| | |
|---|---|
| Project acronym: | BYTE |
| Project title: | Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities |
| Grant number: | 619551 |
| Programme: | Seventh Framework Programme for ICT |
| Objective: | ICT-2013.4.2 Scalable data analytics |
| Contract type: | Co-ordination and Support Action |
| Start date of project: | 01 March 2014 |
| Duration: | 36 months |
| Website: | www.byte-project.eu |

# **Deliverable D1.4:**
# Big Data Technologies and Infrastructures


| | |
|---|---|
| Authors: | Sebnem Rusitschka, Alejandro Ramirez (SIEMENS) |
| Contributors: | Guillermo Vega Gorgojo (UIO), Ahmet Soylu (UIO), Dr. Stefán Péter (NIIF), José María García (UIBK) |


| | |
|---|---|
| Dissemination level: | Public |
| Deliverable type: | Final |
| Version: | 1.1 |
| Submission date: | 2 September 2014 |

## EXECUTIVE SUMMARY

The increasing digitization of content and widespread use of smart devices is resulting in increasing volumes of data. In addition, increasingly the data is generated and consumed in streams, in real-time. Players in online data businesses, where data is the main company asset, such as Search, eCommerce, and Social, were the firsts to face the Big Data challenges: in order to scale their businesses they needed to scale their data management and creation of business insights from data, which comes from a variety of sources, at high speed and cumulates to scales of petabytes. Connectedness, communication, and available bandwidth, also mobile, have been playing a significant role: in the original playing field of Big Data, the Web, as well as in industrial businesses with increasing digitization and automation.

Since the advent of online businesses, which must innovate around data, from the beginning of the millennium *every three years a new wave of Big Data technologies* has been building up: 1) The "batch" wave of distributed file systems and parallel computing, 2) the "ad-hoc" wave of NewSQL with their underlying distributed data structures and distributed computing paradigm, and 3) the "real-time" wave, which allows producing insights in milliseconds through distributed stream processing.

The analysis of current technologies for dealing with variety, velocity, and volume of data – especially how they evolved, gives a way of comparison for capabilities and needs.

There is no single tool, or choice of a platform that will remedy all of the challenges of Big Data business. Looking at *architectural patterns* based on the state of the art in Big Data technologies can assist in making the right bundling of solutions:

- *Schema on Read* allows the cost-efficient storage of all data in its raw form and transforming it as needed by the business users whenever required – instead of transforming the data so that it can be readily consumed by business applications and risking losing data that is currently not used by valuable. Distributed data management and massively parallel processing principle combined with performant commodity hardware makes creating the data schemas on demand, feasible. This pattern allows one to get to know the data and explore business opportunities, however, only in batch mode.
- The so-called *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 (through schema on read) as well as velocity (via scalable stream processing).

Establishing such Big Data computing capabilities allows businesses to excel in *three main application categories*: 1) *operational efficiency* subsumes all applications that involve improvements in maintenance and operations in real-time or a predictive manner based on the data which comes from infrastructure, assets, and end users, 2) *customer experience* applications combine data from usage with other sources where end users are active, 3) *new business models*, especially when applications allow monetization of available data. These applications *require a range of analytics capabilities*. Especially, predictive and prescriptive analytics enable businesses to reap value from investments into Big Data technologies and infrastructure: Whilst predictive and prescriptive analytics enable the extraction of foresight and options for action based on these insights, Big Data can improve predictive models and enable the testing of actions and evaluating their outcome.

Big Data technologies, as of today, enable the cost-effective handling of high-volume, high-velocity data – but variety, still is a challenging problem, for which advanced analytics and semantic technologies are being researched for remedying issues of heterogeneity. The

increasing awareness of private and commercial customers of their need for privacy and confidentiality represents another field of valuable research into Big Data technologies. Veracity, the one most named challenge after the well-known 3V's, requires analytics capabilities to move closer to where data originates and actually be included at each step of data processing from acquisition, to management, to usage such that bad data can be accounted for in extracting knowledge that can be put to action automatically.

## Table of Contents

# 1 INTRODUCTION

"Big Data," the term, can be traced back to 1998[1]. The efforts to understand the information explosion, cope with and make use of it reach even farther back (also see Big Data Definition in Deliverable 1.1 "Understanding Big Data"). 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 deliverable covers the description of main and emerging Big Data technologies, their applications and aims to explain their evolution:

- **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, have also adapted in waves 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, analysis of options, and simulations of outcomes of possible actions such that reliable real-time decision making – even decision automation – may become 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.

  Many other sectors have also long been dealing with the analysis of complex systems and high-dimensional data, such as risk assessment in finance, neuroimaging in health, geo-data analysis for drilling and exploitation in oil & gas. However, only with big data technologies now have these actors the capability to do so faster and more cost-efficiently than ever before increasing the return on data considerably, resulting in more data hungry processes. The "automation" aspect of big data technologies when processing big data for ingestion, storage, and analysis is what makes the difference.
- **Traditional 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 wide-spread adaptation. Two major examples are Hadoop[2] and R[3], which considerably facilitated the spill-over 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[4] of IT software vendors, and R by statistics software vendors SAS and SPSS.

[1] Gil Press, "A Very Short History Of Big Data" in Forbes, 9 May 2013.
http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/
[2] Framework for scalable fault tolerant data storage and computing, see http://hadoop.apache.org/
[3] Statistical computing software , see http://www.r-project.org/
[4] IBM, Microsoft, Oracle, and SAP

- **The future holds the merger of the data management and the data analytics waves with semantics and standardization**: Rather recent experiences, e.g. with analytical algorithms that are hard to parallelize, lead to the merging of distributed computing platforms with analytics libraries and hiding the complexities of distributed data analysis algorithm execution from the user[5]. In order to realize what Analytics 3.0[6] 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. Especially, the interpretability of massive amounts of data will need to improve through semantic knowledge modelling and re-modelling with discovered knowledge from data analysis. 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.

