SIEMENS

EU BIG – European Big Data Public Private Forum

Big Data Story

Demystifying Big Data with Special Focus on and Examples from Industrial Sectors

BIG White Paper | 26 March 2014

Prof. Dr. Sonja Zillner, Sebnem Rusitschka, Michal Skubacz

Siemens AG
Corporate Technology
Research & Technology Center
Business Analytics & Monitoring

"Any organization thinking of simply applying existing information governance practices to Big Data will likely fail." – Gartner Inc.

Acknowledgement

This research has been supported in part by the Big Data Public Private Forum, a project that is co-funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 318062. The responsibility for this publication lies with the authors.

The writing of this whitepaper about would never have been possible without the great support of many of our colleagues. Therefore, we would like to thank the following persons for their valuable contributions, feedback, and input: Dieter Bogdoll, Hermann Friedrich, Thomas Hahn, Gerhard Kress, Vijay Kumar, Steffen Lamparter, Ulrich Löwen, Maximilian Viermetz, Thomas Peter Mahr, Michael May, Philipp Pott, Mikhail Roshchin, Ariane Sutor, Volkmar Sterzing, Ulli Waltinger, Stefan Hagen Weber.

SIEMENS

Executive Summary

What's the big deal about 'Big Data'?

Big Data – a paradigm shift that changes the way business is done?

Although 'Big Data' has become one of the most overused buzzwords of 2013, for many organizations the idea of massive data collection and usage remains a mystery. So what's so special about Big Data?

The common denominator of all discussions is that Big Data is made up of three V's: *Volume* points out the large amount of data that can be stored, *Velocity* indicates the rate at which this data is created and analyzed, and *Variety* refers to the heterogeneous data assets that require new forms of processing to enable enhanced decision making. On the business side, it is assumed that Big Data technologies can transform the way of how organizations do businesses by delivering a new kind of performance.

Why is the development of Big Data strategies so challenging?

Most organizations are unsure how to start or carry out Big Data initiatives. Managers are skeptical about substantial investments in Big Data because they often are not sure about the use of data they already have assembled, and thus, have difficulties to catch the inherent business opportunities. The reasons for skepticism are mainly triggered by practical concerns: *"How to do it?"* A three dimensional approach helps to overcome this skepticism:

Business: How to identify promising new business approaches? Big Data applications indicate a shift in the logic of how business is done. For instance, multi-sided business models that create value through the interaction of multiple stakeholders are replacing the traditional one-to-one transaction.

Data: Which data sources can be transformed into competitive data assets? Without analyzing the data, its value is often unknown. Only by systematically exploring existing and new data sources, a better understanding about its value can be developed, which again is needed to discover new business opportunities.

Technology: How to select the right bundle of technologies matching varying business needs? Big Data technologies evolve fast; the interplay of technology components is quite complex and currently lacks standardization. Therefore, IT governance processes need to be reviewed and strategically aligned to support building up of capabilities with new and disruptive technologies.

Investment decisions face the classic "hen-egg" problem

Due to its disruptive impact on the underlying businesses, Big Data cannot be an IT-driven initiative only. It requires much higher integration with business practices as well as a fundamental understanding about available data assets and their future potential. Focusing on IT procedures and processes only, bears the risk of building up Big Data capabilities that CANNOT be transformed into business value because the business cases are missing or not straight-forward or the access to the data assets cannot be provided.

However, without spending resources for the exploration of new data assets – which again requires some initial Big Data IT infrastructure - the value of data remains unknown. Without knowing the potential value of the data, business scenarios cannot be developed. Without understanding the business opportunity, it is difficult to identify the most suitable technology stack. So the challenging question is: *"Where and how to start?"*

Building up of Big Data capabilities: The key success factor

Successful Big Data players have demonstrated how to follow a strategic approach of incremental steps of Big Data capability building in three dimensions. They focus on either technology, or on data, or on business aspects at once when fostering up Big Data projects – and do not push through the valley of drought until all three capabilities are in synch. However, their developments of further projects build upon the capabilities gained through the first projects.

Implementing Big Data projects is hence a continuous learning spiral that helps the organization to *build up Big Data capabilities in three dimensions:*

- Business: by investigating the business potential of known and unknown user needs
- Data: by building up more and more data assets and gaining insight about their value
- Technology: by setting up a cost-efficient, flexible, and scalable technology stack

The building up of Big Data capabilities is a stepwise process with each Big Data initiative pushing at least one of the three dimensions to the next level. The projects then accumulate to Big Data portfolios in areas that were deemed strategic in the first place.

Big Data portfolio approach

Big Data project portfolios enable the appropriate resource allocation for the incremental building up of organizational competences in one of the three dimensions of data at a time, technology, data, and business, which are all needed to cope with the disruptive potential of Big Data. Besides the generation of new business scenarios (short-term impact), Big Data project portfolios facilitate the continuous improvement of the competitive position of an organization (long-term impact) in terms of data availability and technical capabilities.

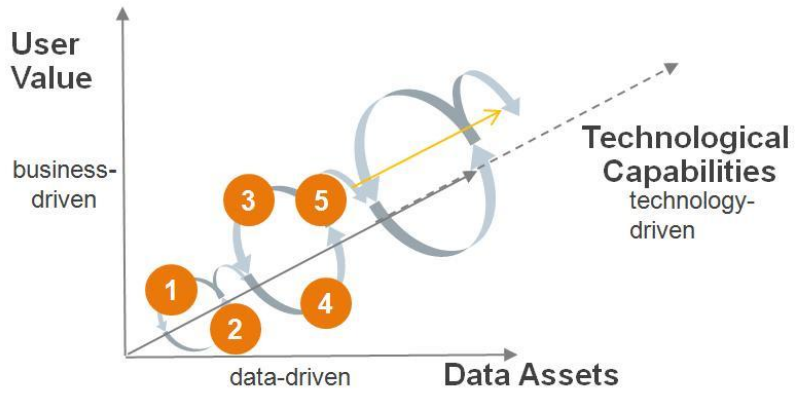


Figure 1 Big Data project portfolio

The inherent uncertainties of Big Data projects can be managed by gaining more insights about the given opportunities. Thus, organizations should focus on the strategy of building up Big Data capabilities. It is about starting small, learning by doing, exploring, trying multiple times and opportunistic adaptations.

The Business Perspective: What is the added value and the business approach?

Successful Big Data organizations do not stick with one business pattern or business model, but explore the various business opportunities. Often, best practice solutions from one industry are transferred to other industries.

Classification of Big Data business patterns and business models

Therefore, a classification of the various Big Data *business patterns* as well as business models helps to systematically explore future business opportunities:

- Four Big Data business patterns differentiate to which extent companies are bound by the digital versus the physical world: A) the *Virtual Business* pattern is based on the digital world only (e.g. Google Search), B) the *Digital-driven Business* patterns is a hybrid approach that uses the digital world to access the market or resources but relies on the physical world to implement the solution, whereas C) the *Data-enhanced Business* makes use of data (analytics) to improve their traditional business. In addition, we find a fourth business pattern that may be orthogonal or complementary to the other three business patterns: D) *IT infrastructure (SW&HW)* providers that support other companies with technology know-how to foster their Big Data applications.
- Four variations of business models can be distinguished: A) The *Optimization and Improvement* of existing businesses relies in general on available data sources, whereas B) the *Upgrading and Re-valuation* is based upon the integration of additional (often external) data sources. C) *Monetizing* describes the realization of new business opportunities that make use of available data sources, whereas D) *Breakthrough* business encompasses new ventures that rely on new data sources, which are often realized with new partners or even within new value networks.

Identifying business cases

Each Big Data business case should answer three questions: Who is paying for the solution? Who is driving the solution? Who is benefiting from the solution?

However, visionary - sometimes even disruptive - Big Data scenarios, such as the implementation of outcome-based healthcare delivery or synchronized wide area monitoring, protection, and control of power systems, rely on the effective interplay of multiple stakeholders in an ecosystem as well as on high upfront investments ensuring the digital data and technology availability. This often leads to the situation, that too many variables remain unknown for a single partner to calculate the business case. In order to overcome this situation, organizations need to focus on transitional/intermediate scenarios that help them to built up the Big Data capabilities that will be needed for implementing the visionary scenario but already allows them today to generate income with innovative business models.

Big Data business ecosystems

The impact of Big Data applications increases exponentially, the more data from various data sources can be integrated and analyzed. Therefore, the implementation of Big Data applications requires the collaboration of multiple – often competing – stakeholders on various levels: a) for sharing the data assets b) for building up a technology portfolio by strategic partnering with IT infrastructure providers and c) for establishing value networks generating new business. The successful governance of Big Data business ecosystems needs to reflect on the interests and strategies of all players involved.

Organizational implications

Big Data applications can only be realized if organizations have the right people with the right skills on board. In addition, organizational structures are needed, which foster the collaboration between employees from different domains as well as emphasize the importance of data governance within the organization.

The Data Perspective: How to build up competitive data assets?

The access to data is an important asset ensuring the future competitive position of organizations. However, the preprocessing and refinement of raw data is needed before it can be used within Big Data applications. And any pre-processing activity involves investments and costs. In addition, at the time when Big Data projects are started, it is often not yet known which data sources will be needed in pre-processed format. Therefore, the additional storage of raw data and successive building up of data assets should be seen as a long time process that is aligned to concrete business opportunities and technical capabilities. The strategy for continuously building up data assets will rely on dedicated processes that address the various data governance aspects.

Data Governance challenges differ according to the data types

Depending on the type of data source, classic data government tasks, such as data management, data quality, data privacy and security as well as the transparency of data usage are facing different challenges: In general, *own data* sources have the advantage that they are under the control of the organization, but often require continuous investments to ensure their technical availability. The usage of external data requires trustful partnerships or transparent licensing conditions, whereas the access to crowd and community data relies on the existence of an active community that is producing and governing data at large scales.

Technical data access strategies

The technical availability of data needs to be ensured, i.e. data needs to be digitalized and made available through communication, as well as interfaces, enabling a seamless data exchange. For decision making in cyber-physical systems, data quality is important to keep in mind as it can drive or hinder the envisioned business opportunity. However, through model- and data-driven analytics even dirty or missing data challenges can be addressed.

From a data sharing policy to data ecosystems

The concept of data ownership influences how and by whom the data can be used. In particular, it is important to distinguish between private, personal and non-personal data, operational data and longitudinal data. In addition, many trends in consumer industries, such as Open Data, Data Marketplaces, or Linked Data, have the potential to trigger a paradigm shift towards an open and sharing economy also within the B2B industries, so-called Data Ecosystems.

Technology Perspective: How to thrive Big Data technological capabilities?

Current big data technologies need to be adapted and further developed to give industrial businesses the cutting edge advantage in the digital transformation of their businesses. It is not just about technological capabilities – but how to grow them with existing business and sometimes even data opportunities.

How to choose your tools: Not everything is nails, and there is more than a hammer

There is no single tool, or choice of a platform that will remedy the challenges of Big Data business. Looking at *architectural patterns* will assist in making the right bundling of tools. However, there is one common paradigm to all the different big data architectural patterns: Moving algorithms closer to where the data is.

Schema on Read and the Active Archive

Active archive, sometimes also called Data Lake, allows the cost-efficient storage of all data in its raw form and transforming it as needed by the business users whenever required. Distributed data management and massively parallel processing principle combined with performant commodity hardware makes the generation of a data schema on read feasible. An active archive allows one to get to know the data and explore business opportunities.

Lambda Architecture

The lambda architecture is based on a set of principles for architecting scalable Big Data systems that allow running ad-hoc queries on large datasets. It consists of multiple layers dealing with volume and variety (via an active archive) as well as velocity (via stream data computing).

In-field Analytics

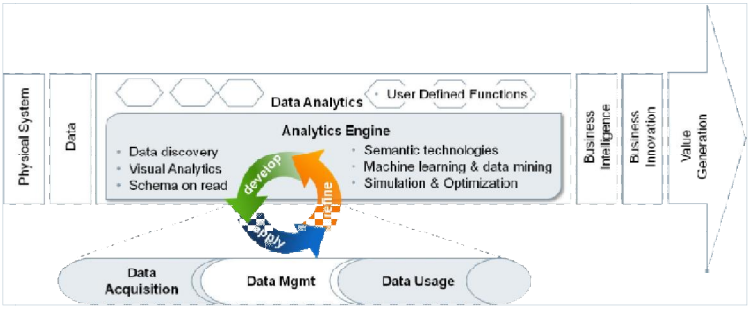
The lambda architecture is described from an enterprise perspective. In cyber-physical systems, such as industrial automation or energy automation, field devices represent important data acquisition as well as computing resources. In-field analytics extends the lambda architecture by encompassing the data acquisition layer, in order to deliver faster insights for industrial businesses.

The domain specifics of handling Big Data in cyber-physical systems require the industrial stakeholders to look beyond the state of the art of architectural patterns, e.g. *In-field analytics*. Nevertheless, how Big Data natives have arrived at these architectural patterns reveals some clues as to how the industrial businesses with vertical IT providers could mimic the building up of technological capabilities without experiencing major disruptions by the technological advances.

"Analytics inside:" With Big Data, data management alone does not suffice

Industrial data sources have considerable variety: Not only are all data types that are also relevant for online businesses becoming increasingly available, but also there are huge streams of high-resolution data from sensors and intelligent electronic devices embedded into the cyber-physical systems. A Big Data refinery pipeline is needed that reaches from data acquisition to data management and data usage.

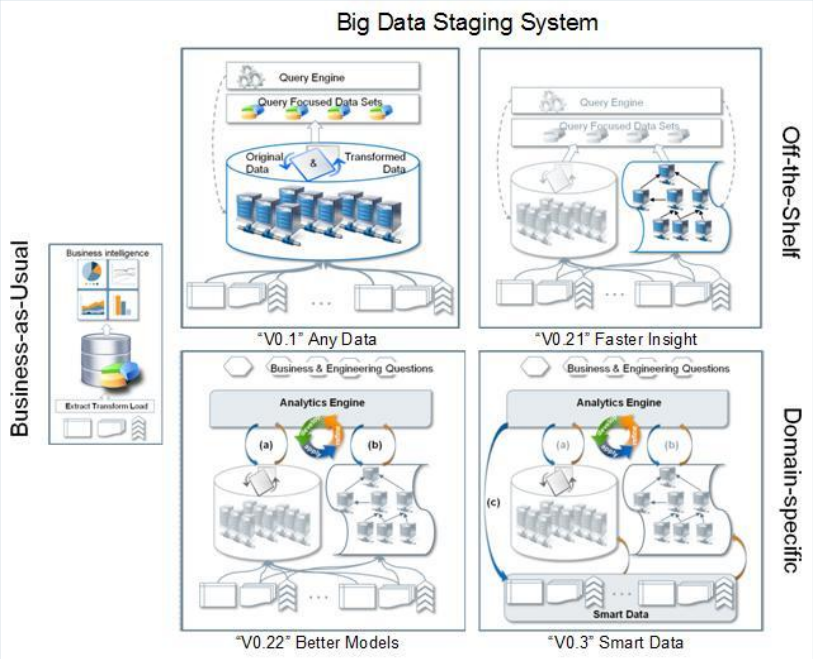
Each step of the pipeline refines the data, and the methods of refinement vary depending on the data type. Although data usage is the last step in the refinement process, the business and engineering questions formulated in this step are the starting point for choosing the technology stack. At the same time the type of the required data sources determines what type of refinement methods will be cost-efficient. Potentially millions of insights per second await industrial businesses, when the breadth and variety of data sources can be refined and used for fast or even automated decision making. Such a capability requires the seamless integration of analytics into each step of the Big Data refinement pipeline.



The Analytics Engine for integrating advanced analytics into the data refinery pipeline is designed such that both flexibility and extendibility are still feasible: Through data discovery, visual analytics, machine learning, information retrieval, and data mining, the incremental understanding of the data becomes possible. Once the appropriate data and analytical models are developed that portray this understanding, schema on read can be utilized to apply the improved models onto the data refinery pipeline. This Analytics Engine will assist in implementing the domain- and device-specific adaptations to Big Data management in a cost-efficient and innovative manner.

"Agile is too slow." – Always Improve Your System

Especially big companies need to be aware that even "agile is too slow" for Big Data and that constant updating becomes routine. However, as the re-engineering or replacement of legacy systems in many industries is not possible, one will need to create a parallel staging system for Big Data technologies. A parallel staging system enables the assessment and adaptation of innovative technologies within all steps of the data refinery pipeline.



The setup of a parallel Big Data staging system in which technological advancements, data opportunities, existing and new business questions can be examined is a living iterative approach. The different architectural versions can be dubbed *"Any Data," "Better Models and Faster Insight,"* and finally *"Smart Data."* Any Data and Faster Insights are powered by off-the-shelf solutions for massively parallel processing and stream data computing. The development of Better Models and extracting relevant data by increasing data interpretability through these models is domain-specific. Especially, in industrial applications generating Smart Data by applying in-field analytics based on domain and device know-how will enable real-time insights for decision automation. At the end, the competitive advantage will lie in acquiring these capabilities in a cost-efficient and innovative manner.

Processes and IT governance practices need to be reviewed to enable the realization of such a Big Data parallel staging system that can flexibly substitute technology components based on their efficiency to translate data opportunities into business value.

Contents

No table of contents entries found.

1. Big Data: A Revolution on Various Levels

"'Big Data' is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." – Gartner Inc.

A revolution of technologies triggers a revolution of business

1.1. Definition of Big Data: 3V's with Many Different Extensions

The 'Big Data' concept indicates a shift in technology that changes the way how business is done. With the availability of scalable and affordable technologies for the storage and processing of large data sets, new business scenarios become possible.

3 V's are common denominator for defining Big Data.

Although "Big Data" is one of the most popular buzzwords of 2013, its meaning and impact on future businesses often remains unclear. If one aims to define the concept, the so-called "3Vs" – Volume, Velocity and Variety - of Gartner's Big Data definition is a good starting point, as they had been one of the first definitions for Big Data and is now seen as the common denominator within all major industries:

- *Volume* points out the large amount of data that can be stored,
- *Velocity* indicates the rate at which this data is created and analyzed, and
- *Variety* refers to the heterogeneous data assets that require new forms of processing to enable enhanced decision making and insight discovery

From a technical perspective, "Big Data" encompasses any techniques establishing the basis for handling potentially large and complex datasets in reasonable time, including the capturing, processing, storing, analyzing and visualization of data (NESSI, 2012).

The various industries add further V's to reflect their specific business needs.

From the business side, one expects that Big Data technologies will transform the way of how organizations do businesses by delivering a new kind of performance. In this way, the technical advances described by the 3V's will trigger new business opportunities that again will evoke new technical requirements. Those expectations can be observed within the daily discussions and publications on internet, press or media where the definition of Big Data is becoming a moving target[1] that goes along with many different flavors incorporating various technical as well as business related aspects. The various industrial sectors that are already reviewing themselves in the light of the Big Data era, add further V's to reflect their specific data processing challenges and to adapt the Big Data paradigm to their particular needs. Many of those extensions, such as data privacy, data quality, data confidentially, etc. address the challenge of data governance, some of the added extensions, such as business value, even address the fact that the potential business value of Big Data applications is yet unexplored and not well understood.

The used definitions indicate a paradigm shift towards recognizing data as important productivity factor.

In sum one can say, that the way how the concept of Big Data is defined, described and discussed indicates that the Big Data concept is no longer perceived as a merely technical trend but seen as paradigm shift towards recognizing "data" as competitive asset. For instance, recent studies, such as (BITKOM, 2012), emphasize the value of data by comparing its relevance for economic growth with the already established productivity factors, such as capital, labor force and resources.

[1] http://www.opentracker.net/article/definitions-big-data

1.2. Origin of the Big Data Concept

"Big Data is what happened when the cost of keeping information became less than the cost of throwing it away." - George Dyson

Scalable technologies for data storage are now available.

About ten years ago, the availability of scalable data storage solution was still a big major technical issue for reusing available data. As of today, scalable technologies enabling the efficient management, storage and analysis of large data set are available as well as affordable.

At the same time, in nearly any industry data volumes are significantly growing, in some domains even exploding. The reasons for this data explosion are well-known: with the advent of internet based services, intelligent products, Internet of Things, sensors, RFID, etc., more and more data is produced, collected and stored 24h and seven days in the week

Large Internet players have been the first movers that transformed the growing data assets into promising innovation opportunities.

Large internet player, such as Google, Amazon, Facebook and Twitter, had been the first movers that managed to handle the growing data assets and transformed them into promising innovation opportunities with continuously growing business value. By establishing technologies and algorithms that were capable to address the data volume challenges at internet scale, first ad-hoc solutions had been available.

Open source community pushes the technical progression of Big Data.

In the meanwhile, several of the underlying technical solutions have been migrated to open source community. The involvement of the open source community helped to advance the technological capabilities in a very fast manner as well as transfer them to other industries and domains. A vivid and dynamic ecosystem of IT infrastructure providers, developers, users, etc emerged. This was the starting point of the Big Data trend.

1.3. The Big Data Hype is Driving Increased Investments

"64% of enterprises surveyed indicate that they're deploying or planning Big Data projects. Yet even more acknowledge that they still don't know what to do with Big Data." – Gartner Inc.

Big Data hype continues to drive increased investments.

The Big Data trend is generating plenty of hype with the drivers of the hype being the new and promising technological opportunities. One assumes that Big Data can transform the way how companies do business by delivering a new kind of performance. And this assumption is enforced by a large number of success stories and studies. For instance, (McAfee and Brynjolfsson, 2012) showed within their research that companies using Big Data could demonstrate productivity and profitability rates being 5% to 6% higher than their competitors. In addition to those success stories from Big Data innovators and early adopters, more and more companies have started to invest in Big Data technologies. A study from Gartner (Gartner, 2013) documented concrete numbers indicating that the Big Data adoption rate in 2013 shows reliable substance behind the hype.

Most organizations are unsure how to precede with Big Data initiatives.

However, most organizations are unsure how to start or precede with Big Data initiatives. (Barton and Court, 2012) found out that a large number of managers are skeptical to vote for substantial investments in Big Data because they are convinced that their organizations are simply not yet ready. In addition, organizations do often not fully understand the data they already have, or they might have already spent large amounts of money for improved IT systems that now do not mesh with the business processes and needs.

Investments in technology are not enough.

In addition, often the main investments are limited to the technical infrastructure, which bears the danger that companies lack the organizational and personal capabilities to make use of the Big Data potential. Parallel to the technical infrastructure, organizations need to build up competences in data governance and data analytics that allows them to investigate and discover the new business opportunities. As any data governance and data analytics tasks have a strong overlap with the established business processes, those competences should always be developed and kept inside the company.

1.4. Big Data Entails Disruptive Potential

Big Data is not only a technical revolution, it also revolutionize the way how business is done.

A detailed analysis (Zillner, 2013) about Big Data related business opportunities across the various different sectors, such that Manufacturing, Healthcare, Transport, Finance, Energy, Media, and Public, indicated that Big Data applications are not only based on revolutionary new technology (revolution in technology) but also significantly transform the way how business is made (revolution in business):

Technology Revolution: By combining multiple technological advances of various disciplines, the status quo of data processing and management could be revolutionized. *D*ecentralized networking and distributed computing for scalable data storage and scalable data analytics, semantic technologies and ontologies, machine learning, natural language processing and other data mining techniques have been the focus of research projects for many years. Now these techniques are being combined and extended to address the technical challenges of the so-called Big Data paradigm.

Business Revolution: Big Data fosters a new dimension of business opportunities: It is a promising method for businesses to target and bind their customers, analyze their preferences and to develop the most effective marketing and pricing strategy, to improve the efficiency of the processes and workflows or to develop new data-based products and services. The range of new business opportunities is wide and manifold. The new technical opportunities have a revolutionary – sometimes even disruptive – impact on the existing industrial business-as-usual practices. The consequences are obvious: New players emerge that are better suited to offer service based on mass data. Underlying business processes change fundamentally. For instance in the healthcare domain, Big Data technologies can be used to produce new insight about the effectiveness of treatments and this knowledge can be used to increase quality of care. However, in order to foster such Big Data applications, the industry requires new reimbursement models that reward quality instead of quantity of treatments. Similar changes are required in the energy industry: energy usage data from end users would have to share beyond organizational boundaries of all stakeholders such as energy retailers, distribution network operators, and new players such as demand response providers and aggregators, energy efficiency service providers. But who is to invest in the technologies that would harvest the energy data in the first place. New participatory business value networks are required instead of static value chains.

2. Strategic Big Data Management

2.1. Three Challenges: Business, Data and Technology

Its transformative potential makes Big Data such a powerful concept. Big Data can be used to provide new insights that help deliver advanced products, intelligence-based user experiences, improved efficiency, etc. in a way that has never been possible before. In order to realize such opportunities, a number of challenges need to be addressed: technological complexity, data availability and governance as well as the alignment of business and technology. But the successful tackling of those challenges needs to continuously align technology, data and business aspects.

Three aspects make the implementation of Big Data applications very challenging: a) missing business cases, b) high investments needed for building up data assets and c) technical complexity.

