, as well as services embedded in business processes, where Customer Identity Resolution can be performed in real-time.

> ***Example: Transactions in Retail.*** Many merchants operate several web shops. So, when a new customer places an order in one web shop, then it is interesting to know whether he is already a “good” customer in another web shop. Indeed, this is an excellent moment for improving customer loyalty. Perhaps, it is a “bad” customer that is already on the black list of the enterprise or of a credit rating organization. Perhaps, this customer even uses a false identity when shopping. Customer Identity Resolution in real-time can now reveal the true customer identity, and before the transaction will be terminated, preventive actions can be taken to minimize risk related to this customer.

The example underpins another benefit of Entity Identity Resolution: It is now more than just a way of cost savings by elimination of duplicates in direct marketing applications, because if applied in CRM, it now also can improve customer profitability and customer loyalty, or if applied in risk management, as well as prevent fraudulent transactions.

Entity Identity Resolution services also come with an additional advantage: They can be used in both ways, as on premise services and as SaaS in the context of cloud computing. Therefore, Entity Identity Resolution services can be quickly deployed and tested, and in a pilot project, it can cost and benefits can be well examined.

Methods of Entity Identity Resolution.

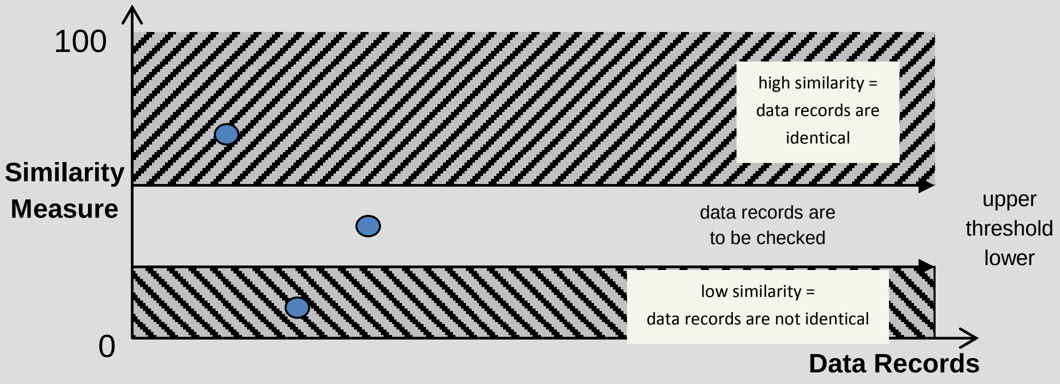
Entity Identity Resolution started with simple character string comparisons and match codes. Today, more and more mathematical methods are used, especially fuzzy logic combined with linguistic approaches: Country and language specific knowledge bases complement mathematics (fig. 44). These knowledge bases are open and can be trained over the course of time. This allows to adapt and to fine-tune broader knowledge bases. For a more detailed description of these methods, we refer to vendors like Uniserv[34].

Entity Identity Resolution can cause errors. If two data records belonging to different entities are assigned to one entity, we talk about a type 1 error (“false positive”). If two data records belonging to the same entity are not assigned to that entity, we talk about a type 2 error (“false negative”). The following diagram illustrates the situation:

In this diagram, the upper threshold defines the minimal similarity so that different data records are assigned to one identity. The lower threshold defines the maximal similarity so that all data records with a lower similarity are assigned to different identities. Data records that show a similarity between the lower and upper threshold are to be checked manually. If

[34] See http://www.uniserv.com/en/company/blog/detail/article/processes-and-algorithms-in-address-management/

the upper threshold is to low, we will get more false positive assignments, and if the lower threshold is too high, we will get more false negative assignments. During the course of time, experience with the upper and lower threshold can be gained, and the thresholds can be empirically optimized.


100
Similarity Measure
0
high similarity = data records are identical
data records are to be checked
low similarity = data records are not identical
upper threshold lower
Data Records

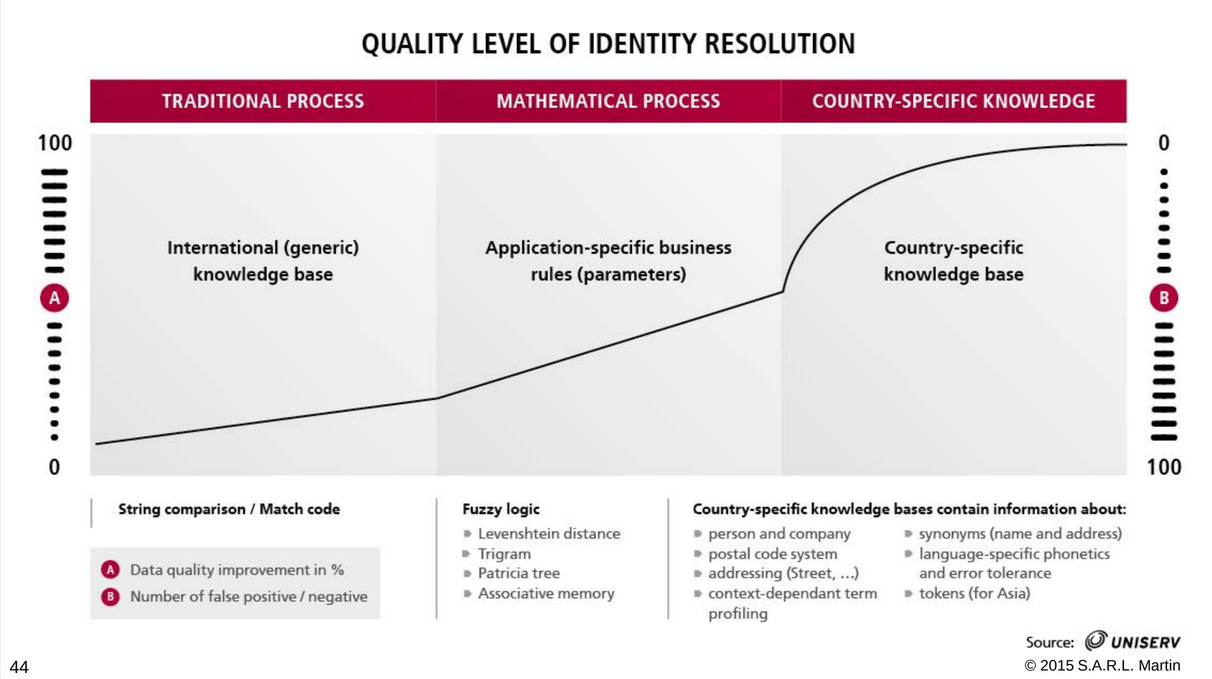


Figure 44: Quality of traditional methods like match codes and character string comparisons is typically no more sufficient for Entity Identity Resolution. But quality can be improved by mathematical methods of fuzzy logic. For reaching towards a quality level close to 100%, we additionally need country and language specific knowledge bases.

6.8 Information Governance

Finally, governance is a critical success factor for information management. Information management needs an appropriate organization with clearly defined roles and responsibilities. It requires rigorous and the right processes and policies, and finally the right technology and platform to run information governance. Best practice information governance can industrialize information management in the sense of “lean” information management processes.

When establishing information governance, the processes and policies for the described tasks of information management have to be modeled and implemented. Furthermore, a continuous improvement process is to be set up. Just as business processes are controlled by performance management, the governance processes are also monitored and controlled by a performance management on the strategic, tactical, and operational level. The experience gained and lessons learned are used for the continuous improvement of all processes. This is a critical success factor for an industrialized information management.

Furthermore, the organization of information governance is to be defined: The organizational units, roles and responsibilities must be specified. A **competence center for information management** has proved itself as best practice here providing advantages and benefits. Its organizational structure corresponds to the structure of a BI competence center (see chapter 3.5): A steering committee chaired by an executive sponsor, the team, and the data stewards. The executive sponsor is needed for backing up strategy and policies of information management. The data stewards are based in the specialist departments and are linked to the information governance. The competence center centralizes management of information management strategy, methods, standards, policies (“rules”), and technology. Its leitmotiv is:

> The information management competence center plans, guides and coordinates information management projects and ensures the effective use of resources and technology.

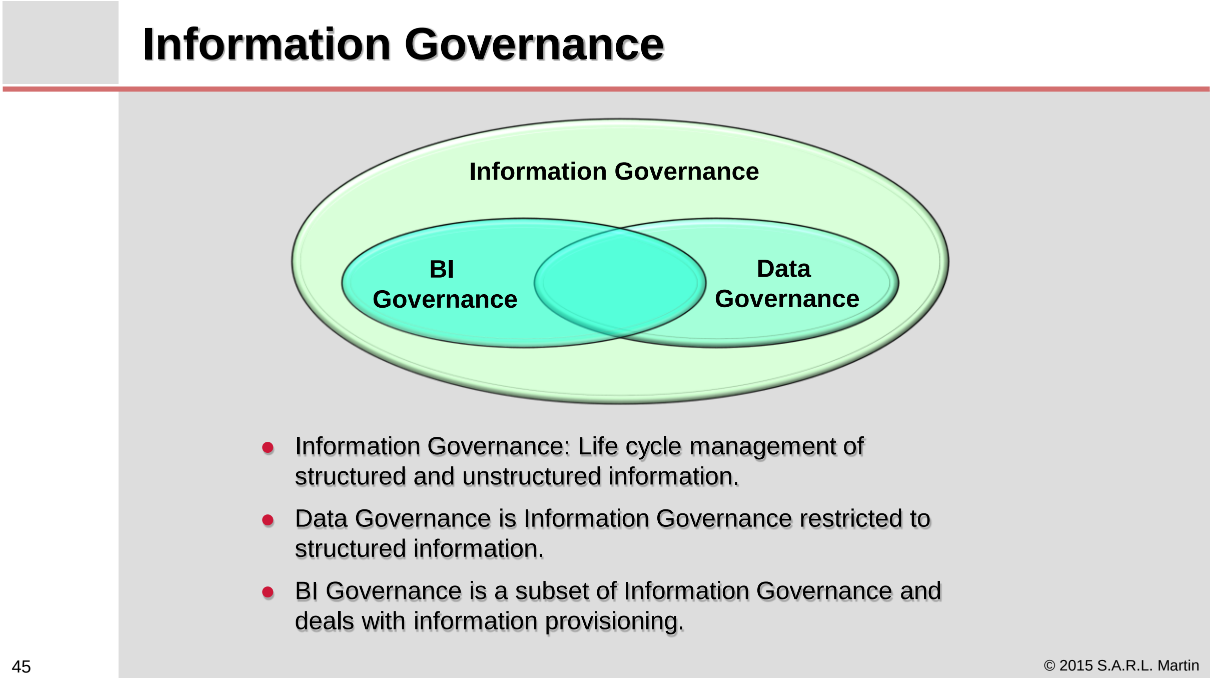
Since information management is the foundation for performance management and analytics as well as for business process management, it is recommended to have both, a BI competence center for the intrinsic tasks of performance management and analytics, and an information management competence center. Here, Information Governance meets BI Governance. Furthermore, there is also „Data Governance“. So, what is what and how can we distinguish the terms and tasks? What are potential overlaps? The most general term is **Information Governance**. Its domain is Information Management. Recall, Information Management means managing the whole life cycle of structured and poly-structured information. **Data Governance** is a subset of information governance that is restricted to the governance of structured data. **BI Governance** is also a subset of information governance. It is related to both, structured and poly-structured information, but BI Governance does not deal with the whole life cycle management, but only with information provisioning and the corresponding tools. Figure 45 shows the positioning of the three governance domains in the context of information management.

Consequently, information governance should cover all governance processes and policies for the following tasks (see chapter 6.1):

- Meta and master data management
- Data modeling
- Data quality management

- Data integration
- Data classification
- Data security and protection
- Content management

Finally, the technology supporting these processes is to be selected. Today, a service-oriented platform is state of the art, because then the information management processes themselves meet the requirements of industrialization, agility, and compliance. Result is lean processes for information management.



Information Governance

Information Governance

BI Governance

Data Governance

- Information Governance: Life cycle management of structured and unstructured information.
- Data Governance is Information Governance restricted to structured information.
- BI Governance is a subset of Information Governance and deals with information provisioning.

45

© 2015 S.A.R.L. Martin

Figure 45: Positioning and overlaps of Information Governance, Data Governance and BI Governance depicted as Venn diagram.

Wolfgang Martin Team's market survey (Martin, 2012) shows that information governance has a rather high importance within enterprises in the German speaking markets. Main drivers are data and process quality. But just under half of the surveyed enterprises use information governance or plan its use. Furthermore, 47% of the interviewed see its enterprise in the planning or starting phase, but just 21% in the final phase of implementation. In smaller enterprises (less than 1,000 employees), the management board has the sponsorship, in larger enterprises (more than 1,000 employees) the CIO is typically in charge. Collaboration between IT and business is nearly ideal: If there is an information governance program, then 80% believe that it is a common IT/business program. But the usage of tools is not yet really satisfying: Just 60% of the interviewed say that information governance tools are used.

In practice, however, it has been discovered that the Governance organization and processes are regarded as a restrictive set of rules which impede flexibility and the agility which is increasingly in demand today. Collaborative methods and tools have proven to be effective here and provided a remedy in the meantime. These are derived from social media: the employees are given a lift and are filled with enthusiasm for Information Governance through a social media work style. Social media promote the participation effect and make a substantial contribution to transparency. And so a set of rules perceived as inconvenient and top down becomes a living bottom up collaboration, in which everyone can communicate and discuss things with each other on equal terms. Modern-day Information Management platforms are already provided with such collaborative tools to a large extent. This should always be considered when selecting a suitable platform and given a high weight in the appraisal because you have the best chance of sustained success in today's digital world with efficient Information Governance – and all the requisite Compliance specifications are met at the same time.

Take Away: Information Management needs information governance for creating trusted data. Information governance gets together people, strategies, processes and organization. Establishing a good governance, lean information management becomes doable: Industrialization of information management means to achieve the best degree of efficiency in the sense of costs and assignment of resources.

6.9 How Data Stewards and Data Scientists act in concert

We have already seen that data stewards are responsible for information governance. They have to take care that information governance is respected by everybody in the organization and is applied in all actions. The goal is to guarantee that in the whole organization, information is correctly, consistently and universally captured, managed, used, and applied. They should not perform this task in a police style of work, but on the contrary, as a service provider for the specialist departments. This includes several tasks. Data stewards should have a leading role in developing data definitions. They should support data profiling for detecting errors in data and for estimating the impact of those errors. They also should promote the usage of data as well as take care of data security. Another task consists of controlling the adherence of rules and of monitoring data quality. Finally, they are also involved in setting priorities for data quality actions.

The tasks of data stewards should be prioritized in the context of business strategy and business goals. A software company, for instance, should focus the activities of data stewards on customer data that is used by sales and marketing processes that are currently prioritized by the business strategy. A general focus on customer data management would be too broad and would typically not show the expected outcomes.

Data stewards can even start their work in the organization if information governance is not yet ready or spelled out. A small mixed team of staff from the specialist departments and IT can begin with formalizing data definitions as well information management processes and usage policies. Staff in this team should be excellent communicators and facilitators as well as smart negotiators. Full time delegation into the team is not a must, but sufficient time for data steward activities should be allocated. Hiring new employees for these tasks is not recommended, since successful data stewards should really know the organization and its informal information channels.

The team then gets networked in the organization. Its duty is also to establish a working relationship between the specialist departments and IT. From an organizational point of view, the team with its contact persons in IT is allocated in the competence center of master data management or information management. But in parallel, data stewards still keep their domicile in the specialist departments. Such a competence center has a C-level sponsor, and is headed by a co-leadership consisting of a data steward acting as primus inter pares, and its counterpart from the IT department. It has a budget, is responsible for the information governance program, and develops its methodology.

Data steward programs sometimes meet cultural problems. Let us assume, a certain specialist department is not willing to recognize problems about data quality, because data is only captured in this department, but used in another department. Thus, the interest in investing in data quality is low. In such cases, data stewards can provide transparency, show the problems impacting the whole organization, and outline how they can support this department with data quality activities. But if such an attempt is not successful, then it’s the sponsor’s task to intervene, and perhaps to apply a formal change management to fix the problem.

Big Data Management. Data stewards have the responsibility for enterprise and departmental data. Now, we are in the era of Big Data. How does Big Data management influence the traditional tasks and goals of data stewards? Let us first have a look at the requirements of Big Data analytics. Big Data analytics needs new skills and roles that are best allocated in an expanded BI competence center. In some organizations like Amazon, eBay, Facebook, Google, Twitter etc. that have already gained experience in Big Data, such new roles have been created. They set up the Big Data analytics teams. We have already seen these roles in chapter 3.7.

How does the BI competence Center with its data scientists collaborate with the information management competence center with its data stewards?

Here, the role of the data hygienists provides the interface. On the one hand, data hygienists have certain tasks of data stewards within Big Data initiatives, but on the other hand, clear differences become apparent. In an organization, data stewards have the responsibility for information governance that rules over all enterprise data. In Big Data initiatives, correctness and completeness of data is set up per project. There is no universal information governance, but project related data governance defined by the data scientists.

This could even mean that certain Big Data project run without governance at all. Data scientists argue that a cleansing of the data from various Big Data sources could bias analytical results, because data could be tampered by assumptions of the data scientists. Outliners, for instance, could contain important information that would be useful for data discovery. Therefore, Big Data analysts quite often start their investigations without hypothesis on data depending on the context and the goals of a Big Data analysis.

This clearly breaks the well-known rules of data quality management. Two things are to be considered:

- Big Data analytics makes decisions based on milliards of data records. The assumption is N = all, i.e. “all” data is analyzed; no data point is left out. Trends are not derived from simple “X=Y” comparisons, but from all “X=Y” comparisons. Decisions are then based on aggregates for depicting trends and estimating probabilities of events. Big Data clusters are not supposed to generate precise numbers (as in finance), but to reveal potentials based on the analyzed data. Big Data analytics is not about deterministic facts, but about fuzzy logic.
- The speed of data production also requires a new thinking about data quality. What is the benefit of data quality when, for example, the hourly data production is of the order of several terabytes? If some of this data is “bad” data due to hard- or software problems, we will momentarily lose some precision, but when new data streams arrive in the subsequent hour, the momentary imprecision will be at least balanced out or even totally irrelevant.

Data velocity and volume reduce the impact of erroneous data in some data records. This is another important feature of Big Data analytics: It’s not about specific data points, but it’s about meta trends. In Big Data projects, the underlying and guiding principle is that data has to serve the purpose of the project, whereas in an organization, data has to serve the long-term business strategy.

Take away. The task of data stewards is information management of enterprise and/or departmental data under the rules of information governance. The goal is to provide high quality, consistent, and easy to access data for the specialist departments. Furthermore, data stewards have the responsibility to continuously adapt information governance to business strategy.

Data Scientists have the task to turn Big Data in “Big Value”. They are responsible for the methodology of Big Data analytics as well as the communication of analytical results to the board and to the whole organization.

There is a role for data stewards in Big Data projects. It is the role of data hygienists that depends here on the project goals, but not on business strategy. In Big Data analytics, data stewards play the role of a SWAT[35] team, i.e. a tactically acting special team, but not the strategic role as in the organization. Again, they are service providers, but in Big Data initiatives, they are called-in by data scientists. Here, they do not act autonomously.

[35] SWAT (= special weapons and tactics) is a term used in the American marketing terminology.

7 Latency matters

Performance management must address all operational, tactical, and strategic aspects of monitoring and controlling of processes and services in a seamless way. Consequently, analysis, monitoring, and controlling must be synchronized with the speed of the related processes (see fig. 9 and 10). When speed of a process is high and when seconds or fractions of seconds matter, then real-time technologies like **Business Activity Monitoring (BAM)** and **Complex Event Processing (CEP)** come into play. You easily find use cases in customer interactions, production and logistic processes. The goal is to detect problems and risks related to an operational process instance. Identified problems and risks should automatically trigger counteractions for process control. Today, such an **operational intelligence** is one of the basic principles of the internet of things. It corresponds to the closed loop model of process control as already discussed in chapter 3.2. We will follow up this discussion in chapter 7.1.

If time becomes critical for analysis, new database technologies like **analytic and NoSQL databases** play an increasingly important role. When data volumes are growing faster than the performance of traditional relational databases, then the analysis of detailed data, for instance, is no more practicable since it simply takes too long. Gartner already said in its report to the Magic Quadrat for Data Warehouse Database Management Systems 2010: “As in 2009, Gartner clients still report performance-constraint data warehouses during inquiries. Judging from these discussions, we estimate that nearly 70% of data warehouses experience performance constraint issues of various types.”

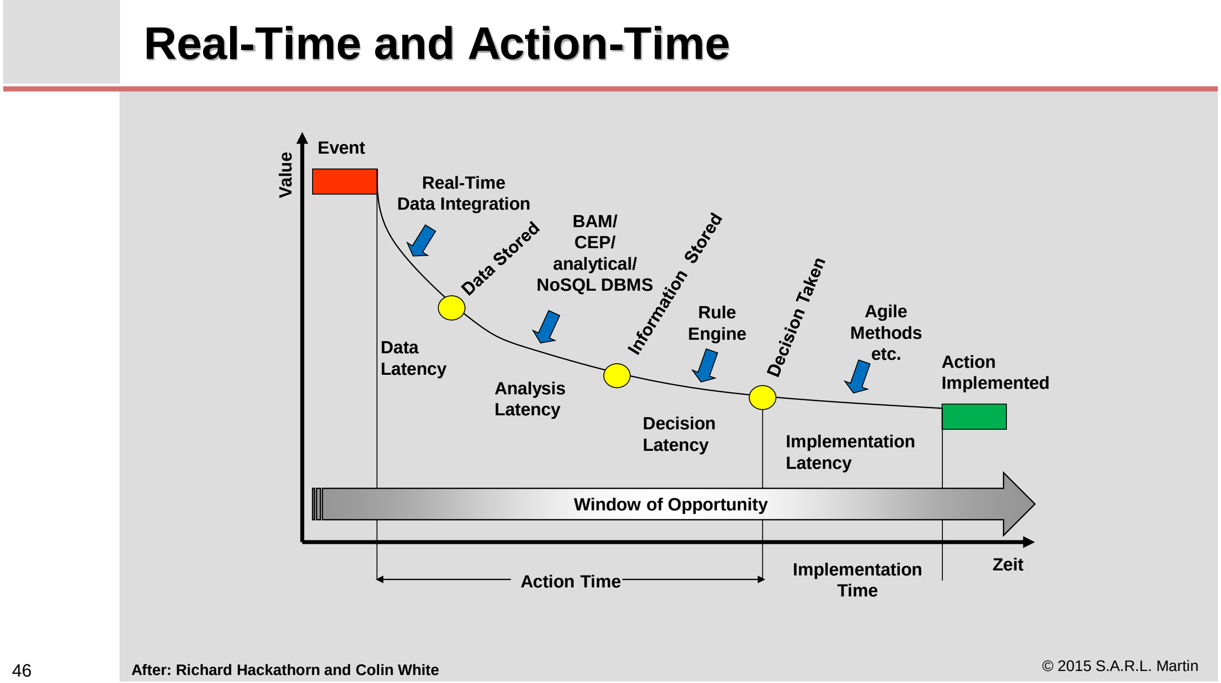
This is why new methods and technologies for data management are evolving and gaining traction to get Big Data under control. Indeed, analytic databases and NoSQL database and data management systems are now established side by side with traditional relational database management systems. They have been designed for processing large and very large databases exposed to a high number of various queries by a large to very large number of users for providing up-to-the-second results: Analysis and the analytic processes are considerably accelerated. The basic concepts of these so-called analytic databases exist already since more than 20 years. But it was not before 2010 that customer and market demand started to boom. We expect the boom to continue, and we even see a further increase in demand for analytic databases in 2016/18. In particular, a strong driver is customer experience management (CEM) when it is all about to deepen customer knowledge through social media data, location data and other Big Data sources. We will discuss analytic and NoSQL database management systems in chapters 7.2 and 7.3 as well as their market evolution in chapter 10.6.

Finally, it should be noted that all these types of database management systems can also be consumed “as a service”, i.e. via cloud computing in all its flavors: private, public or hybrid clouds.

7.1 Business Activity Monitoring und Complex Event Processing

Automation of processes - whenever possible and meaningful - is one of the key principles of process orientation. Consequently, metrics for performance management of automated processes become more important. The purpose is to automatically detect deviations and abnormalities of operational process instances, and – if necessary – to automatically control processes for minimizing risks and maximizing chances. This is done by metrics driving rule (or: decision) engines for triggering actions.

But not all processes can be or should be fully automated. Human interactions are always needed in case of exceptions, escalation management, authorization, entry of triggers (self-service), and when working in teams (collaborative services). Therefore, a process typically consists of a combination of automated sections and manual interactions. Now, when the identification of alerts and exceptions becomes time critical, human interactions may become too slow. This is when latency matters and action time becomes critical (fig. 46). The action time model shows three critical phases, data latency, analysis latency, and decision latency.