The findings in this deliverable are mainly synthesis of the analysis of arguably established Big Data technologies. Hence, we motivate as much as possible through references to the technologies developed by Big Data natives, their open source adaptations and new open source innovations, as well as by references of big data applications that could be found online. The uptake[7] of Big Data technologies in 2012 and 2013 has made such analysis possible.

The deliverable is structured as follows: a short description of Big Data evolution is given, in which mostly the aspects of volume and velocity are motivated by references; variety follows by the sheer number of different domains and the many sources of data that are being digitized one by one in the various sectors. A more detailed and broad definition of Big Data is given in Deliverable 1.1. In Section 3 of this deliverable we are focusing on the evolution of Big Data technologies, by studying the releases of Cloud services, and open source solutions as well as major milestones of Google Cloud Platform, Amazon Web Services, and Apache Hadoop. We detail in this Section the underlying distributed storage and computing paradigms. Section 4 details Big Data analytics applications, by focusing on analytics application classes, business applications classes that benefit from such capabilities – and also provide some motivated examples from different industrial sectors. Section 5 wraps the discussion of Big Data technologies and applications by touching upon some of the emerging Big Data trends and needs. We give a short summary and conclusion in Section 6.

## 2 EVOLUTION OF BIG DATA

Digitization has revolutionized the creation, exchange, and consumption of information. Since digital storage became more cost-effective for storing data rather than paper[8], the information grows exponentially.

The first comprehensive study by P. Lyman and H.R. Varian in 2000 found[9] that in 1999, the world produced about 1.5 exabytes of unique information: “A vast amount of unique information is created and stored by individuals” what it calls the “democratization of data”

[5] E.g. Machine learning libraries on top of Hadoop such as Mahout, see https://mahout.apache.org/ or on top of Hadoop Distributed File System such as Spark MLib, see https://spark.apache.org/mllib/

[6] Davenport, Thomas H., *The Rise of Analytics 3.0 - How to Compete in the Data Economy*, 2013. http://www.strimgroup.com/sites/default/files/images/PDF/Davenport_IIA_analytics30_2013.pdf

[7] As explored with Google Trends, which depicts the popularity of a topic both search/interest and content-wise, see https://www.google.de/trends/explore#q=Big%20Data

[8] Morris, R.J.T., and B.J. Truskowski, “The Evolution of Storage Systems,” *IBM Systems Journal*, July 1, 2003.

[9] Lyman, Peter and Hal R. Varian, “How Much Information?”, *Project Web Site*, no date. http://www2.sims.berkeley.edu/research/projects/how-much-info/

and that "not only is digital information production the largest in total, it is also the most rapidly growing." A repetition of the study in 2003 found[10] that the world produced about 5 exabytes of new information in 2002 and that 92% of the new information was stored on magnetic media, mostly in hard disks.

In 2007 IDC starts the first ever annual study that forecasts the amount of digital data created and replicated each year[11]: The information the information added annually to the digital universe is doubling every 18 months. A similar survey series from EMC finds in its 2014 release that[12]

- *In 2013, two-thirds of the digital universe bits were created or captured by consumers and workers, yet enterprises had liability or responsibility for 85% of the digital universe.*
- *In 2013, only 22% of the information in the digital universe would be a candidate for analysis, i.e., useful if it were tagged (more often than not, we know little about the data, unless it is somehow characterized or tagged – a practice that results in metadata,* i.e. data about data*,); less than 5% of that was actually analyzed.*
- *In 2013, while about 40% of the information in the digital universe required some type of data protection, less than 20% of the digital universe actually had these protections.*
- *Data from embedded systems, the signals from which are a major component of the Internet of Things, will grow from 2% of the digital universe in 2013 to 10% in 2020.*
- *From 2013 to 2020, the digital universe will grow by a factor of 10 – from 4.4 trillion gigabytes to 44 trillion.*
- *By 2020, the useful percentage could grow to more than 35%, mostly because of the growth of data from embedded systems.*

Many of these aspects, regarding origin and ownership of data, usability and interpretability of data, are also focal points in the following sections, when big data technologies and infrastructure (Section 3) as well as big data its applications (Section 4) are discussed. However, most of the issues are only touched upon by what can be called emerging big data technologies and needs (Section 5).

# 3 EVOLUTION OF BIG DATA TECHNOLOGIES

The 3V's – volume, velocity, variety – is an often-cited (and expanded with more V's to point out the challenges of Big Data, such as veracity, value, etc.) definition from Gartner. However, it is not the entire original definition[13]:

> *Big data is high-volume, high-velocity and high-variety information assets that demand* ***cost-effective, innovative*** *forms of* ***information processing*** *for* ***enhanced insight*** *and* ***decision making****.*

The original definition very definitely points out that it is not just about the realization of information explosion, but about technologies that ensure value can be generated from the masses of data. We discuss these technologies in the following subsection and how they

[10] Ibid.
[11] Gantz, John F. "The Expanding Digital Universe", *An IDC Whitepaper,* March 2007. http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf
[12] EMC Digital Universe with Research & Analysis by IDC, *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*, April 2014. http://www.emc.com/leadership/digital-universe/2014iview/index.htm
[13] Garnter, "Big Data", *IT Glossary:* no date. http://www.gartner.com/it-glossary/big-data/

evolved. We also discuss how they enhance data analysis capabilities in an innovative and cost-efficient manner such that the challenges of veracity and value can be taken up (Section 4).

### 3.1 A SHORT HISTORY OF BIG DATA TECHNOLOGIES

In 2002 digital information storage surpassed non-digital for the first time. Google (1998), the web search engine, and Amazon (1994) the eCommerce platform, are so-called first Big Data Natives: Their business is entirely based on digitization and online data. In order to scale their business they always needed to scale their data management and knowledge retrieval processes first[14,15]. There are two decisive factors in the wide-spread adoption and fast evolution of Big Data technologies:

1) Research & Development in these Big Data native businesses are very close, and very close to the research and open source community.
2) Each paper on the cost-efficient innovative information processing techniques has been accompanied by open source adoption within an ever growing ecosystem called Hadoop[16].