The investigation of Big Data opportunities of various industrial sectors (Zillner, 2013) showed that not the availability of technology but the lack of business cases as well as the high effort needed for processing raw data is hindering the implementation of Big Data applications:

Business: How to identify promising business approaches?

Without knowing the value of the data sources, a business case is difficult to find.

In general, new ventures require a convincing business case to receive funding. However, at the beginning of Big Data projects, *business cases are often unclear and difficult to find*. Big Data business cases rely on the value of data and the potential insights that can be discovered, In other words, without knowing the value of the data, its business potential remains unclear. If one aims to explore the business opportunities of Big Data, one needs to gather beforehand large volumes of data.

Moreover, one needs to keep in mind that Big Data applications indicate a shift in the logic how business is done. For instance, multi-sided business models that create value through the interaction of multiple stakeholders are replacing the traditional one-on-one transaction. In addition, the impact of Big Data applications often relies on the aggregation of not only one but a large variety of heterogeneous data sources beyond organizational boundaries.

For the development of new business approaches, one also needs to consider the dynamics of the underlying ecosystem. In particular, the stakeholders' individual interests and constraints – which are quite often moving targets – have to be analyzed carefully in order to establish the basis for the effective cooperation of *multiple stakeholders* with potentially diverging or at first orthogonal interests.

Data: Which data sources can be transformed into competitive data assets?

Big Data applications need data, and the processing of raw data needs investments.

Without analyzing the data, its quality is often unknown. The *primary value* from Big Data does not come from the raw data, but *from the data processing and analysis of it* (Davenport, 2013). In order to generate new insights, provide guidance by decisions, or develop new products and services that emerge from Big Data analytics, one need to enable the seamless access and sharing of various heterogeneous and distributed data sources. In many of the traditional industries, such as the healthcare sector, the access to the needed data is only possible in a very constrained manner. Often, large amounts of data are only available in unstructured or non-standardized formats which significantly constraints the automatic analytics of data. In addition, legal frameworks ensuring the transparent sharing and exchange of data do not only need to be in place but also be implemented within the organizations. In other words, only by systematically exploring new data sources, a better understanding about its value can be developed which again is needed to discover new business opportunities. For being able to discover and implement promising Big Data applications, advanced payments for establishing appropriate data storage and processing capabilities are needed.

Technology: How to select the right bundle of technologies?

Big Data technological advancements open new business opportunities in much shorter cycles.

To keep up with the pace, organizations will need to build an environment for gathering know-how and skills. Big Data technologies evolve fast, i.e. in the last decade, approximately every three years, a new frontier of technological capability regarding innovative and cost-effective information processing has been pushed: Distributed file systems and mas-

sively parallel processing, followed by scalable, highly available key-value stores and NoSQL databases, culminating in stream processing of real-time data, visualization and ever advancing analytics, such as predictive and prescriptive analytics on massive amounts of data. During this time of fast technological advancements in the Big Data arena, it is likely that the right choice today is out-dated in a few years. Moreover, the interplay of technology components is quite complex and currently lacks standardization.

Additionally, industrial stakeholders, as opposed to Big Data natives, might be forced to use Big Data technologies in ways they were not designed to – and eventually come to a point at which adaptations to the domain specifics are required to be cost-efficient. At that point, it will be decisive to have gathered sufficient know-how and skills. IT governance processes need to be reviewed and strategically aligned to support building up of capabilities with new and disruptive technologies.

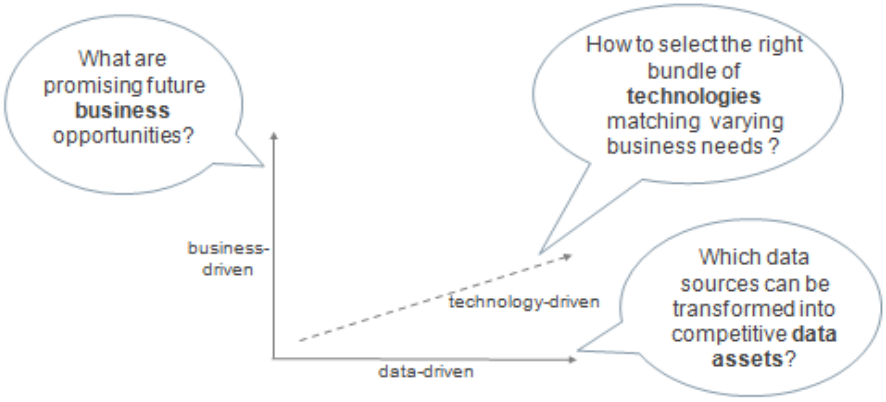


Figure 2 Three challenges in Big Data Management

Investment decisions face the classic "hen-egg" problem; a transparent Big Data capability building strategy can help to overcome this situation.

Business cases are often not yet known in advance, but are discovered while working with the data, technology and customer. Due to its disruptive impact on the underlying businesses, Big Data cannot be a pure IT department initiative only. It requires much higher integration with business practices as well as a clear understanding about available data assets and their future potential. Focusing on IT procedures and processes only, bears the danger of building up Big Data Technology capabilities that cannot be transformed into business value because the business cases are missing or the access to the data assets can't be provided.

However, without spending resources for the exploration of new data sets – which again requires some initial Big Data IT infrastructure - the value of data remains unknown. Without knowing the potential value of the data, business scenarios cannot be developed. Without understanding the business opportunity, it is difficult to identify the most suitable technology stack. So the challenging questing is: Where and how to start?

Implementing Big Data projects is hence a continuous learning loop and helps the organization to *build up Big Data capabilities* in three dimensions:

- Business: by investigating the business potential of known and unknown user needs
- Data: by building up more and more data assets and gaining insight about their value
- Technology: by setting up a cost-efficient, flexible, and scalable technology stack

It is about starting small, learning by doing, experimenting, trying multiple times and flexible adoptions. For any Big Data technology project that gets started, it should be clear in which of the three dimensions – business, data, or technology - one aims to build up capabilities.

2.2. Incremental Building up of Big Data Capabilities as Best Practice Solution

Successful Big Data player have demonstrated to follow a strategic approach of incremental step, each focusing alternating on one -- either technology, data or business -- aspects, to foster their Big Data businesses.

Companies who have successfully implemented Big Data capabilities as well as business offerings show a very typical pattern of how they start and continue initiatives: they continuously switch between the *technical*, the *data* as well as the *business* focus in order to built up their Big Data capabilities and value chain.

Google, for example, which began as research projects started with a technology focus to set up the search engine. After some years, the Google founders allowed simple additions and thus, addressed for the *first time the business aspect*. The main focus was still on usability in order to attract a growing number of internet users. The growing number of users was of importance due to two reasons: First, the user data could be used to develop value-added services by means of analyzing the user behavior. And second, with the increasing number of users, the impact of advertising business improved. Beside the ongoing focus and investments in improving the scale, performance and usability of their technology, Google continuously made investments for building up their data assets: For instance, in 2003 Google acquired Pyra Lays, a pioneering and leading web log hosting website. This acquisition secured the company's competitive capability to use information aggregated from blog postings in order to improve the speed as well as relevance of articles that were contained in their companion product Google News. Other examples are the Google Books library project that aims to build up a searchable collection of several major libraries by scanning and pre-processing of each single pace of the original books. Or investments of Google in the still venture-founded start-up company "23 and me" [2] that provides rapid and very low-price genetic testing. Due to the very low price of the genetic testing, it is not likely that the business model relies on a service with costs. Rather, it is very likely that "23 and me" aims to mainly build up a promising data asset consisting of a large collection of genetic data which can be used for developing personalized healthcare products.

Another example is an major US Airline that could significantly improve its ETA (Estimated time of Arrival) in several steps. First, publicly available data sources about weather, flight schedules and other factors were combined with proprietary data of the airline company, such as data from radar stations. In a second step, the airline could improve its forecast by comparing the data about an airplane that was about to land with all previous times planes approached the airport under similar conditions. Source: (McAfee and Brynjolfsson, 2012)

2.3. Big Data Project Portfolio Approach

Big Data innovation project portfolios enable the appropriate resource allocation for the incremental building up of disruptive competences in the data, technology and business domain.

Beside the generation of improved or new business scenarios (short-term impact), Big Data projects also allows to continuously improve the competitive position of an organization (long-term impact) in terms of data availability and technical competence. Therefore, the decision about resource allocation for Big Data projects should be aligned with the overall strategy of the company. The reasons for starting a Big Data project might be quite different from project to project: One might primary focus on the generation of more income (business-driven projects), or on improving its access to data sources by pre-processing internal data or by partnering with external parties (data-driven projects). Alternatively, one might priories to advance the own technology competences (technology-driven projects). Although all three aspects need to be addressed to some extent by each Big Data project, the commitment as well as expectation regarding each of the three aspects will differ.

[2] In December 2013, 23andMe suspended health-related genetic tests for customers in order to comply with the FDA warning letter.

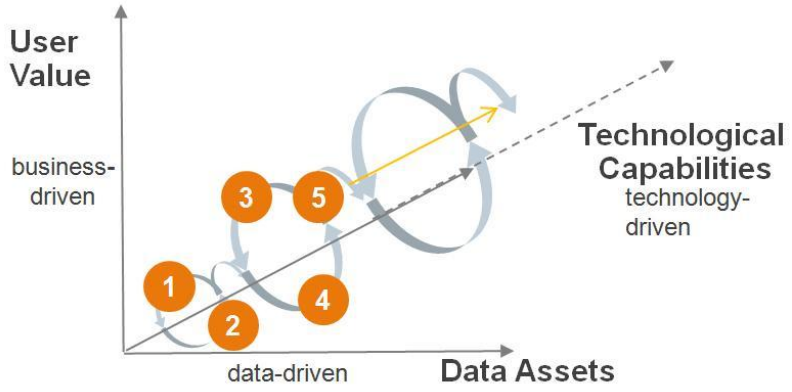


Figure 3 Big Data innovation project portfolio

By establishing a Big Data innovation project portfolio (Zillner and Krusche, 2012), it becomes possible to specify the expectation regarding each Big Data project in a transparent manner: Is the generation of income the project's main focus? Or do we mainly focus on improving the access to (additional / new) data sources? Or is our primary objective to build up more technological competences?

It is recommended to start with simple projects and iteratively get more and more complex in terms of integrating more data sources, using more advanced technology components, enhancing the partner network and generating more user value and innovative business models. This can be compared to spiral movements that allows you to continuously build up strategic competences in the three areas data, technology and business. In addition, a innovation project portfolio approach will help you to balance short-term interest with long-term investments.

3. Business Perspective: Revolution of Big Data Economy

3.1. Big Data Business

"The world is filled with companies that are marvelously innovative from a technical point of view, but completely unable to innovate on a business model." (Clayton Christensen)

Successful Big Data organizations do not get stick with one business pattern or business model, but explore the various business opportunities. Often, best practice solutions from one industry are transformed to other industries. Therefore, a classification of the various Big Data business patterns as well as business models helps to systematically explore future business opportunities:

3.1.1 Big Data Business Patterns

Big Data Business Patterns define the strategic fit between the (newly built) Big Data capabilities and the core competences of the organization.

Big Data Business Patterns answer the question how (existing or newly build) Big Data capabilities (will) relate to the core competency of the organization. By indicating the degree to which companies are bound to the digital versus physical world, Big Data Business Patterns can be used to describe how Big Data is used to generate new business value and thus allows to enhance the relatively strength of an organization relative to others.

We distinguish three main business patterns to describe how organizations are using Big Data to make business: 1) Virtual Business 2) Digital-driven Business / Services and 3) Data-enhanced Business. In addition, we define a fourth business pattern: 4) IT Infrastructure (SW & Hardware) which is orthogonal to the first three mentioned patterns as it supports companies with knowhow in technology to foster their Big Data business (patterns).

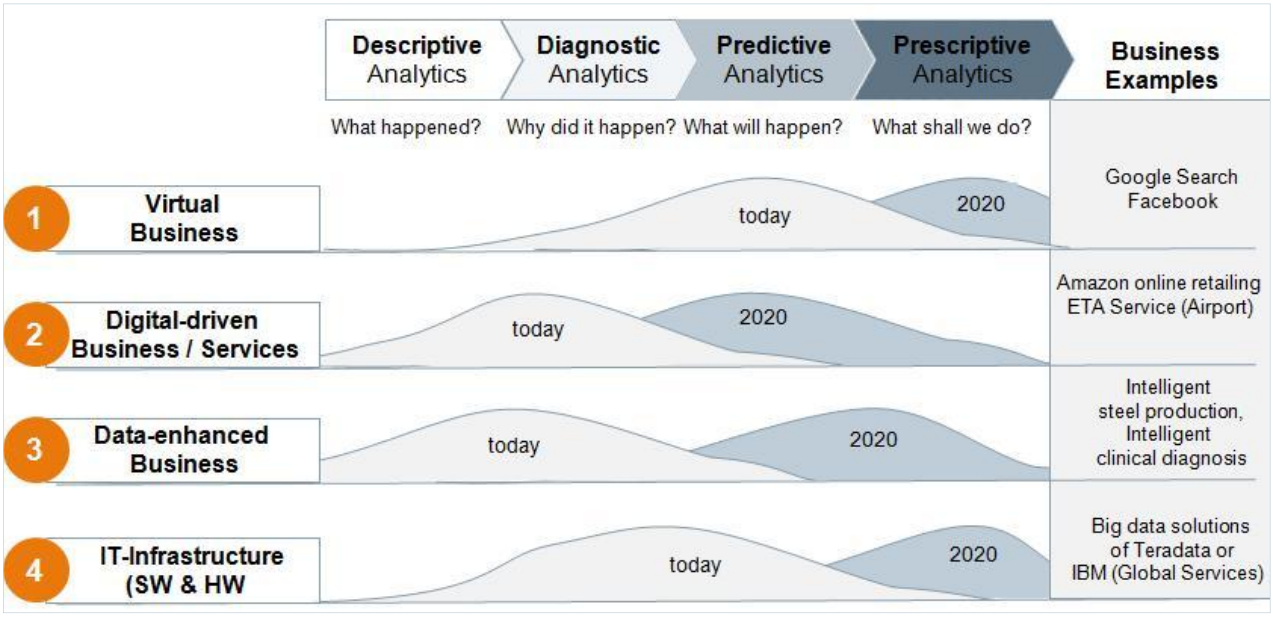


Figure 4 Different Big Data business patterns in the context of Big Data (expert estimation)

1. Virtual Business

- *What:* Virtual Business provider collect, integrate and aggregate various sources of data in order to provide data products in the virtual world (internet, web, etc.).
- *How*: Often this business pattern is combined with dedicated investments in community building in order to establish a foundation for data creation.
- *Example*: Google provides search functionality for millions of user by collecting and aggregating data about internet sources, user behavior, etc.

- *Value Creation:* Virtual businesses often rely on multi-sided business models that create value through the interaction of multiple players rather than the traditional one-on-one transaction. Thus, often not the party who is receiving the value (i.e. single user who is finding the appropriate content) is paying for the solution, but a third party that, for instance, benefits from the attention of the user.
- *Technical complexity*: As of today, the majority of virtual business provider make use of descriptive and diagnostics analytics. In the future, the full spectrum of available data analytics (up to complex and prescriptive analytics) will be used to ensure reliable services. For instance, prescriptive insight might be used to automatically trigger actions for balancing critical user loads.

2. Digital-driven Business / Services

- *What*: Digital-driven business or service provider use the digital world as platform to reach a large, often global, market that allows them to offer services in the real world with an promising economic of scale. Their value added is of digital nature which is tightly connected with a physical value chain.
- *How:* Similar to virtual businesses, digital businesses need to establish a large community of users in order to a) increase the market scale as well as b) to collect comprehensive data sets needed for realizing enhanced customer experience and personalized services with added value.
- *Examples*:
  i. By collecting sensor data of single or several gas turbines, third party providers are able to offer *predictive maintenances services*. Gas turbine operators are interested in such services as they help them to prevent future damages or increase the overall output /performance.
  ii. Amazon uses the Internet as platform to achieve maximum leverage out of its fixed assets, such as its numerous logistics centers.
- *Value Creation*: Digital-driven business or services generate an added value that helps to significantly improve the original business / service and thus strengthen the customer binding, establish an USP of the business / service or make the market more attractive to a larger customer group.
- *Technical complexity*: As of today, the majority of digital-driven business / service providers make use of descriptive analytics to improve their real-world processes. In order to establish the basis for predicting critical events, digital business providers will adopt more and predictive analytics in the next years.

3. Data-enhanced Business:

- *What*: Data-enhanced businesses make use of data insights in order to improve the performance of traditional businesses and services. The physical processes and business is enriched with intelligence guidance / behavior.
- *How*: By combining data insights and IT processes with physical resources in new ways allows generating additional value and revenue. Data-enhanced businesses aim to improve and transform processes, business models and the customer experience by exploiting the optimal interplay between systems, people, places, or things.
- *Example*:
  i. Advanced analytics is used to extract meaning from medical images in order to support clinicians within their diagnosis process.
  ii. By collecting and aggregating the data produced by existing automation and IT system of a steel plant in an integrated manner, it becomes possible to achieve the optimal steel production system performance.
- *Value creation*: Data insights are used to objectively measure financial and operational results and to make improved decisions which helps to increase the efficiency and quality of established processes and products as well as to generate new offerings.
- *Technical complexity*: Usually, data-enhanced scenarios need to tackle first the challenge of data collection, processing and aggregation of new internal as well as external data sources by using descriptive or diagnostic algorithm. As of today, many traditional businesses lack internal expertise in data governance, which often leads to a slowed down implementation progress of Big Data solution. As soon as companies have built up sufficient data assets, advanced data algorithm can be used to even calculate predictive insights.

4. IT-Infrastructure (Software & Hardware)

- *What*: The number of organization offering Big Data IT-infrastructure, such as Teradata or IBM (Global Services), or Big Data technology consulting that help other organization to build up their Big Data infrastructure and

needed technology competences covering the complete portfolio of Big Data technologies. Or consulting companies, such as Accenture, that support other companies in integration the various IT infrastructures.

- *How:* Domain knowhow within various industries help IT-infrastructure providers to generate added value and customized offering for their- in general business - customers. The typical customers are companies that aim to transform their traditional business into data-enhanced or digital-driven business.
- *Value creation*: Domain knowhow in combination with IT competences helps to speed up the implementation of data-driven businesses.
- *Technical complexity*: The technical complexity provided by IT-infrastructure providers evolves with the business needs of their customer. As of today, mainly diagnostics and predictive algorithms are provided. This will change in some years, when more prescriptive algorithms will be requested by their customers.

3.1.2 Big Data Business Approaches

Business model decisions should not be driven by economic calculations only, but consider the value opportunities for building up data asset and technology capability.

The role of business models is to capture value from early stage technologies, such as Big Data applications. By implementing a business model, one might unlock the latent value from a technology and the access to a particular data set. This again can establish the basis for new or alternative business models later on.

As the data economy relies on the efficient interplay of technology, data availability, partners and suitable business models, the development of technology roadmaps needs to be extended by three aspects: a) by a dedicated data strategy b) a partnering strategy as well as c) a sequences / roadmap of business models that help to push the own competitive position by building up accessible data pools (the new currency for partnering) as well as technology competences.

Within the data economy one can distinguish four different types of business approaches

We find various approaches to generate business value with Big Data technologies: One can generate business value by using existing data sources or by integrating further – sometimes even new - data sources. Those offerings can be realized by a single organization or within an ecosystem of partners. In addition, the added-value might help to improve existing products and services within an established market or even be used to generate new businesses and sometimes even new markets.

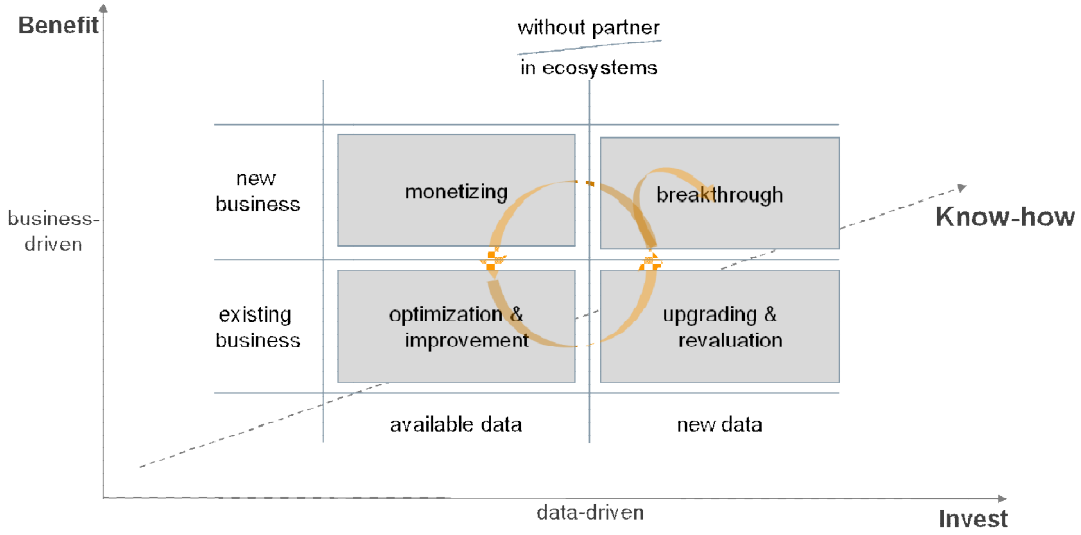


Figure 5 Four variants of business models in the data economy (inspired from (BITKOM, 2013))

1. Optimization and Improvements

- *What*: Available data sources are used to optimize and improve existing businesses.
- *Examples*

    - Healthcare domain: Administrative and financial data in hospital settings are analyzed to increase the efficiency of the underlying administrative processes, such as scheduling tasks or the utilization of resources.
    - Energy domain: Sensor data of gas turbines are analyzed to predict future damage as well as identify the cause of deviations in the process, in the material, etc. (see " Example: Gas Turbine" in Appendix 1).
- *Value Creation*: Optimization and improvement of existing processes and businesses helps to reduce costs or to improve performance.
- *Technical Complexity*: In general, technologies for analyzing business processes and product performance are available and only need to be adapted to the particular context and requirements.
- *Data Complexity*: In general, data assets are available but their technical access needs to be ensured. Depending on the type of data source, i.e. sensor data, private data, etc., the respective data governance challenges have to be addressed.
- *Business Complexity*: Businesses aiming towards the optimization and improvement of established processes are in general a good starting point for data-enhanced offerings. By investigating available data sets, new insights regarding improvement potentials can be discovered while working with the data.

2. Upgrading and Revaluation

- *What:* By employing new data sources either by transforming internal raw data sources into processable format (e.g. by semantic labeling of the content of medical images) or by integrating external data sources (e.g. weather forecast information) new offerings are developed.
- *Examples*:
    - Energy services domain: Data- and knowledge-based services enable the discovery of new insights about trends that help to increase the overall business performance. Analytics applications are applied to all kinds of data (e.g. product data, market data, competitor data, web data, customer data, financial data, etc.) to detect and respond to product, event, personnel, competitor, customer, and market trends (see "Example: Unified Service Intelligence" in Appendix 1).
    - Healthcare domain: The radiologist workflow can be improved significantly by establishing means for the seamless navigation between medical image and dictated radiology report data. Content information of medical images as well as dictated medical reports gets semantically described and linked by metadata. (see "Example: Seamless Navigation between Medical Image and Radiology Report" in Appendix 1).
    - Industry Automation: The data of LHC (Large Hardron Collider) automation and control components and systems from Siemens (WINCC OA) are collected (offline and online) for automated system health check and diagnostic in order to prevent future damage which helps to significantly reduce overall maintenance cost (see "Example: Alarm Management and Event Monitoring (Data Analytics Framework)" in Appendix 1).
    - Global Production Chain: The intelligent integration of global supply chain management information (e.g. via object tracking information) into the production planning processes allows to increase robustness and efficiency of the global value chain (see "Example: Intelligent Integration of Supply Chain Management into Production Planning" in Appendix 1).
    - Smart Grid Systems: Optimized energy production through the interactive planning and optimizing of the top-level design of MicroGrids in collaboration with the user. The interaction relies on visual result analysis that enable the user to detect patterns in large and heterogeneous data sets, such as weather data, power demand data, time series, energy capacity data, etc. (see "Example: Visual Query Analysis for Designing Efficient Smart Grid Systems" in Appendix 1)
- *Value Creation*: Here the underlying idea is to upgrade existing business processes as well as services by making use of additional data sources. By aggregating multiple data sources, insights about process performances, operational and financial measures, as well as guidance for business decisions can be provided.
- *Data Complexity*: The integration of new data sources implies efforts and investments for handling the various data governance challenges. Beside the challenge of accessing external data sources, one might face the challenge of overcoming internal data silos (sometimes even organizational silos) as well as the challenge of preprocessing raw data that is only available in unstructured format, such as images, videos, dictated text, etc.
- *Business Complexity*: The described data governance challenges might imply high investments such that a long-term business strategy is needed.