Real-Time and Action-Time

46

After: Richard Hackathorn and Colin White

© 2015 S.A.R.L. Martin

Figure 46: Time can be critical, if an event occurs, because an event comes with a window of opportunity. This presents the maximal available time for reacting, taking decisions, and implementing actions in order to profit from the event or to avoid damage caused by the event. The action time model describes the phases from the occurrence of an event until actions are decided. It decomposes action time into data latency, analysis latency, and decision latency, and it shows by which analytic-oriented approaches, action time can be minimized. Action time is succeeded by implementation time that is necessary to implement the taken action. We now need an agile business process management for rapidly changing and adapting business processes, or agile software development methods for rapidly modeling and implementing eventually necessary new business services.

Data Latency means the time that is necessary to capture all data that is needed to identify an event. This is addressed by real-time data integration. For a discussion, we refer back to chapter 6.1.

Analysis Latency means the time that is necessary to analyze all relevant data and to deliver decision relevant information. This is either addressed by Business Activity Monitoring (BAM) and Complex Event Processing (CEP) solutions or by deploying analytic or NoSQL database or data management systems: Analytics must be now available in real-time.

Since analysis latency depends on the complexity of events, we first discuss the various types of events in order to understand the different kinds of BAM and CEP solutions and their constraints to analysis latency. We follow the approach taken by Luckham (2002).

- ***Simple events.*** These are events where all data that is necessary to detect the event are available when the event happens. We have already seen examples like *product availability* and *deliverability*. In these examples, the goal of BAM is to compare such a metrics with a threshold to launch actions for control. Data and analysis latency strongly depends on data volume. If we consider events happening in Big Data, the latency time is reduced by using analytic or NoSQL databases. This will be discussed in chapter 7.2 and 7.3.

  Analysis latency can also become an issue when using predictive models. In many situations, it is not possible to derive the predictive in real time. This is the reason why modeling of predictive models by data mining processes was strictly separated from applying predictive models in operational processes (see also chapter 5.6). Therefore, the predictive model was periodically remodeled with the speed of supposed changes (e.g., week, month). New approaches and technologies make a break through. Predictive models can be made self-learning by adaptive algorithms. They match dynamically to the changing process context. Such an adaptive predictive model is always on-line and maps to the presence based on the actual data driving the adaptive algorithm. It presents a low latency solution for analysis latency. In fractions of seconds, adaptive models can be recalculated. As an example, we can think of adaptive, dynamic models for intelligent customer interactions in call centers or in web shops.

- ***Event streams.*** These are a continuous time sequences of events (also called “time series”) like machine or sensor data in the internet of things. More generally speaking, the challenge is the analysis of observation data. Timing, for example, corresponds to the arrival times of events in a BAM or CEP tool or by time stamps. Monitoring and controlling of traffic in all kinds of networks is typically based on event streams. Examples can be found in telecommunications, information processing, air traffic, ground traffic etc. For BAM and CEP tools, there are different application domains to be distinguished.

  - *Simple pattern recognition.* BAM tools for this type of problem are based on time series analysis. The goal is the forecast, i.e. the prediction of the outcome of the next event. Typical examples are sales forecasts as well as forecasts of stock prices or peaks in consumption.

  - *Complex pattern recognition.* Events streams can be conditional. They could happen at different locations at different times and influence each other. Now, CEP tools based on multivariate time series analysis come into play. Examples are concurrent and collaborative processes like sales promotions of several competitors in one and the same market. Task for a CEP tool could be to track the effectiveness of its own marketing activities, to measure the impact of its promotions and to derive marketing strategies for defense or attack based on CEP.

- *Pattern abstraction.* Subsequent events could be detailed events of an event on a higher abstraction level. CEP tools are now used to detect and identify the typically higher value abstract event based on the evaluation of the single and isolated detailed events per semantic reasoning. For example, consider the analysis of buying signals of a customer. Customers send signals about their readiness to buy a certain product, in particular if the investment exceeds a certain level like buying a car or a house. A CEP tool should now detect buying readiness as soon as possible given received buying signals so that sales get a window of opportunity towards a competitor.

BAM and CEP tools for event streams are based on special fast algorithms (e.g., matching algorithms and other semantic methods). Together with traditional time series analysis and rule engines, they are the basic building blocks of operational intelligence.

Decision Latency means the time required to take decisions based on decision relevant information. Indeed, when time matters, decisions cannot be taken anymore by humans. Then, decision taking must be automated by decision engines. Rule (or: decision) engines are based on rule engines. Rules can be generated bottom up via predictive models. Such a set of rules can be rather complex. For instance in e-commerce, intelligent customer interactions use predictive models that are derived from various data sources like real-time and historical surfing properties, buying patterns, buying history, catalogue information, sales strategy, and other external conditions like time of the day, day of the week, and seasonal information to get recommendations with high relevance. In many cases, the set of rules is simplified and reduced to one single parameter, a score. Decision rules and scores are identified by using various data sources, and previously detected data structures and patterns.

Rules may be also specified by experts in a top down approach. This is a certain revival of the old expert systems popular in the late 80s and early 90s. Ultimately, rules engines can be modeled by a combination of predictive models with expert rules. Decision engines have been discussed in detail in Martin (2003-B).

Implementation latency means the time that is necessary to implement the taken actions. We now need an agile business process management for rapidly changing and adapting business processes, or agile software development methods for rapidly modeling and implementing eventually necessary new business services. But such a discussion is no more within the scope of this White Paper.

7.2 Analytical Databases

Analytical databases use special data management methods and technologies for storing and processing large to very large volume of structured data with highest performance. Analytical databases are not new at all. They exist already since the late 80s, but they did not succeed or various reasons. Now, Big Data changes the game. What makes analytical databases different from traditional relational databases, and why are they so well equipped for Big Data? Indeed, there are various new methods and technologies that typically can be even combined. That makes the difference. Let us start with column oriented databases. Traditional relational databases are row oriented. This creates some problems when the data volume is large or very large. We now look first into the disadvantages of row oriented relational databases and into the advantages of column oriented databases.

As an example, let us take a customer data record. It may have about 1,000 attributes, but we do have as any records as customers, perhaps some millions or even more. If you try to select customers according to some attributes, you have to read all records. But "reading" puts a problem to relational databases, because they are not designed for fast reading, but for transactional data processing. A relational database is best (and fast) when there is an index pointing to a set of data for creating, reading, updating and deleting data.[36]

For ad hoc queries on relational databases, indices and aggregates are needed for achieving and guaranteeing fast up to the second responses. But consequently, a query must be known before launching the query so that a database specialist can build the well-tuned indices and aggregates for enabling the query. Unfortunately, this is expensive, because well trained specialists are needed, and this is too slow: If business has a new idea for a query, IT has to build the indices and aggregates first. As we all now, this takes time, too much time in many business scenarios when time is critical for making decisions. If you try to start a query without pre-built indices and aggregates, IT operations can be delicately impacted. Indices and aggregates bring another disadvantage: They consume space and can inflate the size of a database by a two digit factor. This slows down again the performance of the database. Finally, there is a situation, where users do not put queries anymore, because they do not have any chance to get the answers. So, business gets frustrated, and the potential knowledge needed by the business stays unexploited in the database. Information turns into a mere cost factor. Knowledge about customers, market, competitors and risks are underused. This is a quite typical situation in many organizations.

Column oriented databases put things right. In a column oriented database, each column can be put in its own data set. When reading the value of an attribute, the next value to be read is not the value of the subsequent attribute, but the next value of the same attribute. In other words, rows and columns of a table have been interchanged. Intuitively, this works well, since analytics typically evaluates few attributes of very many records. Reading is dramatically reduced, because by interchanging rows and columns there are at most as many records as attributes. Consequently, since the number of attributes is rather low in comparison to the number of records, the performance gain is high. But there is a drawback: Writing of records into a column oriented database is expensive. But methods like applying difference data sets can improve writing performance and partially balance out this disadvantage.

This basic property of column oriented databases provides another advantage: Indices and aggregates are no more needed. The database remains lean, and the less data is to be read, the more performance is gained. Furthermore, compression techniques can be applied to further reduce the data volume to be read, and special methods enable fast relational

[36] This is the "CRUD" principle, "create, read, update, delete".

operations on the compressed data. For example, multiple values can be replaced by some fixed or variable length code that can be translated into the original values by a dictionary. A sequence of identical values can be transformed in runtime coded sequences. Sorted integer value can be stored in some few bits by taking the difference to the respective precursor or to a local minimum. Compression of a column oriented database provides a lot of advantages. It considerably reduces the size (up to 80% and more) and improves the performance gain.

Further increase in performance can be achieved by parallel processing across clusters and by in memory processing. This is an approach for both, row and column oriented databases. Intelligent algorithms automatically spread the data across all servers of a cluster so that queries can use all hardware resources in an optimal way. The design of parallel processing solutions avoids any manual database tuning that is always needed in traditional databases. Indices, compression, and distribution of data across the nodes are done automatically. Other algorithms identify server failures and make sure that recovery procedures restore the system within seconds. Parallelism enables fault tolerance.

But all these concepts and methods for optimizing performance have also some draw backs. Rigorous transaction processing according to the ACID[37] principle is to some extend no more possible. This is why we do not call these systems “database management systems (DBMS)”, because a DBMS must include rigorous and reliable transaction processing.

There are different types of analytical databases. There are parallelized traditional databases that are still row oriented. Typically, they are offered as appliances, i.e. a specialized, optimized and bundled hardware and software solution. A second type consists of analytical databases that are column oriented, but to a large extend independent of hardware. Finally, a third type consists of column oriented databases with in memory processing. They are offered as appliances, mainly with a specialized hardware. A fourth type includes all other special solutions like “database images”, object oriented databases or special data appliances optimizing the communication between server and storage. For a classification of vendor solution, we refer to chapter 10.3.

Analytical databases are excellent solutions to today’s customer problems: performance, scalability, and cost. The main advantages are:

- Ad hoc queries can be specified in a fully flexible way and are speeded up by a factor of up to 100 times.
- The fast and unlimited access to information significantly enhances user satisfaction and creativity. Now, all data in all details can be analyzed. Analytics is empowered: Better decisions can be made based on facts.
- IT is unburdened, because analytical databases are highly automated. Special knowledge about database design and tuning becomes much less important and critical.

Finally, three things should be made clear:

- Analytical databases make physical database design and tuning nearly obsolete, but a logical business oriented design is still needed as well as information management.

[37] ACID (atomicity, consistency, isolation, durability) describes a set of properties that guarantee reliable execution of transactions.

When deploying analytical databases, master and meta data management, data quality management, information governance and other information management tasks continue to be critical success factors.

- Analytical databases do not replace traditional relational databases for transaction processing, but present a new generation of databases especially equipped for analytic tasks.
- In the meantime, a new class of data base technologies is emerging: They manage operational and analytical data in one and the same database, for instance, *Oracle Exadata*, *Kognitio WX2, and SAP HANA* can do both, high performance analytic and transaction processing. We discuss such technologies in the following chapter 7.3 on NoSQL technologies.

7.3 NoSQL Database and Data Management Systems

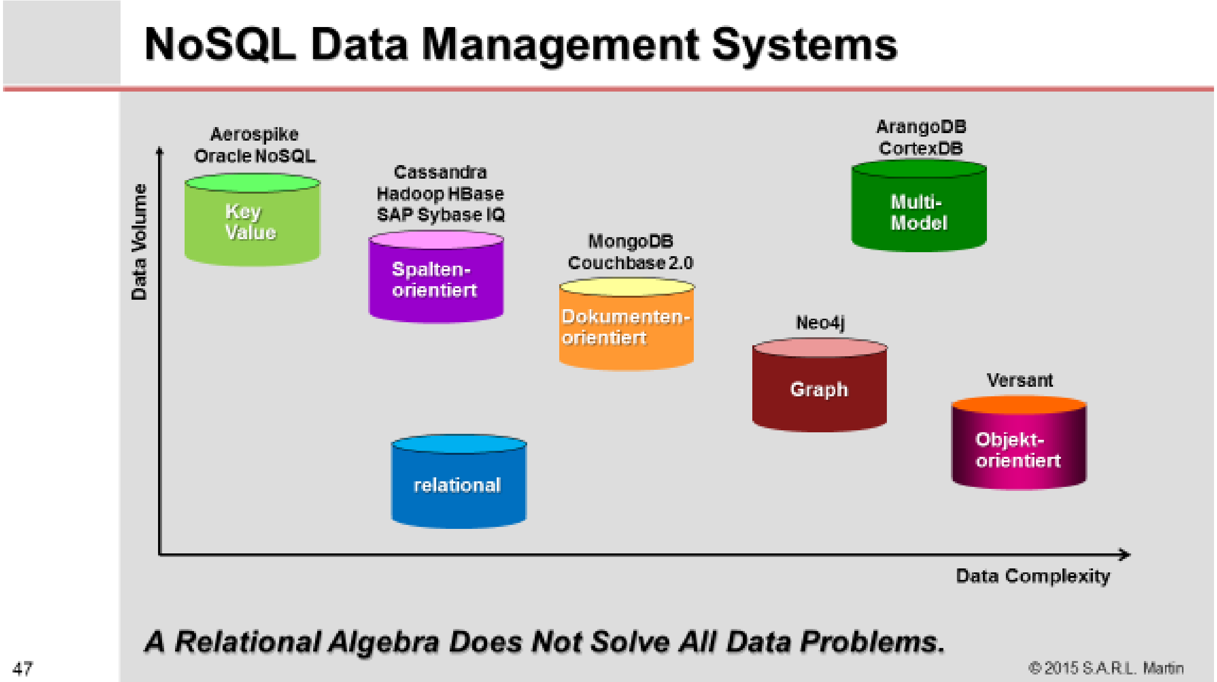
NoSQL data management systems enable managing and processing of poly-structured data. Thus, they complement the traditional relational DBMSs. It is therefore no longer the case that relational databases are the only option. The various methods and technologies used by NoSQL are not all new. Indeed, some of the NoSQL concepts and methods have been in use for many years. Now, Big Data is putting a new focus on them. NoSQL database technologies can be classified as shown in Figure 47.

Object-oriented DBMS. Already in the 90s, they offered alternatives to the relational model. They come with a fundamental principle that can actually be found in nearly all NoSQL data management systems: They do not use a schema and apply alternative techniques for defining how data is stored. Furthermore, they use other protocols than SQL for the communication between application and data management system. Similar to analytic data management systems, the architecture of many NoSQL data management systems is designed and built for scalability. They use distributed managing and processing of large data sets via clusters of standard systems.

Graph-oriented DBMS (or: Entity Relationship DBMS). They use a representation of data through nodes (entities) and relations between the nodes, resp. entities. Thus, nodes replace the traditional data records, and nodes are linked to each other via defined relations. Information about nodes and their relations is stored via properties (attributes). Graph-oriented DBMS especially provide advantages, when networks have to be mapped. An actual example consists of social media networks where the focus is on friends (relations). Graph-oriented DBMS have been first gained interest in the age of Computer Aided Software Engineering (CASE) in the late 80s.

Document-oriented Data Management Systems store "text" of any length together with other poly-structured information and enable search for document content. Stored documents do not necessarily contain the same fields. XML DBMS are document-oriented data management systems with semi-structured data.

Column-oriented Data Management Systems. According to the here used classification, they belong to the class of analytic data management systems. This shows that analytic and NoSQL data management systems are not disjoint to each other: There are analytic data management systems that still rely on the relational model, and there are column-oriented systems that are NoSQL.



NoSQL Data Management Systems

A Relational Algebra Does Not Solve All Data Problems.

47

© 2015 S.A.R.L. Martin

Figure 47: Classification of NoSQL database technologies and positioning according to data volume and data complexity. The mentioned products only play the role of (typical) examples. Multi-modal database technologies combine various NoSQL methods in one database technology. They are "all-rounders". In chapter 10.3, there is a detailed list of column oriented database management systems. This subclass of NoSQL systems is also a subclass of analytical systems. For an exhaustive list of NoSQL database technologies we refer to http://nosql-database.org/.

Key Value DBMS. They are based on keys pointing to values that in its simplest form are any character string. Key value DBMS are also not new, they stem from the UNIX world where they have been used as embedded DBMS like dbm, gdbm and Berkley DB. A Key value DBMS processes data either via in-memory or via on-disk. Their strength is fast search.

Multi Model DBMS. Recently, new NoSQL DBMS came to market that combine two or even several NoSQL technologies. Therefore, they make up a new NoSQL category that addresses not only one of the niche markets, but a much broader spectrum of applications. By the combination of various NoSQL DBMS technologies, they also combine advantages of the different NoSQL technologies. Multi model DBMS are all-rounders and provide answers to today's market needs: presentation of relationships between data objects, joint storage of analytical and transactional data in one and the same DBMS, high performance, distributed databases with eventual consistency and an ecosystem for all tasks. Furthermore, some of the multi model DBMS also provide a time dimension: Data is stored together with its validity ranges so that the DBMS can also track structural changes in time.

NoSQL – when relational database technology reaches its limits. Big Data drives relational database technology to its limits. It is not only the big volumes of Petabytes of structured data that limit the use of relational database technology, but in particular management of non-structured data, of data with complex semantic structures, and of real-time data streams. The latter are best application domains for NoSQL technologies. Their advantages are in particular:

- ***Elastic scaling.*** In contrast to relational database technologies, NoSQL technologies have been designed and built for elastic scaling from the very beginning.
- ***Managing of big data volumes.*** NoSQL systems typically manage data volumes that are higher by one power of ten than the volumes that can be handled by leading relational database technologies.
- ***Simpler and better management.*** NoSQL technologies have been designed for the ease of management. Typically, management functionality includes: automatic repair and distribution as well as easier data models enabling a more efficient tuning.
- ***Economics***. NoSQL technologies run on low-cost, commodity hardware. Cost per terabyte of NoSQL is considerably lower than cost of relational technologies.
- ***Flexible data models.*** Changing data models is much easier with NoSQL databases than with relational. For instance, NoSQL key value stores, document stores and multi-model databases allow applications to define any structure for any data elements. Even the somewhat more rigorously defined column-oriented NoSQL databases like Cassandra or HBase allow adding new columns with big efforts.

Finally, we shall discuss Hadoop, another NoSQL approach, and put some spotlights on Hadoop in the following chapters.

7.4 Hadoop and Spark – Technical Answers to Big Data Challenges

Hadoop was originally initiated by Lucene inventor Doug Cutting. In January 2008, it became a top level project of the Apache foundation. Its goal was cost effective and scalable data management and analysis of poly structured data, i.e. traditional structured data (like tables) as well as non-structured data like text, image, audio, video etc. Hadoop is going to be a standard in Big Data management. It works like an operating system for data. Originally, it consisted of three components:

- the storage layer HDFS (Hadoop Distributed File System),
- the programming framework MapReduce for parallel processing of queries, proposed by Google,
- a function library.

Hadoop includes the **HBase**, a scalable analytic data management system for very large data sets within a Hadoop cluster. HBase is an open source implementation of Google’s “Big Table”. Hadoop also includes database management system **Accumulo**, a key value store database.

The center of Hadoop is the storage layer **HDFS.** It stores data in 64Mb blocks: This supports well parallel processing and is excellently equipped for reading very big data sets. The disadvantage is that such storage does not support transaction processing or real-time analytics. HDFS has built-in redundancy. It is designed for running across hundreds or thousands of low cost servers, where one can expect that some of them will fail again and again. Therefore, in the Hadoop standard settings, each data block is stored three times.

Finally, new data is always added, never inserted („no insert“). This increases the speed for reading and storing data, plus it also increases the reliability of the system.

MapReduce (MR). It was first implemented by Google in its column oriented BigTable based on the Google file system. It is a programming framework for parallelizing queries. The result is a tremendous increase in performance when reading and searching very large data sets. MR is no programming or inquiry language. Programming within MR can be done in various languages like Java, C++, Perl, Python, Ruby or R. MR program libraries do not only support Hadoop, but also other file and data management systems. Some data management systems use MR programs as *in-database* analytic functions that can be invoked by SQL commands. Unfortunately, MapReduce can only be used in batch. It is not modeled and built for real-time and/or interactive processing. YARN and related approaches have very much improved the situation, and in the meantime, they have replaced MapReduce.

Hadoop YARN (Yet Another Resource Negotiator) is a cluster management technology. It is an essential building block of Hadoop version 2. YARN plays the role of a large-scale distributed operating system for Big Data applications. The YARN engine can be considered as the further development of the MapReduce approach. YARN replaces MapReduce as an optional plug-in. This enables alternative processing. For example, the MapReduce batch-model can now be replaced by more interactive components like Apache Storm or by services like Apache HBase.

Hadoop tools. Hadoop is complemented by High Level Query Languages (HLQL) like Hive, Pig, and JAQL. Hive is a Data Warehouse infrastructure originally developed by Facebook. Hive includes the HLQL "QL" that is based on SQL. Since Hadoop programming environments lack of skilled and experienced resources, an HLQL like QL is very welcome, because it allows developers to continue to work in their well-known SQL environment. In the meantime, tools enabling SQL on Hadoop are becoming more and more frequent and are very popular. Hortonworks, for instance, improves Hive by Stinger whereas Cloudera offers direct access to HDFS via Impala. IBM and MapR also offer alternative query engines to Hive.

Pig is another HLQL. It is a procedural language, and in comparison to MapReduce, it eases parallel execution of complex analyses. It is easier to program in Pig, and Pig programs are better documented. Pig also offers certain (not yet really mature) automated optimization of complex arithmetic operations. Furthermore, Pig is open, and can be extended by customized functionality. For managing Hadoop applications, there is Chukwa for real-time monitoring. Ambari, Avro, and ZooKeeper support data and system management.

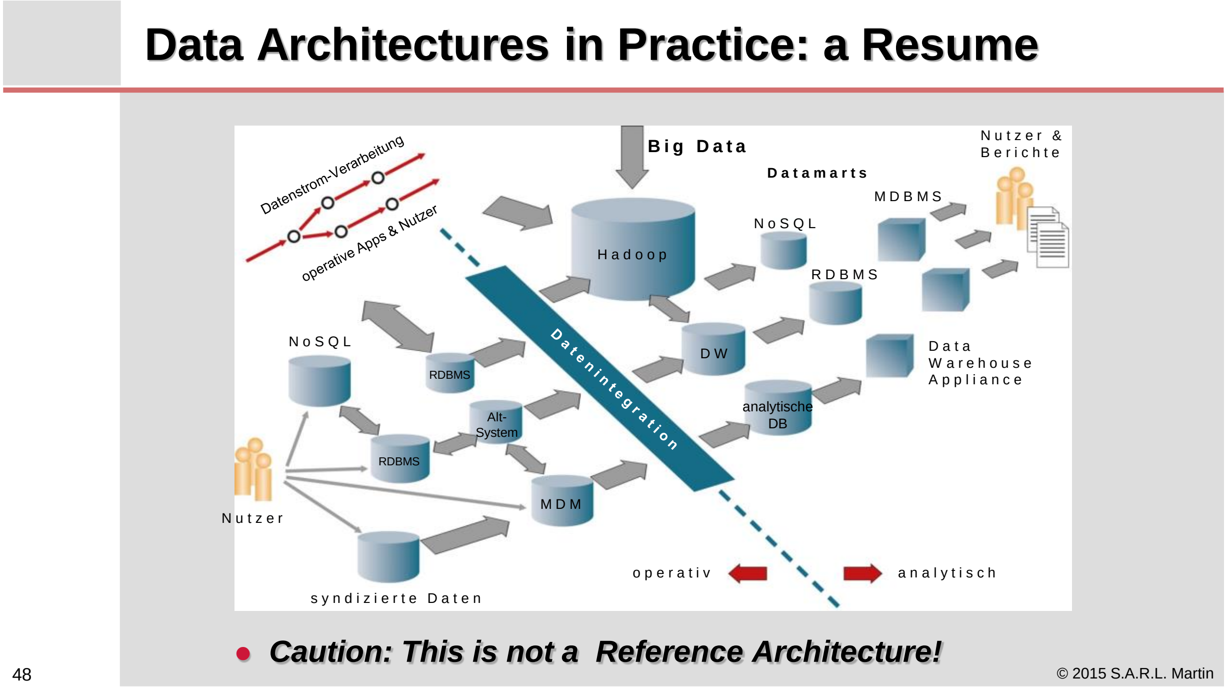
Spark is another open-source platform of the Apache Software Foundation. It is framework of cluster computing originally developed in 2009 by researchers at the University of Berkeley for accelerating processing of Hadoop systems. Since 2013, it is an Apache Software Foundation project, since 2014 a Top Level Project, and since June 2015, IBM supports Spark massively by round about 3,500 developers. The distributor of Spark is Databricks.

Spark can be understood as a fast and general engine for large-scale data processing sitting on top of data stores like Hadoop, NoSQL databases, Amazon Web Services (AWS) and relational databases. It supports in-memory computing, but also traditional disk-based processing of data when data volumes exceed available system memory.

Spark acts as an application programming interface (API) enabling programmers to manipulate data through common applications. It comes with a few ready to go applications plus an SQL query engine, a library of machine learning algorithms, a graph processing engine and a data streaming processing engine.

The actual release 1.4 of Spark now supports the R programming framework and the Phyton language. A data frame API extends Spark SQL and the data frame library by statistical methods. Apache Spark runs on Hadoop 2 clusters via the YARN resource manager, but can also be used as a stand-alone solution, for instance as an Amazon Elastic Cloud (EC2) service.