Two major milestones in the development of Hadoop also added confidence into the power of open source and Big Data Technologies. Only two years after its first release, in 2008, Hadoop won the terabyte sort benchmark[17]. This is the first time that either a Java or an open source program has won[18]. In 2010 Facebook claimed[19] that they had the largest Hadoop cluster in the world with 21 PB of storage for their social messaging platform.

Following these observations, Figure 1 depicts the major development stages of Big Data technologies with some but not all exemplars. The idea is to highlight the different stages in which Big Data technologies evolved, reaching from batch processing in 2000 to real-time analytical processing a decade later. In the accompanying Table 1, the different characteristics, short-comings, and further exemplars of each stage are detailed a bit further.

[14] Brin, Sergey and Lawrence Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, no date. http://infolab.stanford.edu/~backrub/google.html
[15] ACMqueue, A Conversation with Werner Vogels, *Interview Transcript*, 30 June 2006. http://queue.acm.org/detail.cfm?id=1142065
[16] See Apache Hadoop Web Site https://hadoop.apache.org/ for further information
[17] See the Sort Benchmark Home Page at http://sortbenchmark.org/
[18] News entry on Hadoop web site can be found at https://hadoop.apache.org/#July+2008+-+Hadoop+Wins+Terabyte+Sort+Benchmark
[19] Dhruba Borthakur, “Facebook has the world's largest Hadoop cluster!”, *Blog Post*, 9 May 2010. http://hadoopblog.blogspot.com/2010/05/facebook-has-worlds-largest-hadoop.html

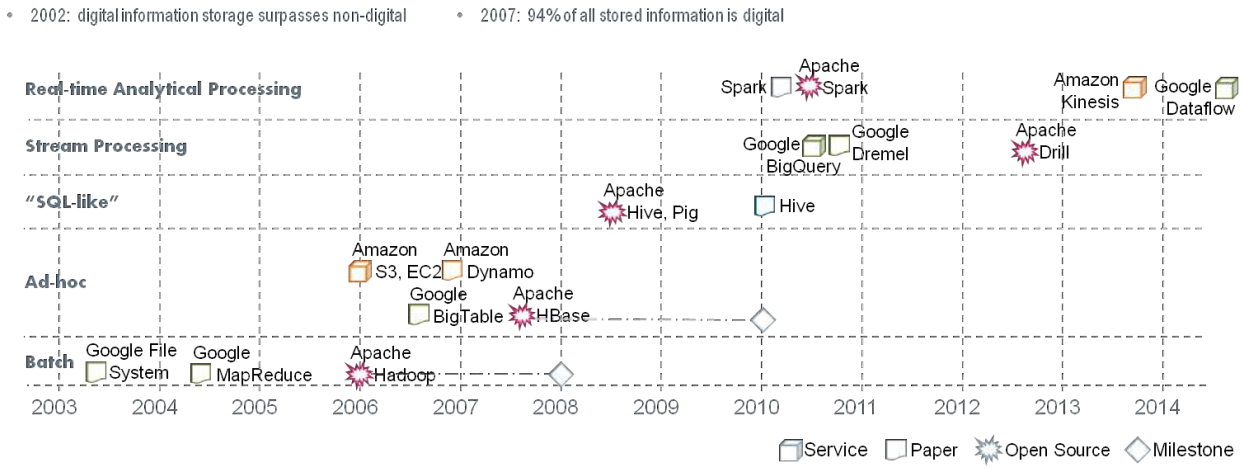


**Figure 1: From Product to Paper to Open Source – the Evolution of Big Data Technologies (source: Siemens AG Corporate Technology[20])**

| Stage | Characteristics / Shortcomings | Exemplars |
|---|---|---|
| **Batch** | Distributed file systems (DFS) allow for fault-tolerant, horizontally scalable storage of data across commodity hardware, because the storage nodes do not share memory but virtually connect their memories through networking.<br>MapReduce is a programming model for distributed parallel processing of data stored at each node in a networked cluster. The model is very efficient for parallelizable batch tasks on massive amounts of unstructured data.<br>Although MapReduce abstracts from the complexities of distributed programming it still requires programming skills. It is most efficient over very large batch tasks, and less efficient for iterative, ad-hoc queries.<br>As such MapReduce originally used for offline analysis or background tasks such as indexing websites. The combination of DFS and parallel computing is optimal for "write once read many" applications (file systems in general are good for sequential data access, but not random read/write access) | Google File System & MapReduce,<br>Apache Hadoop HDFS & MapReduce |

[20] includes info about development of growth of data stored digitally http://www.martinhilbert.net/WorldInfoCapacity.html

| Stage | Characteristics / Shortcomings | Exemplars |
|---|---|---|
| **Ad-hoc** | Online analyses of data or iterative tasks require the capability of ad-hoc querying specific ranges of data within the masses. This requires random read/write access to the data.<br>NoSQL databases arose as a remedy to this shortcoming of DFS. There are two most popular types:<br>1) Columnar databases that build upon DFS, such as BigTable or HBase. Data is stored column-based, i.e. across rows, which makes it very easy to add columns, and they may be added row by row, offering great flexibility, performance, and scalability. This allows efficiency in the face of volume and variety.<br>2) Key-value stores that essentially are a distributed data structure called Distributed Hash Tables, which enable very efficient key-based access to data. Amazon's Dynamo is DHT-based, whereas Facebook's Cassandra has borrowed concepts from both Dynamo and BigTable.<br>NoSQL databases are good for really big amounts of databases (several to 50 TBs)[21] with unstructured or semi-structured, or key-value type, e.g. time-series data.<br>Their one weakness is that they don't support SQL-like querying. | Amazon Dynamo,<br>Google BigTable,<br>Apache HBase,<br>Apache Cassandra |
| **SQL-like** | Programming talents are hard to come by.<br>That's way almost all of the NoSQL databases now offer an SQL-like interface, e.g. CQL of Cassandra.<br>There are also SQL-engines that can connect to any NoSQL database.<br>Newer "NoSQL" DBs typically are born with an SQL interface, that's why they are dubbed "NewSQL"<br>With the SQL-like interface and the inherent capability of sort and organize massive amounts of data, this stage allowed big data technologies to be used like data warehouses (at least for storing, and offline analyses). | Apache Hive, Pig,<br>PrestoDB,<br>H-Store, Google Spanner |