3. Monetizing

- *What*: By exploiting available data sources, complete new business scenarios, offerings and value streams are realized.
- *Examples*
  - Clinical Research: Patient cohorts for clinical programs can be identified more easily by using information extraction and advanced analytics on top of clinical data. In consequence, the feasibility of the clinical trials and, thus, the planning and designing of clinical programs of the pharmaceutical domain can be significantly improved. (see " Example: Clinical Research" in Appendix 1)
  - Information Service: BLIDS[3] is a lightening information service, which offers energy provider, industry, insurances, or event organizer precise information about registered lights. The service aggregates weather data from more than 8 countries and more than 145 measuring stations in Europe as well as enables the user-adapted representation of content.
  - Smart Energy Profiles: Increasingly, metering service providers, which service renewable decentralized energy resources and new types of demand, such as electric vehicles, can bundle the characteristic information of power feed- in and energy usage into smart energy profiles and sell these for profit. Currently the whole energy market operates with standard load profiles, which are inefficient.
  - Energy Automation: Intelligent electronic devices deliver real-time high-resolution data on power network parameters. When transmission network operators install these data sources near bigger renewable energy resources, and utilized advanced analytics, they can resell the information gained on the characteristics of the wind park and the networks area back to the operator of the park for operational efficiency increase on both sides.
- *Value Creation*: Secondary usage of data, i.e. the user benefiting from the collected data is outside the original context in which the data were produced and collected.
- *Data Complexity*: Important here is to clarify whether the usage of the available data sources for other purposes is legally allowed (more details on this in Section Data Sharing Policies)

4. Breakthrough

- *What*: Big Data applications can lead to *breakthrough* scenarios that rely on collaborative ecosystems establishing new value networks by aggregating existing with complete new data sources of various stakeholders.
- *Examples*
  - Healthcare*:* Public Health Analytics application rely on the comprehensive disease management of chronic (e.g. diabetes, congestive heart failure) or severe (e.g. cancer) diseases that allow to aggregate and analyze treatment and outcome data which again can be used to reduce complications, slow diseases' progression, as well as improve outcome, etc.
  - Energy-efficiency: Efficient energy use is highly dependent on energy automation – down to the device-level in a private or commercial user. Learning systems, which adapt to the preferences or business criteria of the energy user, along with the efficient data exchange between retailers, energy market, as well as the network operator and the actual devices with the energy-efficiency service providers, are required. Finally, smart meters and the metering service provider enable the billing of such complex but efficient energy usage.
- *Value Creation*: Fundamental change of the established value generation logic.
- *Data Complexity*: Heterogeneous data sets of various partners need to be exchanged and shared.
- *Business Complexity*: The implementation of Big Data applications with breakthrough / disruptive potential is the most challenging business approach. Usually it relies on the interplay of various partners that have managed to establish an effective collaboration. New data sources are aggregated and used in order to develop new products and services. By addressing a new market (segment), breakthrough applications – as the name is already indicating – have the potential to revolutionize established market settings. It is likely, that new players emerge that are better suited to provide data-based service than the established player and that the underlying business process change fundamentally.

[3] offered by Siemens

3.2. Identifying the Big Data Business Cases

The road leading to a goal does not separate you from the destination; it is essentially a part of it. – Charles de Lint

A business case needs to answer four questions

Any innovative technology that is not aligned with a concrete business case including the associated responsibilities is likely to fail. This is also true for Big Data solutions. Hence, the successful implementation of Big Data solutions requires transparency about the following four questions:

a. Who is paying for the solution?
b. Who is benefiting from the solution?
c. Who is providing the data?
d. Who is driving the development?

For instance, the implementation of health data analytics solutions for improved treatment effectiveness by aggregating longitudinal health data requires high investments and resources to collect and store patient data, for instance, by means of a dedicated EHR solution. Although, it seems to be quite obvious how the involved stakeholder, such as patients, payors, government, or healthcare providers, could *benefit* from aggregated data sets, it remains unclear whether the stakeholder would be *willing to pay* or *drive* such an implementation. In addition, as the sharing of personal health data is subject of high security and privacy constraints, one needs to clarify under which conditions the healthcare provider who produced and thus own the data can and are willing to share the patient data.

Therefore, it is important to understand that those responsibilities might be distributed across organizational boundaries.

Several aspects make the development of business cases for Big Data applications challenging.

For instance, the implementation of data analytics solutions in the healthcare domain using clinical data requires high investments and resources to collect and store patient data, for instance, by means of an EHR solution. Although, it seems to be quite obvious how the involved stakeholder could benefit from the aggregated data sets, it remains unclear whether the stakeholder would be willing to pay or drive such an implementation.

If the business approach is mainly targeting the optimization and improvement of existing offerings, the identification of business cases is a well-known exercise. However, if the business approach is aiming towards a collaborative setting within new market and business domains, the calculation of business cases easily becomes a challenging task with many unknown variables that often cannot be even influenced by the organization:

- *Big Data applications rely on high investments for ensuring data availability*: The collection and maintaining comprehensive and high quality data sets requires not only high investments, but often takes some years time until the data sets are comprehensive enough for producing good analytical results. For instance in the medical domain, one would need to collect large-scale, high quality and longitudinal data in order to gain reliable insight about the progressing of diseases over time. As such high and long-term based investments often can't be covered by one single party the conjoint engagement of multiple stakeholders might be needed.
- *Collaboration with partners with diverging interests*: As the impact of Big Data solutions increases the more data sources can be aggregated, the effective collaboration of multiple stakeholders with potentially diverging or even opposing interests needs to be established. In addition, the stakeholders' individual interests and constraints might even change over time. In the German electricity market liberalization, a new market role, the metering service provider, was created in 2010. That role is responsible for harvesting the energy usage data and could foster whole new branches of business. A range of stakeholders will require the energy usage data: retailers, network operators and new players that offer energy related services like demand response. However, in order to establish the basis for an effective collaboration, the interests of the various stakeholders need to be reflected when developing the business case. Especially, the ambiguous regulatory framework on what are the rights and responsibilities regard smart data usage prevents the existing and possible new players in utilities business to take on the new role of metering service provider.
- *Technological capabilities as well as its cost are a moving target*: Not only technological capabilities but also their cost factors are changing fast. Computing power and memory space per unit costs are still progressing exponentially according to Moore's Law. Additionally, innovative and cost-effective forms of information processing that is the main characteristic of all Big Data technologies, decreases the cost and update cycles of technologies considerably. Thus, the cost factor of technological investments needs to be accounted for in the overall calculation.

Having explained why the development of use cases for Big Data applications is very challenging, we need to emphasize that the lack of a business case should not hinder investments in Big Data Projects. Instead, the organization should focus on the continuously building up of Big Data capabilities (as described in Section "Big Data Project Portfolio Approach") in order to gain more insights about underlying business opportunities as well as to build up internal competences and data assets needed for future business cases.

3.3. Big Data Business Ecosystems

The majority of Big Data business will take part in business ecosystems.

Business ecosystems can be defined as "a dynamic structure which consists of an interconnected population of organizations. These organizations can be small firms, large corporations, universities, research centers, public sector organizations, and other parties which influence the system" (Peloneimi and Vuori, 2004). Business ecosystems allow organizations to access and exchange many different aspects of value, resources, and benefits.

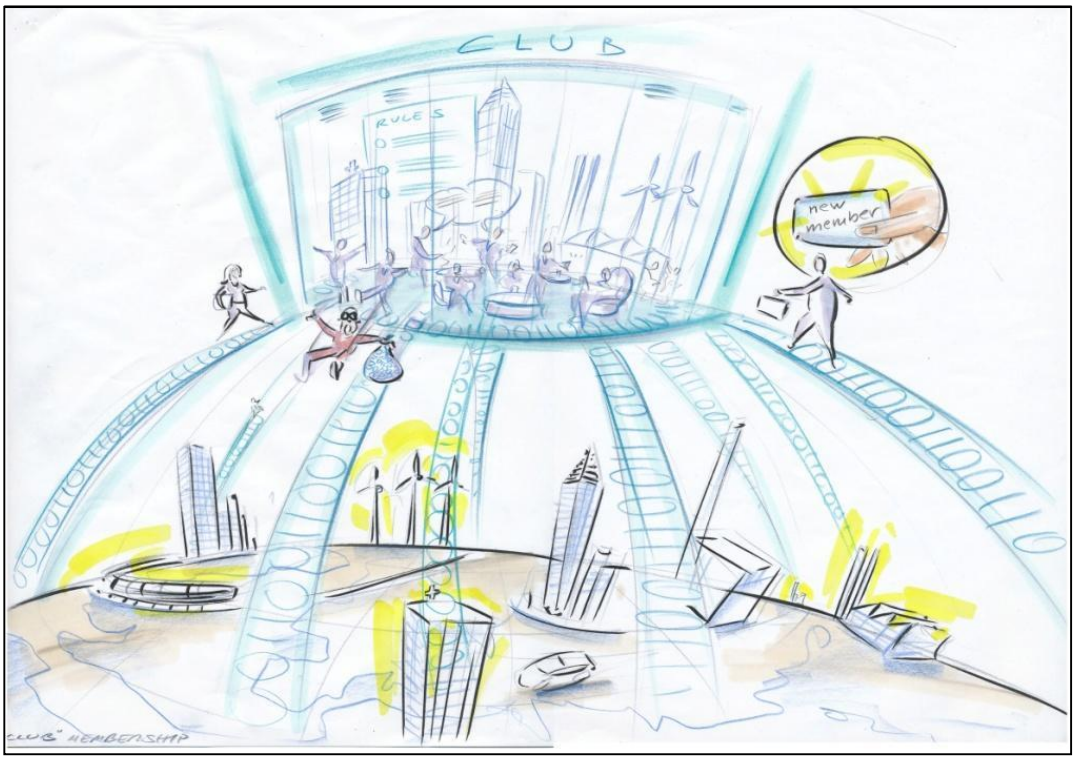


Figure 6 Data = THE entry ticket to become member of the "club of successful business partners"

Thus, successful ecosystems ("Club of winner") can help whole economic sectors as well as single players to prosper and develop. However, the governance of ecosystems relies on a balanced give and take. Looking at the various types of data and actors in the data ecosystem, will help to illustrate the underlying incentives and roles. Within Big Data business ecosystems, the access to data becomes a very important asset ("the entry ticket to the club").

3.3.1 Why Big Data Business Ecosystems are needed

In the data economy, business ecosystems are needed on various levels

Data Sharing Ecosystem:

The impact of Big Data applications increases, if the multiple data sources from the various stakeholders of an industrial sector are integrated. For instance in healthcare, by aggregating the administrative data, financial data with clinical data, it becomes possible to gain insights about the outcome of treatment bundles in terms of resource utilization. Thus, cooperative settings for the sharing of data are needed.

In order to establish sustainable data sharing ecosystems, it is important to understand

- which data source(s) each actor can potentially provide,
- what his or her sharing incentives are, and

- which requirements (e.g. privacy standards, 'opt out' ability, business models etc.) need to be in place in order to enable/foster the sharing of data.

For those who are providing data, mechanism must be developed to ensure transparency and control about data usage as well as some added value that is motivation enough to provide the data. Individuals might want to receive improved offerings, services with added value or better prices. Companies are interested in data to improve their knowledge about the consumer in order to customize their offerings, increase the customer binding, or optimize their pricing strategy.

Technology Portfolio Ecosystem

Big Data applications rely on a stack of technology component. The single component can either be bought or licensed from external providers[4], can be developed in-house, or further progressed within a reliable collaboration of complementing partners (e.g. a vertical player with an horizontal player).

However, the most critical question for each organization is the following:

How to get access to the needed resources while keep control of the key resources and the company's USP?

For example, the open source community supports both the spreading of Big Data technology into different industrial businesses as well as the spreading of know-how and experience in the broader workforce and at universities. The most prominent framework is Hadoop, an open source development based on publications on Google's Distributed File System (GFS) and their MapReduce framework. MapReduce is a powerful abstraction that enables the automated management of massively parallel processing on data that resides in a distributed file system spread over huge clusters of commodity hardware. This combination of technologies is the starting point that considerably changed the economics of computing. Ever since the launch of Cloudera in 2008, the professional support for Hadoop favored the adoption by more traditional industries[5], such as the utilities. In the physical world other companies started using Hadoop to cost-effectively handle the massive amounts of data that became available to them in fields ranging from entertainment[6] to energy management[7] or satellite imagery[8]. In the online business nearly every popular site makes use of Hadoop to analyze the mountains of data they are generating about user behavior and their own operations: Facebook[9], eBay[10], Etsy[11], Yelp[12] , Twitter[13], Salesforce.com[14]. IT incumbents, IBM, SAP, Oracle, and Microsoft – the Big Four – also also integrated Hadoop[15] into their solution suites after the ecosystem around it established.

Value Networks in a Business Ecosystem

The data-driven economy relies on value networks.

In the data-driven economy, value streams are no longer bi-directional but involve several players exchanging different types of value. The party who is benefiting from a value-added service, no longer needs to be the one who is paying for the service. Such value networks already exist in the internet environment.

Most of the established players providing data-base solutions, such as Google, eBay, YouTube, Facebook, ITunes, etc. are building up a growing user community by offering them free services, which allows them increase their income as each advertising companies is paying a fee per click or user.

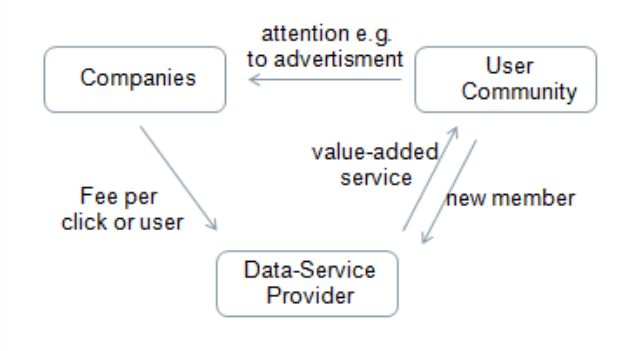


Figure 7 Value streams of Internet-based Services

[4] see business pattern "IT-Infrastructure provider on page 22
[5] http://blog.cloudera.com/blog/2009/06/smart-grid-hadoop-tennessee-valley-authority-tva/
[6] http://gigaom.com/2012/09/16/how-disney-built-a-big-data-platform-on-a-startup-budget/
[7] http://gigaom.com/2012/10/11/the-rent-is-too-damn-high-but-big-data-means-the-power-bill-isnt/
[8] http://gigaom.com/2012/04/17/satellite-imagery-and-hadoop-mean-70m-for-skybox/
[9] http://gigaom.com/2012/06/13/how-facebook-keeps-100-petabytes-of-hadoop-data-online/
[10] http://gigaom.com/2012/01/31/under-the-covers-of-ebays-big-data-operation/
[11] http://gigaom.com/2011/11/02/how-etsy-handcrafted-a-big-data-strategy/
[12] http://gigaom.com/2012/12/02/pinterest-flipboard-and-yelp-tell-how-to-save-big-bucks-in-the-cloud/
[13] http://gigaom.com/2012/03/07/how-twitter-is-doing-its-part-to-democratize-big-data/
[14] http://gigaom.com/2012/09/17/5-ideas-to-help-everyone-make-the-most-of-big-data/
[15] http://www.informationweek.com/software/information-management/big-data-revolution-will-be-led-by-revolutionaries/d/d-id/1107781?

3.3.2 Risk and Opportunities

Business ecosystems bring risks and opportunities.

Business ecosystems include both cooperation and competition. Therefore it is an environment of risk and opportunities. Before entering a business ecosystem, the likely risks and opportunities of the various cooperation strategies should be analyzed in detail. For instance, is it recommended to share my data with my suppliers who are also working with my competitors?

According to the category of business ecosystem, different types of risks might be involved (Smith, 2013). For instance:

- Complexity of relationship management (between actors and keystone)
- Control (centralized or decentralized)
- Co-opetition (simultaneous cooperation and competition)
- Potential loose of intellectual property rights
- Centralized control of key resources, such as data sources
- The risk of business-model replication

Similar, different opportunities are associated with business ecosystems:

- Access to complementary data sources enables improved analytical insights
- Access to new markets and customer
- Risk sharing
- Investment sharing
- Access to competency and technology
- Increased time-to-market

3.3.3 Governance of business ecosystems

Business ecosystems are characterized by a number of loosely connected partners who depend on each other for their mutual effectiveness and survival.

The actors in business ecosystems usually differ in the degree they can influence the other members of the ecosystem (Iansiti and Levien, 2004):

- For instance, keystone actors shape and coordinate the ecosystem largely by dissemination of platforms that form a foundation of ecosystem innovation and operations. As keystone player profit from the number of members in the network, they usually take responsibility for their well-being and survival.
- In contrast to this is the dominator strategy that attacks the ecosystem by absorbing and integrating external assets into internal operations. However, empirical findings show that companies who have dominant influence over the business ecosystems' key resources and at the same time do not take responsibility for the long-term health of the associated (often small) ecosystem partners easily jeopardize the economic basis of the whole ecosystem.
- A large number of organizations follow a niche strategy. Those companies emphasize differentiation by focusing on unique capabilities and leveraging key assets provided by others.

No governance strategy is better than any other. However, a company should be clear about how to set up his own cooperation strategy within the ecosystem.

3.4. Organizational Implications

Big Data business relies on new competences and new organizational structures.

3.4.1 New Competences

New competences and skills are needed to discover the yet unknown business value in large-scale data repositories.

The successful implementation of Big Data projects requires expertise from many different fields and domains. Data needs to be made available, stored, integrated and managed. Analytical formulas and applications need to developed, and statistical integrity needs to be ensured. The connection between business problem and data insights has to be made. This requires domain expertise, business modeling, visualization, etc.

Typically, this wide range of competences is rare to find in single individuals. Thus, a team of people with different skill sets is required.

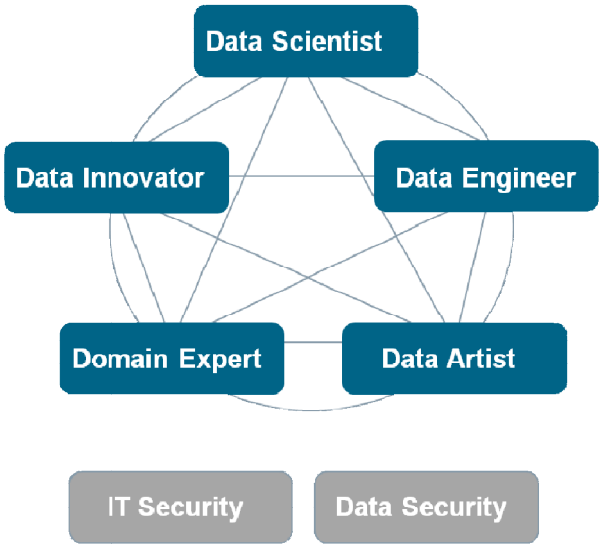


Figure 8 New Competences required for Big Data applications

The following competences are needed:

- Domain expert or subject matter expert (SME) knows the domain and user needs.
- Data Innovator identifies & evaluates new business models.
- Data scientist plays with the data.
- Data engineer develops and implements new algorithms.
- Data artist visualizes the outcome.

In addition, expertise in the domain of data governance and data privacy and security are needed.

- Data governance expert manages the continuously growing data assets.
- Data privacy and security expert takes care about the underlying privacy and security requirements.

In order to build up the needed new competences, organizations will need to provide additional training and education for employees in the area of Big Data technologies, as well as foster the effective cooperation between people coming from different disciplines.

The responsible managers need to understand the spirit of Big Data project in order to efficiently handle the involved uncertainty.

However, the success of Big Data projects relies on the degree the rest of the organization can embrace the underlying uncertainty involved. Big Data means spending effort without knowing what to expect. Big Data is about experimentation without plan. Not only the team member of a Big Data project but also the manager must be capable to apply the principles of scientific experimentation and design while developing their future business. The knowledge of how to build intelligent hypothesis relies on the principles of experimental testing and design, including the selection of appro-

priate populations and samples. The more familiar managers are with those principles, the more effective they can handle the underlying uncertainty of Big Data projects.

3.4.2 Organizational Structures

Organizational structures can foster synergies and cross-sector learning in several areas.

In order to enable cross-sector synergies, to bundle resources as well as foster cross-sector learning, new organizational structures that promote the flexible cooperation between the required competences are needed.

Several variants of organizational structures for the various purposed are possible:

1. *Data Governance:* central versus decentralized
2. *Competence Building:* central competence center versus decentralized initiatives
3. *IT-Infrastructure:* central versus decentralized
4. *Business Initiatives:* central versus decentralized

The centralizing of activities allows streamlining and standardizing the overall Big Data approach which establishes the basis for future cross-sector collaborations. Moreover, synergies and the bundling of resources become possible. However, the side-effects are that the individual business units get confronted with the loss of control over their assets as well as with increased efforts for communication and coordination in order to negotiate and implement the common strategic direction.

4. Data Perspective: Know Your Data

The access to data is an important asset to secure an organization's future competitive position.

The explosive growth in the quantity and quality of data has induced a significant opportunity to generate business value. Depending on the type of data source one is aiming to use for Big Data applications, one is facing different challenges and, thus, needs to follow different strategies for accessing the data.

Three different kinds of data pools can be distinguished: a) own data, b) external data and c) crowd and community-based data:

a. Own data sources are any data sources that are controlled by the own organization. There are two different ways how the data can be generated:

   - *Data sources that are generated within the established processes* and are now exploited for – so called – secondary usage. For instance, any data that gets produced by machine or devices. For those data, technical availability of the data is the main challenge.
   - *Data sources that are generated by dedicated pooling strategies.* For instance, the Google-funded start-up "23 and me" was aiming to build up a repository of genetic testing data by offering a genetic testing data for a very low-prices. Here the main challenge is the design of a scenario that fosters the data collection processes.

b. External data sources are any data sources that are in control of other partners – sometimes even competitors – or data sources that are open source or publicly available. Here, one needs to distinguish

   - *Closed data sources* are owned by another organization or legal body. The main challenge here is:
     - How to establish partnerships establishing the trusted basis for data sharing?
     - What are the underlying legal frameworks that need to be implemented
   - The access to *open data* sources need to address the following questions:
     - Which data sources are available under which licensing condition?
     - How to technically integrate those data sources?

c. Crowd and community-based data sources can only be produced when an active community is producing data on a large scale. So the main challenge here is to build up user communities that produce the data of interest. Usually, some kind of added value or incentive is needed to foster the activity level of the community, and thus the amount of data that gets generated.

4.1. Technical Data Access

The technical availability of data is key – and still a key challenge in many industrial sectors.

In order to realize Big Data applications, one needs to enable the seamless access to the various data sets. As of today, in many of the traditional industries, such as healthcare, energy or manufacturing, the access to data is only possible in a very constrained and limited manner. In order to improve this situation and for establishing the basis for the wide-spread implementation of Big Data applications in those sectors, dedicated technical data access strategies needs to be implemented.

4.1.1 Technical Availability of Data

It is important to understand that in many of the traditional industries, data is not always accessible in digitized form, or that there is no broadband communication available to the data source to retrieve it on a continuous basis. Even when data is available, its type or format does not allow it to be used by business intelligence or other applications right away. There need to be a range of analytics and preprocessing done before the data is transformed into a usable format. With Big Data, of course, these analytics and preprocessing need to be streamlined and automated otherwise cost-efficiency could not possibly be achieved.

Data Digitalization and Communication

Only data that is available in digital as well accessible remotely can be used as input for analytical solutions. However, in many industries the data is only partially available in formats that can be used as input for Big Data applications. Thus, further investments for communications, pre-processing and processing by analytics are needed.

- Example Healthcare Sector: As of today, only a small percentage of health-related data is digitally documented and stored. The extent to which healthcare IT systems, such as EMR systems, are in place varies significantly across and even within countries. The result of a study accomplished by Accenture (Accenture, 2012) showed that, for instance, in average 55% of health provider in Germany use Healthcare IT within primary care settings and 60% in secondary care settings. In UK, the numbers are quite different: 63% of providers rely on healthcare IT in primary care settings but only 15% in secondary care. For more details, we refer to the above mentioned study.

  In general, clinicians, physicians or nurses lack time to spend extra effort for documenting findings, observations, diagnosis, etc. In addition, the usage of IT-technology often impairs the focused patient-doctor encounter. However, there is a substantial opportunity to create value if more data sources could be digitized with high-quality as well as made available as input for analytics solutions.

  In addition, large amounts of health data are provided in unstructured formats (e.g. medical images, clinical reports, sensor data, videos, mp3-files or communications on the social web). The market research institute IDC estimates that in the coming year 90% of health data will be provided in unstructured format (Lünendonk, 2013). The seamless processing of unstructured data sources requires semantic annotation by using labels that rely on standardized and commonly used vocabularies or ontologies. For instance, indication-relevant content of medical images can be extracted and labeled by means of segmentation algorithms and indication-relevant content of medical report can be captured and semantically labeled by means of IE technology.
- Example Power Transmission Sector: The transmission network has always been monitored well, because the economic damage of a disturbance to the power system or single equipment is very high. However, the measurements data and its communication are from the past century. Only now is the GPS-synchronized high-resolution power network parameter and switch state data from so-called intelligent electronic devices becoming available. Together with high bandwidth communication a new area of big data in power systems will be starting through the wide-spread installations of IEDs[16][17].