A big advantage of Spark is its capability to link various traditional data sources as well as big data sources, plus its support for various types of analytics. Spark considerably improves the performance of data-based applications and radically eases the development of learning "intelligent" apps. Therefore, Spark has a good chance to become the unifying technology for big data applications.



Data Architectures in Practice: a Resume

- *Caution: This is not a Reference Architecture!*

48

© 2015 S.A.R.L. Martin

Figure 48: Hadoop and other NoSQL technologies force a further development of traditional data architectures. On the one hand, they complement architectures for operational and analytical data, and on the other hand, they question the future role of a data warehouse. Furthermore, a coexistence between relational and NoSQL technologies is required.

Hadoop and NoSQL technologies make up a big challenge for information management and data architectures. The introduction of these technologies into an enterprise IT landscape ends up in a rather complex structure schematically depicted in figure 48. It is important to note that the shown architecture is not a reference architecture, but an architecture actually existing in many organizations. It can be considered as a starting point, and in the following chapters, we shall outline how “real” data architectures can be built given this point of departure.

Despite the innovative approach and despite all presented functionality and advantages Hadoop already offers today, there are some critical aspects of Hadoop that should not be neglected.

- Hadoop data management still has gaps. In particular, data integration and data management functionality like high availability, disaster recovery, and load management are incomplete or even lacking.
- Hadoop was originally designed for batch processing. Real-time functionality is just arriving, but approaches taken by the various Hadoop distributors differ considerably causing portability problems of Hadoop solutions.
- Hadoop has not yet any notable system built-in mechanisms for data protection. For instance, there is no built-in data encryption.
- Tools are missing for operations and maintenance of Hadoop. A lot of in-house development is required!

But the situation is somewhat better when implementation and operations of Hadoop are cloud based. Hadoop cloud offerings include pre-manufactured, rather easy adaptable Hadoop architectures. Additionally, they come with cost advantages because they allow pay-for-use models. This is also the best way for medium organizations to benefit from Hadoop.

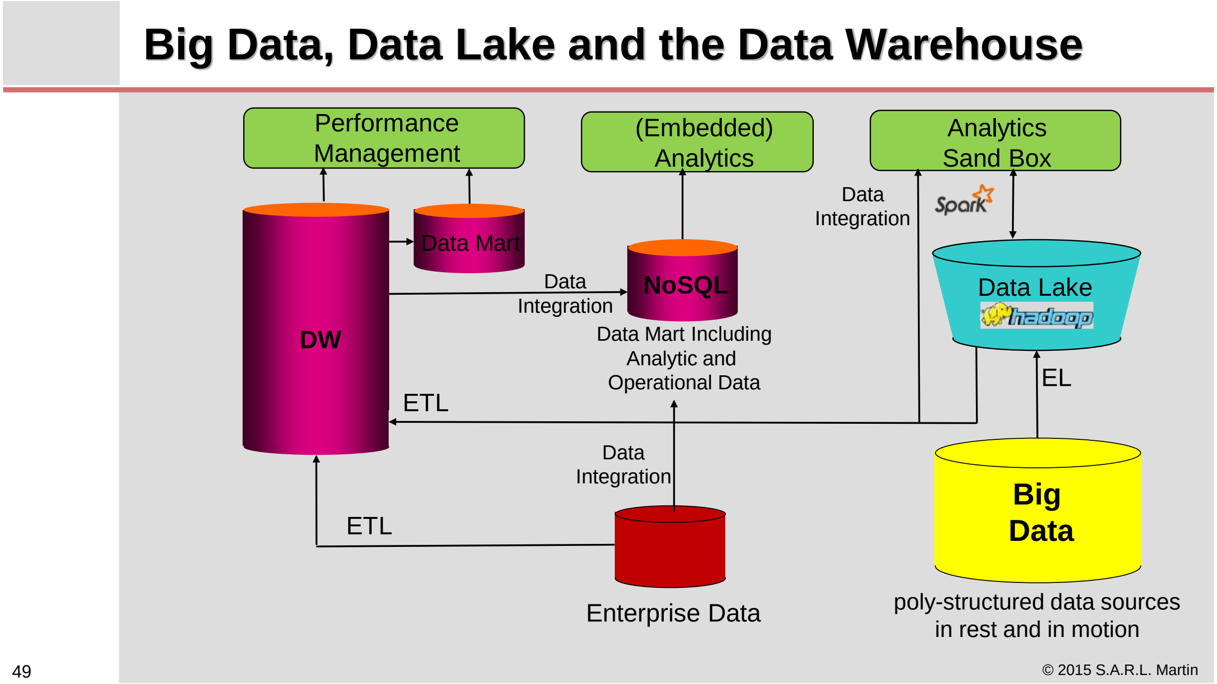
Note. Despite the fact that Hadoop is based on technologies and concepts developed by Big Data enterprises like Facebook and Google, and the fact that Spark is massively supported by IBM, Hadoop and Spark are still in their youth. Bottom line is that deploying and operating Hadoop and Spark needs experienced and skilled resources. But actually, there is still a big lack of such resources. Today, this is definitely a blocker for moving towards these new big data technologies. Furthermore, a lot of functionality is still missing and has to be added by in-house development.

7.5 Hadoop, Spark, NoSQL and the Data Warehouse

Up to now, the data warehouse is still considered to provide the single point of truth, i.e. it is the central platform for all analytics as well as tactical and strategic performance management. In contrast, Hadoop and Spark are new platforms supporting analytical workloads that go beyond those supported by a data warehouse. In addition, more and more vendors provide direct access to Hadoop so that traditional BI tools can now access Hadoop via SQL on Hadoop. Indeed, SQL on Hadoop is making progress.[38] Will the data warehouse and its relational technology become obsolete and be replaced by Hadoop? What will be the role of Spark?

[38] see http://www.b-eye-network.com/blogs/vanderlans/archives/2014/02/the_battle_of_t.php, accessed on December 18th, 2014.

The market place has not yet a final answer to this question. A point of view has been taken by Steve Miller: “Big Data and the data warehouse serve different masters. DW has historically revolved on performance management, while Big Data obsesses on analytical products for data-driven business.”[39] We share his point of view. In fact, the requirements of performance management and analytics to data and data management are quite different.


Big Data, Data Lake and the Data Warehouse

49

© 2015 S.A.R.L. Martin

Figure 49: For a foreseeable future, Hadoop and Spark will not replace the data warehouse, but extend data warehouse architecture. This new extended data warehouse architecture will include NoSQL technologies and guide the transformation of existing data landscapes (cf. fig. 48) towards architecture.

The data warehouse will remain the central platform for tactical and strategic performance management as well as source and feed for various data marts. However, an increasing number of new types of analytical workload will run on Hadoop, Spark and some NoSQL databases. This will in particular be the case if high performance and extreme scalability are required and/or data structures are highly complex or both analytical and transactional data is needed in a data mart. Data integration will combine the later with the corresponding operational systems. In particular, it enables operational performance management in real-time as well as to embed analytics into business processes.

Analytics will increasingly adopt sand box principles and run on Spark. A data lake acts as an intermediate data store (replacing old ODS technologies) typically based on Hadoop for providing a pool of relevant data from big data sources. Spark then offers not only a common analytics platform, but also the integration of a data warehouse and a data lake: Certain analytical results must be transferred to the Data Warehouse, and extracts from the data lake will supply the Data Warehouse. (DW = data warehouse, ETL = extract, transform, load, ODS = operational data store, PM = performance management)

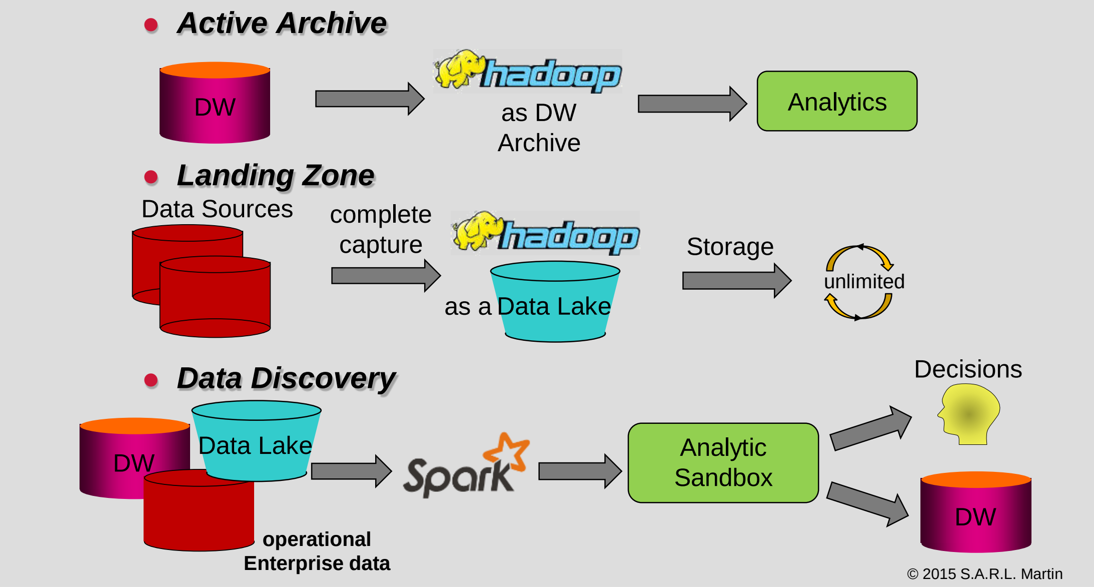
Performance management needs high quality meta and transaction data. This is all about structured data in rest – with the exception of operational performance management. Indeed, operational performance management has always used special data marts. – Performance

[39] see http://www.information-management.com/blogs/big-data-vs-the-data-warehouse-10025458-1.html

management is based on a well-defined data model. Changes to the data model go with the speed of changing business models or slowly running business processes. This is still in the range of days, not hours or even faster. The data warehouse data model is therefore (still) rather stable. Furthermore, the data volumes used in performance management rarely exceed the terabytes range whereas Hadoop can scale to handle petabytes. To summarize, relational data base technology is still sufficient to meet the needs of tactical and strategic performance management, and a data warehouse offers a proven and performant environment for managing important, business critical analytic data. For example, a data warehouse is more than adequate for compliance.

There can be exceptions to this general rule, for instance if a data warehouse does not include the level of detail that is needed for a particular performance management task. Then, the traditional solution is building a data mart with the required level of detail by using data integration and linking the necessary operational data to the corresponding data warehouse data. Data quality services can be embedded so that we end up with the necessary quality of data. If furthermore, operational and analytic data need real-time processing, for instance for embedding analytics into business processes, then NoSQL technology is best suited for such data marts. Operational planning, disposition, and operational risk management are typical use cases (see Martin, 2014).

Hadoop/Spark as a Complement to the DW



- *Active Archive*

DW → hadoop as DW Archive → Analytics

- *Landing Zone*

Data Sources → complete capture → hadoop as a Data Lake → Storage → unlimited

- *Data Discovery*

DW, Data Lake, operational Enterprise data → Spark → Analytic Sandbox → Decisions, DW

50

© 2015 S.A.R.L. Martin

Figure 50: Hadoop and Spark complement the data warehouse in three scenarios: active archiving, data capture, and as a foundation for sand box style data discovery.

Analytics comes with different needs and requirements. It uses structured and unstructured data as well as data in motion. The requirements to analyze data streams is increasing. Analytics also deals with high to very high volumes of data. It uses various Big Data sources and source types and combines them with enterprise data. Furthermore, the purpose of analytics is to gain new, unknown insights into customers and markets as well as to enrich customer interactions as well as production, logistics and other business processes by embedded predictive and operational analytics. One of the best practices to address these challenges is a sand box approach to analytics. Independently from compliance and governance restrictions and requirements, new ideas as well as new methods and technologies can be better tested and of course also rejected by such sand box concepts. Furthermore, sand boxes can be put into the cloud. Then a sand box provides high speed and agility at less cost. For such a type of analytics, the Cloudera's Enterprise Data Hub (EDH) concept makes sense. The idea is to put all relevant data from Big Data sources into a **data lake** (typically based on Hadoop) and use for instance Spark's data integration framework to link that data to the DW and to other enterprise data. Figure 49 depicts the discussed scenarios. It also shows a migration path towards data architecture combining and integrating existing data silos as shown in figure 48.

The idea of a data lake is to store data in its original form preferably without preprocessing, and to apply ETL processes not before data is needed for analytics. It's the concept of „late binding". We save cost at data capture and storage, but we have to pay the full bill when data is needed. But in the end, there is no cost and time saving. Consequently, a data lake needs certain semantic structures for processing its data in a more efficient and cost effective way. At least, we need:

- At data capture of a big data source, a minimum of meta data is indispensable. We need information about who, what and when data is stored and where it comes from. The use of free text for documentation would be the simplest solution, but a wiki would do better. Furthermore, we need specifications for building trust into data lake data, for instance, specifications about completeness and authentication of data.
- A lot of future work will be saved, when data lake meta data is immediately linked with the business vocabulary, and when semantic links are created linking the various terms. In the end, these are active cross references that can be mapped at best by graph methods and technologies.
- A data lake should also be part of information governance. For instance, certain requirements about data security are to be checked, in particular the need for encryption of data or access rules corresponding to the roles of various data lake users.

The leading Hadoop distributors Cloudera and Hortonworks follow this approach, and position Hadoop as a data warehouse complement, not as a replacement. There are in particular three scenarios where Hadoop with its low-cost data storage unburdens and improves a data warehouse also in the sense of cost savings (fig. 50).

- ***Data Warehouse archiving.*** Hadoop offers cost effective and nearly unlimited data storage. If Hadoop is used as a data archive for a data warehouse, it provides an additional advantage: Archived data stored in Hadoop can at any time be used for analytical tasks like ad-hoc queries or time series analysis. Hadoop removes the burden of revival of data that has been archived in traditional storage media. Such a revival typically required interventions from IT. Therefore, an archive managed in Hadoop is an active archive.
- ***Landing zone.*** Data from various sources can be completely captured and stored through Hadoop in a data lake at low cost, until it is used for analysis or transferred into a

data warehouse. As in the case of active archiving, data in a data lake is nearly unlimited and always ready to go for analytical tasks. Filling Hadoop with data is different from filling a data warehouse. We used ETL processes to fill data warehouses, but we use EL processes to fill Hadoop. The steps extract (E) and load (L) transfer data into Hadoop. A schema design follows in Hadoop or Spark, and the transformation (T) step follows, when data from various sources is to be integrated for an analysis. If we follow the guidelines of building minimal semantic structures for data lakes, then the T step gains speed and agility at less cost.

- ***Data Discovery via sand boxes.*** Enterprise data and data from Big Data sources can be jointly and cost effectively stored in a data lake. Here, all data is ready for all purposes of data discovery, for instance, for pilot analyses. Results can then be transferred into the data warehouse or can be continued to be analyzed.

The term "**data lake**" has been proposed by James Dixon, CTO at Pentaho. In his blog mentioning the term data lake for the first time, he writes: "If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.[40]" This approach is currently under some critique (for instance by Gartner[41]), but anyway, it makes a lot of sense when considering that a data lake includes a certain set of policies and rules that turn the somewhat washy term into a data architecture. Andrew C; Oliver gives quite a good reasoning.[42].

Take Away. For a foreseeable future, Hadoop and Spark will not replace the data warehouse, but extend the data warehouse architecture. The data warehouse will continue to be center and single point of truth for tactical and strategic performance management. But ETL processing and analytical workloads will increasingly move to Hadoop and Spark using the principles of a data lake, notably if they involve analyzing non structured data or data in motion.

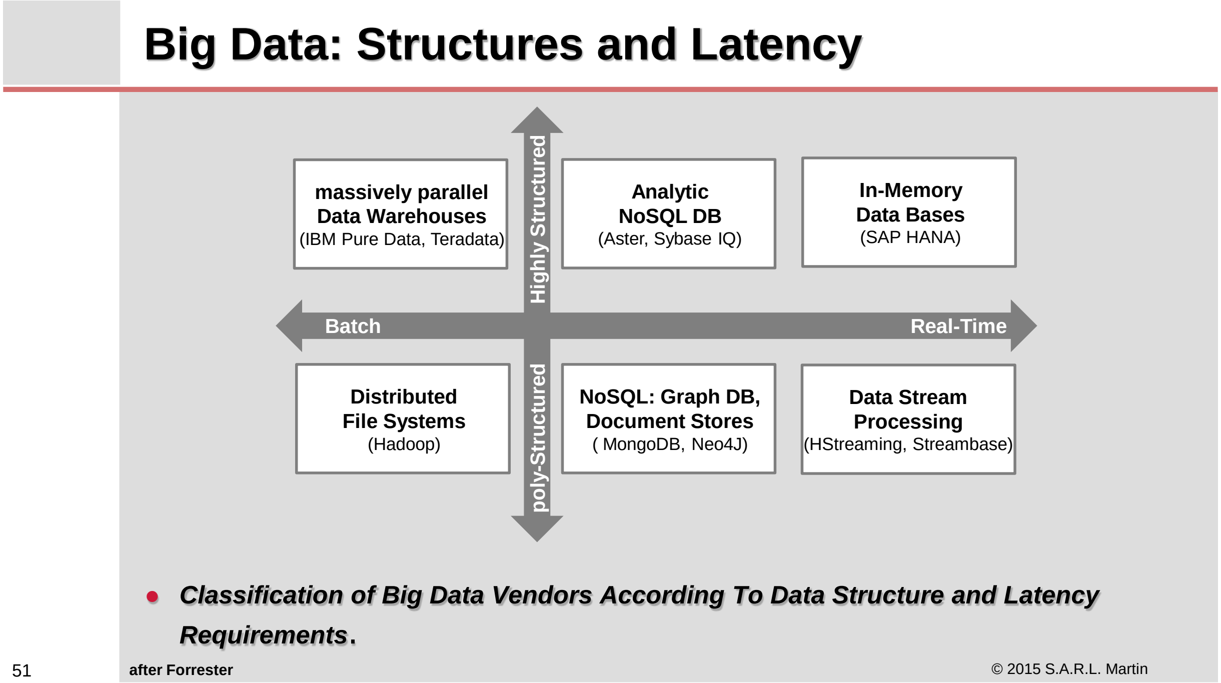
7.6 Big Data: Data Structures and Latency

Vendor offerings for Big Data Analytics finally can be classified according to different data structures and latency requirements. Figure 51 illustrates this classification via the two dimensions complexity of data structures and processing in batch or real-time. "Real-time" can have various meanings as already discussed in chapters 2.3 and 7.1. It is either related to low latency access to available data ("analytics in real-time") or to processing and querying of data streams ("real-time analytics").

[40] see http://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/, access at dec. 17th 2014.

[41] see http://www.gartner.com/newsroom/id/2809117, access at dec. 17th 2014.

[42] see http://www.infoworld.com/article/2608618/application-development/gartner-gets-the--data-lake--concept-all-wrong.html, access at dec. 17th 2014.



Big Data: Structures and Latency

Highly Structured

poly-Structured

Batch

Real-Time

massively parallel Data Warehouses
(IBM Pure Data, Teradata)

Analytic NoSQL DB
(Aster, Sybase IQ)

In-Memory Data Bases
(SAP HANA)

Distributed File Systems
(Hadoop)

NoSQL: Graph DB, Document Stores
(MongoDB, Neo4J)

Data Stream Processing
(HStreaming, Streambase)

- ***Classification of Big Data Vendors According To Data Structure and Latency Requirements.***

51 **after Forrester** © 2015 S.A.R.L. Martin

Figure 51: Big Data solutions classified by data structures (highly structured and poly-structured) and latency requirements (batch and real-time). The mentioned vendors just play the role of proxies for the members of this class. For a more detailed classification of analytic databases, we refer to chapter 10.3.

Let’s look in more detail to the four quadrants of Figure 51:

- ***Batch and highly structured.*** Solutions in this quadrant are based on a massively parallel architecture and a highly scalable, virtual infrastructure. Such an approach reduces considerably cost of storage and highly improves efficiency of processing of traditional data warehouses. Leading vendors are for example Oracle with Exadata, IBM with Pure Data (formerly Netezza), and Teradata.
- ***Real-time and highly structured.*** Here, solutions focus on analytical processing in real-time and data mining for predictive analysis. If the issue is “just” quick analysis (“analytics in real-time”), then analytic NoSQL database management systems are good solutions, but if the issue is “real-time analytics”, then in-memory databases are best solutions, because they jointly manage analytical and transaction data in main memory instead on disks. Furthermore, they gain velocity by radical reduction of I/O time when accessing data. Finally, they offer a better estimation of performance than disk-based data bases. Leading vendors are on the one hand SAP with Sybase IQ and Teradata with Aster, and on the other hand Oracle with TimesTen and SAP with HANA.
- ***Batch and poly-structured.*** Solutions in this quadrant are based on a software framework including typically a distributed file system, a processing engine for large volumes of data, and applications for managing the framework. A prominent example is given by Hadoop.
- ***Real-time and poly-structured***. If again the case is analytics in real-time, then NoSQL technologies like graph or object-oriented data management systems are well

appropriate. The technology for treating real-time analytics is event-stream processing for managing multiple event streams and getting meaningful insights. We already discussed such technologies in chapter 7.1: Complex event processing for identification of complex patterns in multiple events, event correlation and abstraction. Leading vendors are on the one hand MongoDB, Neo4J, and on the other hand HStreaming, Streambase, and Splunk.

8 Ethical Aspects of Analytics

8.1 Big Data – Big Brother

Big data puts a spot on analytics. Analytics becomes more important than ever: We gain details in a not yet known precision; we get insights into yet unknown and unexpected structures. In Healthcare, progress in development of new medical treatments is triggered by the identification of unknown relations through testing hypotheses generated by big data analytics. Real-time information about location and navigation data produced by smartphones of commuters can help city planners to design better transportation systems. The number of shipped UPS parcels indicates economic regression and rebound. Big data analytics makes all these useful things happen. Big data analytics is not only about adding value to businesses, but also about smarter life for everybody.

As always, there are certainly risks and drawbacks in big data analytics. One of the most discussed risks has been PRISM, the spy out program of the US agency NSA (National Security Agency). This program for instance, collects and analyzes data about phone calls and web sites. The goal is national security and counter-terrorism, i.e. totally legal in the US. In Europe, not only data protection officers think somewhat differently. Do consumers or citizens really want that businesses or governments know everything about them? Is big data analytics not a straightforward way to a “Big Brother”[43] world?

The Big Data Troika: data protection and security, privacy, and ethics. The idea of big data is to combine various data sources for analysis. But this is already a violation of fundamental European data protection principles. Where is data protection, where is privacy, where is ethics in using and exploiting data?

Actually, there are three different schools having rather different thinking about data protection and privacy, a US-American, a Chinese-Asiatic, and a European one.[44]

> *Example, the US-American view.* Fred Cate, director of the Center for Applied Cybersecurity Research at Indiana University, believes that the application of big data analytics at PRISM data is a must. The FBI and CIA have always mapped social networks to figure out who is talking to known terrorists. Big Data makes that process faster and more detailed. “There is an old joke about the FBI investigating a lot of pizza delivery places,” Fred Cate said. “People in hiding tend to have food delivered, and make a lot of calls for pizza.”[45]

[43] The term “Big Brother” has been created by George Orwell in his novel “1984”. It denominates the dictator of the totalitarian state “Oceania” that perfected government surveillance and public mind control. See http://en.wikipedia.org/wiki/Nineteen_Eighty-Four (access July 31th, 2013).

[44] Boston Consulting Group, DLA Piper: Earning Consumer Trust in Big Data: A European Perspective, ComputerWeekly.Com http://bitpipe.computerweekly.com/fulfillment/1430306410_422, accessed June 15th, 2015.

[45] “Big Data’s big deal - The power of pattern in collective human behavior”, Farah Stockman, The Boston Globe, http://www.bostonglobe.com/opinion/2013/06/18/after-snowden-defense-big-data/7zH3HKrXm4o3L1HIM7cfSK/story.html, accessed June 19th, 2013.

Whereas in Europe, the question of data protection and data security is principally regulated[46], and that of privacy is under discussion, the question of data ethics is not yet well understood and less considered. But data ethics is at least as important as the two other questions.

> „Consumers need to be convinced that their personal data is adequately protected and used fairly; trust is the key to maintaining a fruitful relationship."[47]

In other words, application of big data analytics needs big data ethics. The goal of big data ethics is to develop criteria for good and bad actions and behavior as well as to assess and to communicate motifs and consequences.

The dark side of big data. According to Victor Mayer-Schönberger and Kenneth Cukier[48], collection of data is not really big data's dark side. The real danger is not so much spying on a single individual, but rather the transformation of an individual into a piece of a vast pattern that can be used to predict events.

> "What we need to fear from Big Data is not necessarily old-fashioned surveillance, but probabilistic predictions that punish us not for what we have done, but what we are predicted to do." (Mayer-Schönberger, 2013)

As good as it may sound to use data to prevent crimes by predicting who will commit them in advance, that kind of activity crosses the line and subvert the most fundamental European principles of justice. Even so, companies already do something similar: Credit cards companies block a credit card if a purchase is made too far afield from the usual routine. Banks and insurances determine levels of risks using algorithms that benchmark individual risks with those of millions of other people in the same customer segment. In Germany, a credit information agency had the idea to use data from the profiles of friends in social media for identifying individual credit risks. Public opinion screamed, and they stepped back from the contract. But do we know whether meanwhile they aren't using another contractor?