[21] Evolutio Team, "Row-Based Vs Columnar Vs NoSQL", *Blog Post*, 2 October 2012. http://evolutivoteam.blogspot.de/2012/10/row-based-vs-columnar-vs-nosql.html

| Stage | Characteristics / Shortcomings | Exemplars |
|---|---|---|
| **Stream Processing** | Increasingly digital data creation and its usage have become in stream.<br>That is why most NoSQL databases now offer an in stream-processing solution or can be extended to use for in-stream processing with fault-tolerant distributed data ingest systems (such as Flume, Kafka).<br>There are also standalone stream processing frameworks, which can be faster.<br>In contrast to DFS, streaming has also many commercial offerings. IT incumbents like IBM were even in the forefront of stream processing, coming from the domain of event processing.<br>However, big data analytics requires a more holistic view than the condensed version of "events" – which requires a great deal of prior knowledge, what an event is. | Hadoop Streaming,<br>Google BigQuery,<br>Google Dremel, Apache Drill,<br>Apache Flume/HBase,<br>Apache Kafka/Storm,<br>Samza |
| **Real-time Analytical Processing** | Analytics talents are hard to come by.<br>That's why the SQL-like movement quickly evolved into also transferring OLAP-like processing into the Big Data technology landscape. OLAP, is a model that has been enabling the efficient querying of traditional data warehouses, and of data structures such as multi-dimensional cubes engineered specifically for analytics.<br>Built on top of big data structures, there are now libraries for machine learning and analytics that utilize these massive data processing capabilities for real-time analytics processing.<br>Apache Spark is the first open source (potentially) real-time streaming platform, that allows complex machine learning. Interestingly it is also the first time that open source has preceded any services from Google or Amazon. | Apache Spark,<br>Amazon Kinesis,<br>Google Dataflow |

**Table 1: The different stages of Big Data technology evolution, characteristics and some exemplars of each wave**

## 3.2 COMMUNICATION AND BANDWIDTH

In 2008 Bret Swanson and George Gilder estimated that[22] U.S. IP traffic could reach one zettabyte by 2015 and that the U.S. Internet of 2015 will be at least 50 times larger than it was in 2006. Cisco's annual report forecast in 2008 that[23] "IP traffic will nearly double every two years through 2012" and that it will reach half a zettabyte in 2012. The latest report[24] from June 2014 suggests:

- *In 2013 mobile networks alone carried nearly 18 exabytes of traffic*

[22]Swanson, Bret and George Gilder, "Estimating the Exaflood - The Impact of Video and Rich Media on the Internet – A 'zettabyte' by 2015?", 29 January 2008. http://www.discovery.org/a/4428
[23] Cisco, "Cisco Visual Networking Index – Forecast and Methodology, 2007–2012", 16 June 2008. http://newsroom.cisco.com/dlls/2008/ekits/Cisco_Visual_Networking_Index_061608.pdf
[24]Cisco. "Cisco Visual Networking Index – Forecast and Methodology, 2013–2018", 10 June 2014. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

- *In 2013, a fourth-generation (4G) connection generated 14.5 times more traffic on average than a non-4Gconnection. Although 4G connections represent only 2.9 percent of mobile connections today, they already account for 30 percent of mobile data traffic.*
- *Globally, smart devices represented 21 percent of the total mobile devices and connections in 2013, they accounted for 88 percent of the mobile data traffic. In 2013, on an average, a smart device generated 29times more traffic than a non-smart device.*
- *Smartphones represented only 27 percent of total global handsets in use in 2013, but represented 95percent of total global handset traffic. In 2013, the typical smartphone generated 48 times more mobile data traffic (529 MB per month) than the typical basic-feature cell phone (which generated only 11 MB per month of mobile data traffic).*
- *Globally, there were nearly 22 million wearable devices (a sub-segment of M2M category) in 2013 generating 1.7 petabytes of monthly traffic.*

Nonetheless, within the Big Data discussions a component that is often neglected is the communication infrastructure. It is the task of the communication infrastructure to provide the means to achieve a timely exchange of information between the different components. While no generic absolute values are defined, the amount of bandwidth and permitted latency for the communication in Big Data scenarios presents an important challenge.

Not many years ago, before Big Data came into play, big research centres for the biomedical industry and radiometric astrophysics used to physically move large amounts of mass storage devices in order to deliver information from one place to another. Under the informal name of “Sneakernet”, and due to the huge density of current information in data storage devices (16 Petabytes per cubic meter using commercially available consumer hard disk drives), companies today still ship storage devices to other offices or to data centres.

Under the premise of Big Data, huge amounts of data have to be communicated in real time between the different layers. Taking the lambda architecture (see 3.4.2) as a basis, the communication is not only required between the batch, serving and speed layers, but also between the components included in each of these layers: This implies providing communication between data sources, data storage units, computing units and actuators, all of them geographically distributed.

For the real time functionality of a Big Data system, new data will be constantly received and analyzed. For these real time analytics and decision making, the traditional Content Delivery Networks will be of limited effectiveness, as it is an artificial external construction to propagate the information. The distributed file systems which are part of the Big Data architectures, for example the lambda architecture, are better suited for this job, as they have direct knowledge of where each piece of data is required.