Data Exchange

In many industries, we are facing the challenge of data silos. In order to overcome data silos, technical means for data exchange between data sources are needed. Data exchange is the process when a data item that is structured according to a source schema is transformed into a data structure of a target schema. As in many industries adequate data exchange formats or interfaces are missing, the data sharing is still very difficult to accomplish:

- Example Healthcare Sector: As of today, large amounts of health data is stored in data silos and data exchange is only possible via Scan, Fax or email. Due to inflexible interfaces and missing standards, the aggregation of health data relies – as of today – on individualized solutions and, thus, on high investments. In comparison to the degree of healthcare IT adoption, the adoption of seamless healthcare information exchange is far less advanced (Accenture, 2012). On average for instance, in Germany less than 26 %, in UK less than 46% and in France less than 27% of the healthcare provider use healthcare information exchange technology.
- Example Smart Grid Segment: The entire market communication requires each market role to exchange data that is needed to retail and bill energy usage. Due to regulations, but also due to the lack of technology utilization data exchange takes days. Due to aged standards, data, even time-series data on consumption is exchanged via email and has to undergo many inefficient steps until it can be used by applications. This is one of the reasons why the electricity business is yet far away from being a real-time business. However, since technology that could remedy this situation has been in successful use in other businesses, an era of low risk efficiency increases awaits the smart grid segment – especially regarding efficient data exchange.

[16] Hype Cycle for Big Data, 2013, https://www.gartner.com/doc/2574616/hype-cycle-big-data
[17] The Digital Tsunami, http://www.truc.org/media/2774/The%20Digital%20Tsunami.pdf

4.1.2 Data Quality

Data quality should not become an end in itself: The impact and value of data quality depends on the context the data will be used, i.e. the use case and the underlying data source.

In the context of Big Data applications, there are two assumptions towards data quality: Either one IS concerned data quality or one IS NOT concerned about data quality.

Many Big Data applications are simply looking for patterns in data. Those applications rely on the assumption that the majority of incoming data is not clean, but simply due to the large amount of random and disconnected data, for instance in social data streams, one can gain valuable insights for some businesses. *Those applications do not depend on high data quality standards.*

However, if one is interested in results and insights that enable sound business decisions, one needs to apply the same data quality standards that one applies to traditional data sources. The quality of data needed refers several data characteristics, such as consistency, accuracy, reliability, completeness, timeliness, and validity.

- For instance data sources, such as RFID tags or sensor data, provide more structuring in comparison to the unstructured content of social media platforms. In general, those data sets should be reasonably clean, although you may expect to find some errors with the data, e.g. due to some missing data, such as missing time intervals or defunct sensors.
- Then, there exist domains, such as the healthcare domain, that require the completeness and comparability of data sources in order to gain reliable insights by means of data analytics. For instance, the features and parameters used for describing the patient health status need to match completely in order to enable the reliable comparison of patient data.

As Big Data applications rely on various different data sources, one is often facing the challenge of inconsistency when integrating the data. Although referring to the same item, the different data sources might label it in very different manner. In order to avoid inconsistency when integrating different data sources, one needs to spend efforts for aligning semantically equal items to each other. If a large number of data sources needs to be integrated in a consistent manner, the usage of standardized labeling of items by means of agreed ontologies or data modeling standards is recommended.

4.2. Data Sharing Policies

The legal foundation for data sharing requires a deep understanding which type of data we want to use for what purpose.

Although the access to data is getting more and more a competitive asset, adequate rules, norms and frameworks ensuring accountable and trusted flow and sharing of data are still missing. Dedicated technical infrastructures, legal processes for data sharing and communication as well as the implementation of suitable organizational processes that enable the secure and transparent data sharing are needed. However, in this discussion it is important to understand that several characteristics of data make the establishment and implementation of rules and legal frameworks quite challenging (World Economic Forum, 2012):

- Digital data can easily be copied infinitely and distributed globally, thus most of traditional trade barriers are no longer applicable.
- By using data, it doesn't lose its value; i.e. data can be reused to generate further value.
- The connection of two data items allows to generate a new piece of data which again might bring new potentials to generate value.

To get at the heart of the issue of data sharing, we need to unpack a number of key concepts in the current debate. The current discussion focuses mainly on two notions: one of *ownership of data* versus one of *privacy of data*. However, as those two notions differ according to the type of data involved, we will first distinguish the different categories of data.

Sharing data policies rely on the category of data as well as the intended usage. In order to clarify the data privacy constraints, at the beginning of each Big Data project the following questions should be answered to select the appropriate data sharing policy:

1. *Is the data personal data?*
2. *How are you planning to process the data?* Are you planning to collect the data, to store the data, to share or to change the data?
3. *Where are you planning to process the data? W*ill this be inside or outside the EU.
4. *With whom are you planning to exchange the data?* Is your partner located within or outside the EU?

4.3. Categorization of Data

To get a better understanding about the underlying opportunities and constraints of data sharing policies, we need to distinguish different categories of data:

- Private and personal Data: Most private and personal data remains unused due to unclear legal situation
- Operational Data: The usage of operational data within Big Data applications bears promising business opportunities, however the sharing of data is often hindered due to missing incentives and the fear for loss of control
- Historical and longitudinal data: Although being an important input for identifying trends and making predications, the usage of historical and longitudinal data faces high data privacy and security issues.

For more detailed description about legal constraints, please see Appendix 4 Data Sharing Policies.

4.3.1 Ownership of Data

The concept of data ownership with respect to data-driven business aspects is difficult to define.

The notion of data ownership refers to both the possession of and responsibility for information. In other words, ownership of data implies power as well as control. The control of information includes not just the ability to access, create, modify, package, derive benefit from, sell or remove data, but also the right to assign these access privileges to others (Loshin, 2002). Thus, data ownership strongly influences the underlying business opportunities and legal situation.

The term data ownership is a misleading concept.

The concept data ownership is a misleading concept and often misinterpreted. This is due to the fact that the basic concept "ownership" which refers to 'having full property right' to something does not apply in the context of data management. Otherwise a data owner could take the data and sell it. However, this is not was the concept of data ownership is about, as data cannot be "owned" in a traditional sense.

In consequence, the question who "owns" the data and the challenge of defining the concept of "data ownership" in an appropriate manner has triggered the attention of researchers, such as (Al-Khouri, 2013), as well as of prominent Think-Tanks, such as the World Economic Forum or the Münchner Kreis. In order to foster the sharing of data, trust between all involved stakeholders needs to be established. In this context, the implementation of transparent commonly agreed data governance principles is seen as important lever to establish trustful collaborations.

The following fundamental principles guide the development of sustainable data ecosystems were established by the World Economic Forum (World Economic Forum, 2012):

- *Accountability:* Who is accountable for the appropriate security mechanism that help to prevent
  - theft of data
  - unauthorized access to data
  - usage of data in a way that is not consistent with agreed upon rules and permissions
- *Enforcement:* How to establish mechanisms those ensure that organizations are held accountable for the accountability obligations through a combination of incentives or – if needed – penalties?
- *Data permission:* How to enable flexible and dynamic permission of data usage that incorporates necessary context information and paves the way for value-creating usage? How to ensure that different stakeholders might have different rights to use the data?
- *Shared data commons*: How to preserve the value to society from the sharing and analysis of anonymized datasets as collective resource?

For more detailed description about legal constraints, please see Appendix 4 Data Sharing Policies.

4.4. Impact of Open Data Trend

Many trends in end user industries, such as Open Data, Data Market places, or Linked Data, have the potential to trigger a paradigm shift towards an open and sharing economy also within the B2B industries.

The idea behind *open data* is that data should be freely available to everyone to use and republish and distribute it without being limited by copyrights, patents or similar control mechanism.

Open data can origin from everywhere; however as of today only data in science and government is published openly on a larger scale.

- *Open science data*: The overall aim is to speed up the scientific activities by publishing observations and result of research work. The idea of open science data exists already since the 1950s, however with the rise of the Internet the publishing and sharing of data becomes much easier, cheaper and thus much more attractive.
- *Open government data:* The open data barometer revealed that 55% of a sample of 77 global countries have formal open data policies. In addition, the study highlights the importance of open data policies in boosting business activity and innovation (World Wide Web Foundation and Open Data Institute, 2013).

There are some arguments against open data

- *Data Sharing Constraints*: Privacy concerns may restrict the access to data specific user groups and need to be managed accordingly.
- *High efforts involved need to be refunded*: As the collection, cleaning, managing and disseminating of the various data sources relies on labour and cost-intensive processes, the party providing those services should be fairly refunded.
- *No incentive to enrich raw data*: Access to raw data is often only half the battle, as many applications rely on additional processing of raw data for being able to effectively use it. For instance, clinical documents need to be semantically annotated by means of Information extraction algorithms before they can be reused within clinical Big Data applications. If the access to the data is unrestricted, no-one may have an incentive to invest in the processing needed to make data useful.

Some of the currently discussed arguments against open data are already addressed by open market places.

Open market places:

- With data becoming a strategic asset for future businesses, it is not surprising that data is recognized more and more as tradable goods. The number of data providers that are setting up platforms for selling, buying or trading data are increasing continuously. A good overview of data marketplaces is provided in (Schomm et al., 2013):
- Open market places are platforms through which data can be purchased, sold or shared. They provide added value in various ways:
  - *Improved discovery and comparability of data*: Data originally being scattered around numerous websites, can now be accessed in an integrated, much more comfortable manner.
  - *Reduced cleaning and formatting issues*: Data from various providers often rely on different access mechanism and formats. By offering a single access mechanism, open market place platform provide customer easier access to the data.
  - *Economy of scale*: Only through the broad access to data and the increased customer base, the provisioning and publishing services can be offered in a cost-effective manner.
- There exist many different categories of open market places, such as web crawler, customizable crawler, search engines, raw data vendor, complex data vendor, matching data vendor, enrichment –tagging services, enrichment – sentiment services, enrichment – analytics services, data market places.
- Through the seamless and affordable access to open data sources, other businesses can more easily align their own data bases with external data and thus compute more insights. For instance, by combining the weather data with energy consumption data, improved prediction can be calculated.

5. Technology Perspective: Revolution by Techno-economic Paradigm Shift

This chapter covers the evolution of Big Data technologies as driven by Big Data innovators and early adopters. Some of the aspects from the evolutionary path are extracted that may assist the majority of Big Data adopters in mastering the challenges of choosing the right tools. The different architectural patterns for mastering Big Data systems are also briefly discussed. One of them, the so called lambda architecture is the latest common sense in the Big Data domain and also a good fit for industrial businesses. However, especially the domain specifics of handling Big Data in cyber-physical systems require the stakeholders to look beyond the state of the art: In-field analytics is only one example that is discussed in this context. Nevertheless, how Big Data natives have arrived at this architecture reveals some clues as to how the industrial businesses and vertical IT providers could mimic this evolution without experiencing major disruptions by the technological advances. The section about choosing the right technologies is only a primer for setting up the stage to enable the adaptation of technologies for industrial businesses. Domain-specific adaptation will be the key to success, as it is clear: the evolutionary process is also based on disruptive selection of the fittest technologies for a specific domain.

5.1. Survival of the Fittest – A Short History of Big Data Technologies

"Those associated with storage devices long ago realized that Parkinson's First Law may be paraphrased to describe our industry—'Data expands to fill the space available'..." - I.A. Tjomsland (1980)[1].

"Big Data," the term, can be traced back to 1998[18]. The efforts to understand the information explosion, cope with and make use of it reach even farther back. The underlying technological paradigms of Big Data such as distributed data storage and parallel processing or distributed computing, machine learning and data mining, or information extraction date back to the 60s and 70s. This section covers the evolution of Big Data technologies first by simply putting the technological advancements in a time plot (see Figure 9). The chronology visualizes multiple interesting facts about Big Data technologies:

- Waves of Big Data processing: The year 2000 marks a tipping point in the agility of technological advancements. Since then every three years a new wave of technological capabilities evolves: waves of i) batch parallel processing, ii) incremental distributed processing, and iii) real-time stream processing of data.
- Waves of Big Data analytics: The methods for actual value generation from data, i.e. data analytics, always require an adaptation to make use of the full potential of the technological advancements, such as cheaper storage and faster computation – hence also three waves of analytics are identified: i) analytics of the past, ii) diagnostic and predictive analysis, visual and exploratory analytics, and iii) prescriptive analytics, i.e., fast analytics that is enriched with knowledge from the past as well as predictions about the future such that reliable real-time decision making becomes feasible.
- Big Data natives in the driver seat: Online data businesses, the so-called Big Data natives such as Google, Yahoo! (Search), Amazon (eCommerce), and Facebook (Social) have been pushing the technological boundaries because the scalability of their business model is directly related to the scalability of their data management capabilities.
- Incumbents are the early adopters, open source drives fast adaptation: With respect to both IT and analytics, the open source movement has been playing a key role for Big Data technologies to mature and gain widespread adaptation. Two major examples are Hadoop and R, which considerably facilitated the spillover of Big Data processing and Big Data analytics into other businesses. Both have been adopted by the major incumbents of their fields, e.g. Hadoop by the Big Four, and R by SAS and SPSS.

  Interestingly, at the beginning of each wave though, there is always one IT incumbent that started it off with the right choice of technologies: e.g. Teradata which developed a data warehousing solution on massively parallel processing architecture already in 1984 to enable scalable batch processing of data; IBM that launched its InfoSphere Streams and SAP HANA with its in-memory database to enable real-time processing of data.

[18] http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

- The future holds the merger of waves and standardization: Rather recent experiences, e.g. with analytical algorithms that are hard to parallelize, lead to developments such as Mahout or RHIPE. These are again open source results of mergers of analytical libraries with the frameworks for distributed storage and processing of data. In order to realize what Analytics 3.0 promises, i.e., millions of insights per second and decision making in the blink of an eye, probably even more streamlined solutions will need to emerge. Once such a high return on data becomes reality it is foreseeable that industrial businesses and the IT industry will heavily concentrate on the standardization of these technologies.

Open source adaptation of the new technologies and the fast growth of a stable ecosystem caused a spill-over to many other traditional businesses – cost-efficient and innovative forms of data processing are available, but require new technical skills, which are rare.

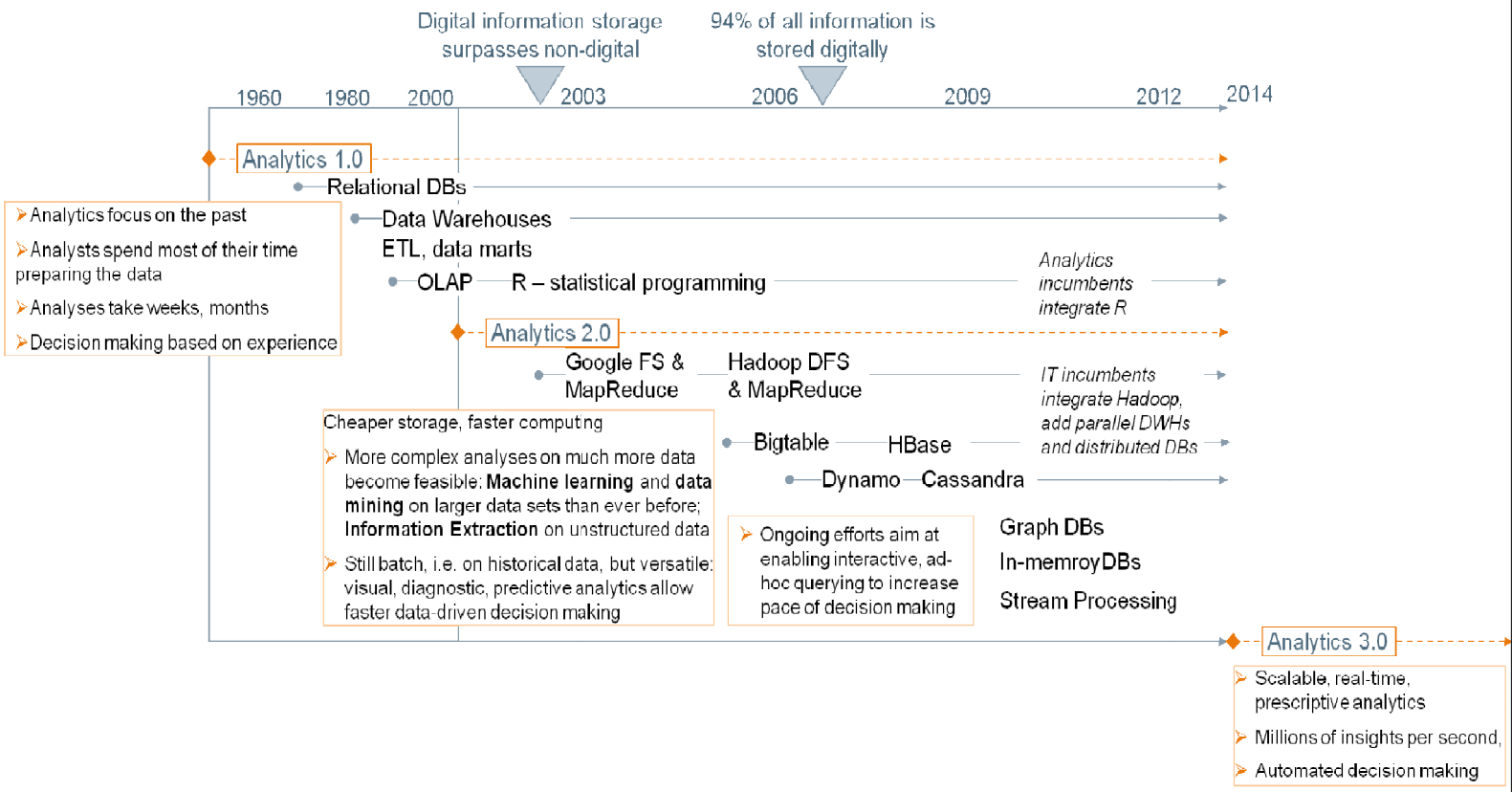


Figure 9 The time plot shows how the Big Data natives push technological boundaries to drive the evolution of Big Data technologies (adapted from Davenport, 2013)

The underlying paradigms to Big Data and Big Data analytics such as distributed computing, parallel processing, machine learning, or information extraction reach back to the 1970s. But they are adapted both to cope with the challenges of increased digitization, i.e. masses, speed, and variety of data, and to make use of technological advancements in computing and storage at each point in time.

Before the first wave of Big Data technologies, 1960s to 2000, there already was the dawn of the information age: increased digitization of businesses and processes, data explosion, and the handling of data as a production good which lead the way to data warehousing. With data warehouses, as the name suggests, started the collection of all "valuable" data, i.e., data that was known to be required for e.g. financial reporting. In those days the analytics could only be focused on the past, because the processing of data, the refinement into so called special purpose data marts took weeks or months. The analysts had to spend most of their time preparing the data for analysis rather than analyzing the data. The entire process of acquiring data and making it available for analysis, as well as the analyses took several weeks and months. So the actual decision making was still based on experience. This is still true today for businesses that do not have much digitization. For online data businesses on the other hand, which started to gain grounds with web search

and ecommerce in 2000, the cost-efficient management of massive amounts of multi-structured data became business critical. In 2002 amount of digital information storage surpasses that of non-digital storage for the first time[19].

Since the advent of online businesses, which must innovate around data, every three years a new wave of Big Data technologies builds up:

1. The "batch" wave of distributed file systems and massively parallel processing via MapReduce,

2.The "ad-hoc" wave of NoSQL and key-value stores with their distributed data structures and distributed computing paradigm, and

3. The "real-time" wave which allows producing insights in milliseconds through stream processing.

The first wave of "batch processing:" So the first wave of Big Data technologies, distributed file systems and massively parallel processing by means of the MapReduce framework was pioneered by the search engines Google and its open source pendant Nutch. Nutch later spinned off Hadoop, which was developed based on the scientific paper about Google's solution. In 2008, Hadoop was acquired by Yahoo!, but remained open source. Both Google and Yahoo! developed very genuine strategies to balance IPR and open source support (see discussion on Technology Portfolio Ecosystem in section "Big Data Business Ecosystems"). The cost of a byte did not only drop due to hardware related innovations, in storage and silicon – but especially through productizing decades long research on elastic and fault tolerant networked computing. Complex analyses from the fields of machine learning, data mining, or information extraction on much more data than ever before became possible. Such technological advancements were also a necessity because Big Data natives already had reached the capacity what commercially available solutions could deliver: In their 1998 paper on the prototype of Google, Brin and Page stated at that time, that their prototype was managing 26 million web pages, and with the cutting edge of what hardware, memory, and networking could provide they could scale to a maximum of 100 million pages[20]. Only two years later Google served an index of one billion web pages and hit the one trillion mark in 2008[21]. During the same time, Amazon started the development of its massively scalable data processing solution called Dynamo, when the company experienced a number of outages in midst of the 2004 holiday shopping season. The CTO, Werner Vogel, attributes these outages to the fact that "Amazon's scaling needs were beyond the specs for their [the vendors] technologies and we were using them in ways that most of their customers were not."[22]

The second wave "ad-hoc processing:" The next wave of so-called key-value stores and NoSQL was set off by online companies that needed a more near real-time take on Big Data. The Web had been transformed into "Web2.0" with masses of user generated content, reaching from blog entries to product reviews, status updates and comments in social networks. Google's Bigtable with open source pendant HBase, Amazon's Dynamo, and Facebook's Cassandra are the most prominent among the technological improvements in this second wave. Whilst parallel processing on a server cluster allows for efficient batch processing the concept of a distributed database allows efficient random access to data. As the name of the original Bigtable suggests, the essence of such non-relational data store is to "host very large tables with billions of rows and X millions of columns."[23] Key-value store is a more popular name for the efficient data structure of distributed hash tables. Distributed hash tables have been a popular technology for peer-to-peer file sharing, and the enabler of disruptive voice over IP client Skype, for example. Such distributed data structures and algorithms expose scale-free characteristics, which became a necessity for internet-scale applications. There are well over 100 NoSQL databases, which also include other non-relational types of databases such as graph databases. This second wave also brought the common sense that "one size does not fit all," i.e., there are solutions that are optimized for fast writes and others that allow fast reads but not both equally efficient. There is always the tradeoff of between consistency, availability, and partitioning tolerance – also known as the CAP theorem[24].

The new economics of computing marks the tipping point for Big Data.

In 2007, already 94 percent of storage capacity was digital[25]. The online business around digital content had grown by the addition of social features by Facebook and Twitter. Consequently, taking on the challenge of cost-efficient data management has been their core business[26]. Around the same time, in 2007, several of the underlying technical solu-

[19] http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/
[20] http://infolab.stanford.edu/~backrub/google.html
[21] http://googleblog.blogspot.de/2008/07/we-knew-web-was-big.html
[22] http://www.allthingsdistributed.com/2012/01/amazon-dynamodb.html
[23] http://hbase.apache.org/
[24] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.9136&rep=rep1&type=pdf
[25] http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/
[26] http://oreilly.com/web2/archive/what-is-web-20.html

tions have been open sourced or delivered as-a-service, which changed the economics of computing[27]. The availability of considerably more cost-efficient computing power may well be the starting point of the Big Data trend. Companies in industrial businesses, but at the verge of increased digitization, could at once afford managing much larger amounts of data, such as in the media sector[28], or only a few years later in the energy sector[29].

The cost-efficient availability of mass data triggered the second wave of Big Data analytics.

Once huge amounts of data could be sifted through in a relatively cheap manner using distributed file systems and parallel processing paradigms, the focus of stakeholders shifted to operationalizing analytics at very large scales. Machine learning, data mining, and information extraction on bigger and multi-structured data sets was now possible; but also required knowledge about distributed, parallel computing. The result is the development of libraries and frameworks to move analytical libraries onto the newly adopted computing paradigms, e.g. Mahout, the analytical library on Hadoop in 2009. Shortly after, the open source programming language for statistical analysis, R, gained popularity in the Hadoop community as a way of programmatically extending analytical capabilities on top of Hadoop. The incumbents of the statistical tools' domain added distributed computing capabilities around 2010: SAS introduced grid computing and IBM offered a cloud-based solution for SPSS Statistics Server about a year after their acquisition. In 2011, R, overtook SAS and Matlab in popularity for the first time[30].

Mahout but also R+Hadoop and RHIPE are the first mergers of analytics with Big Data in a Big Data analytics framework. The open source and programmable nature of R made it the number one choice of the Big Data community.