In the age of Big Data, the issue is not only how to protect and to secure data and how to shield privacy. Privacy, in an old fashioned sense, is even gone. This was already noted long time ago by the Silicon Valley captains. They should know, because they tried hard to kill data privacy. 15 years ago, Scott McNealy said, "You have zero privacy anyway – get used to it."[49] So, the "good" question is now: How can we be sure that big data analytics will be applied and used only for good for all of us, the citizens and the consumers, and in conformity with basic rights? Indeed, data ethics is all about this.

Big Data Ethics. But there are not yet real answers to this question. The Americans have certain consent about data collection. For them, this question does not really matter. Indeed,

[46] Warwick Ashford: EU Data Protection Regulation to be finalised by end of 2015, ComputerWeekly.com, http://www.computerweekly.com/news/4500248164/EU-Data-Protection-Regulation-to-be-finalised-by-end-of-2015, accessed June 16th, 2015.

[47] Boston Consulting Group, DLA Piper: Earning Consumer Trust in Big Data: A European Perspective, ComputerWeekly.Com http://bitpipe.computerweekly.com/fulfillment/1430306410_422, accessed June 15th, 2015.

[48] Big Data: A Revolution That Will Transform How We Live, Work, and Think, http://www.amazon.de/Big-Data-Revolution-Transform-Think/dp/1848547919, accessed June 19th, 2013.

[49] see InfoWorld http://www.infoworld.com/d/consumerization-of-it/maybe-just-maybe-users-can-win-the-privacy-war-213222?source-twitter, accessed April, 16th, 2013.

data collection is wide-spread and done without our control. Sometimes, it is even necessary. For instance, mobile telephony vendors need such data for running their systems. Location-dependent services do not run without location data. Finally, we all give voluntarily data to companies like Google and Facebook that in return give us information that is hardly available or accessible. Most of us also give away their navigation data to Apple or Google. In other words, if data collection is prohibited we should be ready to give up all advantages coming from collected data. Are we ready to give up these habits?

Furthermore, certain customer types do want to get an added value through big data analytics in the context of customer experience management. It's a rather new issue triggered by the digitalization of the world and of organizations: Real and virtual world melt in the mobile internet.

> *Example: mobile commerce.* In a shopping mall, locations of customers can be identified with a precision of up to 2m through triangulation of Wi-Fi signals. This allows to initiate customer contacts and to use them for customer interactions, for instance a recommendation for a customer to visit a nearby contact point. Based upon geocoding, all relevant customer contact points (boutiques, branch offices, agencies, etc.), can be used to identify the "best next point of local contact". The customer could also be given a stimulus to visit a particular contact point, e.g. with a special offer coupon. Here, at the latest, is where the customer can be identified and all information from the CRM/CEM system made available locally for further cross- and up-selling purposes – a possibility previously only known in web shops. Furthermore, the customer profile can be enriched with spatial information providing a movement profile. It is the basis for an improved local customer approach and interaction – the best opportunity for positively intensifying the customer experience. Furthermore, the success of hit ratios *(visits to the contact point and redemption of coupons)* can be measured and hence, can be controlled and optimized.

Such a scenario is definitely perceived as horror for certain types of customers, whereas others are not only open for such experiences, but even delighted. A new challenge for customer analytics opens up: How to identify customer segments that have affinity or non-affinity to such scenarios. Indeed, in the US, retail marketers research actively solutions through an experimental business approach[50].

In my eyes, such an approach is a consequent application of big data ethics: Use and exploit data for the benefits of customers, respectively, apply big data analytics only if customer agrees and desires and will get a similar advantage as the business. Then, it's a real "win-win". Such an ethical principle in dealing with customer interactions corresponds to Schopenhauer's message of his hedgehog parable[51]: not too close and not too far.

It implies a new principle of customer experience management. As CRM is evolving towards CEM, the golden rule of CRM:

- Treat your customers the same way as **you** want to be treated

[50] Stefan Thomke, Jim Manzi: The Discipline of Business Experimentation, Harvard Business Review, December 2014 https://hbr.org/2014/12/the-discipline-of-business-experimentation, accessed June 15th, 2015.

[51] Arthur Schopenhauer: Die Stachelschweine, see http://gutenberg.spiegel.de/buch/arthur-schopenhauer-fabeln-und-parabeln-4997/1, accessed July 15th, 2015.

turns into the platinum rule of CEM:

- Treat your customers the same way as **they** want to be treated.

Consequently, knowing his customers even better than ever is key, hence customer analytics is key! On the contrary, in case of doubt, do not apply such procedures, or: think of your big data ethics.

Big data ethics as a foundation of trust. Therefore, the focus of all discussions about the big data troika should more focus on big data ethics. Question is how to establish good and comprehensible rules for use and applications of big data analytics. "If I use that data to save your life, you are not going to care how it was collected," said security expert Fred Cate. "But if I use that data to track you down, then it is going to bother you." A report from the World Economic Forum "Unlocking the Value of Personal Data: From Collection to Usage"[52] goes into the same direction. It recommends a major shift in the focus of regulation from data collection toward restricting the use of data. "There's no bad data, only bad uses of data," says Craig Mundie, a senior adviser at Microsoft, who worked on the position paper.

Unfortunately, European legislation still neglects the question of big data ethics. It still focusses on regulating mere and sheer data collection. But, we are all concerned, and we have to find a compromise between data protection hardliners and a non-existing and insufficient responsibility and ethics of data barons of companies that offer data-centric products and services. Such a compromise should provide us with acceptable and reasonable rules and policies about who how when why and for what profits are reaped with our data. But this goal is still far away.

8.2 Governance in Dealing with Big Data Analytics

In chapter 3.3 governance and BI governance and in chapter 6.8 information governance have been defined and discussed. In the context of ethical aspects of analytics, due to big data analytics, governance gets important additional tasks and goals.

When starting with big data analytics of customer-related data, businesses should implement a training program in customer rights, data protection and security, as well as general data handling policies. In the era of big data, this should be part of an overall information governance program: Information governance now becomes even more important. Customer data must be handled with sensitivity. Therefore, organizations need rigid rules for handling and exploiting of data. These rules have to be communicated, trained, and must be lived. Furthermore, big data brings more and more employees into contact with customer data. It creates a huge amount of additional training. But only few organizations have yet taken actions. There is a considerable backlog.

Information governance of big data also requires knowledge and documentation which data is collected, which data is acquired, who is responsible for what data, where the sources are, and where data is exploited. These are just standard tasks of information governance, but

[52] see New York Times http://www.nytimes.com/2013/03/24/technology/big-data-and-a-renewed-debate-over-privacy.html?smid=tw-share&_r=2&&pagewanted=print (access June 21st, 2013).

organizations are rather lazy in keeping track of this information about data, its collection, and its usage.

It should be a good practice that there is a responsible C-level person for information governance and its add-ons for big data analytics. A "**Chief Data Officer (CDO)**[53]" is needed. A CDO should have the full responsibility to leverage the data asset in the organization. They also need to report directly to the business, because IT does not feel the pain of data problems, but the business does. Consequently, a CDO is also responsible for the creation of an analytical culture in the organization. Indeed, understanding and appreciation for the role that data plays in the business is essential. The last part of it is that the organization has to recognize that data work must precede any process modeling and development work: Collection, preparation, and processing of data come first. If not, then any of the good initiatives for speeding up application development, for instance by agile methods, will fail, because agile methods only work really well when predefined data components in a library are ready to be accessed.

Definition: A chief data officer CDO) has the responsibility for enterprise-wide information management and governance. He/she should be on the C-level. The position of a CDO is related to the tasks of a CIO, but separated. As a rule, the CDO should report to the Chief Marketing Officer (CMO) or to the Chief Execution Officer (CEO). He/she has the responsibility to manage data as an asset for the organization and to optimize adding value by data and information. He/she assures that the right data is collected, analyzed, and used for decision making. He/she also assures that ethics for analytics are developed and applied in the context of compliance.

The position of a CDO gets definitely necessary when big data analytics is to be deployed. Shawn Banerji, Managing Director of Russell Reynolds Associates, an executive search firm, believes that in 2015, 50% of all fortune 500 companies will have a CDO. In 2012, just 5% had one.

Let us stress one task of a CDO: He/she has the responsibility to manage data as an asset for the organization. Therefore, a CDO should also head up information governance empowering him/her to develop an analytical culture and ethics for analytics. This culture and these policies have to be developed in close cooperation with the CMO and CIO.

Finally, a CDO also has the responsibility for customer communication about usage of customer data. Businesses should clearly and correctly communicate to its customers which data is collected and how it is exploited and applied. Such a transparency creates trust, and in the end even competitive advantages. Furthermore, customers should have the option to determine how data collected about them should be and should not be applied. This seems to be a good approach to switch off big brother attitudes as also discussed in the previous chapter. But it looks like we still have to go a long way.

[53] see Peter Aiken in http://searchdatamanagement.techtarget.com/news/2240185340/Does-your-C-suite-need-a-chief-data-officer-Peter-Aiken-thinks-so, accessed June 26th, 2013.

9 Résumé: Performance Management and Analytics versus traditional BI

9.1 *Performance Management versus traditional BI*

Performance management has its roots in old decision support approaches from the 70s. In the meantime, performance management became a considerably broadened model in comparison to traditional business intelligence.

- Performance management is a top down model that begins with business strategy. Business Process Management links process analysis and design with cross-functional and cross-departmental process flows and performance management. Process performance metrics are created at the same time processes are designed, and they are linked via roles to the organizational structure. The foundation is a professional information management.
    - Business Intelligence (BI) was bottom up and not process-oriented.
- Performance management is based on an information supply chain model that permanently synchronizes the provision of information with the need for it.
    - Business Intelligence was restricted to an information-providing model (Bill Inmon "Information Factory").
- Performance management is a closed-loop model that controls and monitors business processes at operational, tactical and strategic levels.
    - Business Intelligence only supported decision making, but not action taking. The operational aspects of Business Intelligence were not covered by a coherent approach.
- Performance management metrics are forward-looking. Predictive analytics enable the identification of problems before they appear. Nevertheless, traditional retrospective metrics can remain useful.
    - Business Intelligence was retrospective (based on the past). The focus was on analysis and diagnosis. Potentials of predictive models were not exploited.
- Performance management enables transparency by means of BI governance. Everybody gets in right time exactly the information required in the context of his/her processes.
    - Business Intelligence tools did not provide the information consumer with sufficient information. Either one had information that was not accessible (on occasion even hidden or held back), or you had an absolute flood of data ("information for the masses"). This spoiled acceptance.
- Performance management is based on analytic services that are published, consumed, and orchestrated in the context of a SOA (service oriented architecture). This also enables a seamless transition of traditional deployments of performance management via

on premise solutions to cloud-based solutions, mobile included.

- Business Intelligence was a tool-related approach, based on proprietary technologies. This resulted in stove piped information silos causing redundancies and inconsistencies.

9.2 Analytics versus traditional BI

The purpose of analytics is analyzing large and very large data sets ("Big Data analytics") for supporting the processes of decision-making through facts and acquired knowledge. Data sources to be analyzed are historical ("analytical") and operational ("transactional") data. Besides static reporting, traditional BI already included some components of analytics like ad-hoc queries and OLAP analysis. In the course of time, analytical components like statistical procedures and data mining were added. In the previous chapters, we have shown how these components evolved to today's analytics, and which components have been added up to Big Data analytics.

Diagnosis, identification of correlations and interferences as well as detection of interrelationships within complex systems were the tasks of first approaches to analytics in the context of traditional BI. Focus was put on comprehending the past. It was a retrospective view.

Today, objectives of analytics also include comprehension of complex systems in the future. Focus is put now on a forward looking view based on the comprehension of the past. That’s the idea of predictive analytics. It follows the lines of analyzing the actual state of a system, understanding the interrelationships in this system and deriving trends and tendencies about the future development of the system. A good example is given by proactive maintenance of equipment, engines and systems as already discussed.

In the meantime, predictive analytics evolved to prescriptive analytics. First, it worked as the foundation of buying recommendation systems in retail. It used customer behavior and other customer attributes as well as the context of a customer interaction as input to the machine-based decision process. Prescriptive analytics is also well established in insurance and health care enabling individual treatments. Take IBM’s Watson as a good example. Watson recommends individual treatments to doctors who have the last word to make decisions based on these machine generated recommendations. So, a human being is (still) the intermediator between a machine and a recommended individual treatment;

To summarize, we can define analytics as the discipline including all methods and technologies for data discovery, statistics, predictions, and optimization. In this sense, analytics supports performance management by provisioning the right information in right time and to the right spot for enterprise and business process monitoring and controlling. Consequently, in an enterprise, an analytics platform has to be established first. It then serves as the foundation of a performance management methodology as a management concept.

We have also seen that reliability of analytics depends on quality of data, effectivity of methods and tools, and last but not least, on competency of human workers in their various roles within analytic processes.

There is another drawback: We need well-trained managers and staff members for applying, comprehending, and communicating analytics and analytical results. Tom Davenport (Henschen, 2010) even said that the lack of sufficiently analytically trained staff members was one of the reasons for the ongoing finance and debts crisis: All financial and trading systems are highly automated and analytically absolutely top-equipped. But, it is the lack of people that were (and still are) in a position to pursue all metrics and all analytical results, to interpret them, and to explain them to top management. The same holds for Big Data analytics. The needed data scientists do nearly not exist in the market. We still have to go a long way until we will come to a solution.

10 Players in the Performance Management/Analytics Market

10.1 Trends in Performance Management and Analytics

During recent years, the performance management and analytics market ("BI market") has shown an over average performance with sometimes even two digit growth rates. We believe that the following five statements deliver good insight and good arguments about reasons: the extensive rearrangement of the BI market, the evolution of BI role, methods and tools, the transition to performance management and analytics as well as the growing penetration of the specialist departments by tools.

Statement 1: The market for Business Intelligence meanwhile disintegrated and rebuilt itself. For several years already, we have observed increasing merger and acquisition activities in the market. The clou happened in 2007 when three mega acquisitions took place: Oracle/Hyperion, SAP/Business Objects and IBM/Cognos. No real big independent BI vendor exists any more (only exception: privately held SAS). In consequence, there is no independent BI market any more, but it has been absorbed by the BPM/SOA, respectively ERP II market. Indeed, the big four in the BPM/SOA market are all leading BI vendors, and this holds also for the ERP II market. The market leader Infor already gives a good example. This market change did not happen by surprise, we anticipated this trend already in 2006 (expert opinion in is-report 3/2006). But in the new, extended market, the remaining small and independent BI vendors can very nicely occupy interesting and lucrative niches. The on-going process and service orientation, Big Data as well as the cloud computing trends (including BI as a Service, the new IT provisioning model for consuming external services in a SOA) empowers a best of breed model for vendor selection more than ever, because integration is no more a challenge, but is a given. As a consequence, the outlook for the smaller players is excellent due to this market move.

For instance, Big Data opens new market potential, and Data Discovery is becoming a new core area. It is driven by mid-sized vendors like Datawatch, MicroStrategy, Qlik, Tableau und TIBCO Spotfire plus the two open source heavy weights Jaspersoft (acquired by TIBCO) and Pentaho that meanwhile moved into the Data Discovery market segment and also address big data analysis. The traditional Big Four have missed to address this segment in due time and run the risk to lose market share. Furthermore, by the acquisitions of Vertica and Autonomy, another big player entered the BI market in 2011: HP. Finally, TIBCO followed in 2014 with the acquisition of Jaspersoft: TIBCO now became a full-range BI vendor. From now on, we have to speak about the Big Six.

Finally, new markets for "intelligence" spun off the former BI market, for instance, content intelligence, customer intelligence, financial intelligence, competitive intelligence, and social media intelligence (text analytics). These emerging markets offer new growth opportunities for new and/or repositioned vendors. Indeed, the disintegration of the traditional BI market does not mean death of the market, but a restart with many opportunities for all players. The market for analytics and performance management regained high competitiveness.

Statement 2: Analytics and information are ubiquitous. During recent years, BI changed and evolved a lot. BI became operational and mobile, BI was finally put into the context of business processes, and BI was extended to a ubiquitous closed loop model for monitoring and controlling business processes. Finally, the old paradigm that BI only works on top of a data warehouse was shown to be too restrictive and insufficient. Operational data sources became as important as traditional data warehouse data. The data warehouse stopped to be the single point of truth, and a new challenge was created: data integration extending traditional ETL. So, ETL processes continued to be necessary, but we needed more and had to end up with a true “Information Management” enabling the transition from traditional BI to performance management. This “new” BI became an indispensable part of everybody’s everyday business operations. Indeed, the mobile internet is empowering all field workers and services to benefit from BI just as office workers.

In the digital world, it is no more sufficient to apply analytics to enterprise data only. Big Data offers enormous potentials to be opened up for businesses. Consequently, the traditional data warehouse will still be the source for performance management, but will be replaced by Hadoop as the foundation for analytics. Enterprise data can be enriched by social media data – but mind the impacts of data protection laws. Social media enriched enterprise data enables better and more targeted marketing and individual customer communication. The mobile internet provides localization data. Now we can put information not only in the context of time („real-time“), but also in the context of space. Spatial and time related information has a much higher value than information per se. The convergence of information, time and space triggers new and innovative processes people could up to now even not think about: space related customer communication, for example in m-commerce, or operational information services providing information about time-tables, delays, traffic jams, weather hazards etc. that can be used to innovatively redesign processes. Other examples include mobile services for localization, registration, and authentication or for payment. To summarize, we can rightly say: information and analytics are ubiquitous.

Statement 3: BI arrived at C level. Indeed, performance management and GRC are within key responsibilities of the board. In the past, BI was sold to the IT. Many BI projects simply suffered due to this ill position. The creation of value by BI was very difficult to show, sometimes even abnegated. Huge data warehouses caused costs, but nobody really liked to use all the existing data. This is now changing. The value creation of BI in the context of business processes is beyond dispute. GRC turns the office of the CFO to the control room of the enterprise. Even the role of the CFO is changing. The CFO undergoes a metamorphosis to a CPO, the chief performance (and analytics) officer with full responsibility for GRC. End of 2010, we could see another upcoming and strong driver: smartphones and tablets. Starting with board level, these devices became a “must have”, and mobile BI is one of the most attractive applications. Consequently, via the mobile internet, BI becomes established on board level and is no more delegated down to assistants and clerks.

Statement 4: BI-based decisions are increasingly taken in the specialist departments. Originally, BI was driven by IT, and just some few selected experts in the specialist department, in particular in controlling, had access to BI-based facts. But in the meantime, the number of BI users is increasingly growing. The establishment of BI competency centers and of Self-Service BI made it happen. Ergonomics of the tools is getting better by service orientation („mashing up“), by better visualization, higher degree of automation and more intuitive features. On the one hand, the tools are well improving, but on the other hand, users in the specialist departments get more and acquainted with digital tools. „Digital natives“ who today have better computing equipment at home than many enterprises can provide, are now entering the business and make up a new community of computer experienced users. Furthermore, if BI competency centers offer help for self-help, then analytics can be applied by a broad user community in a rather autonomous way. The old dream of the 80s comes true: specialist departments can deploy their reports and analytics by themselves. Self-Service BI today is reality. What matters is the interplay of organization and technology. BI governance is the most critical success factor.

Consequently, decisions about BI programs and platforms moved step by step into the specialist departments. In the year 2013, for the first time, IT spending by specialist departments exceeded budgets allocated for BI within the IT. This has been shown in a study by IDC, and after that study, this process is believed to continue[54].

The continuously increasing market dynamics is another driver for this trend. Pressure on organizations is growing, and fast and flexible changes of business processes, even of business models are required. Agility, smart acting in markets, and innovation are in the focus of management. Here and now, IT is challenged. But in many organizations, IT cannot deliver what is expected. IT budgets have been shortened since years, and cost savings have been the most important goals of CIOs. The lion’s part of the shrinking IT budgets went into traditional infrastructure and into the maintenance of existing systems. IT budget for new developments was nearly not existing. In the end, in the eyes of specialist departments, IT become a blocker of innovation. **Shadow IT** was the answer. Specialist departments took self-help and provisioned themselves with the needed IT solutions. They believe in getting a better service now, and in faster fulfillment of their requirements. In the end, their goal is improved agility, because missed opportunities do not only cost money, but also losses of competitiveness.

Social media concepts and tools empower new collaboration and new approaches to BI governance. We believe that this is the domain where most significant changes will be seen in future. Since some time, Facebook style solutions are intending to penetrate business. But we believe that the Facebook approach may be not sufficient for an enterprise, since it is completely built on Marc Zuckerberg’s purely human related social approach. This will be not sufficient for comprehensive corporate management and governance. We believe that the more media oriented approach by Jive, SAP Jam, Microsoft Yammer and others is more appropriate, because it is more based on communication than on social principles. This will better fit to businesses as an innovative approach to internal and external enterprise communication. But today, a traditional enterprise culture makes up a big blocker, because such a “democratized” communication model is seen as alien.

Statement 5: The digital enterprise and analytic technologies. The transformation into a digital enterprises driven by four IT mega trends: mobile, cloud, social, and Big Data. It is interesting to note that these four trends are intertwisted, and all four trigger an ongoing increase in analytical technologies in digital enterprises.

[54] siehe *silicon.de*, Zugriff am 07. Juli 2014.

The mobile internet produces a high volume of data with high velocity, the very Big Data. It is localization and navigation data of all mobile devices. In the mobile internet, space, time, and information converge. Today, we know exactly where and when a customer, a product, or any device resides, because each smart phone and any mobile device produce data. Furthermore, the convergence of space, time, and information creates a new world: the **Internet of Things (IoT).** Smart meters, sensors, image recognition technologies and payment through NFR (near-field communication), for instance, are and will be embedded in various (mobile) devices and make up the IoT. In the end, “mobile” will no more be restricted to mobile telephones or tablets, and cellular mobile technology will go beyond cellular mobile networks. Communication will additionally use NFC, Bluetooth, LTE, and WLAN, and will be soon integrated in many new devices like watches, glasses, medical sensors, intelligent posters, home entertainment systems, and in cars. All this will further stimulate the production of data.

Mobile also drives the cloud, because the mobile internet works according to the principles of cloud computing. Each app we use works this way. Cloud computing is also very tightly related to Big Data, because cloud computing is an IT provisioning model that due to its elasticity, flexibility, and cost advantages is well-equipped for Big Data and Big Data analytics. For example, many vendors of analytic databases already offer a DWaaS (data warehouse as a service). We can assume that this trend will continue.

Mobile drives social, because social works best with everybody always and everywhere reachable. Social again drives Big Data: It provides more and more and even completely new data about people. For instance we can now identify relations between people via social media data.

This makes Big Data indispensable among the core competencies of digital enterprises. The volume of digital content increases exponentially. More than 90% of this information is poly-structured data (photos, videos, audios, and data from social media and the internet of things). This data is full of rich content, and organizations are more and more interested to gain valuable insight into Big Data. Consequently, we can be sure that analytical technologies will be soon main stream, and will make up an absolutely critical success factor for digital enterprises.

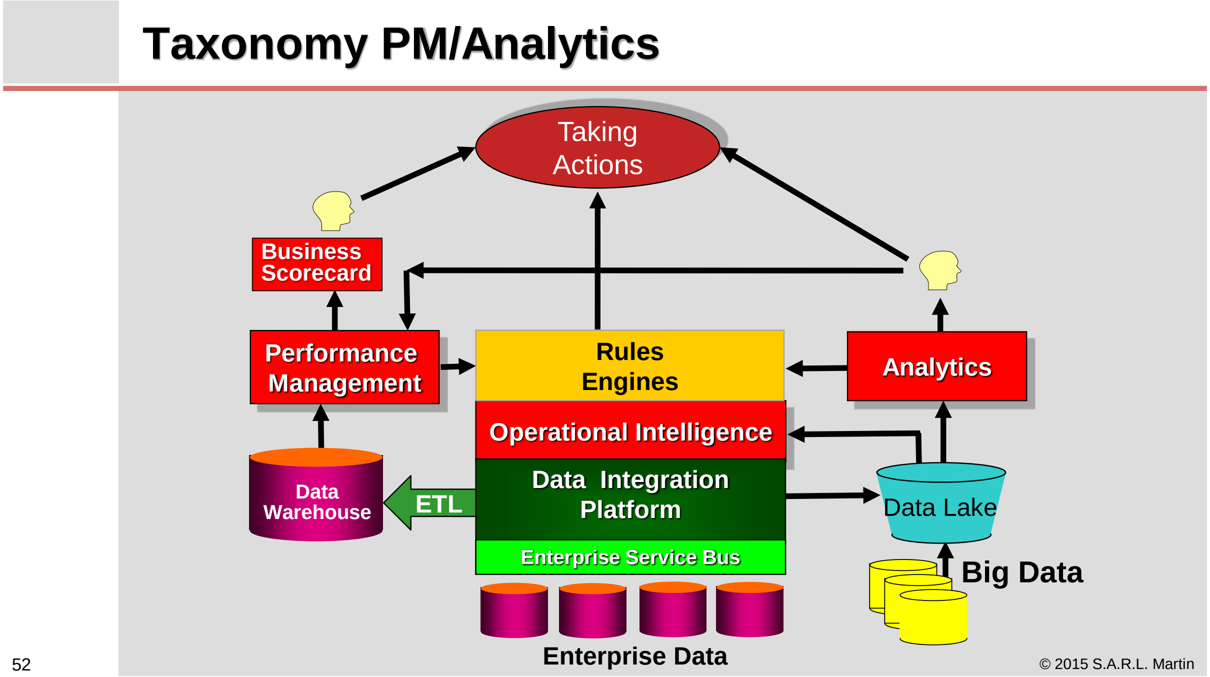
Big Data does not only mean big volumes, but also a data production in “big” velocity. Ten/fifteen years ago, we talked about the transformation into a real-time enterprise. Today, we see that usage and mastering of real-time is one of the essential facilities of digital enterprises, and certainly a driver for real-time analytics. Its success is not limited to better targeted customer communication, but even essential in the internet of things. It empowers the application and usage of machine learning. Algorithms for self-learning, self-healing, and self-adjusting trigger process automation, improve productivity, cut cost, and lower risks. A good example is at the verge of reality: the driverless automobile. One of its key technical pre-requisites is in-memory processing, because computing power is essential for processing all incoming signals in real-time. This is a new huge area for analytical technologies.