Another partial solution is the use of Wide Area Network (WAN) Optimizers. As the current communication protocols have been developed over 40 years ago, the mechanisms used weren’t originally thought of for use with the relatively high bandwidths and communication distances (manifesting as communications latency) such as the ones found today. The shortcomings that current communications protocols suffer are well known. Through the optimization of these protocols, bringing advantages to throughput, bandwidth requirements, latency, and congestion, WAN optimizers achieve a significant increase in the connection speeds. These optimizers lead to very large improvements in intercontinental communication links (i.e. 5x-10x of bandwidth between Europe and Asia, 2x-3x of

bandwidth between Europe and America), but have little or no effect when communicating inside the same country. Another feature of WAN optimizers is the ability to cache content locally, so that sending and receiving of information can be done faster. A local index is kept with commonly sent information contents, so that a pointer to the cache can be sent instead of the full packet contents. This ability can reduce the communication latency, as well as reduce the bandwidth required (thus reducing congestion). Unfortunately, as it looks for patterns in the communication, it will not work with encrypted links, and it will bring no improvement if frequently changing sensor data is being forwarded. It should be mentioned that, for WAN optimizers to work, such optimizers are required at both ends of the communication channel.

The challenge of supplying Big Data with all of its communication requirements can be divided in two parts: local communication and non-local communication. Local communication refers to the exchange of information inside the same server, rack, or data centre. Non-local communication identifies the data exchange which takes place outside of the data centre. Even though the communication requirements set by Big Data to both local and non-local may be the same (in regards to bandwidth, latency and reliability), the possibility of fulfilling these requirements usually differs.

In the case of non-local communication such as the one required for operational optimization of cyber-physical systems, the volume and variety of data as well as the velocity at which data will be communicated may present a bottleneck in the Big Data architecture. As a simple example, let’s take the braking system of a car. A typical ABS braking system requires the sensors for the wheel speed, the current mechanical power been forwarded to each tire, lateral acceleration, steering wheel angle, centrifugal forces and yaw rate. These sensors generate information hundreds of times per second and this information is analyzed at the same rate. The current wireless communication channels built into the vehicle are not able to provide enough bandwidth to forward every single point of information to the outside world in real time.

A good approach to solve such bottlenecks is to utilize the in-field analytics paradigm discussed in Section 5.4. In-field analytics takes into account that some decisions have to take place locally, as they must happen in a time frame in which the communication and the analysis can be done quickly enough. The local intelligence is enriched by the extensive knowledge discovery facilitated by big data computing in the enterprise level.

Such bottlenecks can be further reduced by the use of data compression before the transmission of the information. The most common class of data compression is lossless compression, which allows the original data to be perfectly reconstructed from the compressed data. Through the generation of statistical models for the input data, bit sequences can be mapped to the input data in such a way that probable data produces shorter output than less probable data.

The second class of data compression algorithms are lossy data compression, which permits the reconstruction only of an approximation of the original data, though typically improving the compression rate. Lossy data compression is often used for the compression of video and images, while still maintaining the most important elements of the original data, so that it is still useful for further analysis.

One big change that Big Data is pushing in communication networks is the further use of peer-to-peer communication (see also further discussion in next section). As data processing can be partitioned between different processing units to accelerate the processing speed and also improve resiliency in the case of failure, the parallel communication also speeds up communication significantly. Through the concurrent sum of several communication links,

the partitioned data is able to arrive to a destination quicker than when the data is coming from a single source. The reason for this is that the weaknesses of the old network communication protocols affect multi-source communications less than a single source.

## 3.3 DISTRIBUTED STORAGE AND COMPUTING PARADIGMS

In distributed computing, each processor has its own physical memory, i.e. distributed memory. Information is exchanged by passing messages between the processors. Due to these characteristics, distributed systems can achieve linear or even logarithmic scalability depending on the networking protocol[25]. An important goal as well as a major challenge in designing distributed systems is the so-called location transparency. Location transparency enables the development loosely coupled applications.

Parallel computing, Cloud computing, and stream computing are some of the distributed storage and computing paradigms behind Big Data technologies. In the following, a short explanation is give with examples in each Subsection.

Distributed storage and computing are concepts researched and developed since the 1950s as well: The concept of virtual memory was developed by German physicist Fritz-Rudolf Güntsch[26]. Storage that is managed by the integrated hardware and software, which abstract from the details (i.e. location transparency) allows for data processing without the hardware memory constraints. These distributed storage and computing paradigms, collectively dubbed "big-data computing," will transform the way we consume data similar to how search engines have transformed how we access information[27].

### *3.3.1 Nosql Databases, Distributed File Systems and Parallel Computing*

Volume and velocity of data requires a flexible data management stack which supports multiple types of storage depending on the various data types and usage scenarios. The following range of technological capabilities is commonly put to use to optimize cost and performance of data management:

Traditional data management solutions handle highly structured, mainly relational data, often of transactional value or of descriptive value, such as:

- Data warehouses manage highly valuable data, which is not only well-defined, structured, but also cleansed and integrated for special and known, i.e. recurrent purposes, such as financial reporting.
- Transactional DBs provide storage for highly structured data in business transactions, which enable the accountability of business and hence are per default valuable.

*Massively Parallel Processing* (MPP) is a computational approach that has been in use since 1984[28] for performance increase in massive and high-end relational data warehouses.

---

[25] E.g. by implementing an efficiently distributable data structure such as Distributed Hash Tables, for an original discussion see Stoica, Ion, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. "Chord: A scalable peer-to-peer lookup service for internet applications", *In Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, 2001, ACM, pp.149-160.