Technological opportunity now allowed the analysis of data bigger and faster than ever before. This highly increased the quality of analytical answers, especially regarding descriptive, visual, predictive or diagnostic analysis, but the impact of analytics had not changed significantly – yet: analytical insight came still from historical data. Just recently, the Big Data analyst community started to build on top of the "ad-hoc" querying capabilities of the new technology. Analytical application, which immensely benefit from faster response times, such as data discovery and exploratory analysis of Big Data or visual analytics and descriptive analytics, are now applied as widely as batch analysis of Big Data.

Message: Big Data analytics requires distributed computing, with inherent features such as fault tolerance through redundancy and scalability through decentralization.

The third wave "(near) real-time processing:" The technology race is far from the finish line. Currently, the real-time analyses of streaming data as well as the visualization are the technological frontiers where innovators are pushing the envelope. Two offerings from IT incumbents, SAP HANA and IBM Streams can be said to have been at the starting point of the third and current wave of (near) real-time data and cost-efficient stream processing. The former is an in-memory database that makes use of decreasing cost of memory per server. The latter is based on stream computing that makes use of distributed computing topologies over a network of servers. Stream computing is also the underlying paradigm for Twitter's Storm and LinkedIn's Samza that were subsequently introduced. Both use Apache Kafka, which is a distributed messaging framework that enables fault tolerance and scalability. In Big Data analytics scenarios both fault tolerance and scalability are fundamental enablers.

Stream computing enables the cost-efficient analysis of massive amounts of data in real-time as it streams into the enterprises. Real-time processing opens up the technological opportunity of generating millions of insights per second through advanced analytics[31]. An increase in responsiveness can not only be gained through stream computing and streaming databases. Special purpose databases that better represent the data for the specific purpose of analysis have also emerged. Graph databases for the analysis of network data, mainly used for interactions analysis in social network data, are one example of such efficiency increase by domain-specialization.

The different notions of "real-time" relate to the age of data and how long the response to a query takes.

It is becoming increasingly more feasible to combine the different timeliness and data freshness requirements of analytics to generate and reuse insights as input to further analysis steps. Figure 10 is an example from power systems' wide area monitoring protection and control systems. These systems are still in piloting phase today, partly because digitization of the power system automation has not yet reached its tipping point. A so-called wide area monitoring system (WAMS) delivers analyzed information for the control center of a power system. The data from thousands intelligent

[27] http://www.roughtype.com/?p=1189
[28] http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun/
[29] http://gigaom.com/2009/11/10/the-google-android-of-the-smart-grid-openpdc/
[30] http://blog.revolutionanalytics.com/2011/02/r-overtakes-sas-and-matlab-in-programming-language-popularity.html
[31] http://iianalytics.com/resources/archived-webinars/analytics-3-0-the-era-of-impact-april-2013/

electronic devices in the field can be analyzed as soon as they arrive at the enterprise backend. As such a WAMS enables near real-time analysis for visualization and alarming capabilities, which is called interactive or exploratory analysis. Together, the analysis of streaming data and historical data can yield diagnostic and predictive insights, i.e., why did a fault occur and what is most likely to happen as a consequence. The wide area protection and control, however, require not only actionably insight, but also that this insight is generated fast enough such that there is still time to react. Prescriptive analytics is what yields answers on the viable options according to the current situational understanding and what is known about system characteristics through predictive analysis. Prescriptive analysis within the required reaction timeframe for a complex and dynamic system such as the power network is probably one of the most demanding Big Data analytics scenarios today. It is foreseeable that with increased digitization and communication, the different notions of "real-time" will be better exploitable by the various advanced analytical capabilities. Many of the advanced analytical techniques such as diagnostic and predictive analysis mainly rely on historical data. However, Prescriptive analytics, i.e., the analysis that returns options to act upon, is most useful when it produces fast answers on reliable fresh data.

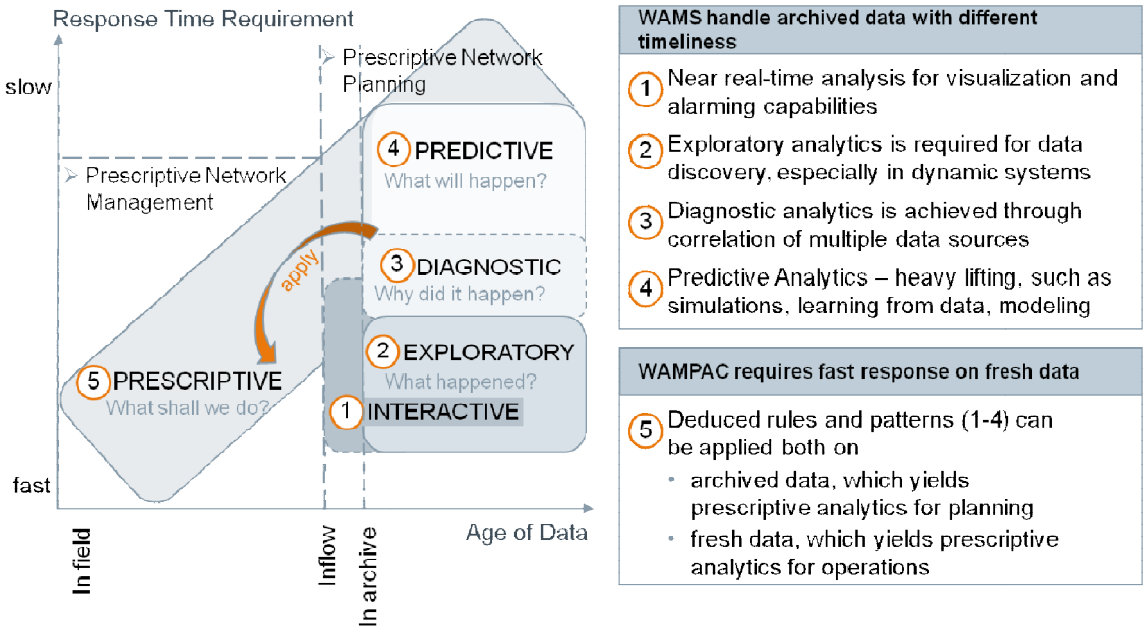


Figure 10 Different notions of "real-time" and advanced analytics, in the example of a Wide Area Monitoring Protection and Control (WAMPAC) system for power networks

Technology has evolved through Big Data natives and early adopters, the open source ecosystem enables spill-over into traditional businesses. The momentum of Big Data is still very high – and technologically there is room for improvement, but the real economies of scale for Big Data technologies will only be realized once there are best practices standardizable patterns.

The Big Data technologies have evolved by Big Data innovators and early adopters – but they are not matured – and as such they do have a latent potential for disruption: Without established standards larger corporations will need to review existing processes and IT governance to keep up with the technological advancements. The cost of the new governance will influence the appraised cost-efficiency of Big Data technologies when utilized in bigger scenarios and organizations. A whole new dimension to this discussion opens up when the Vertical IT segment is considered. Deep domain know-how on the peculiarities of cyber-physical systems is required to be able to keep the cost-efficiency features of Big Data technologies for operational intelligence. Further discussion of these aspects will be given in the Section "The Challenge of Choosing the Right Technology." The following section discusses the architectural patterns on the evolution path of Big Data technologies.

5.2. Architectural Patterns

The extraction and review of architectural patterns is important for the further discussion on data usage, specifically data analysis. For the longer part of history, data analysis has been set equal with business intelligence, and data warehousing with special purpose data marts have been the way of preparing data for reporting (see Figure 11).

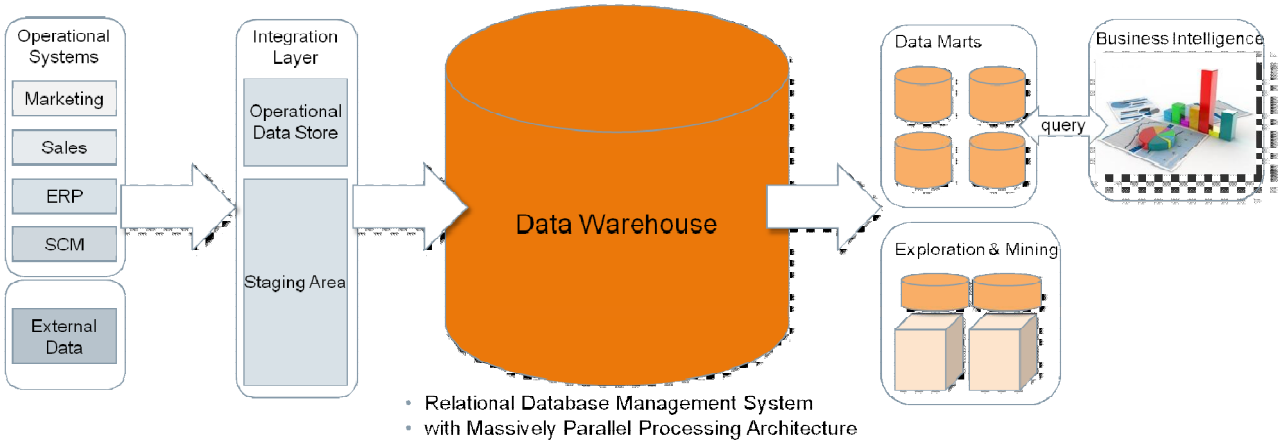


Figure 11 Typical data warehouses employ "schema on write," also known as ETL, extract-transform-load. The integration of the variety of data sources is feasible as long as data is well known and business is static.

However, similar to the mainstreaming of distributed computing concepts to handle the masses of data, advanced concepts from machine learning and data mining reach a tipping point. Analytics ceases to be a task at the end of an ETL (extract transform load) process, but increasingly becomes embedded in to the very core of data handling.

This section on architectural patterns reviews the technological advancements from the perspective of how advanced analytics of data is enabled.

5.2.1 Schema on Read

Schema on read is not a new concept either – but with distributed file systems with parallel processing frameworks such as MapReduce, schema on read becomes pivotal for cost-efficiently handling unstructured data.

The trend of moving from ETL to ELT started more than a decade ago when more and more relational database management systems (RDBMS) became cluster-capable. Running a transformation process step on multiple servers results in considerable speed up. With traditional ETL, every addition of a new data source meant going back to the drawing board. If there was an error in the ETL logic, correction takes weeks depending on when the error was discovered. If the logic has been wrong for months then all those past transformations have to be redone, whilst continuing the current work. Parallelizing this process over a cluster of servers eases this pain by firing up additional servers to handle increased work load – which is also known as elasticity in the distributed computing domain. Hence, the transformation of the ETL step moved into the RDBMS, resulting in the loading first and transforming later logic of ELT. However, the increasing masses of data and, especially in the online businesses, the unstructured nature of data resulted in a degrading query performance of RDBMS because of the heavy load of transforming the unstructured data stored in relational databases as binary large objects.

This is where a distributed file system like Hadoop has advantages over relational databases to better perform ELT with unstructured or multi-structured data. The data can be stored very cost-efficiently in its original form. Through the schema on read capability one or multiple data models can be efficiently applied with the MapReduce parallel processing framework – on read, or within the transformation process. The advantages with respect to challenges related to Big Data include, but are not limited to:

- The data integration logic becomes part of the application logic and can be adapted as the variety of data sources grows, which increases the agility of analytics.
- Multiple data models can be applied, which is especially useful for multi-structured data or for the exploration of unstructured data and data discovery.
- Unknown features of the data that may become relevant in future applications are still accessible, and can be harvested once the appropriate data model is learnt.

5.2.2 Lambda Architecture

"*The lambda architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer*" – Nathan Marz[32].

2013 is the year when some consolidation into the Big Data technology stack arrived: The so-called lambda architecture. A key aspect is that this architectural style acknowledges the very different challenges of volume and velocity. The data handling is split into so called speed layer for real-time processing of streaming data and the batch layer for cost-efficient persistent storage and batch processing of the accumulations of streaming raw data over time. The serving layer enables the different views for data usage. Figure 12 is an adapted depiction of the lambda structure to show also the architectural sub-patterns that enable the required characteristics of each layer:

1) The batch layer is a cost-efficient active archive of big raw data based on parallel processing frameworks. The cost-efficiency is a defining characteristic in dynamic complex businesses with massive amounts and various sources of data. That data needs to be affordably stored, archived, and processed in its raw format. Massively parallel processing systems, including cluster-capable RDBMS for structured data or distributed file systems with parallel processing frameworks such as MapReduce for unstructured data, are viable choices for this layer.
2) The speed layer is a flexible, fault tolerant topology of servers for stream computing. The pendent of schema-on-read in stream computing is the ability to flexibly configure a distributed topology of servers that can process the incoming streams of data according to the logic required for presentation of the user defined business questions. The topology management of a cluster requires a distributed messaging solution for cost efficiency: the distributed messaging solution realized the flexibility, durability, and fault tolerance characteristics of the stream computing topology. The actual stream computing logic, i.e., the assigning of processing tasks to each of the nodes and the hand-over logic is based on distributed computing for cost-efficiency reasons.
3) The serving layer prepares access to the views defined by business questions. The views are generated both by the batch and speed layer according to the business questions that are formulated by the users. Whilst the views from the batch layer consists of pre-computed and recomputed data sets, the views from the speed layer are continuously updated meaning the views require a data handling solution that support both fast random reads as well as writes. For this purpose linearly scalable distributed data management systems, such as distributed DBs on top of distributed file systems or key-value stores based on distributed hash tables are the right state of the art choice.

The lambda architecture is based on the fact that with Big Data there is no one size fits all: Especially volume and velocity call for a well orchestrated architecture, which balances the trade-off between speed and cost of byte optimally with the current state of the art.

[32] http://www.manning.com/marz/

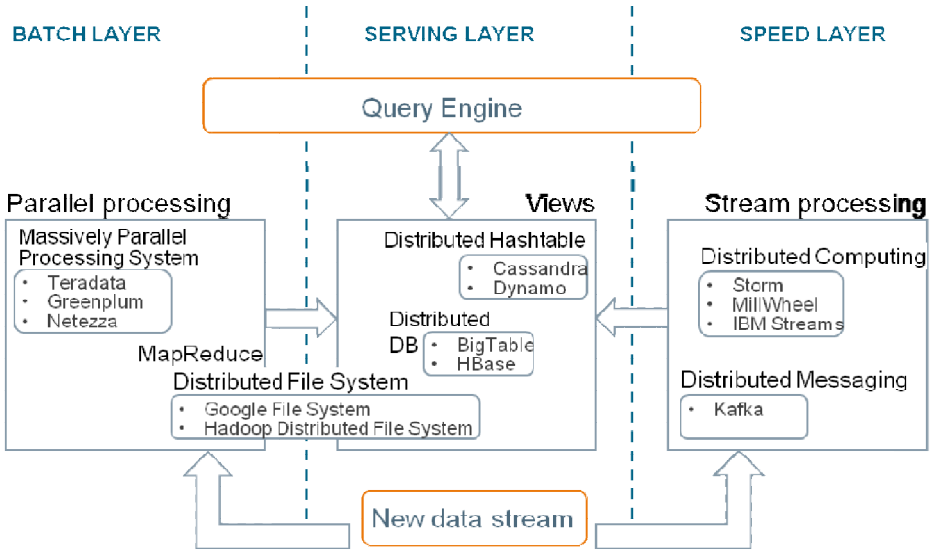


Figure 12 The lambda architecture with three layers: batch, speed and serving layer; adapted to display the architectural sub-patterns for each layer; data analytics requirements not explicitly covered

The Lambda Architecture is also a suitable pattern for the industrial businesses to build upon, but it is incomplete with respect to the data acquisition and usage challenges of industrial businesses, e.g. energy automation, are not covered.

For the discussion in this white paper, the different requirements for data usage must be emphasized: Especially in industrial settings there is a lot of heavy lifting involved, e.g. simulations and forecasts, optimization etc. They come with a whole new set of challenges that must be absorbed by the Big Data technologies. Many of the machine learning and data mining algorithms are not straight forward to parallelize. The serving layer will need to provide a flexible and extensible abstraction, which we call more specifically the *Analytics Engine*. This layer also needs to serve other analytical tasks such as manual data discovery, or automated pattern matching and as such satisfy the different timeliness requirements. Additionally, it must be noted that the data acquisition, which is implicit in the lambda architecture, is an essential part in industrial settings and cyber-physical systems. The lambda architecture only considers the inflow and in-archive data, and hence the real-time capabilities can only be near real-time. The fresh data in-field, which is available in cyber-physical systems, has no presentation in online data businesses yet and hence in the original lambda architecture. The next section briefly introduces the in-field analytics architectural pattern, which has the potential of delivering real-time analytics in cyber-physical systems.

5.2.3 In-field Analytics

Intelligent electronic devices in cyber-physical systems are both the data acquisition as well as the value delivery point. As such it is an evolutionary process to include these in the real-time value generation from Big Data in industrial businesses.

Industrial businesses must not only cope with and reap value from the increased digitization of business processes, but also from the increased digitization of the very products and services they offer. Especially in cyber-physical systems such as energy automation and industrial automation, but also increasingly in intelligent transportation systems or value-based healthcare, the intelligent electronic devices in the field are a crucial part of the system. Those devices are equipped with ever increasing computing capabilities, and are both data acquisition points as well as value delivery points. As discussed before, prescriptive analytics is not only about extracting actionable information through analytics, but also about analyzing options to act upon. In cyber-physical system the time span for action is in the sub-second range, meaning that the speed layer of the lambda architecture is not sufficient. Although the speed layer is engineered such that millions of insights can be generated within milliseconds, the time count only starts after the inflow of data into the enterprise. The latency between the actual capture of a situation in the data in the field and the streaming data analytics at the enterprise level would detriment the timeliness that is needed for prescriptive analytics in cyber-physical systems.

With respect to the different notions of real-time, the lambda architecture is certainly suitable to use prescriptive analytics for real-time business decision making or for economic planning of cyber-physical systems. For operational optimization of cyber-physical systems this is not the case, because prescriptive analytics must be applied on fresh and accurate data as close to the event as possible, both in terms of time and space. It may still not be hard real-time, but certainly

within hundreds of milliseconds from the event. For achieving such timeliness, the in-field analytics is shown in Figure 13 as only an example of how the architectural patterns will continue to evolve and be adapted to the different business needs. The major addition will be the explicit addition of a data acquisition layer with following characteristics:

- The data acquisition layer synchronizes patterns from the long-term memory of business analytics and monitoring. The key aspect will be the realization that once patterns and dynamic rules are extracted through advanced analytics in the enterprise backend, these pieces of systemic knowledge will be planted into the intelligent electronic devices in the field. With this knowledge and distributed computing capabilities in-field analytics can yield fast insights from fresh data, which will be needed for prescriptive analytics in operational management of cyber-physical systems. The main purpose of distributed computing here is the forming of locality and domain aware computing overlays for assuring data redundancy and data quality as well as performing prescriptive analytics in the field.

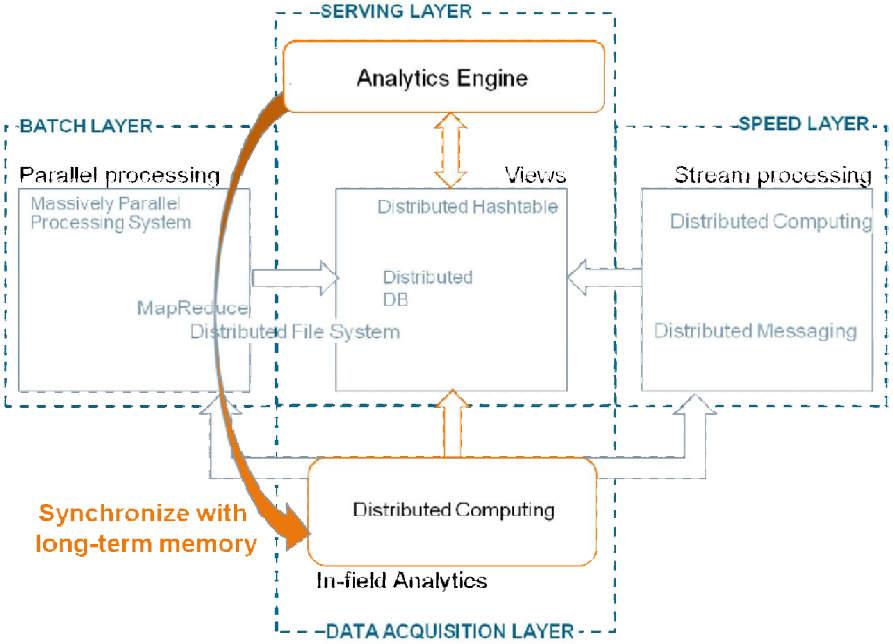


Figure 13 Cyber-physical systems include computing resources in the data acquisition layer - In-field analytics can tap into this potential and deliver insights fast enough for actual prescriptive analytics

This differentiation from Big Data processing scenarios in online data business to scenarios in cyber-physical systems is further discussed in the next section in the context of the Big Data refinery pipeline. As already mentioned briefly here, the data acquisition and data usage phases in industrial cyber-physical businesses differ considerably from online data businesses. Applying advanced analytics is not only a means to an end, but is embedded into all phases of data refinement. For in-field analytics, as an example, the synchronization of extracted knowledge is essential and requires an Analytics Engine. This layer can be understood as a toolbox, and will further be detailed in the next section on choosing the right tools.

5.3. The Challenge of Choosing the Right Technology

"Agile is too slow" – Tom Davenport[33]

One important aspect is that there are no right or wrong technologies, but the right choice of technologies. Most importantly, during this time of fast technological advancements in the Big Data arena, it is likely that the right choice of today might be out-dated in a few years. Additionally, industrial stakeholders, as opposed to Big Data natives, might be forced to use Big Data technologies in ways they were not designed to – and eventually come to a point at which adaptations to the domain specifics are required to remain cost-efficient. At that point, it will be decisive to have gathered sufficient know-how and skills.

[33] http://digitalcommunity.mit.edu/community/featured_content/big-data/blog/2013/07/16/tom-davenport-reports-on-big-data-in-big-companies

Especially big companies need to be aware of that even "agile is too slow" for Big Data and that constant updating is becoming routine[34]. The architectures and underlying business processes for governance will need to be reviewed continuously to assess technological advancement and adaption if they increase user value or data opportunity substantially – or increase cost-efficiency.

Although theoretically being around for many years, Big Data technologies found their way into mainstream quite fast: from being applied in niche domains, such as file-sharing, to being put to use for competitive advantage in all large online businesses within a single decade. As discussed before, the technological advancements, the waves, have been building up approximately every three years. However, industrial businesses have a modus operandi: This means systems, and customers that have not been changing nearly as fast, leading to a greater the Big Data technology push than the market pull from industrial businesses. Nonetheless, even if not every wave will be adapted and find its way directly into industrial businesses, they will be experimented with and analyzed. The suitable adapted sets of technologies with the right balance of innovation and return on data will most probably find their ways into the day-to-day business of more traditional industries. As with industries that have undergone digitization, such as telecommunications and the media, once the tipping point of digitization is reached, technological adaptation can be very fast driven both by new players and incumbents alike. The section on "perpetual beta, or always improve your system" gives a discussion of possible setups to be agile enough for the change that is certain. Before such a discussion can be started within organizations, it is important to understand what Big Data really means in industrial businesses and how the value generation is affected by the very nature of cyber-physical systems that are at the core of digitized industrial businesses.

5.3.1 Big Data Refinery Pipeline

"Analytics inside"

The Big Data refinery pipeline for industrial domains consists of three main phases: data acquisition, data storage, and data usage. Data analytics is implicitly required at all steps, and is not a means in itself. During data acquisition analytics supports the data cleaning and ensures data quality: For sensor data, analytics can enable efficient event-specific data compression and increase data quality through anomaly detection at acquisition time. For data usage, business and engineering questions are formulated and translated into suitable models. These models among other aspects imply which types of mass storage and data handling are cost-efficient for the specific purpose.

Each step of the pipeline refines the data, and the methods of refinement vary depending on the data type.

- Data Acquisition Phase: Each step of the pipeline refines the data, and the methods of refinement vary depending on the data source. At the data acquisition step, security and privacy or confidentiality policies can be applied; the data can be checked for quality features, such as missing data or implausible data in time series. Depending on whether data is generated during an automated process, or by a human, e.g. repair crew in the field or customer call agent, data is highly structured or unstructured; it is acquired as continuous streams of high frequency samples or sent irregularly. The methods for analyzing the data, acquired in very different ways and formats, would reach from content analytics and natural language processing for unstructured data to signal and cross-correlation analysis on structured time series data.

Industrial data sources have considerable variety: Not only are all data types that are also relevant for online businesses increasingly becoming available, but also there are huge streams of high-resolution data from sensors and intelligent electronic devices embedded into cyber-physical systems.