Take Away: Performance Management and Analytics today, the “new“ BI in digital enterprises must follow the principles of “simplicity, mobility, extreme analytics and collaboration”. The critical success factor is: Everybody in business should follow these principles. Besides analytics in real-time for rapidly analyzing large data sets, real-time analytics for controlling and automating of processes through embedded analytics is even more important in future.

10.2 Taxonomy of the Markets for Performance Management and Analytics

Let us now move to the market players. From the three phases of action time (fig. 56), we can derive **taxonomy** for classifying the players (vendors) in the market (fig. 52). Key players in the different categories are listed in chapters 10.3 to 10.5. More details on specific, selected vendors are published in part 2 of this white paper, where in each white paper we mapped the vendors’ architecture and strategy to the vision and reference architecture developed in this part 1.

Part 2 – Available white papers (March 2015): arcplan, BOARD, Clueda, Cortex, Cubeware, EPOQ, IBM, Informatica, geoXtend, Kapow Software, Lixto, Metasonic, Panoratio, PitneyBowes MapInfo, SAP, Stibo Systems, TIBCO/Spotfire, Tonbeller, USU Service Intelligence (see www.wolfgang-martin-team.net)



Taxonomy PM/Analytics

Figure 52: Action time (see fig. 46) based taxonomy of performance management (PM) and Analytics market players. (ETL = extraction, transformation, load)

CPM Toolkits in particular provide the actual state-of-the-art implementation of performance management. The term goes back to Gartner Group. CPM Toolkits are positioned between

CPM Suites and single tools (like spreadsheets). CPM Toolkits provide performance management specific functionality as services: They are open and have standardized interfaces, whereas CPM Suites are built on proprietary technologies and are tightly integrated. This makes implementation of CPM Suites difficult: It is not easy to customize them, and it needs rather high efforts and costs to integrate a CPM Suite into existing applications. CPM Toolkits follow the SOA principles. This loosely coupling of performance management services has several advantages. Service-orientation eases up customization. Services can be used like LEGO building blocks. They are easily invoked by right mouse clicks and are ideal for mash ups. The end user can build its own composite application: By mashing up performance management services, the user creates the analytic workflow and orchestrates the analytic services provided by the CPM Toolkit.

This is why CPM Toolkits address the old requirement that BI tools should enable the business to work autonomously and to build their own reports and analyses without programming (see statement 4 in chapter 10.1). Whereas traditional CPM Suites typically only cover specific requirements by a special module, a CPM Toolkit is an open, but coherent platform offering all objects and functions as services for composing and orchestrating services by mashing up without coding.

10.3 Classification of Data(base) Management Systems

The following listing of vendors is not supposed to provide a complete view on the market. But it is quite comprehensive and puts a focus on the German speaking markets: It includes many local players. This listing does not evaluate vendors in terms of completeness and quality of their products. The classification follows the taxonomy presented in Figure 52.

In the past, Data Marts and Data Warehouses have been typically implemented via relational database technology. This is why we included as well vendors of traditional relational DBMS as vendors of “true” analytic DBMS we discussed in chapter 7.2 (cf. Fig. 53), and NoSQL data management systems as discussed in chapter 7.3. OLAP systems also include ROLAP, whereas ROLAP does not support data persistency, and ROLAP vendors typically use relational or analytic databases for persistency.

OLAP data management systems (MOLAP, ROLAP, HOLAP)

BOARD, IBM Cognos, IBM Cognos TM1,FoundationDB, Infor/Alea, Information Builders, instant OLAP, Microsoft SQL Server, MicroStrategy, MIK, Oracle 11g, Oracle/Essbase, Paris Technologies (PowerOLAP), Quartet FS (Active Pilot), SAP Netweaver BI, SAS OLAP Server, Teradata

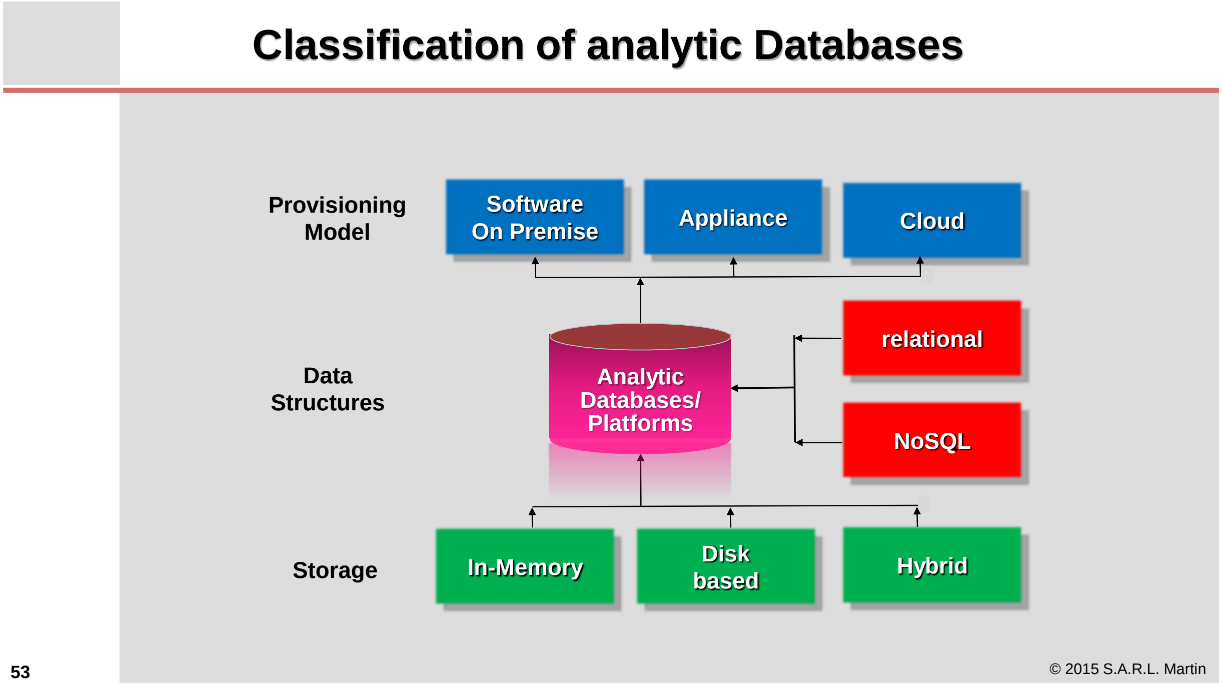
OLAP auf Hadoop: Kylin, Kynos Insights

Open Source: The Bee Project, icCube, Jedox/Palo, Hitachi Data Systems/Pentaho/Mondrian, TIBCO/Jaspersoft

(traditional) relational database management systems

IBM DB2, IBM Informix, Microsoft SQL Server, NuoDB, Oracle 11g, SAS Scalable Performance Data Server

Open Source: Ingres Data 10, Lucid DB, MariaDB, Oracle/MySQL, PostgreSQL



Classification of analytic Databases

Provisioning Model

Software On Premise

Appliance

Cloud

Data Structures

Analytic Databases/ Platforms

relational

NoSQL

Storage

In-Memory

Disk based

Hybrid

53

© 2015 S.A.R.L. Martin

Figure 53: Classification of analytic databases through storage methods, data structures, and IT provisioning model.

Analytic relational MPP database management systems

IBM DB2 (InfoSphere Warehouse), IBM Smart Analytics System, IBM Pure Data[55], Kognitio, SAS Scalable Performance Data Server (together with SAS Grid Computing and SAS In-Memory-Analytics), Teradata, XtremeData

Open Source: EMC/Greenplum, Volt DB

Analytic NoSQL data(base) management systems (no in memory processing)

Amazon DynamoDB, Calpont, CortexDB, HP/Vertica, Illuminate, Kx Systems, Sand Analytics, SAP Sybase IQ, Teradata/AsterData, Vectornova

Open Source: Apache Cassandra, Apache Hadoop HBase, ArangoDB, InfoBright, MongoDB

Analytic NoSQL data(base) management systems (in memory processing)

1010Data, Actian/ParAccel, Amazon Redshift, Anceus Database, Exasol, IBM Smart Analytics Optimizer, SAP HANA

Special data(base) management systems (technology in parentheses)

Actian/Versant *(OODB),* CrossZSolutions *(QueryObject System),* Dataupia *(DW Appliance)*, Drawn-to-Scale *(Big Data Platform on top of Hadoop)*, dimensio informatics *(minimal-*

[55] formerly called IBM Netezza

invasive performance tuning), HPCC Systems *(Big Data Framework à la Hadoop),* InterSystems *(OODB),* Oracle Exadata Database Machine *(Data Appliance with massive parallel grid),* Oracle Exalytics In-Memory Machine *(special technology for CEP),* Panoratio *(database images),* ParStream *(DMBS for real-time Big Data analytics)*, Spire *(Big Data operational SQL DB)*

Hadoop Distributors

Amazon Elastic MapReduce, Cloudera, Hortonworks, IBM InfoSphere Big Insights, Intel Apache Hadoop Distribution, MapR Technologies, Pivotal HD, Talend Platform fir Big Data, VMWare (HVE, Serengeti)

SQL on Hadoop

Citus Data, Cloudera, EMC/Greenplum, IBM, Hortonworks, HP, Jethro Data, NuoDB, Oracle, SAP, Slice Machine, Teradata/Hadapt

Analytic databases provide completely new opportunities: They dramatically improve scalability, performance, and can also help to cut cost running a database. Organizations that want to run complex analyses on very large databases, where many users are executing many different queries that need high performance and scalability with high flexibility and simplified maintenance should consider analytic databases. We believe, an evaluation is a must, and enterprises should not wait any longer.

10.4 Classification of Performance Management/Analytics Vendors

The following listing of vendors is not supposed to provide a complete view on the market. But it is quite comprehensive and puts a focus on the German speaking markets: It includes many local players. This listing does not evaluate vendors in terms of completeness and quality of their products. The classification follows the taxonomy presented in Figure 52.

Performance Management (traditional BI – General Purpose Frontends: Suites, Toolkits, Reporting, Dashboards, and Specialized Tools)

- *the BIG Six:* HP, IBM, Microsoft, Oracle, SAP, TIBCO Analytics
- *world-wide acting PM specialists:* Infor, Information Builders, MicroStrategy, Qlik, SAS Institute
- *Challengers:* Adaptive Planning, Advizor Solutions, Alteryx, Antares Informations-Systeme, arcplan, aruba Informatik, Bime, Birst, Bissantz, Bitam, BiX Software, BOARD, CA/CleverPath, cobra computer's brainware GmbH, Comma Soft/Invonea, Connexia, Cubeware, DataHero, Dataself, Decisyon, Domo, Dell/StatSoft, Dimensional Insight, DSPanel, ElegantJ BI, Evidanza, GoodData, Host Analytics, Indicee, InetSoft Style Intelligence, instantOLAP, Intellicus, Intensio, iQ4bis, kpiWeb, Kofax/Altosoft, LogiXML Looker, MAIA-Intelligence, Menta, MIK, Neubrain, Oco, OpenText/Actuate, Orbis AG, Panorama Software, Paris Technologies, Phocas, PivotLink, Prevero, Prognoz, Pyramid Analytics, Reboard, Salient Management Company, SAMAC (only for IBM iSeries),

Strategy Companion, Tableau Software, Targit A/S, Teleran, TIBCO/Spotfire, Tidemark, Tonbeller AG, Verix, Vizion Solutions, Windward Reports, Yellowfin, Zap Technology

- *Special tools for (agile) BI projects:* Balanced Insight
- *Open Source:* OpenText/BIRT, The Bee Project, Hitachi Data Systems/Pentaho, SpagoBI, TIBCO/JasperSoft

Analytics / Data Discovery

- Accenture Insights Platform, Advizor Solutions, Alation, Armanta, Attivio, Ayashdi, BeyondCore, BiBOARD, Business Intelligence System Solutions Holdings B.V., ClearDataStory, Comma Soft/Invonea, Data Mentors, Datawatch, Dell/Kitea Analytics Suite, Dimensional Insight, Domo, Dremel, Drill, EligoTech, IBM/Cognos Insight, Information Builders, Kofax/Altosoft, Lavastorm, Lyzasoft, MetaLayer, Microsoft/Power Pivot, MicroStrategy, Neutrino BI, OpenText/Actuate, Opera Solutions, Oracle/Endeca Information Discovery, Panorama Software, Precog, Pyramid Analytics, Qlik, Salient Management Company, SAP/Lumira, SoftLake Solutions, Tableau Software, TIBCO Spotfire, Treasure Data, VisualCue
- *Special tools for Cassandra:* Acunu Analytics
- *Special tools for Hadoop:* Alpine Data Labs, Datameer, Dell/Kitenga Analytics Suite, Impala, Kylin, Kynos Insights, Platfora, Teradata/Hadapt
- *Special tools for MongoDB:* Precog
- *Additional tools (approach in brackets):* Ankhor (*visualization of log data*), Datonix *(QueryObject System)*, human IT software *(InfoZoom)*, Panoratio *(Database Images)*
- *Open Source: Hitachi Data Systems/Pentaho, TIBCO/*JasperSoft

Analytics / Predictive Models (Data Mining, Statistics & related tools)

Advizor Solutions, Alpine Data Labs, Alteryx, Anderson Analytics, Angoss, Avail Intelligence, Axtria, Blue Yonder, Context Relevant, Dell/StatSoft, EPOQ, Equbits, Exeura Rialto, Expert Systems, FICO Predictive Analytics, IBM, IBM/SPSS, IBM/Unica, InfoCentricity, Infor/E.piphany, ISoft (Alice), Lityx, MetaLayer, Megaputer, Microsoft, MicroStrategy, MIT GmbH, Mnubo, OpenText/Actuate, Oracle, Pega Systems (Chordiant), Pitney Bowes Software (Portrait Software), Predixion, Prudsys, SAP, SAP/KXEN, SAS, Savi Insight, StatPoint Technologies, Synesis Solutions, Systat Software, thinkAnalytics, TIBCO Spotfire, Teradata, Treparel, Verix, Viscovery

Specialists for machine learning: Bigml, GlassBeam, Skytree, Spark

Open Source: Knime, Microsof/Revolution Analytics, Orange, RapidMiner, Rattle, R-Project, Weka

Software-Packages for R. Which are the user top-ranked software packages for R in H1/2015? We refer to an analysis based on the number of downloads from CRAN Package Repository in the period of January to May 2015. Please note that there are also some more sources for R packages, but with CRAN, you find 6,778 active packages. Statistics based on these downloads are available on KDnuggets http://www.kdnuggets.com/2015/06/top-20-r-packages.html

Analytics / Text Analytics

Alias-i, Anderson Analytics, Attensity, Attivio, Ayasdi, Basis Technology, Business Intelligence Group, Clarabridge, Del/Kitenga Analytics Suite, Digital Reasoning, Expert Systems, IBM, IBM/SPSS, ITyX, Lexalytics, Linguamatics, Megaputer, MetaLayer, OpenText/Nstein, Rocket Software/AeroText, SAP, SAS/Teragram, Saplo, Sentimetrix, Serendio, StatSoft, Teezir, Thomson Reuters/Clear Forest, Treparel, Viscovery, ZyLab

Open Source: Gate, Python NLTK, R (TM module), RapidMiner

Operational Intelligence (BAM/CEP) & Streaming

Axway, Amazon Kinesis, Business CoDe, ClearPriority, Clueda, Datawatch, Domo, Gemstone, IBM InfoStreams, Inetsoft, Informatica, Information Builders, JackBe, Kofax/Altosoft, LogiAnalytics, Magnitude Software/Noetix, Microsoft StreamInsight, Miosoft, Oracle CEP, SAP CEP, ScaleOut Software, SL Corporation, Software AG, SpaceTimeInsight, Splunk, SQLStream, TIBCO, UC4 Decision, Verix, Vitria, VizExplorer, VMWare/Gemstone

Open Source: HStreaming, LinkedIn Samza, Twitter Storm, Yahoo S4

Web Analytics

Adobe/SiteCatalyst, AT Internet, Avail Intelligence, Bango, Bime, ComScore/Nedstat, Enecto, eTracker, Foresee Results, Google Analytics, IBM Cognos Customer Insight, IBM/MarketingCenter, Intellitracker, Lyris, Mindlab Solutions, Nielsen/Glance Guide, Nurago/LeoTrace, Odoscope, Precog, sitespect, Targit A/S, Visible Measures, webtrekk, WebTrends, Wired Minds, Yahoo! Web Analytics

Open Source: eAnalytics, Open Web Analytics, Piwik

Location Intelligence

APOS Systems, Aruba Informatik, BOARD, Cubeware, deCarta, Digital Globe, DMTI Spatial, ESRI, Galigeo, Google Earth, Integeo, mapdotnet, MapQuest, MetaCarta, Microsoft/ VisualEarth, Navteq, Oracle, Pitney Bowes Software, Spatialytics, Space Curve, Tableau Software, Talent Information Systems, TomTom Global Content, TIBCO/Maporama, TIBCO/Spotfire, Universal Mind, Vistracks, Yellowfin

Decision (Rule) Engines

Angoss, Avail Intelligence, CA Aion, Bosch SI/Innovations, Corticon, EPOQ, FICO (Fair Isaac Corporation), IBM, IBM/SPSS, Infor/E.piphany, MicroStrategy, Oracle, Pega Systems, Pitney Bowes Software (Portrait Software), Prudsys, SAP, SAS, StatSoft, thinkAnalytics, TIBCO, UC4 Decision, Versata, Viscovery

Open Source: RapidMiner

Financial Performance Management (Budgeting, Planning, Forecasting, Financial Consolidation etc.)

A3 Solutions, Acorn System, Adaptive Insights, Alight Planning, Anaplan, Antares Informations-Systeme, arcplan, ASRAP Software, Axiom EPM, Bitam, BOARD, Complan &

Partner (EPUS), CoPlanner, CP Corporate Planning AG, Corporater, CSS Computer Software Studio, Cubeware, Cubus AG, Decisyon, Denzhorn, DSPanel, Evidanza, HaPeC, Hologram BI, Host Analytics, IBM/Clarity Systems, IBM/Cognos, IDL Systems, Infor, InformationBuilders, KCI Computing, Longview Solutions, LucaNet, macs software, Microsoft, MIK, Neubrain, Oracle/Hyperion, Orbis AG, Paris Technologies, PMS GmbH, Prevero, Procos AG, Prodacapo, ProfitBase, Prophix, Performance Solutions Technologies (managePro), River Logic, SAP, SAS, Software4You, Tagetik, Targit A/S, Thinking Networks, Tidemark, UNIT4 Coda, Whitebirch Software

Special tools for spreadsheet management and compliance: ClusterSeven

Business Scorecards (Dashboards, Strategy Maps), stand-alone solutions

Active Strategy, Axsellit (Corporater Express), BOC (AdoScore), Business CoDe, Chartio, Communic (Vision.iC), Corda Technologies, Corporater, dMine Business Intelligence, Dundas Data Visualization, eBrains Consulting, Hologram BI, Horvath & Partner, Hyperspace, iDashboards, iGrafix, InetSoft, Jinfonet Software, Klipfolio, macs software, Nevron Data Visualization, Performance Solutions Technologies (managePro), Prelytis, Procos AG, Prodacapo, ProfitMetrics, Push BI, Quadbase, Qualitech Solutions, QPR Software, Rocket CorVu, SiSense, Software AG, statsmix, Stratsys AB, VisualCalc, Visual Mining, UNIT4 Coda

10.5 Classification of Information Management Vendors

The following listing of vendors is not supposed to provide a complete view on the market. But it is quite comprehensive and puts a focus on the German speaking markets: It includes many local players. This listing does not evaluate vendors in terms of completeness and quality of their products. The classification follows the taxonomy presented in Figure 52.

Data Integration – Platforms

- *Leaders:* IBM, Informatica, Oracle, SAP, SAS Institute/DataFlux
- *Challengers:* Actian/Pervasive, Adeptia, Astera, Atacama, Attunity, Axway, CA/Inforefiner, Cirro, Cisco/Composite Software, Columba Global Systems, Comlab/Ares, CortexPlatform, DataStreams, Dell/Boomi, Denodo, Diyotta, eq Technologic, Full360, Gamma Soft, HVR Software, Information Builders/iWay Software, ITyX/Context, Lavastorm, Magnitude Software, Miosoft, MuleSoft, Nimaya, Parity Computing, Paxata, Progress Software, Sclera, SnapLogic, Software AG, Stone Bond, TIBCO, Uniserv, Versata
- *Complementary tools for data integration:* AnalytixDS
- *Specialists for data streams/time series:* DataTorrent, InfluxDB, Infochimps, InitialState, Prometeus, Sigmoid, Spark
- *Special platforms for analytical services:* Cirro, Magnitude Software/Kalido
- *Information Management for Hadoop:* Teradata/Revelytix

- *Open Source:* CloverETL, JBOSS Enterprise Middleware, Jitterbit, JumpMind, Spark, Talend

ETL/ELT

AbInitio, Actian Pervasive, Astera Software, CA/Advantage Data Transformer, Capsenta Compact Solutions, Datarocket, Datawatch, ETL Solutions, IBM, Informatica, Information Builders, iQ4bis, ITyX/Context, Lavastorm, Menta, Microsoft, Open Text, Oracle, Pitney Bowes Software, SAP, SAS, Sesam Software, Software Labs, SQ Data, Syncsort, Theobald Software, Tonbeller AG, Uniserv, Versata

Special tools for DW planning ("pre-ETL"), resp. design: Indyco, Wherescape 3D; *and for managing DWs:* BIReady

Open Source: Apatar, The Bee Project, CloverETL, Enhydra Octopus, Hitachi Data Systems/Pentaho/Kettle, KETL, RapidMiner, Talend

ETL – Specialized Tools: semantic web crawlers/extraction robots

30 Digits Web Extractor, Brainware, Business Intelligence Group, Connotate, Datawatch Denedo Technologies, Fetch Technologies, Kapow Software, Lixto, Silwood, Teezir

Master Data Management

Agility, ARM Advanced Resource Management, Bosch Software Innovations, DataRocket, Enterworks, IBM, Informatica, Magnitude Software/Kalido, Oracle, Orchestra Networks, Riversand, SAP, SAS/Data Flux, Semarchy, Stibo Systems, SyncManager, Systrion, TIBCO, Uniserv, Visionware, zetVisions

Data Classification

Data Global, EMC/Kazeon, FileTek/Trusted Edge, Index Engines, ITyX/Context, Microsoft, Nogacom, Rocket Software/Arkivio, StoredIQ, Varonis Systems

Data Quality

Alteryx, AS Address Solutions, Ataccama, Business Data Quality, Clavis Technology, Datactics, DataMentors, Datanomic, Datras, emagixx, Eprentise Harte Hanks, Human Inference, IBM, Informatica, Innovative Systems, Melissa Data, Omikron, Oracle, Pervasive, Pitney Bowes Software, Posidex Technologies, Scarus, SAP, SAS, tekko, TIQ Solutions, Uniserv, Trifacta, Versata, X88 Software

Open Source: CloverETL, Infosolve Technologies, RapidMiner, SQL Power, Talend

10.6 Big Data Evolution: Market Forecasts

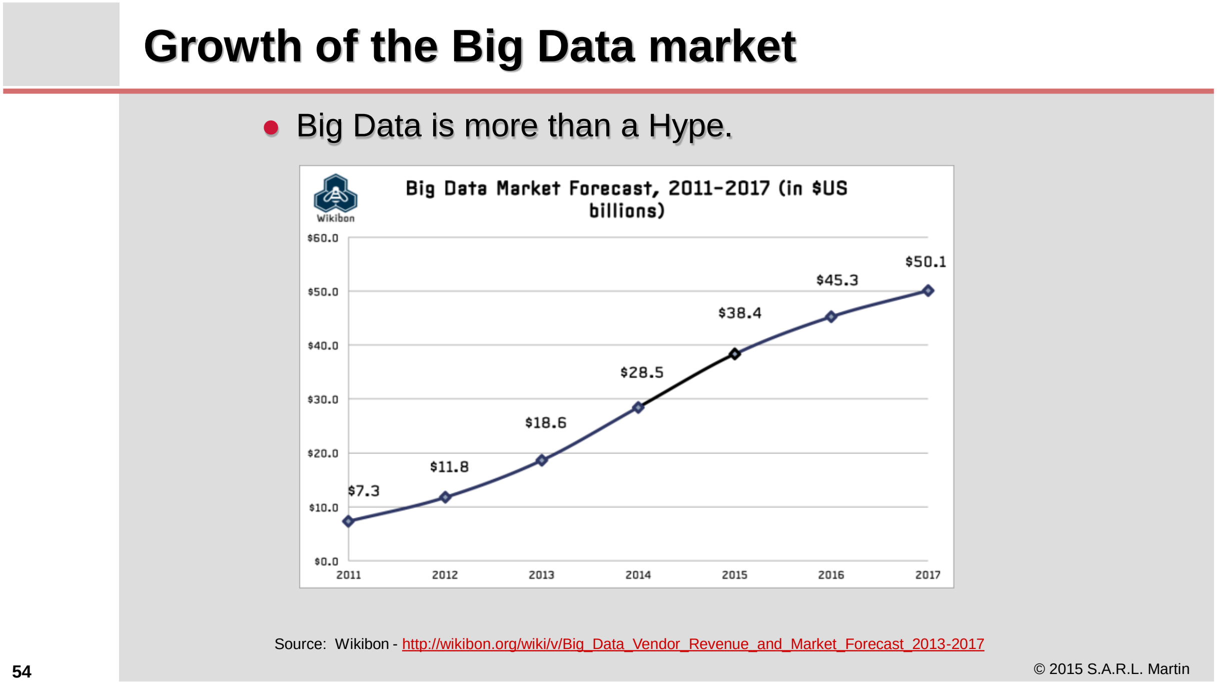
The Big Data market consists of software, hardware, and service providers. Big Data Software includes data and database management systems, and analytics for storing, processing, and analyzing Big Data, i.e.:

- Data and database management systems: NoSQL systems like Hadoop and analytic databases,

- A new generation of data warehouse software and hardware technologies,
- Big Data management: data management (integration, lineage, quality, governance) applied to Big Data,
- Big Data analytic platforms and applications including new concepts for data visualization, location intelligence, text analytics etc., especially focusing on the analysis of poly-structured data.