[26] Jessen, E., "Origin of the Virtual Memory Concept", *IEEE Annals of the History of Computing*, Vol. 26, April 2004, p. 71

[27] Bryant, Randal E., Randy H. Katz, and Edward D. Lazowska, "Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society," Version 8, 22 December 22 2008. http://www.cra.org/ccc/docs/init/Big_Data.pdf

[28] Pereira, Brian, "Marrying Strategic Intelligence with Operational Intelligence", *InformationWeek*, 1 January 2010. http://www.informationweek.in/informationweek/news-analysis/239690/marrying-strategic-intelligence-operational-intelligence

MPP enables processing of data distributed across a cluster of compute nodes. These separate nodes (separate memory and processor in each node) process their data in parallel and the node-level output sets are assembled together to produce a final result set. In these high-end environments it can also be said that typically the utilized hardware is highly specialized, tuned for CPU, storage and network performance. But most importantly, MPP fully supports SQL and the relational data storage model of traditional data warehouses for which it was designed.

With increased digitization, Big Data means that data in their raw formats hardly contain any actionable information. Typical examples for this are time series data from sensors or the vast availability of unstructured data resulting from electronic processes and communication, such as email, etc. At the same time these multi-structured data contain many insights for many different purposes, even purposes which may not be yet known. Hence, the rationale is to manage these sources with the most cost-effective technologies, which again differs by type and intended use of data e.g.:

- *NoSQL databases* commonly describe non-relational data stores, such as Graph DBs, key-value stores, Columnar DBs, etc. Whilst graph databases enable faster results for geo-spatial analysis, Columnar DBs may prove most efficient for feature discovery in time series. There are over 100 NoSQL databases [Reference], the most common types are described below:

  | | |
  |---|---|
  | *Key-Value store* | Key-value (KV) stores use the associative array (also known as a map or dictionary) as their fundamental data model. In this model, data is represented as a collection of key-value pairs, such that each possible key appears at most once in the collection. Distributed Hash Tables are the most commonly used data structures for storing massive amounts of data. |
  | *Columnar DB* | Data in columnar model is kept in column as opposed to rows in relational DBs and row storage models. Since column-based functions are very fast – there is no need for materialized views for aggregated values in exchange for simply computing necessary values on the fly; this leads to significantly reduced memory footprint as well. |
  | *Graph DB* | In computing, a graph database is a database that uses graph structures with nodes, edges, and properties to represent, store, and access data efficiently. |

- *Hadoop*, which in its core consists of a *Distributed File System (HDFS)* and *MapReduce* for processing data in parallel within the distributed file system, is another flavor of managing massive amounts of data. While in principle very close to MPP, MapReduce and Hadoop find themselves deployed to clusters of commodity servers. The commodity nature of typical Hadoop hardware, the free nature of Hadoop software, combined with the programmable self-administration capabilities of the underlying distributed file system means that clusters can grow as data volumes do very cost-efficiently (given the programmatic skills required to use Hadoop).

  Whilst traditionally a "*Historian*" is the name for an archive of time series data, in the world of multi-structured data, the term "*Data Lake*" has been established for describing the cost-efficient archival of any type of historical data. Data Lake needs to have the least cost per byte.

  | | |
  |---|---|
  | *Data Lake* | A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. When a business or engineering question arises, the data lake can be queried for relevant data (Schema on Demand – |

see 3.4.1), and that smaller set of data can then be analyzed to help answer the question. In comparison, a data warehouse contains only preprocessed and selected data for a special purpose defined before the acquisition of data.

The choice depends on the cost-effectivity of a technology for the type and intended usage of data, which may change over the lifetime of a data source.

### 3.3.2 Cloud Computing

Similarly to "Big Data" the concept behind "Cloud Computing" reaches back to 1950: The first mainframe computers had the same rationale behind: multiple users share both the physical access to the computer from multiple terminals as well as the CPU time. This eliminated periods of inactivity on the mainframe and allowed for a greater return on the investment.

Cloud computing is the result of evolution and adoption of existing technologies and paradigms (see 3.3.1). The commercialization of these techniques as services offered on Internet-scale was pioneered by Amazon's Web Service offerings for Cloud storage S3 and Elastic Compute Cloud (3.1). The goal of cloud computing is to allow users to take advantage of all of these technologies, without the need for deep knowledge about or expertise with each one of them. The cloud aims to cut costs, and help the users focus on their core tasks instead of being impeded by IT obstacles. The following is a short list of these technologies and paradigms[29]:

- *Virtualization* software allows a physical computing device to be electronically separated into one or more "virtual" devices, each of which can be easily used and managed to perform computing tasks.
- *Autonomic computing*[30] automates the process through which the user can provision resources on-demand. By minimizing user involvement, automation speeds up the process, reduces labor costs and reduces the possibility of human errors
- Concepts from *Service-oriented Architecture (SOA)* help the user break complex business problems into services that can be integrated to provide a solution[31]. Cloud computing provides all of its resources as services[32].
- Concepts from *Utility Computing*[33] in order to provide metrics for the services used. Such metrics are at the core of the public cloud pay-per-use models. In addition, measured services are an essential part of the feedback loop in autonomic computing, allowing services to scale on-demand and to perform automatic failure recovery
- *Grid Computing* paradigms was improved by addressing the QoS (quality of service) and reliability problems. Cloud computing provides the tools and technologies to build data/compute intensive parallel applications with much more affordable prices compared to traditional parallel computing techniques.

[29] adapted from http://en.wikipedia.org/wiki/Cloud_computing

[30] http://en.wikipedia.org/wiki/Autonomic_computing refers to the self-managing characteristics of distributed computing resources, adapting to unpredictable changes while hiding intrinsic complexity to operators and users.

[31] ACMqueue, op. cit., 2006.

[32] As the most prominent example see Amazon Web Services at http://aws.amazon.com/

[33] http://en.wikipedia.org/wiki/Utility_computing is the provisioning model of on-demand computing such as grid computing; a service provider makes computing resources and infrastructure management available to the customer as needed, and charges them for specific usage rather than a flat rate.