- Data Management Phase: Multiple types of storage may be considered for the same data source type depending on for what it will be used: Graph databases can yield faster results for geo-spatial analysis, key-value stores may prove most efficient for feature discovery purposes in time series. Regarding the generation of value from a variety of data sources the data storage step needs to enable the cost-efficient integration of the different data sources in an extensible way. Flexible abstractions, multiple layers of abstractions for data and analytical models will be required and implemented for the data storage step. Lightweight semantic data models that represent the multiple links within various data sources, such as Linked Data, may provide such cost-efficient abstractions, especially when the data domain is complex and highly interconnected. For high-dimensional data, multi-dimensional data structures such as tensors and space-filling curves support the design of more efficient analytics algorithms, which are capable of exploring multiple relations. These are only some examples of a few methods from machine learning, semantic data technologies, information extraction etc. What will be needed is a flexible toolbox that can be built with increasing business need to gain more value from ever increasing

[34] http://digitalcommunity.mit.edu/community/featured_content/big-data/blog/2013/07/16/tom-davenport-reports-on-big-data-in-big-companies

amounts and variety of data. The next section discusses such an Analytics Engine that enables the creation and linkage of an analytical toolbox with the underlying Big Data refinery pipeline.

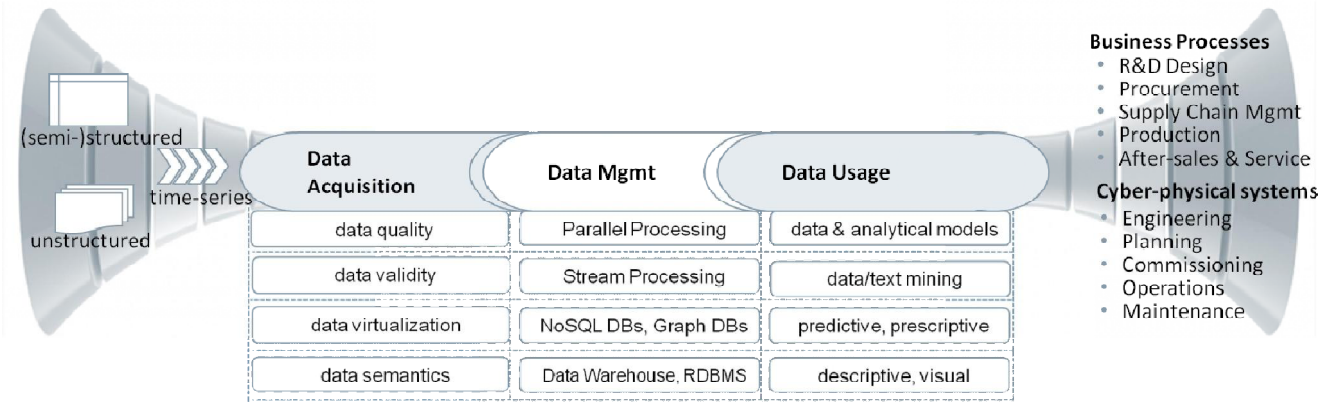


Figure 14 The Big Data processing pipeline refines data at each step, but the methods of refinement vary depending on data source and its intended usage

Although data usage is the last step in the refinement process, the business and engineering questions formulated in this step are the starting point for choosing the technology stack. At the same time the type of the required data sources determines what type of methodologies will be cost-efficient.

- Data Usage Phase: Although data usage is the last step in the refinement process, the business and engineering questions formulated in this step are the starting point for choosing the technology stack. This step represents the reason for setting up a particular connection through the pipeline to the required data sources. In the industrial domain, analytics for data usage, reaches from descriptive analytics such as dashboards, to predictive and prescriptive analytics such as forecasting for portfolio management or simulation of interconnected systems and extraction of operational intelligence. In the B2B segment, the operational intelligence covers both, how product and solutions are delivered, as well as how they are utilized in the customers' operations. This means operational efficiency as main usage scenario, covers aspects of both business processes as well as processes in cyber-physical systems as depicted in Figure 14. Additionally, operational intelligence will tightly be coupled with strategic intelligence, covering areas from financial reporting, competitive and market intelligence etc. Of course, in the industrial domain, too, Big Data usage scenarios rise from strategy, or marketing and sales departments. However, even then they will be tied to the physical product of the company division, e.g. fleet intelligence on gas turbines or magnet resonance devices.

Industrial data usage also requires considerably more precision than data usage in online data businesses. The business value generation is versatile, too, since it not only covers the provisioning of a product but also the usage of that product in B2B as well as B2C settings.

Such connectedness with the sources of data may be the main differentiator from Big Data usage scenarios in the online consumer facing businesses. If data acquisition and analytics do not deliver actionable information with available options fast enough then the value of data may even be negative. Eventual consistency for cyber-physical systems is not an option. But if they do, then the economic welfare is much greater than online businesses can ever deliver.

Due to the major differences in and importance of data acquisition and usage for industrial businesses, adaptation of Big Data architectural patterns and technologies will be necessary.

5.3.2 The Analytics Engine

"How cool is it to imagine a machine that can combine deductive and inductive processes to develop, apply, refine and explain a fundamental economic theory?" – David Ferrucci[35]

Potentially millions of insights per second await industrial businesses, when the breadth and variety of data sources can be refined and used for fast or even automated decision making. Such a capability requires the seamless integration of analytics into each step of the Big Data refinement pipeline.

[35] http://bits.blogs.nytimes.com/2013/05/06/david-ferrucci-life-after-watson/

The Analytics Engine for integrating advanced analytics into the data refinery pipeline is designed such that both flexibility and extendibility are still feasible: Through data discovery, visual analytics, machine learning, information retrieval, and data mining, the incremental understanding of the data becomes possible. Once the appropriate data and analytical models are developed that portray this understanding, schema on read can be utilized to apply the improved models onto the data refinery pipeline. This Analytics Engine will assist in implementing the domain- and device-specific adaptations to Big Data management in a cost-efficient and innovative manner.

The type of data sources and the form of insight delivery – as we also discussed in the Chapter "Data Perspective – Know your Data"– has implications not only on business relevant aspects such as data ownership, but also strongly affects the right choice for the tool chain. For example, efficient acquisition, storage, and usage of time-series sensor data differs from the tool chain required for retrieving knowledge from unstructured data sources such as inspection reports of field crews as briefly discussed in the "Big Data Refinery Pipeline" Section. As formulated in the discussion of the "Lambda Architecture" there is no one type of technology that can efficiently handle both volume and velocity of data. Similarly, there is no one analytical tool that can assist in the analysis of all sorts of data, especially analyze correlations across various data sources.

There have been two somewhat divergent opinions regarding analytics in the Big Data dispute: (1) "Having more data beats out having better models"[36],[37] and (2) Not the data but the model brings forth the benefit of using the data[38]. A model represents the fundamental understanding of a system, including the assumptions. Models are informed by data but not blindly driven by data. In some applications in power systems, for example, models even help in identifying bad data from poorly linked sensors. On the other hand, real-time measurement that capture system characteristics in very high resolutions can also help to improve bad or outdated models of dynamic and complex systems.

Getting insight into a large amount of data at once and making the right decisions needs the combined abilities of computers (fast and accurate) and humans (intuitive and brilliant). In fact Visual Analytics is more than just visualization and can rather be seen as an integrated approach combining visualization[39], human factors[40] and data analysis[41]. With respect to the field of visualization[42], Visual Analytics integrates methodology from information analytics, geospatial analytics, and scientific analytics. Especially human factors (e.g., interaction, cognition, perception, collaboration, presentation, and dissemination) play a key role in the communication between human and computer, as well as in the decision making process (Keim et al., 2006). Figure 15 illustrates the interactive Visual Analytics approach. (Weber et al., 2012) compares a selection of state-of-the-art commercial Visual Analytics frameworks in detail.

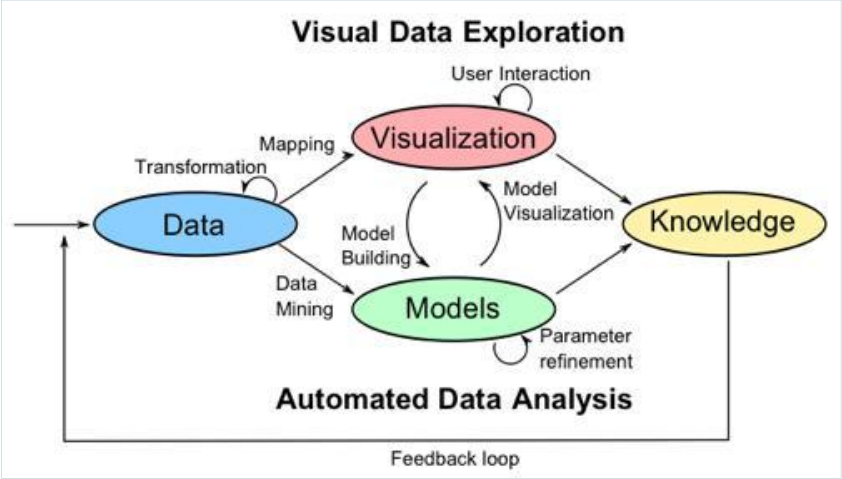


Figure 15: Visual Analytics approach (Keim et al. 2010)

The examination of Big Data in industrial, especially in cyber-physical systems show, that these views must be combined:

- A model for cause and effect analysis and the description of all knowns and unknown knowns is fundamental and is developed by

[36] http://strata.oreilly.com/2012/01/what-is-big-data.html
[37] http://data-informed.com/why-more-data-and-simple-algorithms-beat-complex-analytics-models/
[38] http://readwrite.com/2013/05/20/blinded-by-big-data
[39] http://www.infovis-wiki.net/index.php?title=Visualization
[40] http://www.infovis-wiki.net/index.php?title=Human_factors&action=edit&redlink=1
[41] http://www.infovis-wiki.net/index.php?title=Data_analysis&action=edit&redlink=1
[42] http://www.infovis-wiki.net/index.php?title=Visualization

- Simulation and optimization including the usage of real-time system data
- Correlation of data from various sources capturing the dynamic and static system characteristics as well as its environment is crucial to discover unknown unknowns, by:
  - Data mining and machine learning methods
  - Analysis through data discovery
- Adaptive data models can be realized by combination of semantic technologies and schema on read to describe and deploy the new knowledge about the data.
- Combination of human cognition with computational power by
  - Visual Analytics by data scientists.

These combinations of advanced analytics in an extendible toolbox can be facilitated by an Analytics Engine as depicted in Figure 16:

Not the data but the model brings forth the user value – but having more data and advanced analytical capabilities will enable continuous improvement of models.

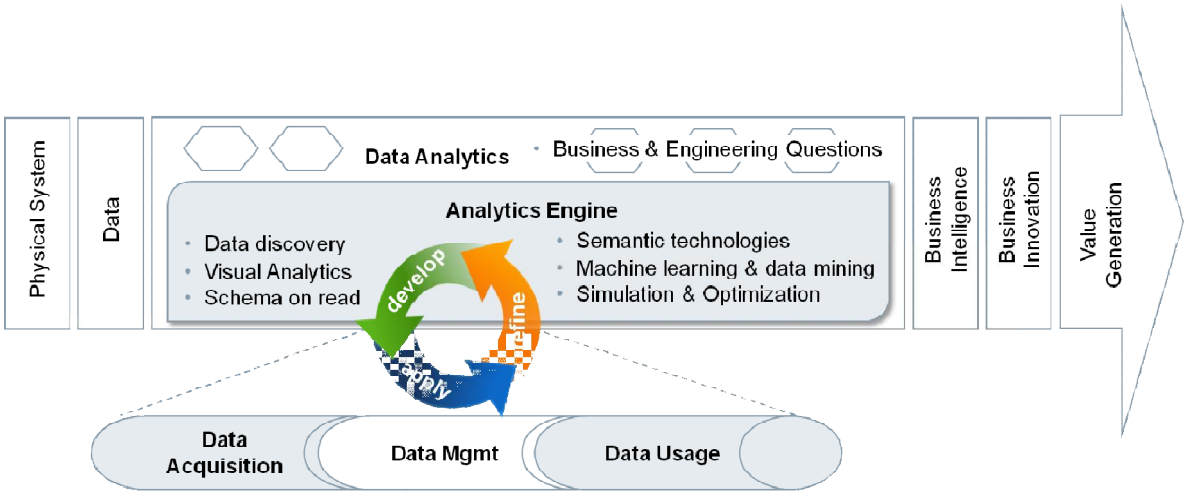


Figure 16: The Analytics Engine will enable the utilization of the required technology stack depending on the requirements of business question formulated for business or operational intelligence and the types of required input data

Schema on demand enables the continuous discovery, enhancement, application and refinement of models that better represent the physical system that the data captures. Such continuous improvement enables never before available precision of automation.

The new economics of computing efficiently enables the loading of data first and modeling it later[43,44]. This flexibility also allows changing the data and analytical models, testing multiple models, and improving it at no additional cost. So one should start with formulating the business or engineering question, identify a suitable yet simple model, observe and test it as more data comes in, and make adjustments as necessary. In order to do this in a cost-efficient way, we add the Analytics Engine between the user defined functions and the scalable mass data storage. The Analytics Engine will enable the utilization of the required technology stack depending on the requirements of business question formulated for business or operational intelligence and the types of required input data. This layer will achieve the mapping between data and the analytical models by what is called Schema on Read. Secondly, the layer not only enables the data discovery and testing of quantitative analysis performed by data scientists and quants but also large scale machine learning and data mining on very large data sets to arrive at better models with lesser bias[45], which one may call "human-machine hybrid computing[46]." The Analytics Engine is an abstraction required to cost-effectively combine deductive and inductive processes to develop, apply, refine and explain the models for data representation as well as for data analytics[47]. With this abstraction layer, we can always improve our representation of data based on our increasing knowledge about the data.

[43] http://tdwi.org/Articles/2013/10/15/Load-First-Model-Later.aspx?Page=2
[44] http://blog.cloudera.com/blog/2013/02/big-datas-new-use-cases-transformation-active-archive-and-exploration/
[45] http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/
[46] http://www.wired.com/business/2013/10/next-big-thing-economic-data/
[47] bits.blogs.nytimes.com/2013/05/06/david-ferrucci-life-after-watson

5.3.3 Perpetual Beta – Always Improve Your System

"Any organization thinking of simply applying existing information governance practices to Big Data will likely fail." – Gartner Inc[48].

Industrial B2B scenarios are typically much slower paced than online consumer facing applications. Hence, industrial players will not re-engineer their legacy systems to reap the value of Big Data, nor is it necessary. However, they will need to create a parallel staging system where they can gradually employ new technologies and acquire experience and skills required to compete in the Data Economy.

Big Data technologies open up a new universe of possibilities – but also bring along a whole new challenge: there are no standards, almost no best practices. ITIL, the set of practices for IT Service Management, is grounded in mainframe technology, where inputs and resources have been well known and defined for ages[49]. Big Data technologies represent a paradigm shift and do bring disruption to these established processes. However, this is only a transitional phase. There will be standard ways of making use of the powerful aspects of Big Data technologies, such as the linear scalability, fault tolerance, and high-performance on commodity infrastructure. However, it is hard to predict how long such standardization will take. The transition phase for databases in the 1960s lasted for more than a decade until standard ways of utilizing the technology, i.e., SQL, as a standard database query language, could be formulated. In the transition phase companies will need to acquire skills and experience with the new technologies. In the Big Data era these are: semantic data integration, streaming data processing, scalable mass data storage and massively parallel processing or distributed computing, and eventually advanced analytics utilizing all three V's of Big Data in cost-efficient and innovative ways.

"When somebody tells me 'I'm very interested in Big Data and I think we can find more information in our data,' that's going to be a long sales cycle," – Eric Baldeschwieler[50]

Similar to the spiral approach of well defined projects in section "Big Data Business Approaches," which ultimately result in an efficient Big Data project portfolio, we propose to emulate the stepwise evolution from batch to stream processing to advanced analytics in a parallel Big Data staging system. Rigorous iterations on this parallel system will be done based on cost-efficiency of every component: A new data source may imply setting up a new tool chain, e.g. for natural language processing for written reports of the field crew or a new user defined function, i.e. a formulated business question, may require the analysis of multiple data sources in new combinations. Such new analytical models and their continuous improvement process should be defined and implemented in the Analytics Engine (see Figure 16).

One cannot start a Big Data project within existing constraints of legacy systems – Instead a mirrored parallel Big Data staging system is required which enables the assessment and adaptation of innovative technologies within in all steps of the data refinery pipeline. Processes and IT governance need to be reviewed to enable the realization of such a Big Data staging system that can flexibly substitute technology components based on their efficiency to translate data opportunities into user value.

But how does one start with setting up such a parallel Big Data staging system? There cannot be a "small pilot," but one must look for options that do not require re-engineering entire processes, which support the business-as-usual. Figure 17 depicts a possible route to setting up a parallel Big Data staging system following an extended spiral model: each version of this architecture must be supported by at least one business case, i.e., either supported by new business question, or new data opportunity. Whether the business case is feasibly will be very much determined by the cost-effectiveness of the technologies employed. So here again the "spiral" of versions is determined by all three dimensions: business, data, and technology.

[48] http://www.dataversity.net/best-practices-for-big-data-goverance/
[49] http://blogs.forrester.com/jean_pierre_garbani/09-06-26-my_issue_itil
[50] http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/2/

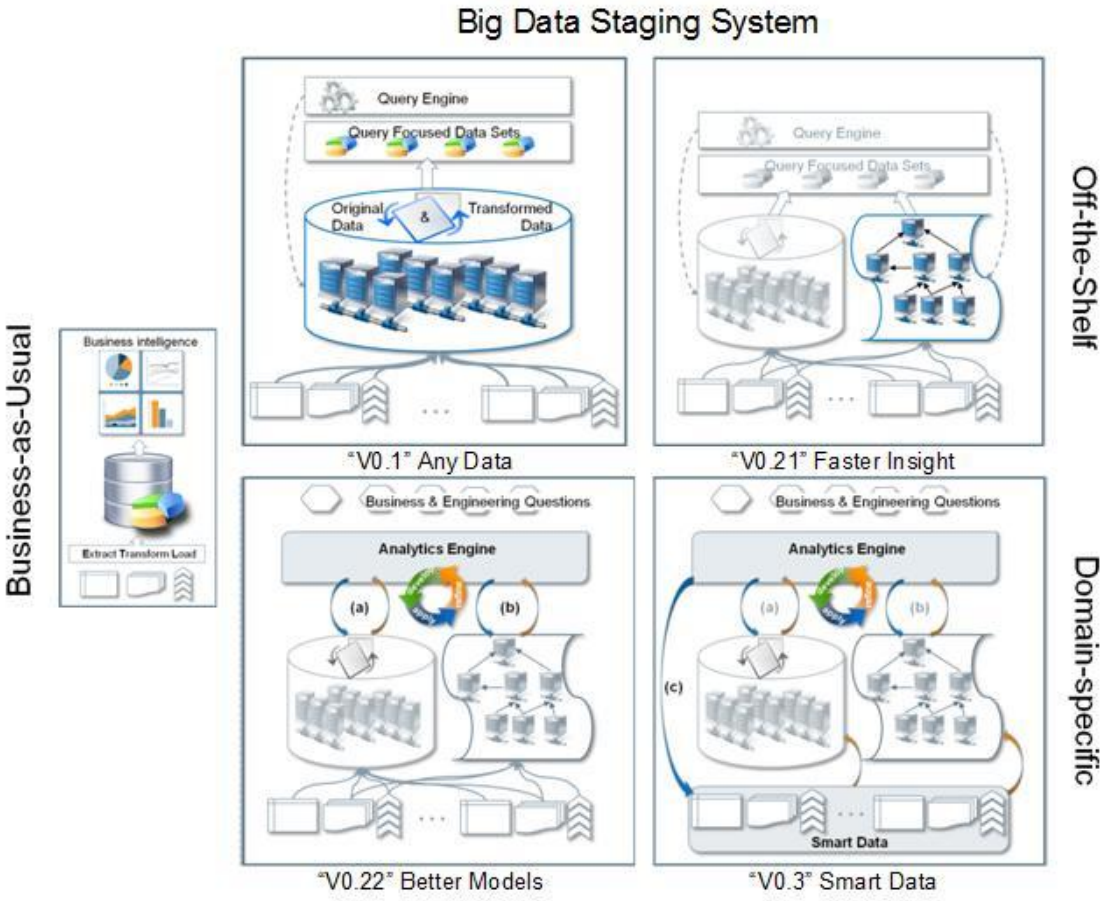


Figure 17 The setup of a parallel Big Data staging system in which technological advancements, data opportunities, existing and new business questions can be examined is a living iterative approach – the Extended Spiral Model. The different architectural versions can be dubbed "Any Data," "Better Models and Faster Insight," and finally "Smart Data"

"Version 0.1" – Any Data

Recall the four quadrants of a "Big Data Business Approach" in Figure 5: The low hanging fruit category is to make existing business better with available data, i.e., by increasing efficiency and visibly saving money. This could be by reducing computation time of analytical tasks, or by improving existing analytical models through the addition of more data points for yielding better predictions. Important is that the mass storage scales linearly with the volume of data and the framework supports massively parallel processing or distributed computing. These will be the important factors that determine the cost-efficiency and ultimately the business case for such a project. This initial step of mirroring the available data from the legacy system to a scalable mass storage in the parallel Big Data staging system should clearly demonstrate the new economics of computing.

"Version 0.21/0.22" – Better Models and Faster Insight

An Analytics Engine could initially employ simple methods for detecting data quality issues, and domain-specific anomalies. Eventually grow to its full potential for mapping flexibly between analytical models and data models, and applying automatically improved models by the schema-on-read feature as discussed in Section "The Analytics Engine." Much domain know-how and abstraction skills will be required to setup an easy-to-use environment, such that business and engineering questions can be formulated as user defined functions to run on the Analytics Engine. Many of the analytical methods are very hard to parallelize and compute in a distributed manner. This, for example, is an open research issue in the Big Data domain (see also Section on The Future of Big Data Technologies and Research).

For the time being, the challenge will be mastered by close cooperation of data scientists, domain experts, and data engineers and resulting innovations as discussed in Section "Organisational Implications". The Analytics Engine will be necessary to realize projects in the categories that create new business with available and new data sources by enabling cost-efficient addition of new user defined functions.

Once analytical capabilities are setup, they will also be used on the data as it streams in real-time. A second iteration of the parallel Big Data staging system should be based on adding stream processing capabilities. The scalable architecture for real-time analytics will be the dynamo behind many of the projects that represent a break-through (see Section Big

Data Business). Especially in industrial automation, such as Energy Automation, the real-time streams of sensor data must be analyzed and interpreted fast enough to leave time for decision making such that (semi-)automated process can be optimized during run-time. Energy efficient factories or plant and substations automation are examples of how complex causations and correlations represented in multi-relational sensor data need to be analyzed in the blink of an eye to present options to an operator or for the next process steps. This is the holy grail of operational intelligence. Such intelligence will not only enable industrial players to improve their product and solution offerings for automation, but also will enable their customers to improve the operations, in planning and at execution time.

"Version 0.3" – Smart Data

As we discussed, data freshness and fast response times are crucial. One peculiarity of industrial automation is that data acquisition is not just carried out by simple sensors, but by intelligent electronic devices[51]. These devices have enough computing power, memory, and networking capabilities in order to be able to realize in-field analytics by applying patterns and dynamic rules on streaming data as it is being captured in the field. A strong vertical IT solution is required that is backed by the above depicted Big Data system with an Analytics Engine to support such in-field analytics. Big Data projects, which utilize the entire refinery pipeline, including the data acquisition infrastructure, will realize cost-effective, innovative forms of information processing for enhanced insight and decision making. This will enable the creation of Big Data portfolios with break-through projects.

Finally, to cope with the challenge that Big Data technologies will show fast advancements – even more so when adopted within the industrial domains – each component in this Big Data parallel staging system is substituted when innovative and more cost-effective technology arrives. In some sense we propose to apply the spiral model for rapid prototyping of software, for the iterative enhancements of the entire architecture. Each technology iteration will then be marked by the business case of the addition of a new user defined function, i.e. business or engineering question, or the addition of a new data source. For each such addition, new technology components will be probed and substituted, if this increases the cost-efficiency of the project. Conversely, if the current state of the art as well as own innovations and adaptations do not enable the cost-efficient handling of a new data source or business question, then the business case for that project probably is not yet mature. However, as discussed in the Chapters Business and Data Perspectives, the technological maturity may not be the only reason for a weak business case: the ecosystem and organizational processes as well as the regulatory framework regarding data access and sharing should always be taken into account.

[51] http://blogs-images.forbes.com/louiscolumbus/files/2012/08/Hype-Cycle-of-Big-Data.jpg

5.4. The Future of Big Data Technology and Research

Multiple paths are possible: Technologies to cope with the impact of Big Data will concentrate on handling volume, velocity, and variety by decreasing cost of computing power, storage, and communication. In addition innovation will drive technologies that reduce the impact of the 3 V's by intelligent data reduction, i.e., smart data, by semantic technologies and smart schemas, as well as smart algorithms.

2013 has been the year that Big Data, the terminology, became mainstream[52]. Hadoop, as the open source pendant to Google's framework, has achieved a substantial user base and ecosystem. Many other open source NoSQL databases such as Cassandra have gathered substantial attention. Many of the IT incumbents advanced their own offering, innovated around their core technologies, or even announced some sort of support for Hadoop, to decrease the cost of data handling considerably.