Big Data services correspond to the traditional BI services like support, training, consulting, application development and system integration services, now related to Big Data. Big Data Hardware includes all types of hardware for Big Data, in particular the now popular data appliances bundled and tuned software and hardware solutions that are sometimes also packaged with the corresponding services.

According to Wikibon (http://wikibon.org/wiki/v/Wikibon:About), in 2013, the Big Data market has grown to $18.6 billion (software, hardware, services), and shall grow to $50.1 billion by 2017 (fig. 54). We have already discussed the reasons for this extreme growth in chapter 2.4. The benefits and the potential that are offered by Big Data analytics directly target the bottom line of many enterprises: increase of revenues, cost savings, increase of competitiveness. Who will be reserved about these opportunities?



Growth of the Big Data market

- Big Data is more than a Hype.

Source: Wikibon - http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

54

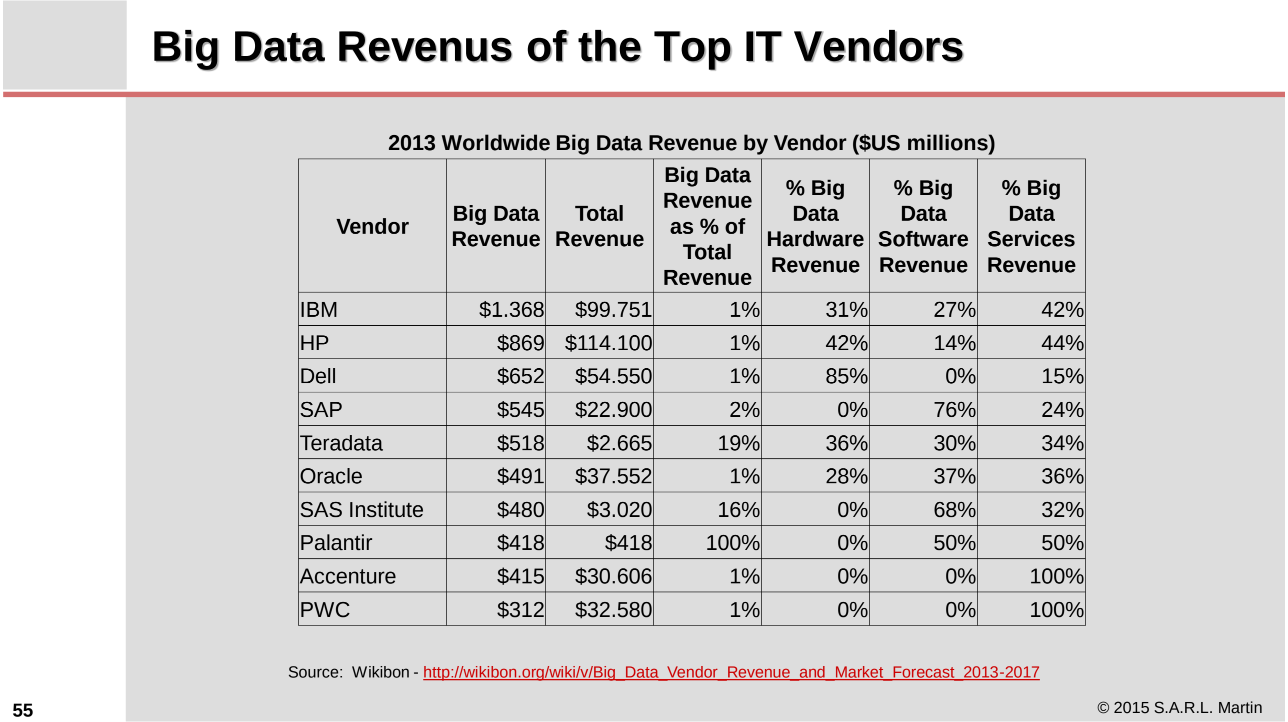
© 2015 S.A.R.L. Martin

Figure 54. Wikibon estimates that the total Big Data market (software, hardware, services) reached $18.6 billion in 2013. This corresponds to a growth of 58% compared to 2012, and this is $0.5 billion more than estimated in 2012. By 2017, the market should exceed $50.1 billion, about $3.1 billion more than estimated in 2012.

What does shape the actual $18.6 billion Big Data market in 2013? As always, the market is dominated by the large IT vendors, and they all preach Big Data. Figure 55 presents the top 10 IT vendors ranked by Big Data revenues according to Wikibon's estimate. The numbers also reveal that actually most of the business is in hardware and services (for instance: IBM), and that – exception of SAS Institute, Teradata, and Palantir – the Big Data revenues of all vendors are rather low compared to their total revenues. Teradata with its 19% and SAS Institute with its 16% Big Data revenue portion are succeeded by SAP with a 2% Big Data revenue portion due to HANA. But this is expected to change rather quickly: The Big Data market as well as SAP with SAP HANA has just started, and we will see many acquisitions of the smaller, innovative vendors by the large IT vendors just as in the BI acquisition wave in the years 2007/08.

***Take Away:* Big Data – the market:**

- The market is still young. We have just started, but a market explosion has already taken place: Big Data has big potential, and it presents a very rapidly growing market.
- Organizations should start Big Data initiatives not later than now! If not, they will miss the chance and loose competitiveness. We recommend identifying benefits and potentials first, and then to start a pilot in dependence of the outcome of such analyses.
- Vendors should build a credible and reliable position as well as a roadmap providing clearly defined customer value and the necessary flexibility for prospering in the Big Data market.



Big Data Revenus of the Top IT Vendors

2013 Worldwide Big Data Revenue by Vendor ($US millions)

Vendor	Big Data Revenue	Total Revenue	Big Data Revenue as % of Total Revenue	% Big Data Hardware Revenue	% Big Data Software Revenue	% Big Data Services Revenue
IBM	$1.368	$99.751	1%	31%	27%	42%
HP	$869	$114.100	1%	42%	14%	44%
Dell	$652	$54.550	1%	85%	0%	15%
SAP	$545	$22.900	2%	0%	76%	24%
Teradata	$518	$2.665	19%	36%	30%	34%
Oracle	$491	$37.552	1%	28%	37%	36%
SAS Institute	$480	$3.020	16%	0%	68%	32%
Palantir	$418	$418	100%	0%	50%	50%
Accenture	$415	$30.606	1%	0%	0%	100%
PWC	$312	$32.580	1%	0%	0%	100%

Source: Wikibon - http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

55

© 2015 S.A.R.L. Martin

Figure 55: Top 10 vendors ranked by Big Data revenue (software, hardware, services worldwide). In comparison to 2012, IBM and HP stayed on spot 1 and 2. Dell moved up to spot 3, SAP to 4, and Accenture to 9. Teradata moved down to spot 5, Oracle to 6. Newcomers are SAS, Palantir, and PWC. 2012 TOP 10 vendors EMC fall to spot 23, Cisco to 13, and Microsoft to 15. Source: Wikibon, (see fig. 54).

10.7 Fundamentals for selecting PM/Analytics Platforms and Tools

The evolution of traditional BI from reporting and analysis of historical data (early “data warehouse“ concepts) to process oriented and operational BI had already started in 2003/04. This is one of the main reasons why the big BPM/SOA platform players (HP, IBM, Microsoft, Oracle, SAP) and leading ERP II players like Infor have complemented their platforms by BI solutions and realigned their boundaries mainly in 2007 (HP not before 2011). Given this market situation, when selecting BI solutions today, users have the choice between three scenarios:

Conservative scenario: Select one of the four platforms of the Big Five as your BI solution. The benefits for users and customers are obvious: It is a one shop stop deal, all is integrated, and all fits (this is what vendor’s marketing says, but it is not always untrue!). But the disadvantages are also obvious: You are locked-in by the vendor, and you are completely exposed to his pricing policies (licensing of new software, maintenance and support). Plus, there is an additional risk that the new owner of acquired BI solutions will not apply the same care and the same amount of investments when further developing, enhancing and improving the software. But a conservative customer will judge: I can live with such disadvantages, I feel confident, since a decision for one of the Big Four contains a minimal risk.

Innovative scenario: The SOA based platforms of the Big Five also offer a great opportunity, because a SOA follows industrial standards. In the IT market, we have never ever before seen such a willingness and real efforts across all major vendors to collaborate and to jointly drive standards (This could be nevertheless improved, but many of the today available standards have been already proven in day to day operations.). That’s said, SOA principles arrange for open platforms. Open platforms act like an open bus any other vendor can jump on: best-of-breed becomes true. This is a great opportunity for small, innovative, and agile vendors to offer and to market complementary niche or alternative functionality to the platform’s base functionality. Innovative BI users now also profit from the open platform and can gain competitive advantages by better BI tools like tools for enterprise search, simulations, text mining, linguistic procedures, network analysis etc.

Low budget scenario: Given the openness of the platforms, enterprises looking for commodity solutions for a commodity price get an opportunity, too. Open Source solutions can be easily deployed in a SOA, because integration efforts are minimized by service orientation. This is the today’s low budget alternative to the BI solutions of the platform vendors. Today, nearly all performance management and analytics functionality is available as Open Source solutions.

Today, there is an upcoming additional opportunity, “**cloud computing**”. It is an alternative to the traditional “on premise” licensing models and will put much pressure on the platform vendors to also offer BI as a Service. Nearly all vendors are already providing cloud offerings. BIaaS has also the potential to bring BI to small and medium enterprises (SME). Up to now, SMEs just use Excel as their tool of choice, but SMEs are as subject to compliance as larger enterprises. This heralds the end of Excel as a stand-alone enterprise tool. The huge SME market is best conquered by an approach like BIaaS and other cloud offerings, but not necessarily by traditional platforms of the (still) leading vendors.

In the meantime, more and more vendors offer a Data Warehouse as a Service (DWaaS). Amazon Web Services has started a new round in December 2012: Amazon AWS Redshift, a DWaaS based on ParAccel. Redshift aims for Data Warehouses in the higher terabyte range up to petabytes. Its price is remarkable: $1.000 per month and per terabyte. Traditional comparable offerings based on on premise solutions cost more by a factor of 10, 20 and even more. This could be a game changer. To be watched!

10.8 Roadmap for Big Data Users

Today's Big Data activities in business can be subdivided into three groups, Agile Big Data, Operational Big Data and "High Resolution Management". This classification also eases the comprehension of vendors' product offerings, and CIOs and CTOs are enabled to better select the right offers for their goals and objectives.

Agile Big Data is the idea that it shouldn't cost a fortune or take years to start getting the value of big data. There are a variety of players who have technology that can quickly allow analysts to understand if a big data set has potential and to start processing it. Data as a service offerings are in particular very attractive for Agile Big Data: They come with an OPEX financial model, i.e. investment turns to a short time expense. Furthermore, they are quickly installed, and if Big Data pilots do not show measurable benefits, they are as quickly de-installed. An agile approach to big data puts the means to analyze big data directly in the hands of analysts or data scientists as much as possible (see chapter 3.5). Most companies that use Agile Big Data have a robust culture of data-driven decision making. The key question of Agile Big Data is: How can we take the ease of use and empowerment of spreadsheets to the world of big data?

Operational Big Data is about automating and streamlining the process of analyzing huge amounts of data into a way that can support decision making and creating intelligent processes. This space is a cooperative battleground between the commercial open source world of Hadoop and the enterprise vendors listed in chapter 10.3 together with Data Discovery vendors like MicroStrategy, Qlik, SAS Institute, Tableau Software and TIBCO/Spotfire. The key question of Operational Big Data is: How can we create an infrastructure so everyone can get the benefit of what we learn from big data?

High Resolution Management is the idea that the management processes and the design and execution of many other business processes should change based on the far more detailed picture of business activity provided by big data. They key question of High Resolution Management is: How can we change the way we manage our businesses based on the high resolution view big data provides?

This model also allows very nicely presenting SAP's Big Data strategy. Due to the position and important role that SAP plays in the German speaking markets, we cite Sanjay

Poonen[56]: “SAP is attempting to create an integrated approach that allows companies to perform analytics, make big data operational, and support applications for high resolution management all in one environment.”

As we have already stated, Big Data technologies are rather young and immature. Big Data roadmaps are just supported by some handful of experiences made. But anyway, we have distilled five main challenges that could help users to make their first steps towards Big Data successful.

1. ***Identifying and hiring of talented people who know Big Data and Big Data Analysis and who have made experiences.*** This is already very tough since experts are rarely available in the market. Thus, specialized consultants should be engaged, because it is easy to rapidly lose much time and money without getting an added value from Big Data. Furthermore, external consultancy should not only provide expertise in Big Data technologies, but also in organizational questions and challenges. Big Data requires new ways of cooperating between IT and business as well as new roles and new skills. We have already presented data scientists in chapter 3.3, for example.

2. ***Selecting technology and tools.*** Here, the external consultant is in charge. But before tackling the question of technology and tools, the definition of a Big Data strategy (Strategy comes first, as always!) should be elaborated, i.e. do we strive for Agile Big Data, Operational Big Data or even for High Resolution Management? For, the selection of the right technology and tools as well as the question of how to provision the technology (cloud or not) depends on strategy.

3. ***Determining the relevance of information for the given Big Data challenge.*** Which information offers a real added value in relation to costs of identifying, extracting, storing, and analyzing? This fundamental question cannot be answered a priori in most of the cases. A method of resolution consists of defining relevance measures. In sentiment analysis for instance, data sources are evaluated according to the frequency of certain terms in a given period. Search engines can be rather supportive for evaluating such statistical data. Furthermore, in such situations, external consultants should help out with their experience. But in the end, the rule is: trying and iterating, or “trial and error”. Here, we are entering virgin soil.

4. ***Continuous thinking out of the box.*** The rule is: Do not make assumptions, do not make hypotheses. The reason is, Big Data analysis is all about hypothesizing and discovering the unknown and unexpected. Testing theses hypotheses makes up the second step. This Big Data procedure is quite different to traditional working when there was nearly no information available, and people were used to work with hypotheses based on knowledge. In Big Data analytics, this is just the opposite: analysis is used to identify hypotheses. This is a new and alternative thinking we are not yet used to.

5. ***Coming to an end and trusting the results of analyses.*** Here we can get on with our discussions of the second of the five Big Data benefits in chapter 2.4: Testing all hypotheses. When we have hypothesized, then a test should be performed quickly, and

[56] Sanjay Poonen is President and Corporate Officer at SAP Global Solutions, see Forbes: http://www.forbes.com/sites/danwoods/2012/01/05/bringing-value-of-big-data-to-business-saps-integrated-strategy/

customers and market should decide whether a hypothesis is false or leads to positive effects. This is exactly what the Big Data forerunners are doing: transforming hypotheses into test environments and measuring the impacts. This can be done in a fast way and be monetarily assessed. In the end, again, this is an iterative procedure according to the “trial and error” principles. Each iteration shows a direct impact on the bottom line, since customers and market are involved. Thus, we have a reliable control of Big Data analytics: We can measure the generated revenue and profit. Now, we also recognize why external consultants are so important in organizational questions: Only if this iterative approach is feasible in the enterprise’s organization and culture, Big Data analytics can bring a measurable added value.

Take away: Big Data roadmap.

- Big Data initiatives should (as always) start with strategy. Strategy should follow the directions of Agile Big Data, Operational Big Data, or High Resolution Management.
- Big Data comes with five challenges. It is not only mastering of technology (as always), but also in particular changing the organizational structure (How to form up a Big Data organization?) and the methods (iteratively hypothesizing and testing).
- Success of Big Data analytics is to be measured and monetarily evaluated by its impacts on customer and market behavior.

11 Summary and Literature

I believe that my definition and my ideas about performance management and analytics across operations, tactics, and strategies will become a standard for continuously evolving metrics-driven management which a prerequisite for mastering today's challenges in the New Normal. I also believe that the discussed reference architecture of analytic services' infrastructures will become a standard for dynamic enterprise specific service oriented architectures (SOA). The mergers that happened especially since 2007 support my hypothesis: The vendors of SOA platforms complete their offerings by service oriented analytics. This Whitepaper will therefore continue to help making strategic decisions on strategies and platforms as well as analytic services including BIaaS, DWaaS and other cloud offerings for IT provisioning.

Performance management and analytics are the right answers to today's challenges running a business: **You can only manage what you can measure**. But it is important to note that measuring does not stop by measuring financial metrics. We have to combine financial with non-financial metrics, and we have to consider operational, tactical, and strategic metrics for composing the relevant KPMs. This is one of the big challenges to master for guiding enterprises into a successful future. Technological innovations in mobile internet and in social media as well as the resulting collaboration methods are now laying the ground and act as the big drivers.

Performance management and analytics will only be successful, if managers and staff members will be trained in analytics. Only if results of analytics are carefully tracked, understood, interpreted, and well communicated to all organizational levels in the language of the business, then analytics will be a success, and performance management will become an indispensable building brick of monitoring and controlling an organization and its processes.

Take Away: Critical success factor for performance management and analytics is to put information into the context of business processes and enriching them with embedded analytics. Information related to processes and integrated into processes fosters compliant and traceable processes. Process managers and executives are empowered and can comprehend the actual position and state of each process and take actions in right time, when problems and risks are arising. This truly enables better decisions, the ultimate goal of performance management and analytics. Therefore, performance management must be part of all workplaces. Analytics provide the foundation for intelligent and smart processes monitored and controlled by performance management. Indeed, BI is turned into Performance Management and Analytics.

Annecy, August 2015

Contact-Address: Wolfgang.Martin@wolfgang-martin-team.net

Literature:

Inmon, W.H., Imhoff, C., and Sousa, R.: Corporate Information Factory, John Wiley & Sons, New York, 1998, 274 pages

Henschen, D.: Analytics at Work, Q&A with Tom Davenport (Interview), InformationWeek Software, 04. January 2010.

Hinssen, P.: The New Normal: Explore the Limits of the Digital World, Lannoo Publishers, Tielt/Belgium, 2011, 208 pages

Lehmann, P., Martin, W., and Mielke, M.: Data Quality Check 2007 – Trends im deutschen Markt, Steinbeis Edition, Berlin - Ludwigshafen, 2007, 34 pages

Luckham, D.: The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems, Addison Wesley Professional, Boston, 2002, 400 pages

Martin, W.: Business Performance Management und Real-Time Enterprise – auf dem Weg zur Information Democracy, Strategic Bulletin, IT-Verlag für Informationstechnik GmbH, Aying/Munich, 2003-A 32 pages

Martin, W.: CRM 2004 – Kundenbeziehungsmanagement im Echtzeitunternehmen, Strategic Bulletin, IT-Verlag für Informationstechnik GmbH, Aying/Munich, 2003-B, 32 pages

Martin, W.: BI 2004 – Business Intelligence trifft Business Integration, Strategic Bulletin, IT-Verlag für Informationstechnik GmbH, Aying/Munich, 2004, 32 pages

Martin, W.: SOA 2008 – SOA basierendes Geschäftsprozessmanagement, Strategic Bulletin, IT-Verlag für Informationstechnik GmbH, Aying/Munich, 2007, 28 Seiten

Martin, W.: Rule-based composition of agile Business Services, White Paper, S.A.R.L. Martin, www.wolfgang-martin-team.net, Annecy, 2008, 21 pages

Martin, W.: Information Governance – Ergebnisse einer Marktbefragung zum Status Quo und zu den Trends 2012, Research Note, S.A.R.L. Martin, www.wolfgang-martin-team.net, Annecy, 2012, 12 pages

Martin, W.: Big Data, Strategic Bulletin, IT-Verlag für Informationstechnik GmbH, Aying/Munich, 2012, 42 pages

Martin, W.: Wie NoSQL-Datenbanken den Einsatzbereich von BI erweitern: BI entlang der Prozesskette, BI Spektrum, Ausgabe 01, 2014, 4 Seiten

Mayer-Schönberger V., Cukier, K.: Big Data: A Revolution That Will Transform How We Live, Work, and Think, John Murray (Publishers), London, 2013.

12 Glossary and List of Abbreviations

In this glossary, we summarize the key definitions used in this white paper.

Analytics denotes the process for gaining information and for deriving a model for utilization of information (e.g. a predictive model as a result of a data mining process) as well as embedding and utilizing this model within a business process. The idea behind is to create an "intelligent" process.

Analytic Databases dramatically improve scalability and performance in comparison to traditional databases. They also can reduce the operational costs of running a database. This is achieved by a combination of known and of new technologies like column orientation, compression, special and intelligent access methods, massively parallel processing as well as in memory technologies.

Agility is defined as the capability of flexibility.

In IT, **Architecture** specifies the interplay of components of a complex system. It describes the translation of business requirements into construction instructions. Hence, architecture has characteristics and consequences.

BI as a Service (BIaaS) means the provisioning of analytic and performance management services via **Cloud Computing**. Cloud Computing[57] is an IT provisioning model based on virtualization. One or several providers offer resources like infrastructure and applications or data as distributed services via the internet. These services are flexibly scalable and can be invoiced according to consumption.

Big Data (formerly: data deluge) denotes the huge volume of data in the web consisting of a mix of structured and poly-structured data linked via complex relationships. There is an immense ocean waiting for being exploited. Data sources in the web are manifold: portals, web applications, social media, videos, photos and much more, all kind of web content. Besides the mere deluge, another problem about Big Data is the variety and multiplicity of sources. Each time the access to new source has to be specified again. Consequently, this causes a time and resource problem. Furthermore, more and more users want to analyze big data in real-time.

BI Governance consists like any governance of four elements, an organizational structure, the BI governance processes, the corresponding management policies and a technology platform. The organizational structure consists of a steering committee directed by the BI sponsor, the BI competency center, and the data stewards. This goes hand in hand with well-defined roles and responsibilities.

Big MDM, see Social MDM

Business Activity Monitoring (BAM) is a subset of operational performance management focusing on processing events (simple events and event streams). The task of BAM is the

[57] Definition of Cloud Computing according to Prof. Dr. Helmut Krcmar, TU Munich, 2008.

identification and evaluation of events for controlling purposes. In addition to BAM Complex Event Processing (CEP) focuses on processing more complex events.

Business Intelligence (BI) means the capacity to know and to understand as well as the readiness of comprehension in order to exercise this knowledge and understanding for mastering and improving the business. In more detail, we define: Business Intelligence is a model consisting of all strategies, processes, and technologies that create information out of data and derive knowledge out of information so that business decisions can be put on facts that launch activities for controlling business strategies and processes.

A **Business Process is…**

a set of activities and tasks carried out by resources
(services rendered by people and machines)
using different kinds of information
(structured & poly-structured)
by means of diverse interactions
(predictable & unpredictable)
governed by management policies and principles
(business rules & decision criteria)
with the goal of delivering agreed upon final results
(strategies & goals)

Business Process Management (BPM) manages the life cycle of business processes. It is a closed-loop model consisting of three phases:

- **Phase 1:** Analyzing, planning, modeling, testing, and simulating business processes.
- **Phase 2:** Executing business processes by cross-application process flows through a process engine on a SOA (service oriented architecture) infrastructure.
- **Phase 3:** Planning, monitoring and controlling of processes and the performance of the ensemble of all business processes.