The path is quite clear that, from 2014 out, IT incumbents, IT suit and solution providers will employ the paradigms beyond relational database management techniques, such as parallelism, or columnar access to data. There will be accepted best practices or even processes that successfully merge the SQL and the NoSQL world. This world will be easier to access than it is today in the midst of its disruptive creation – but it will still be more complex than what data management and data handling has been for most of our digital history.

So when companies build a parallel Big Data staging system, they will have been establishing alternatives to choose from, in order to experience the new economics of computing, its linear scalability. They will have been creating the bases to build the new skills required for the post-SQL world. Most importantly, they will have created the bases to build domain-specific value via Big Data analytics – which is the competitive advantage in the new Data Economy.

2014 is the year when the focus shifted to the new possibilities and challenges, once the masses of data are accessible in an economic way: Businesses are determined to make use and generate value out of the masses of data available to them – by means of real-time scalable data analytics. Insights for decision making, especially in operations and industrial automation, are more valuable the closer they are to real-time. Just knowing that something is wrong – but not having enough time to act upon that knowledge – is not a strong motivator for wanting Big Data or Big Data analytics.

So 2015 is the year when (near) real-time stream processing, visualization, and especially prescriptive analytics will be taken as the next technological frontier. The technological race in this domain is just starting. But with respect to prescriptive analytics, one cannot expect off-the-shelf solutions: Real-time stream data processing will enable having actionable information from data fast; but the options what to do with that information is the result of prescriptive analytics – and that requires the transformation of deep domain knowledge into analytical and data models, flexibly, in order to allow the continuous refinement as the knowledge about the data improves the knowledge about the systems that are being monitored and automated (see Section on "The Analytics Engine").

From this starting point we see two paths along which the Big Data technologies will evolve:

(1) Research and innovation will continue to concentrate on coping with the impact of Big Data: The integration and analysis of various data sources, with different owners and formats, will require "smart schemas," which annotate the data access with respective policies and mappings. The massive volumes of data will continue to foster hardware innovations such as GPU and multi-core computing. Communication advances, will ensure that these masses of data need to travel fast and securely from data acquisition to storage to usage. In dealing with massive data interactive visualization becomes more and more important to leverage the human perception abilities enabling to act upon the findings immediately.

(2) Research and innovation will revolve around reducing the impact of Big Data: with respect to variety of sources and formats, this would mean adding semantics and linkage to reduce unnecessary mappings between schemas that result in lost information. Regarding volume and velocity, in-network processing, for example, would enable the reduction of transferring all raw data to a central location. In many cases, it could already be decided at the source of data, which sequences of raw data are featureless and can be transferred and stored in a highly compressed way.

Smarter algorithms that can be deployed at each step of the Big Data refinery pipeline will assist in solving today's insurmountable problems, such as mining data at very large scale whilst preserving privacy and confidentiality. Homomorphic encryption would allow algorithms to run on encrypted data for example. But today only very simple data manipulations are possible. The farther away we try to look into the future the wider the funnel gets between these two paths. It is likely that research and innovation efforts will oscillate from one side to the other, opening up ever growing possibilities to realize the potential of data as an additional production factor.

[52] http://www.forbes.com/sites/louiscolumbus/2012/08/16/roundup-of-big-data-forecasts-and-market-estimates-2012/

References

Accenture (2012). Connected Health: The Drive to Integrated Healthcare Delivery. Online: www.acccenture.com/connectedhealthstudy

Barton D. and Court D.(2012). Making Advanced Analytics Work for You. Harvard Business Review, October 2012

BITKOM (2012). Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Leitfaden des BITKOM AK Big Data. Online: http://www.t-systems.de/loesungen/leitfaden-bitkom-big-data-2012/1014046_1/blobBinary/Leitfaden_Bitkom_Big_Data_2012_online.pdf

BITKOM (2013). Management von Big-Data Projekten. Leitfaden des BITKOM AK Big Data. Online: http://www.bitkom.org/files/documents/LF_big_data2013_web.pdf

C. Bizer, T. Heath, und T. Berners-Lee (2009). Linked data - the story so far. Int. J. Semantic Web Inf. Syst. 5(3):1–22

Davenport T.H. (2013). The Rise of Analytics 3.0. – How to compete in the Data Economy. E-Book from the International Institute for Analytics. 2013, online available.

Gartner (2013). Survey Analysis: Big Data Adoption in 2013 Shows Substance Behind the Hype online: http://www.gartner.com/newsroom/id/2593815

Iansiti, M. & Levien, R., (2004). The keystone advantage, Harvard Business Press, 2004

Keim D. A.,Mansmann F., Schneidewind J., Ziegler H. (2006). Challenges in visual data analysis, In Proceedings of the Tenth International Conference on Information Visualization, pages 9-16, 2006

Keim D. A., Kohlhammer J., Ellis G. (2010). Mastering the Information Age: Solving Problems with Visual Analytics, Eurographics Association, 2010

Klas Research (2012). BI Perception 2012 – A wave is coming. Perception Report. Klas Research, April 2012

Loshin, D. (2002). Knowledge Integrity: Data Ownership (Online) http://www.datawarehouse.com/article/?articleid=3052

Lünendonk Gmbh (2013). Trendpapier 2013: Big Data bei Krankenversicherungen. Bewältigung der Datenmengen in einem veränderten Gesundheitswesen. online available

McAfee A. and Brynjolfsson E.(2012). Big Data: The Management Revolution. Harvard Business Review, October 2012

McKinsey Global Institute (2011). Big Data: The next frontier for innovation, competition, and productivity.

NESSI (2012). Big Data – A new world of opportunities. NESSI White Paper, December 2012. Online available: http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf

Schomm F., Stahl F. and Vossen G. (2013): Marketplaces for Data: An Initial Survey. SIGMOD Record, March 2013 (Vol. 42, No. 1)

Smith D. (2013). Navigating Risk When Entering and Participating in a Business Ecosystem. Technology Innovation Management Review, May 2013.

Stoffel A., Zhang L., Weber S. H., Keim D. A.. AMPLIO VQA – A Web Based Visual Query Analysis System for Micro Grid Energy Mix Planning. Proceedings of the Eurovis workshop on visual analytics (EuroVA), Vienna, 2012.

Weber S. H., Mittelstadt S., Stoffel A., Keim D. et Al. (2012). Visual analytics for the Big Data era - A comparative review of state-of-the-art commercial systems, In VAST '12 Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pages 173-182, IEEE Computer Society Washington, DC, USA ©2012

World Economic Forum (2012). Rethinking Personal Data: Strengthening Trust, Industrial Agenda, May 2012

World Wide Web Foundation and Open Data Institute (2013). The Open Data Barometer 2013 Global Report, 2013. Online available: http://www.opendataresearch.org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf

Zillner S., Krusche B. "Systemisches Innovationsmanagement: Grundlagen - Strategien – Instrumente", Schäffer-Poeschel Verlag, Stuttgart, 2012

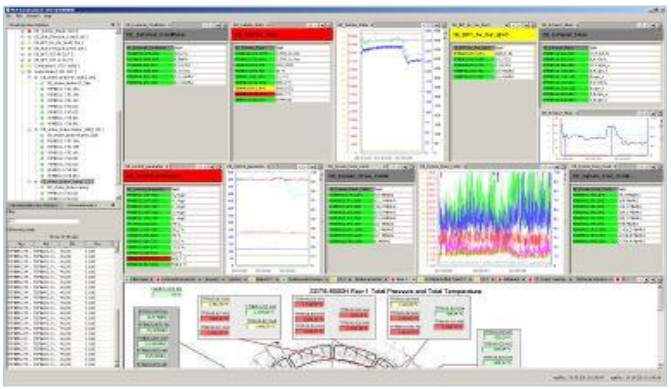
Zillner S., Rusitschka S. Munne R., Lipell H., Vilela F.L., Hussain K., Becker T., Jung R. Paradowsky D., Huang Y. (2013). D2.3.1 First Draft of Sector's Requisites. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4).

APPENDIX

Appendix 1: Big Data Use Case Examples

Optimization and Improvement

Example: Gas Turbine



What is it?

- Sensor data of a single gas turbine or a set of gas turbines of similar type are collected in order to generate and train mathematical models, such as neuronal networks
- The design of mathematical models is driven by concrete functional aspect in mind:
  - For example, *predictive maintenance* analysis is used to prevent future damage.
  - The mathematical models can also be used to improve the efficiency of the gas turbines through the *continuously learning from the data*
  - The model can also be used to design *soft sensors*, i.e. model that are able to calculate sensor values that cannot be measured in the physical world or are too expensive to measure. In order to find out why a turbine is running differently than expected, for instance, one can use the soft sensors to identify the cause of deviations in the process, in the material, etc.

Which data sources are used?

- Numeric data of sensor data that are collected by the gas turbine

How was the data used to create value?

- data in combination with the trained models can be used to produce insight that lead to improved overall system performance

Which technology was used?

- Neural networks
- Reinforcement learning
- Statistical methods

Constraints / issues this raises?

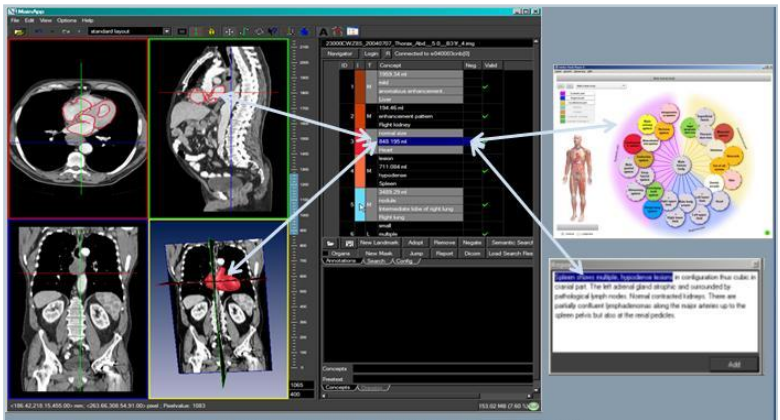
- data might not be available due to data sharing restrictions by gas turbine owner or legislation
- limited quality of data due to size constraints or storage restrictions

Source: Siemens AG, BAM, Learning Systems

Example: Seamless Navigation between Medical Image and Radiology Report

What is it?

- In the healthcare domain, the content information of medical images as well as dictated medical reports can be semantically described by metadata in order to significantly improve the radiologist working environment by enabling the seamlessly navigation between medical images and associated text reports.
- Clinicians and radiologist do no longer waste valuable time by manually switching between two different IT-Systems, such as RIS (Radiology Information System) and HIS (Hospital Information System). Instead, they can seamlessly navigate through image and associated text data in order to better understand the patient case.



When marking text in reports, associated diagnosis are highlighted in lists and images

Which data sources are used?

- medical images
- radiology report

How was the data used to create value?

- content of medical (e.g. findings) was described by semantic labels
- content about findings in radiology reports was described by semantic labels

Which technology was used?

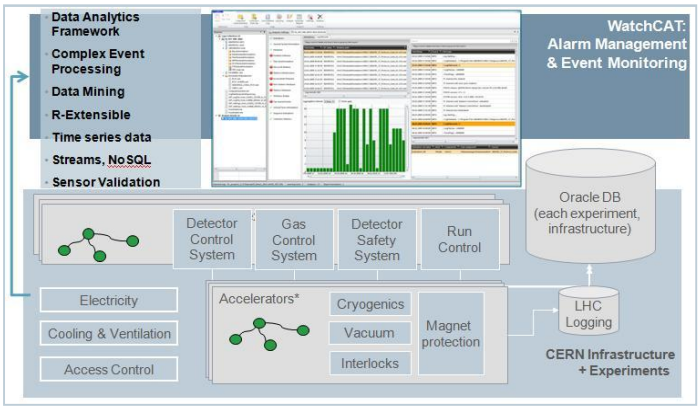
- Information Extraction
- Machine Learning
- Semantic Mapping by using Medical Ontologies (Radlex)

Constraints / issues this raises

- Availability and coverage of medical ontologies: Only concept that are covered by the used medical ontologies can be used for integrating the two different data sources

Source: Siemens AG, BAM

Example: Alarm Management and Event Monitoring (Data Analytics Framework)



What is it?

- The data of LHC (Large Hardron Collider) automation and control components and systems from Siemens (WINCC OA) are collected (offline and online) for automated system health check and diagnostic in order to prevent future damage which helps to significantly reduce overall maintenance cost
- By aligning formalized expert knowledge with automatically inferred implicit knowledge, it becomes possible to integrate heterogeneous data sources that establish the basis for advanced diagnostics.

Which data sources are used?

- Event information from control components and systems
- Process instance data (including sensor measurements)

How was the data used to create value?

- Fault and failure detection and isolation
- Root cause analysis and anomaly detection
- Preventive maintenance insights
- Predictive modeling

Which technology was used?

- Semantic models and declarative rule-based sequence models
- Complex event processing and temporal reasoning
- Learning of sequence patterns and data mining
- R-based statistical algorithm
- DSL 4 CSMC (Domain Specific Language for Comprehensive Sensor Measurement Characteristics)

Constraints / issues this raises?

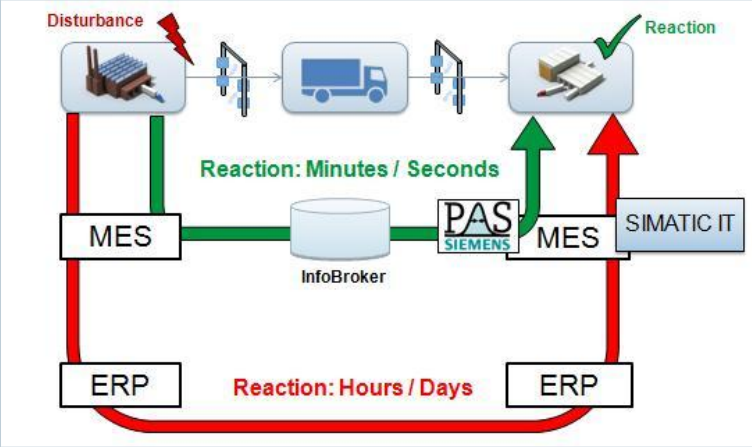
- Missing data requires dedicated strategies for establishing data base
- Asynchrone data needs to appropriately matched

Source: Siemens AG, BAM , KMR in collaboration with CERN

Example: Intelligent Integration of Supply Chain Management into Production Planning (Production Assistance System (PAS))

What is it?

- Increase robustness and efficiency of global value chains by addressing challenges like high rework rates or expensive callback actions that arise from today's lean and highly complex automotive supply networks.
- Production assistance systems continuously analyze object tracking information collected along the supply chain via Auto-ID technology (e.g. RFID, Barcode) to recommend actions for mitigating expected disturbances (e.g. predictive replanning of production sequences) and for optimizing the operational efficiency of the production network.



- The object tracking information is provided to all participants of the supply chain in real-time using the distributed and standard based InfoBroker network. The approach enables early detection of critical disturbances and shortened reaction times compared to the traditional EDI and ERP based solutions.

Which data sources are used?

- Auto-ID information from the value-chain, i.e. which object is observed at which point in time in which context at which location
- Data is collected explicitly by dedicated read-points, e.g. RFID Gates,
- or implicitly by means of third party systems, such as MES, ERP, Controller

How was the data used to create value?

- Timely detection and mitigation of disturbances to increase robustness of processes and plant productivity
- Seamless tracing of parts reduces call back costs in case of quality issues

Which technology was used?

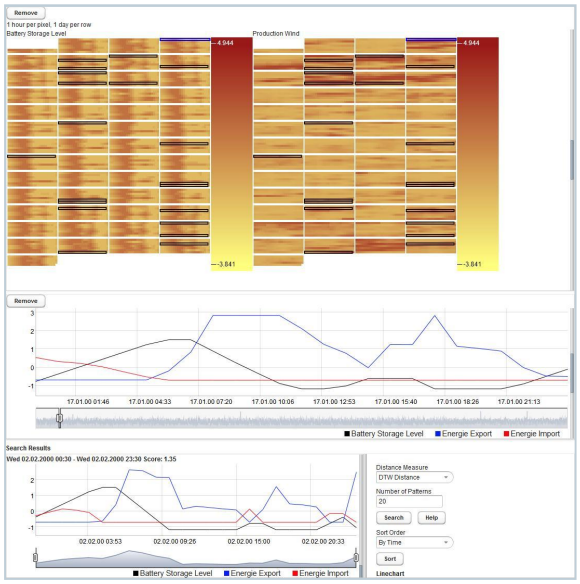
- Event-driven Resequencing
- Regression analysis,
- Dynamic baysian networks
- Monte carlo simulation
- Rule-bases monitoring

Constraints / issues this raises?

- Issue of handling data sharing constraints
- Cost-benefit sharing of the infrastructure not clear yet (who is paying for the Auto-ID infrastructure)

Source: RAN (RFID based Automotive Network) Project (BMWI-funded Project)

Example: Visual Query Analysis for Designing Efficient Smart Grid Systems



What is it?

- Interactively plan and optimize the top-level design of MicroGrids in collaboration with the user
- Optimize energy production per energy type and per time, energy import / export per time step, minimize cost (variable production, import, CO2 certificates), minimize CO2 amount
- Visual result analysis: easy visual detection of patterns in a large dataset and easy comparison across different dimensions to analyze interrelations

Which data sources are used?

- Weather
- Power demand
- CO_2 certificates
- Geo-locations
- Energy capacities
- Energy trading information (import / export)
- Production costs

How was the data used to create value?

- Simulation model
- Geovisualization of renewables
- Visual pattern detection in time series data
- Automatic pattern search

Which technology was used?

- Data mining
  - MultidimeInsional pattern recognition (time boxes with multiple time series)
  - Similarity measures (e.g. Euclidean, Manhattan, Angular, Chebychev, Cosine, DTW, Pearson, Correlation)
- Linear programming
- Advanced visualizations (recursive patterns)
- Time series analysis

Constraints / issues this raises?

- The run time of the simulations depends on the detail level of the model. There is a trade-off between an very accurate model and a fast calculation of the results.

Source: Siemens AG, BAM,IBI-DE in collaboration with University of Konstanz (Stoffel et Al., 2012)

Monetizing

Example: Clinical Research

What is it?

- The overall goal of this use case is to improve clinical research by providing means for identifying patient cohorts for clinical programs. Clinical programs that encompass a whole set of clinical trials and are running for several years can be improved if information about the number of patients with particular characteristics is known beforehand.
- By relying on information extraction and advanced analytics, the needed patient cohorts can be identified much better. In consequence, the feasibility of the clinical trials and, thus, the planning and designing of clinical programs of the pharmaceutical domain can be significantly improved.
- As clinical data is not owned by pharmaceutical companies but healthcare providers and patients, the described Big Data-based application scenario can be realized as “new business” provided by healthcare care providers targeting pharmaceutical companies.

Which data sources are used?

- Clinical data

How was the data used to create value?

- Large sets of patient data are analyzed in order to identify a subset of patient that are fulfilling the criteria for a particular clinical study

Which technology was used?

- Information extraction / semantic annotation
- Data pseudonymization
- Semantic modeling and ontologies

Constraints / issues this raises?

- The sharing of clinical data needs to respect very strict privacy constraint; depending on the legal situation the underlying business model might change.

Source: Berliner Forschungsplattform Gesundheit (BFG), Astra Zeneca

Breakthrough

Example Public Health Analytics

What is it?

- Public health analytics applications rely on the comprehensive disease management of chronic (e.g. diabetes, congestive heart failure) or severe (e.g. cancer) diseases that allow to aggregate and analyse treatment and outcome data which again can be used to reduce complications, slow diseases' progression, as well as improve outcome, etc.
- Example Use Case*: Success Story Sweden*: Since 1970, Sweden established 90 registries that cover today 90 % of all Swedish patient data with selected characteristics (some cover even longitudinal data). A recent study showed that Sweden has best health-care outcomes in Europe by average healthcare costs (9% of GPD).

Which data sources are needed?

- Clinical data, financial data and administrative data
- Outcome data

How can the data be used to create value?

- Several stakeholders could benefit from the availability of broad national and international infrastructure for public health analysis:
  - *Patients:* can access improved treatments according to best-practice knowhow
  - *Clinicians:* usage of best-practice recommendation and informed decision making in case of rare diseases
  - *Payors:* As of today, payors lack the data infrastructure required to track diagnoses, treatments, outcomes, and costs on the patient level and thus are not capable to identify best-practice treatments
  - *Government:* can reduce healthcare cost & improved quality of care

Which technology is needed?

- Information extraction
- Mature analytics (e.g. data mining, machine learning) as well as advanced analytics (e.g. prediction, devise)

Constraints / issues this raises?

- *Clinical Engagement*, i.e. active engagement, i.e. clear responsibility for data collection and interpretation, by the clinical community
- *National Infrastructure*, i.e. common standards, shared IT platform and common legal framework defining the data privacy and security requirements for tracking diagnosis, treatments, and outcomes on the patient level
- *High-Quality Data* that is achieved through systematic analysis of health outcome data of a population of patients and
- *System Incentives* that rely on the active dissemination and usage of outcome data

Source: Big Data Public Private Forum Project (http://www.big-project.eu/)

Appendix 2: Organizational Implications

New Competences are needed

New competences and skills are needed to find business value in large-scale data

Typically, this wide range of competences is rare to find within single individuals. Thus, a team of people with different skill sets is required.

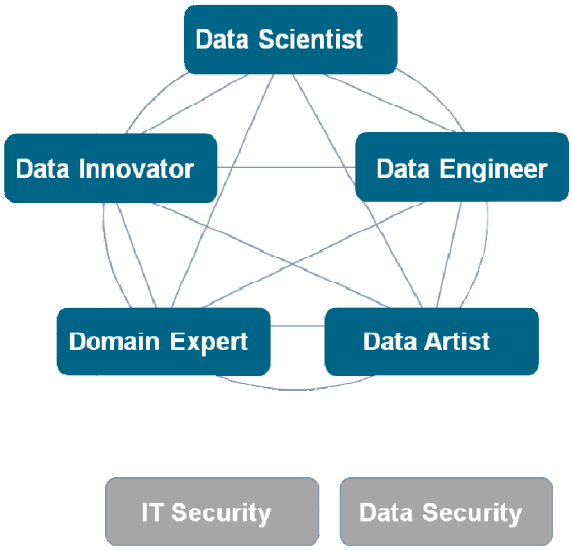


Figure 18 New Competences required for Big Data applications

The following competences are needed:

Domain expert or subject matter expert (SME) *knows the domain and user needs*. He or she is a person in a particular area or topic, for instance the healthcare domain or domain. By knowing about the particularities of the domain, they help the team to develop insights that make a different to the envisioned user as well as inform the group with background information about likely but relevant constraints, such as legislation. In order to facilitate mutual understand within a Big Data team, domain experts should have basis knowledge about technical and IT-related subjects.

Data Innovator *identifies & evaluates new business models*. Many Big Data solutions rely on a multisided business approach with the party primarily benefiting from the solution not necessarily being the one who is paying for it. Data innovators have expertise in business model innovation and help the team to identify, evaluate and develop new business models. In addition, data innovators

Data scientist *plays with the data*. Data scientists have capabilities in exploiting Big Data sources, i.e. they know how to extract meaning from data and create data functionalities or products. Data Scientist have expertise in many fields including mathematics, statistics, data engineering, pattern recognition and learning, psychology, management, computer science, etc. Data scientist investigates data of multiple data sources, spot trends and transforms his insights into a story that can be easily understood by non-practitioners. Due to this wide range of needed competences, as of today data scientists are still rare.

Data engineer develops and implements new algorithms. Data engineers built up the investigations and insights of domain experts, data innovators and data scientist and transform them into new algorithm and applications. They have strong expertise in programming, as well as statistics, mathematics, semiotics and semantics, as well as economics.

Data artist *visualizes the outcome*. Big Data-based insights often rely on complex relationships. Data artists help to publish the complex information to a variety of user groups with a single point of entry. By reflecting the mental models of the particular user groups, they are able to visualize the outcome in an intuitive and comprehensive manner. Data artists have expertise in graphic design, psychology, mathematics, IT and communication.

In addition, expertise in the domain of Data Governance and Data Privacy and Security are needed.

Data governance expert *manages the continuously growing data assets*. Data governance experts take care of the sustainable governance of the data assets. They help to integrate internal and external data sources, to extract information from unstructured data and to efficiently manage the large amount of continuously growing raw and processed data.

Data privacy and security expert *takes care about the underlying privacy and security requirements*. Data privacy and security constraints can be the showstopper for envisioned Big Data applications. Data privacy and security experts support the development and implementation of Big Data projects from the very beginning to ensure that privacy constraints, such as anonynimization or pseudonimization, or constraints regarding the exchange of data, are respected.

In order to build up the needed new competences, organizations will need to provide additional training and education for employees in the area of Big Data technologies, as well as foster the effective cooperation between people coming from different disciplines.

Organizational Structures

Organizational structures can foster synergies and cross-sector learning in several areas.