A **Business Scorecard** is a consistent and comprehensive group of metrics together with a management policy for monitoring and controlling the performance of a group of processes, a business division or even the total enterprise. Consistency of metrics is very important, because metrics should not be contradictory and cause conflicts between roles and collaborative teams working in different contexts. A business scorecard is best defined by BI Governance: It can be derived from the information profile of an organizational role and is implemented through dashboard technologies.

Business Services are components of a business process. They represent functional services provided by service centers, competence centers or traditional departments. Provisioning is either internal or by third parties. They also can be deployed in a SaaS model.

Business Vocabulary (or Business Glossary or -Terminology). It represents the terminology of all business items and business knowledge within an enterprise and within a network of enterprises. It is an indispensable prerequisite for business process management, since processes need a common terminology for modeling and communication across all participants in both, business and IT, but also across relevant suppliers, partners, and customers.

A **Chief Data Officer (CDO)** has the responsibility for enterprise-wide information management and governance. He/she should be on the C-level. The position of a CDO is related to the tasks of a CIO, but separated. As a rule, the CDO should report to the Chief Marketing Officer (CMO) or to the Chief Execution Officer (CEO). He/she has the responsibility to manage data as an asset for the organization and to optimize adding value by data and information. He/she assures that the right data is collected, analyzed, and used for decision making. He/she also assures that ethics for analytics are developed and applied in the context of compliance.

Cloud Computing. In the context of this white paper, we refer to BIaaS.

COBIT is a framework for developing, implementing, monitoring and improving information governance and management practices. It is published by the IT Governance Institute and the Information Systems Audit and Control Association (ISACA). The goal of the framework is to provide a common language for business executives to communicate with each other about goals, objectives and results. The original version of the framework was published in 1996. The name COBIT originally stood for "Control Objectives for Information and Related Technology". The actual version COBIT 5 is based on five key principles for governance and management of enterprise IT: meeting stakeholder needs, covering the enterprise end-to-end, applying a single, integrated framework, enabling a holistic approach, separating governance from management

Competence center. This is an organizational construct for governance. It is a cross-functional business unit that acts as an interdisciplinary team and has the responsibility to advocate the deployment of a defined program (e.g., BI, Information Management) in the business.

Complex Event Processing (CEP) deals with processing of complex events. Task of CEP is the identification, analysis, consolidation and classification of dependent events. Therefore, CEP is the next step beyond BAM. CEP means the methods, techniques and tools for processing events when they occur, i.e. in real time. CEP derives a deeper and more valuable insight into the events by a complex event. A complex event is a situation that can only be recognized by the combination of several events.

Compliance means to act in accordance to all management policies as well as to all legal and regulatory requirements: Everybody acts as he/she should act.

Data Discovery means a new generation of business intelligence (BI) tools that excel by extraordinary user friendliness and flexibility. Furthermore, they use in-memory technologies for internal storing and processing. The big advantage of in-memory technology is performance: Therefore, Data Discovery tools are especially equipped and suited for Big Data analytics. Furthermore, Data Discovery tools come with visualization, interactive intuitive analysis, collaboration and autonomy of users.

Data Time Variance means a time-dependent data management. This is achieved by a bi-temporal data management affording to create a complete history of data from the users and the system’s view: What has been the state of an object at a certain point of time (resp. over a period of time) and what has been the value of data in an application at a certain point of time (resp. over a period of time)? Bi-temporal data management means that both, duration of validity and time of transaction of data are stored.

Data Integration provides information services for analyzing not only data, but also meta and master data, for managing data models, for preparing and profiling data and meta data, as well as it is a platform for ETL- (extraction, transformation, load) services.

A **Data Integration Platform** enables parallel, simultaneous, and if necessary real-time access to operational and analytic data via services in the context of a SOA. Now, the traditional Data Warehouse becomes a component of a data integration platform.

Data Lake. This term goes back to James Dixon, CTO at Pentaho. In his blog mentioning the term for the first time, he writes "If you think of a data mart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples." (See chapter 7.5)

In contrast to a data warehouse, a **Data Mart** is not an enterprise wide solution, but a data warehouse concept for either an isolated partial task in decision support or in the context of data warehouse architecture a task related subset of a data warehouse.

Data Mining is a process that identifies and/or extracts information from a large or a very large amount of (structured) data; information that has not been known before, that is non-trivial, unexpected as well as it is important.

Data Quality should be implemented into operational processes by a TQM (total quality management) program. The four cornerstones of data quality are:

- Quality is defined as the degree of *compliance* with the requirements.
- The principle is: *Preventing* is better than healing.
- *The zero-defect principle* must become the standard.
- Cost of quality is the cost of non-compliance with the requirements.

Data Scientists have the responsibility to advocate the exploitation of Big Data and to make sure that Big Data potentials can be realized. They act as an intermediate between business and IT, and should drive for a continuous improvement of the collaboration across all business functions, IT inclusive. This requires new skills, and a shift in IT: In the age of Big Data, data management is becoming the proper and main task of IT.

Data Virtualization means a virtual (logical) access to data via a data abstraction layer. Access to data is centralized avoiding data replication and duplication.

A **Data Warehouse** is a model for a subject oriented, integrated, time variant and persistent database for tactical and strategic decision support in an enterprise. It is separate from operational IT systems, and via a data integration platform, it can also be used in operational decision support.

Data Warehouse as a Service (DWaaS) means the provisioning of a data warehouse via Cloud Computing. (See also BIaaS).

A **Decision (Rule) Engine** (or: business rule management system – BRMS) implements a particular decision logic. Decision logic is defined by rules.

Decision Management is a process or a set of processes for streamlining and improving actions. Decision management systems treat decisions as reusable assets and use technology for automating decision processes. Decisions can be either fully automated or can present suggestions to human decision makers.

DevOps is the blending of tasks performed by a company's application development and systems operations teams. The term DevOps is broader than BI, but DevOps is perfectly suited for BI programs and projects. It is being used in several ways. In its most broad meaning, DevOps is a philosophy or cultural approach that promotes better communication between the two teams. In its most narrow meaning, DevOps is a job description for an employee who possesses the skills to work as a both a developer and a systems engineer. In addition to traditional development tools, the DevOps toolkit includes configuration management tools, a repository for storing versions of code, indexing tools, and tools for performance monitoring how changes to code affect the environment.

A **Digital enterprise** is an organization that uses information technology as a competitive advantage in its internal and external operations. It is based on a platform of technologies designed to help a business engage with customers across all digital channels (web, social, mobile, print, i.e. omni-channel) in an efficient manner. A digital enterprise has already adopted the principles of industrialization, agility, compliance, and smart acting. It is based on comprehensive service oriented business process management and analytics (cf. fig. 5).

Digitalization is the use of digital technologies to change a business model and provide new revenue and value-producing opportunities. It is the process of moving to a digital business (source: Gartner).

Embedded Analytics is the integration of analytical services into business software and business processes. It allows business users to easily access analytical tools while performing everyday tasks. Real-time access enables more informed and efficient decision making. Analytics functionality embedded within business software includes dashboards, data visualization tools, self-service analytics, visual workflows, benchmarking, static, interactive and mobile reports etc.

ELT see ETL processes

Enterprise 2.0 see social business.

Entity Identity Resolution is about managing entity identity data when merging data from different sources for correctly identifying an entity like product, service, customer, supplier, lead, opinion leader, patient, tax payer, criminal etc.

ETL Processes (extract, transform, load) load and update data marts or a Data Warehouse with internal or external data. „Extract" means unloading of data from various data sources, „transform" means transformation of the extracted data into the Data Warehouse model and „load" means the corresponding load process. ELT processes are an alternative, where the transform and load phases are interchanged. Whereas ETL processes perform the transform phase outside the database, ELT processes transform inside the database which can lead to certain performance gains.

Gamification is the application of game theory concepts and techniques to non-game activities, i.e. to real life situations. The goal of gamification is to engage the participant with an activity he finds fun in order to influence his behavior.

Geocoding is the determination of position of one or several points via allocation of geographical coordinates. Geocoding allocates area codes (district, street, street section) as well as xy coordinates to spatial addresses or IP addresses. Based on address data, geocoding enables the spatial representation of customers on a map or tracking of customers in the mobile internet.

A **Golden Record** presents the unique and well-defined version of all data entities in an organization. It is the ‘single point of truth”, where “truth” means a reference for users ensuring the correct and valid version of a data record. It is the goal of master data management to derive and to deploy the golden record for all entities.

Google BigTable is a distributed, column-oriented data store created by Google Inc. to handle very large amounts of structured data. From the beginning, the technology was intended to be used with petabytes of data. It uses a simple data model that Google has described as "a sparse, distributed, persistent multidimensional sorted map." Data is assembled in order by row key, and indexing of the map is arranged according to row, column keys and timestamps. Compression algorithms help achieve high capacity. Google BigTable serves as the database for applications such as the Google App Engine Datastore, Google Personalized Search, Google Earth and Google Analytics. Google published the documentation of BigTable. This has allowed other organizations and open source development teams to create BigTable derivatives, including the Apache HBase database, which is built to run on top of the Hadoop. Other examples include Cassandra, which originated at Facebook Inc., and Hypertable, an open source technology that is marketed in a commercial version as an alternative to HBase.

Governance means an organization together with a controlling of all activities and resources in the enterprise directed on respectful and durable value creation based on longevity. The result is a compliant management and behavior. It must be ensured that all management policies and guidelines are respected and followed in all activities of all resources – people, machines and systems. In this way, governance ensures compliance.

Hadoop is a development project of the Apache Software Foundation. It acts like a data operation system and consists of three components: the storage layer HDFS (Hadoop Distributed File System), the programming framework MapReduce for processing queries, and a function library. In addition, there are some complements like the data management system HBase for structured data, the High Level Query Languages (HLQL) Hive, Pig und JAQL, ZooKeeper (managing of distributed configuration), Chukwa (real-time monitoring) and others.

In-database analytics is a technology that allows data processing to be conducted within the database by building analytic logic into the database itself. Doing so eliminates the time and effort required to transform data and move it back and forth between a database and a separate analytics application. In-database analytics is beneficial for applications requiring intensive processing - for example, fraud detection, credit scoring, risk management, trend and pattern recognition, etc. In-database analytics also facilitates ad-hoc analysis and data discovery.

Infonomics is a social science describing the study and emergent discipline of quantifying, managing and leveraging information as a formal business asset. Infonomics endeavors to apply both economic and asset management principles and practices to the valuation and handling of information assets. The word is a composite of “information” and “economics.”

Information Management is all about to create ***trusted data*** in the sense of the “single point of truth”. Information management includes data definition (the enterprise terminology), data modeling (the enterprise semantic), meta and master data management (transparency and traceability), data quality management (relevance and accuracy), and data security and protection.

In-Memory Analytics is an approach to querying data when it resides in a computer’s random access memory (RAM), as opposed to querying data that is stored on physical disks. This results in vastly shortened query response times, allowing business intelligence and analytic applications to support faster business decisions. As the cost of RAM declines, in-memory analytics is becoming feasible for many businesses. BI and analytic applications have long supported caching data in RAM, but older 32-bit operating systems provided only 4 GB of addressable memory. Newer 64-bit operating systems, with up to 1 terabyte addressable memory and technologies like SAP HANA have made it possible to store entire data warehouses or data marts in a computer’s RAM. In addition to providing incredibly fast query response times, in-memory analytics can eliminate the need for data indexing and data aggregates. OLAP cubes or aggregate tables are no more necessary. This reduces IT costs and allows faster implementation of BI and analytic applications.

The **Internet of Things** (IoT) is a concept describing the expansion of the Internet when physical objects like end user devices are connected to the Internet. Typical elements of the IoT are embedded sensors, image recognition technologies, payment by NFC (Near-field Communication) etc. Consequently, the term “mobile“ will not be restricted any more to mobile telephones and tablets. Mobile cellular technology will soon be integrated into many new devices, for example into cars.

Knowledge Management. It has two tasks, the person to person transfer of knowledge as well as the documentation of knowledge. In a first approach, it is about bringing knowledge into the heads of all employees, but more important, it is about extracting knowledge from the experts so that it can be shared by everybody.

Location Intelligence means the geographic dimension of Business Intelligence. It combines technology, data and services with domain knowledge. It enables enterprises to measure, to compare, to visualize and to analyze their business data in a geographical context.

Machine-to-Machine (M2M) means any technology that enables networked devices to exchange information and perform actions without the manual assistance of humans. M2M communication is often used for remote monitoring. In product restocking, for example, a teller or a vending machine can message the distributor when a particular item is running low. M2M communication is an important aspect of warehouse management, remote control, robotics, traffic control, logistic services, supply chain management, fleet management and telemedicine. Key components of an M2M system include sensors, RFID, a Wi-Fi or cellular communications link and embedded software programmed to help a networked device interpret data and make decisions.

Management by Exception is a policy by which management devotes its time to investigating only those situations in which actual results differ significantly from planned results. The idea is that management should spend its valuable time concentrating on the more important items.

A **Manufacturing Execution System (MES)** is a system for efficient production controlling. It optimizes order processing and manufacturing processes. It is directly linked to business processes and enables real-time controlling of production. As a rule, an ERP system sits on top of an MES for efficiently allocating resources and production planning.

MapReduce, proposed by Google, is a programming framework and model for distributed processing across many to very many nodes. These nodes consist of low-price, commodity hardware. For more details, we refer to Big Data and Hadoop.

Mash Up means the creation of new content by seamless (re-)combination of existing content. Prerequisite for successfully mashing up is a SOA as an infrastructure.

Master Data means business-oriented meta data that build the foundation for the business vocabulary. Hey describe business structures like assets, products and services, and the business constituents (e.g., suppliers, customers, employees, partners etc.) This provides a single and unique view on all enterprise structures.

Meta Data provide Information about data that is typically managed in a repository. Meta data consists of master data, navigational and administrative data. It describes structure, elements, and properties of the data elements plus the corresponding rules.

Metrics specify how to measure and to manage the performance of processes and / or how to monitor and control business processes. Metrics are derived from enterprise and process goals. Metrics act like sensors alongside a process for proactively identifying problems (e.g. the right-time identification that a planned target cannot be reached) so that the necessary counteractions can be put in place. A metric consists of a performance figure and its scale for interpreting the outcome of the performance figure so that the right decisions can be taken.

NoSQL means “not only SQL“. It is an initiative to drive an alternative approach to data management. It is about databases with a non-relational model. The concept of fixed table schemata and of the join-operation is abandoned. The result is horizontal scalability and excellent read-performance even with very large data volumes à la big data. The NoSQL databases are also called “structured data stores”.

OLAP (online analytic processing) is an analytic method enabling fast and interactive access to relevant information. It provides complex analysis functions and features based on a multi-dimensional data model. In such a data model, metrics are aligned by various dimensions, e.g. revenue in respect to customer, product, region, time period etc.

Operational Intelligence or operational performance management means the application of performance management principles to monitoring and controlling of operational processes in (near-)real-time.

Performance management is defined as a business model enabling a business to continuously align business goals and processes and keeping them consistent. The task of performance management within BPM is planning, monitoring, and controlling of processes and their performance.

Poly-structured data denominates data with unknown, insufficiently defined or multiple schemas, for instance machine-generated event data, sensor data, system log data, internal/external Web Content inclusive social media data, text and documents, multi-media data like audio, video etc.

Predictive analytics deals with extracting information from data and using it to predict trends and behavior patterns. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome.

Prescriptive analytics goes beyond predicting future outcomes (as predictive analytics does) by also suggesting actions to benefit from the predictions and showing the implications of each decision option. Prescriptive analytics not only anticipates what will happen and when it will happen, but also why it will happen. Further, prescriptive analytics suggests decision options on how to take advantage of a future opportunity or mitigate a future risk and shows the implication of each decision option. Prescriptive analytics can continually take in new data to re-predict and re-prescribe, thus automatically improving prediction accuracy and prescribing better decision options.

R is a free programming language already developed in the early 90s. It is published under a General Public License (GPL) of the Free Software Foundation (see http://www.r-project.org). R stands for **The R Project for Statistical Computing**. It is a software for statistical computing and graphics.

Real-time. In the context of a business, it means to have the right information in right time in the right location for the right purpose. The “real-time“ requirement in business does not necessarily mean clock-time. Real-time can be best explained by the availability of information with the speed that is needed.

A **Repository** is a database for managing meta data.

Risk management encompasses all activities of risk identification, minimization, and avoidance.

Self-service Business Intelligence enables users to analyze their data in an interactive and visual way. It has a role specific and customizable user interface and also supports search functionality. It creates autonomy of BI users towards IT. Today, self-service BI is in particular part of data discovery tools.

A **Service** (in informatics) is a functionality typically triggered by a request-response mechanism via a standardized interface and consumed according to an SLA. In consequence, a service is a special instantiation of a software component.

A **Service Level Agreement (SLA)** defines the (legally) binding service delivery terms and conditions for service consumers and producers.

Service Orientation (SO) describes the collaboration between a consumer and a provider. The consumer is looking for a particular functionality (a “product” or a “service”) offered by the provider. Such collaboration works according to the following principles:

- Principle 1 – **Consistent Result Responsibility**. The service provider takes responsibility for the execution and result of the service. The service consumer takes responsibility for controlling service execution.
- Principle 2 – **Unambiguous Service Level**. The execution of each service is clearly agreed to in terms of time, costs and quality. Input and output of services are clearly defined and known to both parties by the Service Level Agreement (SLA).
- Principle 3 – **Proactive Event Sharing**. The service consumer is informed about every agreed change of status for his work order. The service provider is required to immediately inform the service consumer of any unforeseen events.
- Principle 4 – **Service Bundling.** For service provisioning, a service can invoke and consume one or more other services, and can be invoked and consumed by other services.

Shadow IT means usage of IT systems with authorization by the organization.

Smart is a property that means user-focused, autonomously adaptive, forward thinking, self-controlled, context sensitive and always connected (see Henseler[58]).

Social Business (formerly called enterprise 2.0) means the use of social media software platforms in an enterprise or between enterprises and its customers and partners (according to Andrew McAfee). Social Business is all about use and use patterns of social business technology (formerly called Web 2.0) in an enterprise. This is more than just using social business technologies.

Social Business Intelligence denotes the extension of performance management and analytics by social media functionality and collaboration, by knowledge management, by new technologies (web and cloud integration tools, analytic databases, text analytics), and by new application areas (social media performance management, social media analytics).

Social MDM (master data management) hast two meanings. Some vendors like Informatica and SAS define social MDM as merging social media data with MDM. This is sometimes also called **“Big MDM”.** (I also prefer and use this definition). Others define social MDM as the maintenance and quality assurance of master data by social community members: Everybody maintains his/her data, and consequently creates trustful data that can be used by all community members. In the end, this is a master data market.

Social Media Interaction builds on social media monitoring and closes the loop. It means the interactions of an organization with social media participants. An organization now can immediately react to relevant contributions and opinions in social media, and can intervene. This creates advantages in customer service or when introducing new products to market, because a communication can be built and sustained with social media communities.

[58] Prof. Wolfgang Henseler (Hochschule Pforzheim), Keynote „„Schöne, smarte Welt – Wenn die Qualität der Daten der neue Wertmaßstab ist und persönliche Daten immer häufiger zur Währung werden.“ at „Best in Cloud 2014“ (IDG Business Media).

Social Media Monitoring. Its task is to sniff where, when and how in social media, an enterprise, a person, a product or a brand is talked about and discussed. It starts with the identification and extraction of relevant Big Data sources by applying agile web integration tools. The extracted data is then analyzed by text analytics. This provides statistical information about where traces are in the web and how many, and more importantly, it also provides the tonality of all contributions by sentiment analysis.

Spark is another open source platform for big data processing. This framework for cluster computing was developed in 2009 by researchers at the University of Berkeley for accelerating processing in Hadoop systems. Since 2013, it is a project of the Apache Software Foundation, since 2014 a Top Level Project. Since 2015, Spark is massively supported by IBM with some 3,500 developers. Databricks is the distributor of Spark. Spark can be understood as a fast and general engine for large-scale data processing sitting on top of data management systems like Hadoop, NoSQL DBMSs, Amazon Web Services (AWS) and relational databases.

SQL stands for “sequential query language”. It is a set theoretic oriented, standardized query language for relational database management systems.

Text Analytics is a new type of analytics expanding data and text mining into content management and the World Wide Web. It combines linguistic methods with search engines, text mining, data mining, and machine learning algorithms.

Total Quality Management (TQM) means to declare quality as an overall, sustainable goal. It is an integrated, continuous, comprehensive, monitoring and controlling, as well as organizing set of activities across all departments of an organization. TQM was developed by the Japanese automotive industry, and it proved to be a successful model. For TQM, full support of all employees is a critical success factor

Web Analytics means the application of performance management and analytics to web data produced by (human) visitors when surfing on web pages.

Yarn (Yet Another Resource Negotiator). Hadoop YARN is a cluster management technology. It is an essential building block of the Hadoop 2 version of the Apache Software Foundation. YARN now plays the role of a large-scale, distributed operating system for Big Data applications.

List of Abbreviations

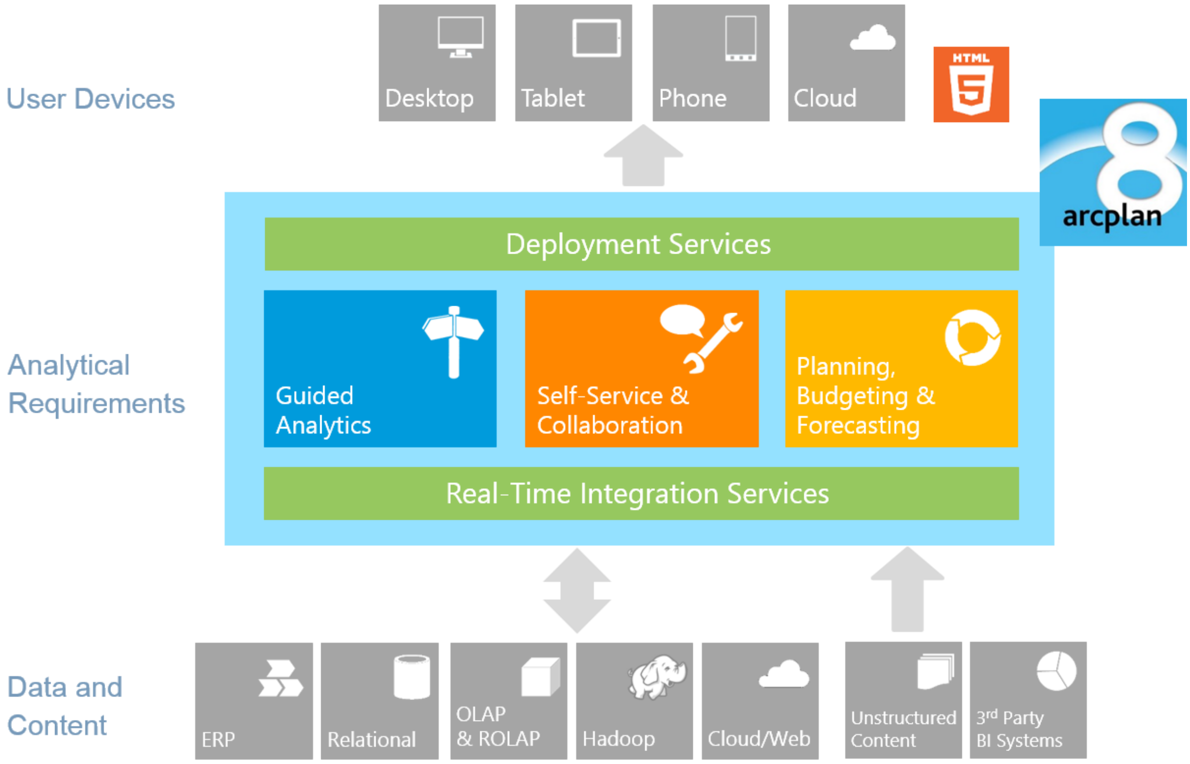
ACID	atomicity, consistency, isolation, durability
B2B	business to business
B2C	business to consumer
BAM	business activity monitoring
BI	business intelligence
BIaaS	business intelligence as a service
BPM	business process management
BYOD	bring your own device
CAD	computer aided design
CAGR	compound annual growth rate
CAM	computer aided manufacturing
CC	competency center
CDC	change data capture
CEM	customer experience management
CEP	complex event processing
CDO	chief data officer
CFO	chief financial officer
CMO	chief marketing officer
CIO	chief information officer
CPM	corporate performance management
CPO	chief performance officer
CRM	customer relationship management
CRUD	create, read, update, delete
DBA	database administrator
DBMS	database management system
DI	data integration
DW	data warehouse
DWaaS	data warehouse as a service
EDH	enterprise data hub
EII	enterprise information integration
ELT	extract, load, transform
ERP	enterprise resource planning
ESB	enterprise service bus
ESDB	enterprise service data bus
ETL	extract, transform, load
GIS	geographical information system
GRC	governance, risk management and compliance
HDFS	Hadoop distributed file system
HLQL	high level query language
HOLAP	hybrid OLAP
IaaS	infrastructure as a service
IoT	Internet of Things
IT	information technology
KPM	key performance metric
LLDM	low latency data mart
M2M	machine-to-machine

MDM	master data management
MDM	mobile device management
MES	manufacturing execution system
MR	map reduce
MOLAP	multidimensional OLAP
MPP	massively parallel processing
NFC	near-field communication
NoSQL	not only SQL
ODS	operational data store
OODB	object oriented database
OLAP	online analytical processing
OLTP	online transactional processing
OPEX	operational expenditure
PaaS	platform as a service
PM	performance management
PPM	process performance management
RDBMS	relational DBMS
REST	representational state transfer
RFID	radio frequency identification
ROLAP	relational OLAP
ROI	return on investment
SaaS	software as a service
SDK	software development kit
SLA	service level agreement
SME	small medium enterprise
SOA	service oriented architecture
SOM	self-organizing maps
SQL	sequential query language
TQM	total quality management
XML	extensible mark-up language

13 The Sponsors

arcplan Information Services

arcplan is a leader in innovative Business Intelligence, Dashboard, Corporate Performance and Planning software solutions for use on any device. arcplan-based solutions deliver timely, contextual and actionable information that empowers businesses to improve business performance while leveraging existing infrastructure. The arcplan platform is used by more than 3,200 customers worldwide to build and deploy analytics applications, dashboards, reporting, and planning solutions that precisely match the user requirements. The arcplan customer base includes companies of all industries.



arcplan product offering by functional requirements

The **arcplan platform -** as illustrated in the graphic above – provides solutions for the following use cases: **arcplan Enterprise**, the company’s flagship product, is a highly flexible BI solution for *guided analytics*, best suited to build analytics applications, dashboards and reporting solutions. **arcplan Edge** is a robust *budgeting, planning and forecasting* solution and combines this with process-oriented workflow and Web-based reporting strengths of arcplan Enterprise and a Microsoft Excel integration.

arcplan Engage complements the arcplan product suite with enhanced *self-service options*, delivering Web 2.0 *Search & Collaboration* features to the BI world, integrates unstructured data sources and offers with arcplan Spotlight and arcplan Excel Analytics ad-hoc analysis functions for Web and Excel respectively.

With version 8 any arcplan application is directly suited for mobile deployment. Using HTML5 and the Responsive Design principle for application development, any arcplan application can be deployed to desktop, tablet, and smartphones, instantly supporting the device specific form factor and interaction functions. arcplan calls this *DORA* – **Design Once, Run Anywhere**.

With the **arcplan Application Designer** all products share a **highly visual development environment** so that customers may successfully meet their project requirements and timelines within the established budget. Complemented by **application templates** provided with the arcplan Application Framework, development processes can be enhanced further.

The **arcplan Application Server** (available for Windows®**)** serves as the engine for all BI and Corporate Performance Management tasks in the organization. It provides **bi-directional connectivity** for all arcplan applications to over **20 data sources** in real-time. Supported data sources are SAP HANA, SAP BW, SAP BW-IP, as well as the Oracle suite (including any Oracle database, Oracle OLAP, Oracle Essbase, Hyperion Financial Management, Hyperion Enterprise), Hadoop, IBM (Cognos TM1, IBM DB2 UDB, DB2 Cubing Services), Kognitio, Microsoft (SQL Server & Microsoft Analysis Services), Salesforce.com, Teradata, Cloudera Impala, LucaNet, any ODBC, OLE DB, XMLA, XML or any Web-service / SOA with an open API.

arcplan's flagship product arcplan Enterprise is also rated the #1 third party tool for SAP BW, Oracle Essbase, and IBM Cognos TM1 in The BI Survey 14 (2014).

More information can be found at: www.arcplan.com

DATAWATCH
GET THE WHOLE STORY™

Datawatch Corp.

Only Datawatch provides a Visual Data Discovery solution that can finally leverage all of the structured and unstructured data within organizations, including real-time sources. Users can unlock valuable insights from data in static reports, PDF files, print spools and EDI streams. And visualize both historic as well as rapidly changing data using real-time data streams from sources like CEP engines, tick or machine data.

Organizations of every size, worldwide use Datawatch products, including 99 of the Fortune 100. Datawatch is headquartered in Chelmsford, Massachusetts with offices in New York, London, Munich, Stockholm, Singapore, Sydney and Manila, and with partners and customers in more than 100 countries worldwide.

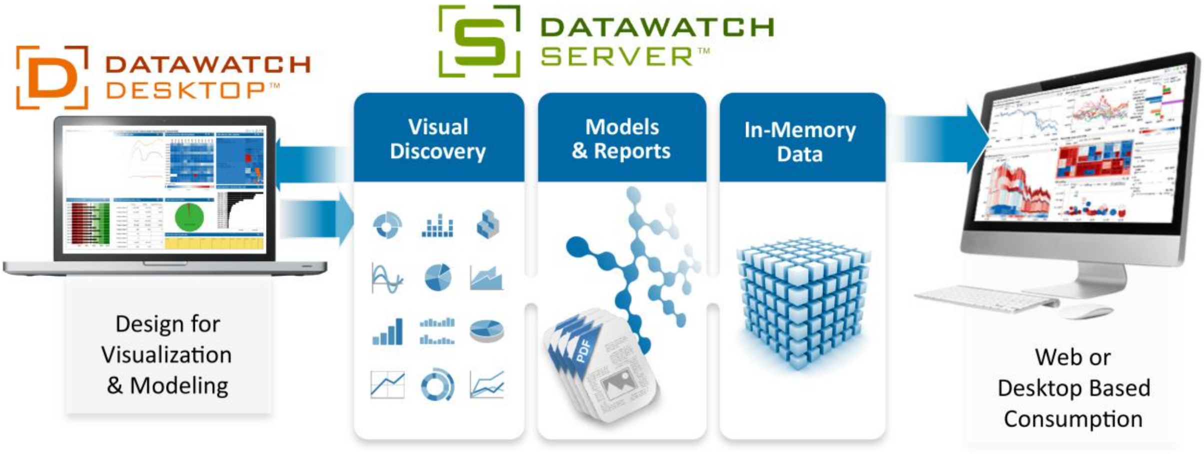
Next Generation Analytics

Datawatch is at the forefront of this rapidly emerging need for Next Generation Analytics. By associating all relevant data in a visually-rich, real-time analytical environment enables businesses to isolate and resolve problems as they occur, perceive hidden patterns, track emerging market trends, and identify opportunities for competitive advantage and improved business processes. Users need more than canned reports and static dashboards. And IT cannot be burdened with another wave of heavy upfront implementation cost and time.

The promise of “Big Data” has driven organizations to rethink their approach to traditional business intelligence. It’s no longer good enough to rely on a relatively small subset of business critical information that eventually makes its way into the data warehouse or delivered by canned reports. To stay competitive, organizations need to harness all of the relevant information to run your business regardless of its type (variety), its size (volume) or the speed in which its delivered (velocity).

Benefits with Datawatch Solutions

Only Datawatch provides a Visual Data Discovery solution that can finally leverage all of the structured and unstructured data within your organization, including real-time sources. Through our modular product design, you can start with deployments of any size and grow your visual data discovery deployment as your business needs evolve. From an individual analyst, to small departments, to fully integrated enterprises, Datawatch will be with you at every step of the way as you transform your business with Next Generation Analytics.



Industries

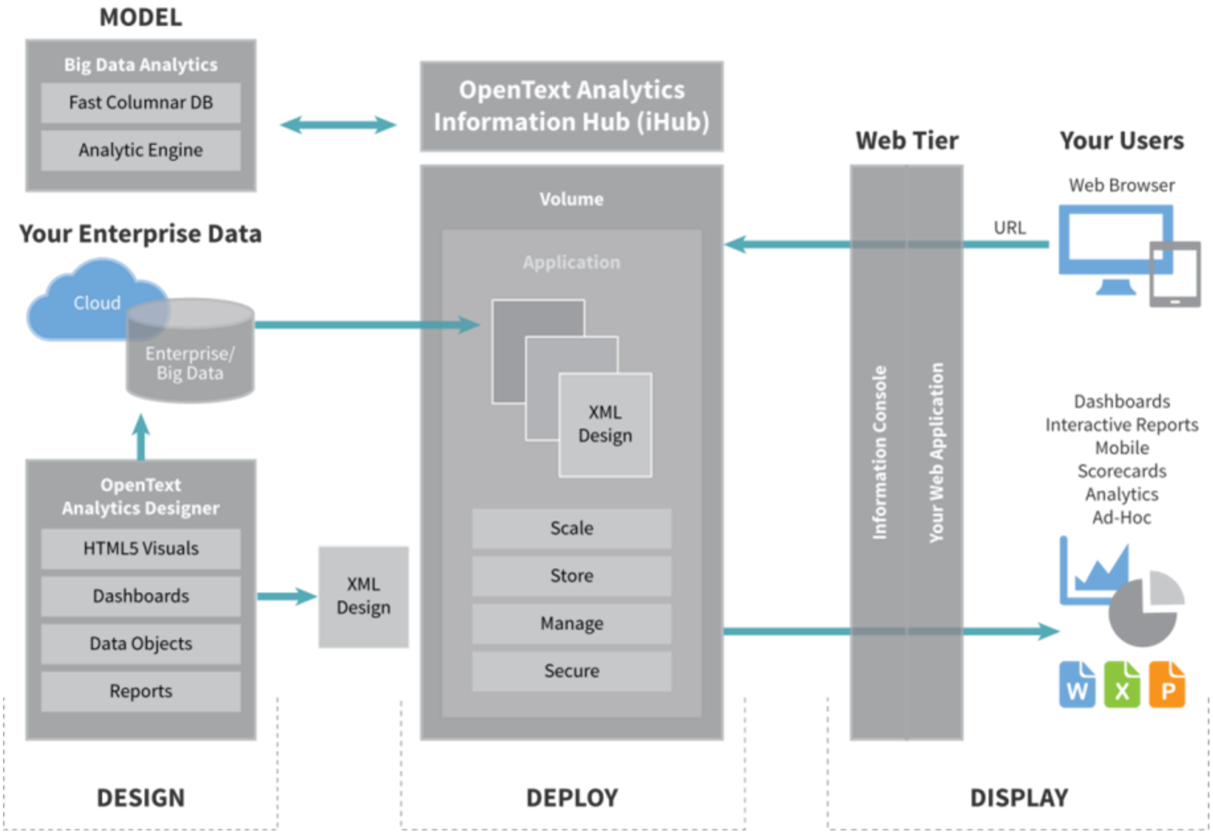
- Capital Markets
- Financial Services
- Manufacturing
- Solutions Center
- Energy
- Healthcare
- Retail
- Logistics
- Telecom

Please find further information on http://www.datawatch.com/

Actuate – The BIRT Company™

OpenText Analytics (formerly Actuate) provides embedded analytics solutions spanning enterprise scale data-driven applications, big data analytics and customer communications. More than 3.5 million BIRT developers and 200 million customers use embedded analytics solutions to build scalable, secure applications that save time and improve brand experience.

OpenText is the leader in Enterprise Information Management (EIM) software to companies across all industries to leverage their business information on-premises, in data centers or in the cloud. More than 100,000 companies use OpenText solutions to unleash the power of their information.



- ➔ **OpenText Actuate Big Data Analytics** turns big data into business intelligence in order to optimize marketing strategies, predict customer behavior and closing cycles, analyze key business processes, and more. By using personalized business analytics, your company will gain valuable customer insights.
- ➔ **OpenText Actuate iHub** is a world-class integrated development environment and deployment platform to create analytics, dashboards and data visualizations that seamlessly integrate with any enterprise and OEM applications.
- ➔ OpenText for **Customer Communication Management** is aimed at CCM Architects who deliver and manage high-volume customer communications by way of pixel perfect statement design, processing and storage.

For more information, please visit www.opentext.com/analytics

Company Profile pmOne AG

Founded in 2007, pmOne AG is a software vendor and consultancy specializing in solutions for Business Intelligence and Big Data. pmOne builds solutions using the technology platforms of Microsoft and SAP combined with its own software cMORE. cMORE helps business users quickly build and efficiently operate scalable reporting and analysis solutions that they can extend to meet their changing needs.

pmOne also sells and implements Tagetik, a leading global software solution for enterprise planning and consolidation. pmOne has 200 employees in 8 offices in Germany, Austria and Switzerland.

Customers and Customer Statements

pmOne AG has successfully implemented solutions at several companies and different industries, i.e.:

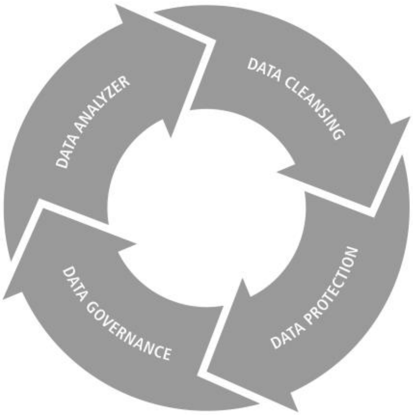
- Dr. August Oetker KG: Group Balance Sheet & Legal consolidation
  - “The rich capabilities at a reasonable price were just as important to us as the advantages of a Web-based solution for our global users.” Dr. Manfred Jutz, Director of Accounting at Dr. August Oetker KG
- AirBerlin: Enterprise Data Warehouse and Reporting
  - “Today we have no collection of isolated Data Warehouse solutions, but actually a comprehensive Data Warehouse for the entire Air Berlin Group.“ Claus Glüsing, Director Controlling Air Berlin
- Nationale Suisse: Consolidation and Reporting
  - „For us, transparency is important. We were impressed by Tagetiks clear structure and guidance through the consolidation process, while simplicity and clarity are maintained.“ Martin Harmann, Chief Accountant, Nationale Suisse
- M+W Group: Consolidation, Planning, Reporting and Treasury
  - "The reconciliation process is now easy and paperless. Subsidiaries accurately report transaction values from their business operations. Balances are consolidated and available for further analysis in the integrated cockpit.“ Renate Jäger, Project Leader, M+W Group
- Heraeus: Enterprise Data Warehouse
  - "We have created a basis for accommodating reporting both at a group level as well as meeting the specific needs of professionals who have to answer business questions on a local level, now and in the future.“ Torsten Kluin, Manager BICC Heraeus

Contact:

kontakt@pmone.com | www.pmone.com

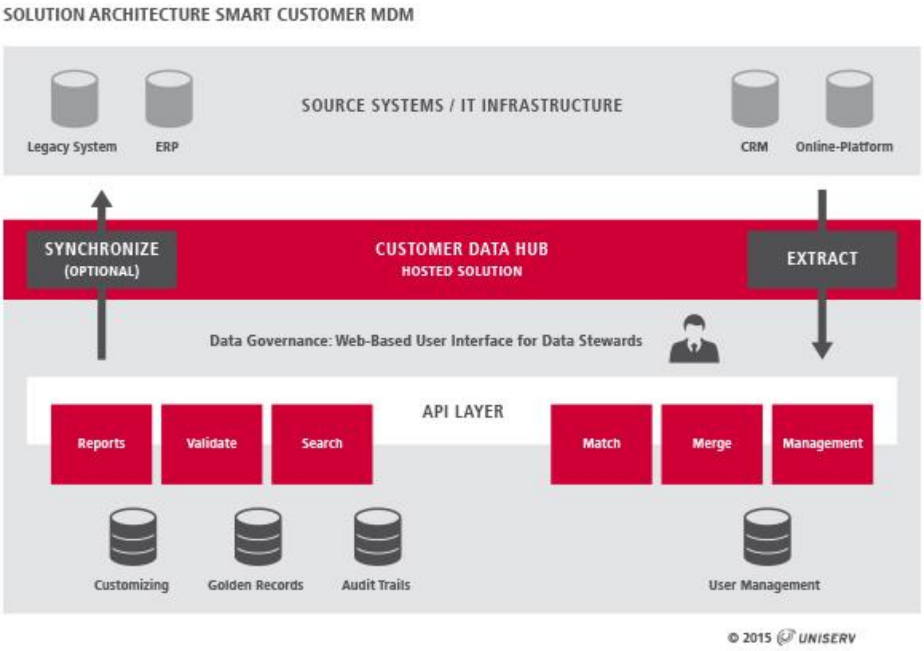
Uniserv GmbH

Uniserv is the expert in successful customer data management. Smart Customer MDM, the MDM solution for customer master data, combines data quality assurance and data integration in a comprehensive approach. In this respect, customer data is at the heart of initiatives for Master Data Management, data quality, data migration and data warehousing. Companies such as Sixt, Deutsche Bank and Volkswagen trust in the problem-solving expertise and quality of Uniserv in the environment of CRM applications, eBusiness, Direct and Database Marketing, CDI applications, Business Intelligence or MDM.



The task of performance management and analytics is the continuous coordination of corporate objectives and business processes in a closed loop to keep them consistent as well as a systematic investigation of the actual situation. The basis for this is the company data in general and the customer data in particular.

In order to be able to work effectively with customer data, it must be understood, maintained, protected and monitored throughout its life cycle. On the technical side, this is precisely what Smart Customer MDM guarantees: The products Data Analyzer (profiling), Data Cleansing (cleansing), Data Protection (first time right) and Data Governance (monitoring) ensure that the customer data is ready for use and, above all, reliable. The basis for this is the so-called Golden Record. It is the key to a 360° view of the customer and therefore to an improved process quality as a result of consolidated and trustworthy customer master data over all functional areas, applications and databases. This means that perfect data is available to the business as raw material for information, the optimum condition for performance management and analytics. To put it in a nutshell: Better Data. Better Business.



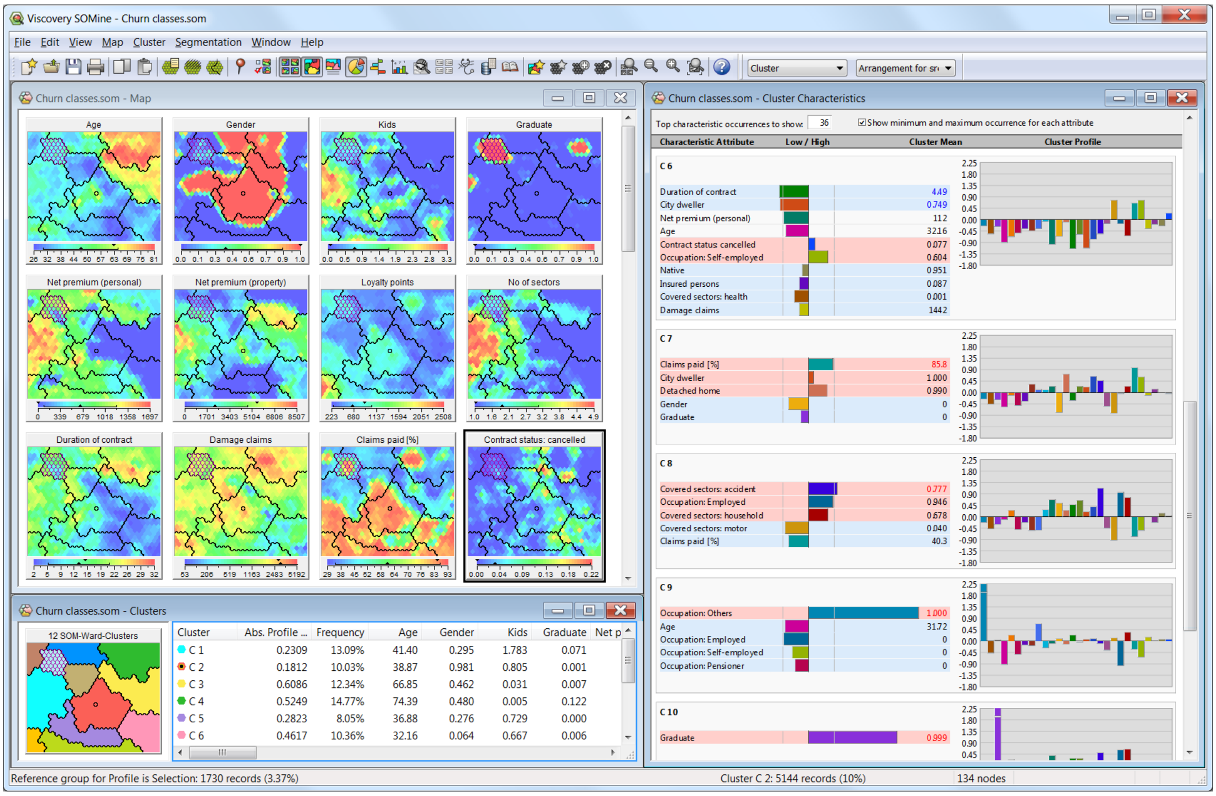
Further information can be found at www.uniserv.com

Viscovery Software GmbH

Viscovery Software GmbH is the vendor of one of the most powerful systems for explorative data mining and predictive analytics worldwide. Founded in 1994 under the name of Eudaptics and, since 2007, a member of the Biomax group, Viscovery has already satisfied more than 800 customers working on all continents.

Users in the fields of banking, insurance, telecommunication, industry, media and retail, as well as at research organizations and universities appreciate Viscovery because of its ease-of-use and its wide range of applications. Viscovery is successfully being used in risk and fraud applications, customer profiling and scoring, as well as for market segmentation, manufacturing optimization and patient-data analysis.

Viscovery SOMine combines the concept of self-organizing maps (SOM) with multivariate statistics for use in explorative data mining and predictive analytics. Its versatile set of easy-to-use tools allows visualization of high-dimensional data sets, recognition of dependences and estimation of attribute values for new objects.



Cluster	Abs. Profile ...	Frequency	Age	Gender	Kids	Graduate	Net p
C 1	0.2309	13.09%	41.40	0.295	1.783	0.071	
C 2	0.1812	10.03%	38.87	0.981	0.805	0.001	
C 3	0.6086	12.34%	66.85	0.462	0.031	0.007	
C 4	0.5249	14.77%	74.39	0.480	0.005	0.122	
C 5	0.2823	8.05%	36.88	0.276	0.729	0.000	
C 6	0.4617	10.36%	32.16	0.064	0.667	0.006	

Top characteristic occurrences to show: 36 ☑ Show minimum and maximum occurrence for each attribute

Characteristic Attribute	Low / High	Cluster Mean
C 6		
Duration of contract		4.49
City dweller		0.749
Net premium (personal)		112
Age		3216
Contract status: cancelled		0.077
Occupation: Self-employed		0.604
Native		0.951
Insured persons		0.087
Covered sectors: health		0.001
Damage claims		1442
C 7		
Claims paid [%]		85.8
City dweller		1.000
Detached home		0.990
Gender		0
Graduate		0
C 8		
Covered sectors: accident		0.777
Occupation: Employed		0.946
Covered sectors: household		0.678
Covered sectors: motor		0.040
Claims paid [%]		40.3
C 9		
Occupation: Others		1.000
Age		31.72
Occupation: Employed		0
Occupation: Self-employed		0
Occupation: Pensioner		0
C 10		
Graduate		0.999

Reference group for Profile is Selection: 1730 records (3.37%)

Cluster C 2: 5144 records (10%)

134 nodes

Viscovery SOMine, excerpt of a churn analysis of insurance customers. On the left, we see the attribute window of the SOM for selected variables. The window below shows the partitioning of the map by 12 clusters, as well as a table with the means of selected cluster attributes. In the middle on the right you see cluster characteristics, i.e. the attributes describing significantly the cluster. The red resp. blue marked variables represent largest resp. smallest profile values in the corresponding cluster. On the right from top to bottom, we see the complete cluster profiles. Users can arbitrarily combine various windows and store any arrangements.

Self-organizing maps are used to represent data distributions in perceptual maps. On this basis, Viscovery provides intuitive tools for visual cluster analysis and profiling, exploration of the data and revealing hidden dependences. Classification and scoring models based on the patented technology enable accurate prediction. For statistical analyses, a variety of functions are available, including descriptive statistics, correlation analysis, PCA, histograms, box plots, scatter plots and frequency tables.

The Viscovery suite offers data scientists as well as business users, regardless of their technical skills, an appropriate and powerful tool. The interactive Viscovery interface allows even users without a statistical background to recognize dependences in the data, perform visual cluster analysis and to derive group profiles. The entire data-mining process – from data import to cluster definition and predictive modeling and further to the creation of applications – is supported by workflows.

Viscovery has demonstrated a particular strength in the explorative analysis of Big Data. As a SOM represents data in perceptual maps, dependences can be identified on the basis of even small samples. The visual analysis immediately reveals interesting data clusters which, in turn, can be used for the formulation of hypotheses. A free of charge test version of Viscovery SOMine can be downloaded at www.viscovery.net/trial-version.

With a strong focus on the visualization of data distributions, Viscovery has for more than 20 years been a leader in data mining technology based on SOM and statistical methods. Already in 2008, Viscovery was the only continental European data-mining provider listed in Gartner's "Magic Quadrant for Customer Data-Mining Applications".

Viscovery SOMine Suite can be ordered via Viscovery’s Web shop. Additional information about Viscovery can be found at www.viscovery.net.

Impressum

Dr. Wolfgang Martin
S.A.R.L. Martin
6 rue Paul Guiton
74000 Annecy - France
www.wolfgang-martin-team.net
N° TVA FR81413866625
RCS Annecy B 413 866 625 00010
NAF/APE 6202A

WOLFGANG MARTIN TEAM
powerful connections