Several variants of organizational structures for the various purposed are possible:

1. *Central versus decentralized Data Governance*: The buildup of strategic data assets can be done in a central or decentralized manner. If a dedicated central organizational unit takes care of the strategic processing and collecting of data sources, the sharing of data across sectors becomes straight forward and easier. By introducing a new dedicated role – the chief data officer/Chief analytics officer – one ensures that activities are aligned by increasing the awareness across the organization. However, additional efforts are needed to standardize and streamline the data collection activities across the organization.
2. *Central Competence Center versus decentralized Competence Building Initiative:* A centralized competence center fosters the ad-hoc composition of cross-sector Big Data teams in a flexible manner and supports synergies and cross-sector learnings.
3. *Central versus decentralized IT-Infrastructure:* By hosting and managing the IT-infrastructure in central manner, costs as well as experiences and expertise with the vast technical progresses can be easily shared.
4. *Central versus decentralized Business Initiatives:* By aligning business initiatives across the various sectors and departments of an organization, allows to share learnings and insights with new business approach, the development of business models as well as the fostering of sustainable data ecosystems and value networks.

The centralizing of activities allows streamlining and standardizing the overall Big Data approach which establishes the basis for future cross-sector collaborations. Moreover, synergies and the bundling of resources become possible. The consequences are that the single business units will lose control over their assets and that as well the communication and coordination effort will increase in order to negotiate and implement the common strategic direction.

Appendix 3: Big Data Pools: Variety is the Key

The main impact can be realized by integrating various and heterogeneous data sources. And each industry is characterized by its own variety of data sources. We illustrate this along two examples:

Example: Multiple Data Pools in Healthcare

The health care system has several major pools of health data which are held by different stakeholders/parties. When planning to use data sources, the ownership of data is of relevance to distinguish the open and the closed world, i.e. between data sets that are open available and data sets that are owned by a particular stakeholder and can only be accessed under restricted conditions.

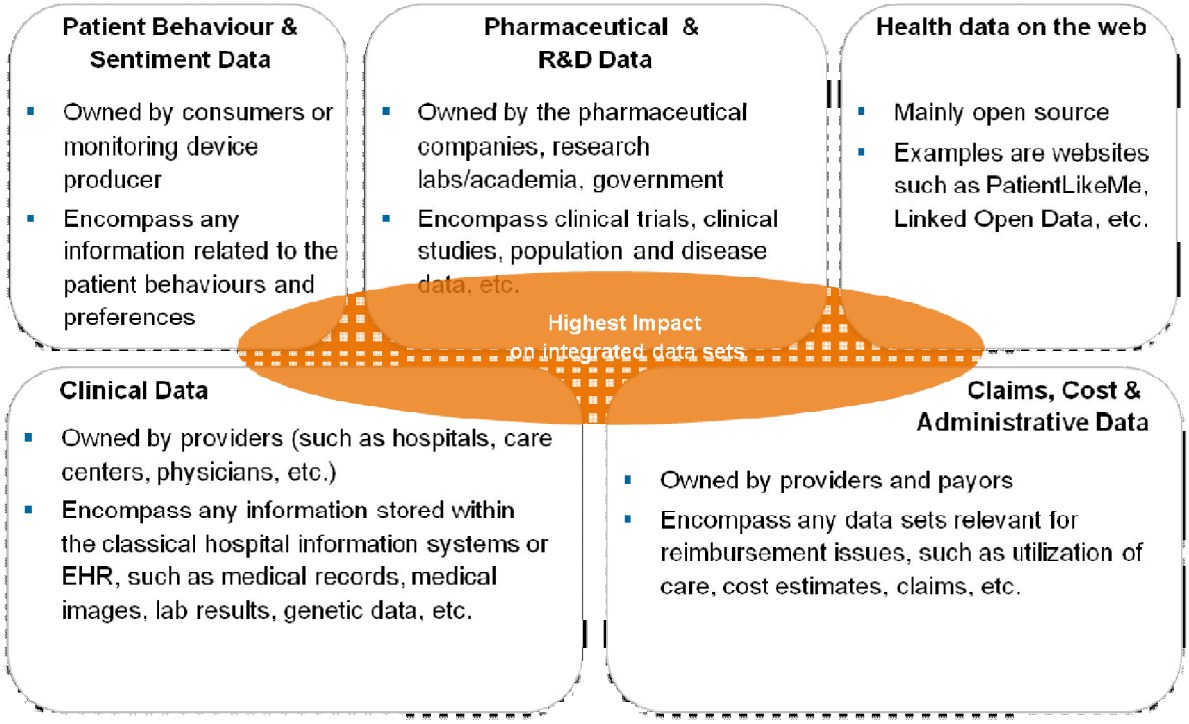


Figure 19 Multiple Data Pools in Healthcare

Closed Health Data Sources

In the category of closed – or not open accessible data – we can distinguish four types of data:

Clinical data that encompass any information stored within the classical hospital information systems or EHR, such as medical records, medical images, lab results, genetic data, etc. Clinical data sets are owned by the provider (such as hospitals, care centres physicians, etc.). As clinical data is personal data and, the data privacy and security restriction are very high. In addition, due to the fact that health is a very sensitive topic, clinical data needs to be of high quality in order to be a valuable input for Big Data applications.

Claims, cost and administrative data that encompasses any data sets relevant for reimbursement issues, such as utilization of care, cost estimates, claims, etc. Those data sets are owned by the provider and the payors. Those data sets are structured data and of high quality, the highest quality of data can be found in the accounting area. However, it is important to keep in mind that claims, cost, and administrative data cover information that can be used to improve admin or financial related processes however do not cover detailed information about the health status of patient.

Pharmaceutical and R&D data that encompass clinical trials, clinical studies, population and disease data, etc. Those data sets are owned by the pharmaceutical companies, research labs/academia or the government. As of today, a lot of manual effort is taken to collect all the data sets for the conducting clinical studies and related analysis. The manual effort for collecting the data is quite high, and same is true for the data quality. However, it is important to note that the data collection strategies are mainly hypothesis driven meaning that the collected data attributes allows to prove or disprove a particular hypothesis. Therefore, the collected data sets are often limited in their selection of attributes and cannot be easily compared with data sets from related patient groups originating from different study focus

Patient behaviour and sentiment data that encompasses any information related to the patient behaviours and preferences and is either documented by the patient (subjective data) or by some monitoring device (objective data). The data is either owned by consumers or monitoring device producer. Due to its bias, subjective data that relies on human interpretation should not be used as input for automatic analytics applications. Objective data from any kind of monitoring devices seems to be very promising for (big data) analytics applications; however, its efficient processing requires semantic interpretation /annotation beforehand.

As each data pool is held by different stakeholders/parties, the data in the health domain is highly fragmented. However, the integration of the various heterogeneous data sets is an important prerequisite of big health data applications and requires the effective involvement and interplay of the various stakeholders. Therefore, adequate system incentives, that support the seamless sharing and exchange of health data, are needed.

Open Health Data Sources

There is a large number of open health data sources already available. The available open health data sources can be divided into data sources that have been composed

- by structured databases, such as NCBI Gene, UniPort, DrugBank, etc.,
- by semi-structured documents, such as Pubmed or Clinical Trials.gov, or
- by medical ontologies and terminologies being collected by the community in dedicated repositories, such as UMLS, OBO or NCBO BioPortal..

In addition, Web 2.0 technologies are used to share health related data on the web. For instance, websites such as 'PatientsLikeMe' getting more and more popular: By voluntarily sharing data about rare disease or remarkable experiences with common diseases, their communities and user are generating large sets of health data with valuable content.

There are several activities and projects driving the availability of open health data:

1. The project BIO2RDF aims to transform life science data bases into RDF- format. It is an open-source project that uses Semantic Web technologies to build and provide the largest network of Linked Data for the Life Sciences. Currently the BIO2RDF encompasses 19 updated databases. In general, those databases that of greatest relevance for the life-science community get integrated first. Currently, the project has no proper funding. However, Michel Dumontier, the initiator and manager of the project, expects that the project will receive a more official status as well as more funding at Stanford University in the future. Currently, each of the people working with him dedicated one day time per week to work on the infrastructure.
2. Linked life data (linkedlifedata.com) is a service that provides access to the full path of health-related data, such as gene, protein, molecular interaction, pathway, target, drug, disease and clinical trial related information. This project was partially funded by Linked Life Data and several EU IST projects.
3. NCBO BioPortal is an open repository of biomedical ontologies that provides access via Web browsers and Web services to ontologies. BioPortal supports ontologies in several formats, such as OBO, OWL, RDF, and other formats. At a first glance, one gets easily overwhelmed by the large number of ontologies. By analyzing them in more detail, however, we found out that only a limited number of the covered ontologies seem to be of relevance for Big Data applications: From more than 330 ontologies of the repository, we could only identify about 20 ontologies that seem to be of higher relevance for big health data applications in terms of their size and coverage of terms. The remaining ones were excluded, as they have been either too small, within an experimental status or addressing a very focused yet important domain.

However, although many open health data source are available, the *licensing situation is often hindering the usage of data* in clinical products or commercial health applications. Thus, one should always make clear where the data is coming from and how the underlying rights and licensing situation is, before one is planning to use a particular data source. For instance, the usage of the data of the OMIM (Online Mendelian Inheritance in Man) database in commercial settings requires licensing and registration. The same is true for the large collections of abstracts of scientific publications of the PubMed repository, i.e. although the content can be easily accessed via web and dedicated interfaces, the ownership of the materials lies with the publishers and there are restrictions on derivates or commercial use of the data.

It is also important to keep in mind that, although a lot of open health data is available, the data can quite often not be used due to the fact that they are either only provided in textual format in a non-standardized manner. For instance, the service of the U.S. national Institute of Health called ClinicalTrials.gov currently lists more than 150 000 clinical studies. However, the data is only provided in textual / unstructured format with no links to external identifiers such as standardized drug codes. Thus the content cannot automatically be processed but requires human interpretation beforehand. In

addition, there exist no standard format for representing clinical studies; it is difficult to compare them in an automatic manner.

Example: Multiple Data Pools in Smart Grid

We reference the network operators' vision for 2020 as stated by the European Electricity Grid Initiative (EEGI)[53]. [Figure8] depicts this vision and outlines how much more data is to be managed and analyzed from various sources:

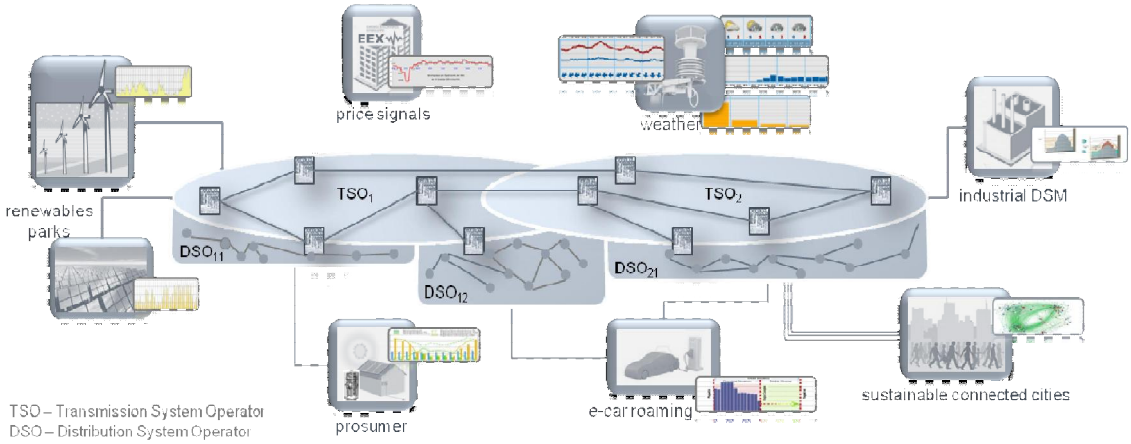


Figure 20 European network operators vision for 2020 can only be realized when real-time insight can be gained about this highly connected and dynamic new power system.

1. Active integration of the efficient new generation and consumption models: Integrate new intermittent renewable resources at the different voltage levels; enable and integrate active demand from end users, enable and integrate new electricity uses, e.g. recharging infrastructure for electric vehicles; enable new business opportunities and innovations for market players.
2. Coordinated operation of the interconnected electricity network: Enable coordination of the operation for the pan-European transmission network; enable the coordination of the operation between transmission and distribution networks.
3. New market rules to maximize European welfare: Facilitate resulting market transactions and possibly continuous changes. It is extremely unlikely that there is a single optimal network topology for all possible market realizations and events over a long time horizon. The operation of the networks must hence be situation-aware, adaptive, and must take also the transmission and distribution assets into account.

In order to realize these necessary scenarios for a smarter grid, following data sources need to be integrated:

Structured data sources from within the Smart Grid segment include:

Smart Meter data can measure usage, feed-in of energy including power, gas, water, heating, and cooling. It can measure the units down to the minute or event-based. Technically a smart meter could also measure power quality parameters of the power network.

Phasor Measurement Units measure all required power network parameters in real-time and GPS-synchronized, in order to be able to have a dynamic view on large-scale power networks over a wide area. The data is extremely high resolution, reaching from 20-120 samples per second. Some devices even sample at higher rates.

SCADA (operational) data includes almost all data that phasor measurement units also collect. However, there is fundamental difference in usage scenarios as well as in the lower frequency with which this data is collected namely only 2-4 seconds. Additionally, this type of data is not time-synchronized, which means there needs to be some pre-processing and ordering before the data can be utilized.

Non-operational data, such as from digital fault recorders is so highly resolved that they are only triggered when a fault happens. The data is then kept in the substations and only retrieved by technicians for post-mortem analysis of the fault. Mainly this is because the communications is not sized to handle burst of output of data at 720 Hz and 8 kHz samples.

[53] http://ec.europa.eu/energy/technology/initiatives/doc/implementation_plan_2010_2012_eii_electricity_grid.pdf

Newer devices that are being researched and commercialized include for example the Digitizer. Digitizer takes sample measurements of current and voltage at 25,000 times per second and applies real-time calculations to derive further parameters that relate to power quality (National Physical Laboratory, 2013).

This ever advancing measurement and sensing technology is at the heart of the Energy sector and will continue to supply the utility control centers with following information: power factor, power quality, phase angles indicating increasing load on a power line or transformer, equipment health and capacity, meter tampering, intrusion, fault location, voltage profiles, outages, energy usage and forecasting, etc.

External data sources with uses in the Smart Grid business include:

- Building energy management software enables building owners to more accurately purchase and usage of energy. The same data can also be used to better size the power network or the purchases on the wholesale market.

- Home automation data, such as data from smart thermostats. EnergyHub for example mined interesting statistics on energy behavior using their connected thermostat, such as that residents in states with cooler average temperatures are setting their thermostats lower than those living in warmer states (EnergyHub, 2012)

- Electric cars will most certainly use communication to increase charging efficiency and other services that will be needed for increasing the comfort of these cars. E-car roaming is assumed to be another application that cannot be realized with vast data sharing among many stakeholders.

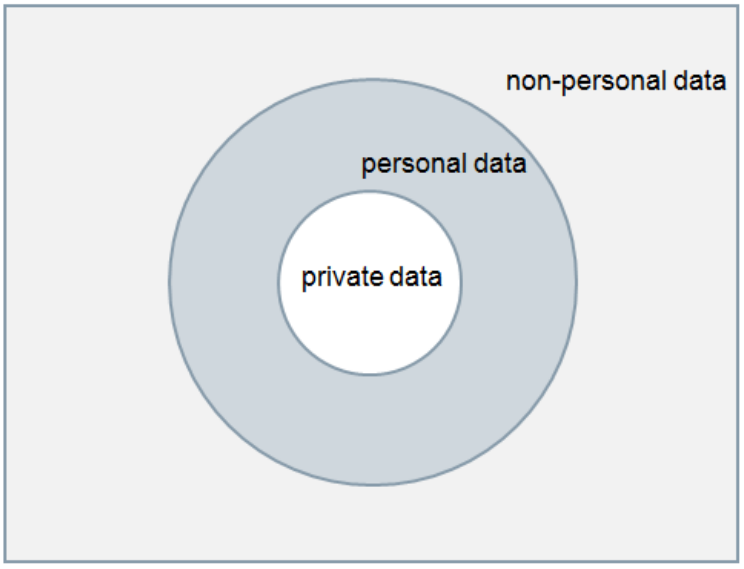
- Most prominently weather data: since energy consumption, and if Renewables are involved also energy generation are highly weather dependent this data source is useful for all stakeholders. Even distribution and transmission network operators use weather data for fault detection and location for example, or calculating the thermal loading of lines. Mining weather data for location-based aspects and forecasting for heating-/cooling-heavy days is among the most prominent uses.

Appendix 4: Data Sharing Policies

Categorization of Data

To get a better understanding about the underlying opportunity and constraints of data sharing policies, we need to distinguish the different categories of data:

- Private and personal data
- Operational data
- Historical and longitudinal data



Private, personal and non-personal data

Personal data is digital data that is created by or about a person. It includes information on a person's identity, age, gender, place of birth, address, hobby, etc. In everyday live, personal data plays a central role in various settings: Clinicians might use health data to improve health care delivery by comparing for instance treatment outcomes of similar patient populations. Or companies might use a wide range of personal data to improve existing and develop new products and services that rely on the individual or context-specific circumstances in order to provide tailored offerings.

Any data that covering name, address, email-address, password, data of birth, bank account number, credit card data, phone numbers or invoice details are classified as personal data. Private data is any information about the religion, ethnical background, political opinion, health data, genetic and biometric data, data about children, location data and personal data in large size, such as Big Data. Depending to the category of data involved, different data privacy constraints needs to be respected.

Most personal data remains unused due to unclear legal situation

However, as of today, most personal data resides in data silos due to missing exchange standards and legal contracts. Often simply due to the lack of effective systems of permission as well as the lack of guidance in implementing established country-specific data governance rules, the secure and trusted data flow and, thus potential value creation, is prevented.

Personal data are associated with high security and privacy constraints in order to protect and secured it against intentional and unintentional security breach or misuse. In this context, anonymization and pseudonymization are very important privacy enhancing techniques. The overall goal is to withdraw and replace the identification data of individuals with pseudo-identifiers: Anonymization aims towards the complete removal of all identifiers from the data, whereas pseudonymization enables the linkage of data associated with pseudo-identifiers independent of time and place of data collection without having to reveal the personal identifiers. In the context of data analytics, pseudonymisation, i.e. the opportunity to refer back to the original person the data is coming from, is quite often needed. For instance, in longitudinal healthcare studies, when data about specific patients that was originally stored within many different and distributed repositories needs to be aggregated in a way that the link to the patient is not lost. However, it is important to mention that the more data about a particular person -- in this example a patient -- is aggregated, the more details are known about this person and, thus, the more likely it becomes that the patient can be de-identified although the original data had been pseudonymized.

To conclude this discussion, it is also important to mention that the personal data ecosystem is increasingly complex, fast-moving and global, for traditional regulatory mechanism to be effective (World Economic Forum, 2012). In order to gain trust of their users, organizations will need to held accountability for protecting personal data in accordance to the established rights defining trusted flow of data. In order to address the various needs of a global market places, it is likely that data sharing polices will be continuously changing, which again will trigger changes in how the reliable data exchange can be implemented on an organizational level.

Operational data

The usage of operational data within Big Data applications bears promising business opportunities; however the sharing of data is often hindered due to missing incentives and fear for loss of control

Operational data is the data collected by any kind of device, machine, process, sensor, etc. In general, this data relates to some business or operational processes and is of relevance for the current business. Operational data in the healthcare setting, for instance, is any data that gets collected for clinical care, for billing purposes, for scheduling tasks, for patient assessments, etc. As the data is needed to accomplish routine tasks, typically the data collection efforts are paid internally. For accomplishing particular routine tasks, the operational data is already today shared between the involved business partners, often even in standardized formats. For instance, health data required for billing is exchanged between payors and healthcare provider by means of standardized ICD and DRG-codes.

Due to several reasons the usage of operational data entails *promising opportunities*:

a. operational data are usually produced in enormous quantity and granularity
b. operational data can be used to understand and improve the underlying processes
c. real-time collection and availability of operational data can be realized more easily

However, the *sharing of operational data is often hindered* due to several reasons:

a. data silos or organization silos
b. fear of loss of control of the party who is owning the data
c. missing incentives for sharing the data.

Longitudinal and historical Data

Data that is no longer needed for the ongoing processes might be of high valuable for other purposes, such as historical analysis, cross-functional processes, etc. In some cases, historical data needs to be store due to some legal constraints. Or dedicated efforts are made to built up longitudinal data storages, such as disease registries, in order to gain further insights about the characteristics of selected populations and about the progression of trends over time.

In general, historical and longitudinal data are stored at dedicated, separated storage solutions, such as data warehouses.

Longitudinal and historical data are of high value because they describe past information that can be used make forecasts, for instance about the company future or about the likely malfunction of a machine.

However, the storage of longitudinal and historical data needs to fulfill particular data security and privacy constraints. In addition, the anonymization of historical and longitudinal data bears challenges, as the more data are collected about one object or individual, the more difficult – sometimes even impossible – data anonymization becomes.

Ownership of Data

Due to several reasons is the term “data ownership” a misleading concept and gets thus often misinterpreted:

1. Data out of context has no value. Data in standalone-mode, i.e. a simple set of characters, have no relevance/meaning and therefore no value. If something has no value, than ownership is a category that does not apply. Is it then more appropriate to think about the ownership of context information?
2. Data ownership stays with the owner of the location: Whenever data is generated, is also gets stored. In general, if one speaks about data ownership, than this refers to the storage process. Consequently, the data ownership stays with the owner of the storage location. For instance in the healthcare sector, patient data ownership refers to the organization the data belongs to, and in general, this is the organization that created the data. In this regard, several questions arise, e.g. Has the data owner, in this case the hospital, the ability to derive benefits from the data as well as to make the data available to third parties? And what happens if the health data is outsourced to the cloud? Although cloud operator won’t have direct access to the patient data, do the cloud providers have the responsibility to ensure same data security and privacy requirements? Etc.
3. Each time data is shared, new data is generated: Data needs to move to generate value. In general, in order to accomplish the daily tasks, data items are shared between business partners. For instance, the billing information is sent by one company and the other company approves it. Thus, data items are are handed over in various transactions, and with every transaction, information is shared and new data is generated, and new data

ownership is created. Or if a Big Data company collects, aggregates, and analyzes large amount of data sets, this company is generating new data, which belongs then to the company.

Glossary

Anonymization	The separation of links between people in a database and their records to prevent the discovery of the source of the records
Big Data application	Big Data applications are pieces of software relying on Big Data technologies that is designed to perform a specific task or suite of tasks
Big Data Business Patterns	Big Data Business Patterns define the strategic fit between the (newly built) Big Data capabilities and the core competences of the organization.
Business ecosystem	Business ecosystems are “a dynamic structure which consists of an interconnected population of organizations.
Data exchange	Data exchange is the process when a data item that is structured according to a source schema is transformed into a data structure of a target schema.
Data governance	A set of processes or rules that ensure the integrity of the data and that data management best practices are met
Data ownership	Data ownership refers to both the possession of and responsibility for information.
Data privacy	Data privacy is the aspect of information technology that deals the ability of an organization or an individual has to determine what data can be shared with third parties
Data quality	The measure of data to determine the worthiness of data for knowledge discovery, decision making, insight generation, or related activities
Data security	data security is the practice of protecting data from destruction or unauthorized access
Data source	A collection of data, for example a data base or a data stream
Data structure	A specific way of storing and organizing data
Data-enhanced Business	Data-enhanced businesses make use of data insights in order to improve the performance of traditional businesses and services. The physical processes and business is enriched with intelligence behaviour.
De-identification	The act of removing all data that links a person to a particular piece of information
Digital-driven Business	Digital-driven business or service provider use the digital world as platform to reach a large, often global, market that allows them to offer services in the real world often to reach an economic of scale. Their value-add is of digital nature which is tightly connected with a physical value chain.
Hadoop	An open source software library project administered by the Apache Software Foundation. Apache defines Hadoop as “a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model.”
Internet of Things	The concept Internet of Thing referts to uniquely identifiable objects (i.e. ordinary devices) and their virtual representation (e.g. data generated by sensors) in an Internet-like structure.
Map / reduce	A general term that refers to the process of breaking up a problem into pieces that are then distributed across multiple computers on the same network or cluster, or across a grid of disparate and possibly geographically separated systems (map), and then collecting all the results and combines them into a report (reduce). Google’s branded framework to perform this function is called MapReduce.
Open data	Open data is the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control

Pseudonymization	Pseuodnymization is a procedure by which the most identifying fields within a data record are replaced by one or more artificial identifiers, or pseudonyms.
RFID	RFID (Radio radio-frequency identification) is a technology that relies on wireless communications to send information about an object from one point to another.
Virtual Business provider	Virtual Business provider collect, integrate and aggregate various sources of data in order to provide data products in the virtual world (internet, web, etc.).

Siemens AG
Corporate Technology
Otto-Hahn-Ring 6
81739 München

www.siemens.com