- *Peer-to-Peer (P2P) Computing* paradigms have been deployed to allow for Internet-scale distributed data management in data centers. Distributed Hash Tables (DHT)[34] is such a P2P pattern that has been adopted within Amazon's Dynamo[35], the distributed data management behind many of the Amazon Web Services, or by Facebook's Cassandra[36]. This type of data storage became commonly known as *key-value store* within the class of *NoSQL databases* (see 3.3.1).

### 3.3.3 Distributed Stream Computing Frameworks

Many of the mature, commercial stream computing frameworks today, such as StreamBase[37] or Apama[38], are complex event processing (CEP) software. CEP or event processing systems (EPS) consume and react to a stream of event data in real-time. High frequency trading is a domain which widely uses CEP to identify, make sense of and react quickly to patterns in streams of event data. In order to incorporate this, traditional CEP requires the definition and management of rules, events, and patterns.

CEP moves event stream processing toward the more complex "big data domain" of generating new knowledge through correlating a variety of (event) data sources. However, the rule engine – core of a CEP framework – can grow unmanageably complex in Big Data scenarios, in which in addition to high-velocity and variety, also volume of data plays a significant role.

Hence, the newer technological solutions in the Big Data domain, such as Apache Storm[39] (originated at Twitter), Apache Samza[40] (originated at LinkedIn), or Apache Spark Streaming[41] are all highly scalable, fault-tolerant, *distributed* stream computing frameworks. Both, CEP and stream computing, store queries and run the data through those queries – as opposed to storing data and running queries against the data such as with typical databases. However, there are several differentiators: e.g.

- distributed stream computing enables horizontal scalability, which is an important factor to tackle massive amounts of data in a cost-efficient manner,
- through the focus on the low level "data-as-is" (versus events) and streaming algorithms (versus rule engine), distributed stream computing also remains more flexible to incorporate machine learning techniques, e.g. data stream mining, pattern matching etc., instead of predefining events and rules.

However, the concept of CEP can be scaled to big data by utilizing a distributed stream computing framework, since the computing model, i.e. sending streaming data through stored queries, is the same.

---

[34] Stoica, et al., op. cit., 2001.
[35] DeCandia, Giuseppe, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels, "Dynamo: Amazon's Highly Available Key-value Store", *In Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles - SOSP '07*, p. 205. doi:10.1145/1294261.1294281
[36] Lakshman, Avinash, Prashant Malik, "Cassandra - A Decentralized Structured Storage System", *In Proceedings of the ACM SIGOPS Operating Systems*, Review archive Vol. 44, No. 2, April 2010, pp. 35-40.
[37] StreamBase was the commercialization of the academic project Aurora at MIT and acquired by TIBCO 11 June 2013, for further details see http://www.streambase.com/
[38] Apama Real-time Analytics company acquired by Software AG on 13 June 2013, for further details see http://www.softwareag.com/corporate/products/bigdata/apama_analytics
[39] For further information see https://storm.incubator.apache.org/
[40] For further information see http://samza.incubator.apache.org/
[41] For further information see https://spark.apache.org/streaming/

## 3.4 BIG DATA ARCHITECTURAL PATTERNS

Big data technologies or computing, as discussed in the previous Section, is mainly built on distributed storage and computing paradigms. Currently, the deployed solutions in use are mainly concerned with overcoming the challenges that high-volume and high-velocity brings with them. Below is a description of the two architectural patterns that allow to

1) cost-efficiently manage high volumes of multi-structured data
2) cost-efficiently manage both the high-velocity incoming data streams and their high-volume accumulation over time, for all kinds of queries on any type of data

Beyond these two established patterns, research and development today is also concerned with the third dimension of Big Data: variety, e.g. through adapting semantic technologies to scale in Big Data environments, and with the many other challenges such as veracity, actionability, privacy as discussed in Section 5.

### *3.4.1 Schema on Read*

Traditional data warehouses as well as relational database management systems rely on what is called "Schema on Write." The structure of the data is determined before it is acquired. This structure, schema, is applied when data is written into the database. However, there are several drawbacks when applied to big data, especially for the purpose of big data analysis[42]:

- Schemas are typically purpose-built and hard to change
- Loss of the raw/atomic data as a source
  Loss of information: If a certain type of data cannot be confined in the schema, it cannot be effectively stored or used
- Unstructured and semi-structured data sources tend not to be a native fit

Schema on Read follows a different sequence: The data is written into storage as is and only schematized after the business and engineering question arises, giving structure to the data depending on the way data needs to be analyzed to yield the appropriate answer. This is the typical data-driven mindset which even led to Big Data. The concept behind this pattern is related to[43] "Extract Load Transform" (ELT) in MPP.

A distributed file system like Hadoop has advantages over relational databases to better perform ELT with massive amounts of unstructured or multi-structured data in batch. 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.

[42] Tom Deutsch, "Why is Schema on Read So Useful?", *IBM Data Magazine*, 13 May 2013. http://ibmdatamag.com/2013/05/why-is-schema-on-read-so-useful

[43] A discussion can be found at http://blog.cloudera.com/blog/2013/02/big-datas-new-use-cases-transformation-active-archive-and-exploration/

### 3.4.2 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[44].

2013 is the year when some consolidation into the Big Data technology stack arrived: The so-called lambda architecture[45]. 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 2 is an adapted depiction of the lambda structure to show also the architectural sub-patterns that enable the required characteristics of each layer:

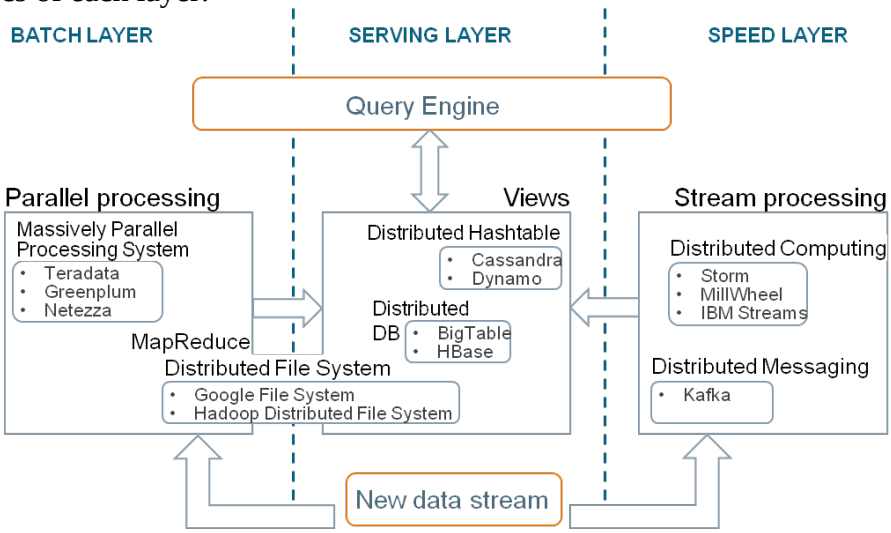


**Figure 2 The lambda architecture with three layers: batch, speed and serving layer; (source: Siemens AG Corporate Technology adapted from Nathan Marz to display the storage and distributed computing paradigms (see 3.2.1, 3.2.3) with examples, for each layer**

1) **The batch layer is a cost-efficient active archive of big data (raw and transformed) 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

[44] http://nathanmarz.com/about/
[45] Marz, Nathan, "Big Data – Principles and best practices of scalable realtime data systems", *Manning MEAP Early Access Program*, Version 17, no date. http://manning.com/marz/BDmeapch1.pdf

to each of the nodes and the hand-over logic is based on distributed computing for cost-efficiency reasons. Stream computing can also be utilized for ETL for analysis and archival of real-time streaming data.

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.

# 4 BIG DATA ANALYTICS APPLICATIONS

The analysis of the potential of the Big Data economy including traditional sectors was a focal work area of the European Project BIG – Big Data Public Private Forum[46]. In the pursuit of collecting the many sectorial requirements towards a European Big Data economy and its technology roadmap, Big Data applications have been analyzed[47]. A finding that is also congruent with Gartner's study on the advancement of analytics[48] (see Figure 3) is that big data applications can be categorized in "Operational Efficiency," "Customer Experience," and "New Business Models."

| | Mfg. | Education | Banking | Insurance | Govern-ment | Energy & Utilities | Comm, Media & Svcs. | Retail | Trans-portation | Health-care |
|---|---|---|---|---|---|---|---|---|---|---|
| Process efficiency/ cost reduction | 1 | 3 | 3 | 3 | 1 | 1 | 1 | 2 | 1 | 1 |
| Enhanced customer experience | 3 | 2 | 5 | 1 | 2 | 2 | 2 | 1 | 3 | 3 |
| Improved customer service | 4 | 6 | 6 | 4 | 4 | 4 | 4 | 3 | 2 | 2 |
| Identifying new product/market opportunities | 2 | 7 | 4 | 2 | 7 | 5 | 3 | 5 | 4 | 4 |

**Figure 3 Excerpt from Gartner's study on business problems to solve with big data analytics (1= top prio). Heat map also indicates that the highest priority business applications powered by big data analytics can be categorized into Operational Efficiency, Customer Experience, New Business Models**

In this Section we give a brief introduction into each category of business applications (4.1), analytics applications (4.2), and give some examples from each sector (4.3) that also was part of the EU BIG sectorial forums.

## 4.1 CATEGORIES OF BIG DATA BUSINESS APPLICATIONS

New data sources come into play mainly through increasing levels of digitization and automation. *Operational efficiency* is the main driver (also see Figure 3) behind the

[46]EU BIG Project web site http://www.big-project.eu/

[47]Zillner, Sonja, Sabrina Neururer, Ricard Munné, Elsa Prieto, Martin Strohbach, Tim van Kasteren, Helen Lippell, Felicia Lobillo Vilela, Ralf Jung, Denise Paradowski, Tilman Becker and Sebnem Rusitschka, "D2.3.2: Final Version of Sectors' Requisites", BIG project Deliverable 2.3.2, 31 July 2014. http://big-project.eu/sites/default/files/BIG_D2_3_2.pdf

[48]Kart, Lisa, "Advancing Analytics", April 2013, p. 6. http://meetings2.informs.org/analytics2013/Advancing%20Analytics_LKart_INFORMS%20Exec%20Forum_April%202013_final.pdf

investments for digitization and automation. The need for operational efficiency is manifold, such as revenue margins, regulatory obligations, or the retiring skilled workers.

Once pilots of big data technologies are setup to analyze the masses of data for operational efficiency purposes, industrial businesses realize that they are building a digital map of their businesses, products, and infrastructures – and that these maps combined with a variety of data sources also deliver insight into asset conditions, end usage patterns etc. *Customer experience* becomes the next target of value generation – just because it is now in their reach to find out more about how their products, services or infrastructure are being utilized and where frustrations might be remedied, how customers can even be excited. Stakeholders in customer-facing segments of the two sectors may start with customer experience related big data projects right away, but soon discover the close interactions with operational efficiency related big data scenarios.

Literally playing around with data makes creative. Some analytical insights reveal that systems are used in ways that they were not planned for, or that there is value in some niche, which previously was unknown, or that there is value in own data for some other organization possibly in another sector. Hence, *new business models* are the third pillar of big data application. Departments such as strategy or business development in the companies also benefit from this data-driven creativity.

## 4.2 CLASSES OF BIG DATA ANALYTICS APPLICATIONS AND TECHNIQUES

Depending on the business or engineering questions ("What happened", "Why did it happen?", "What will happen?", "What shall I do?"), analytics can be differentiated conceptually into four types of analytics applications. The four types vary in the feedback they deliver, the input needed, and the level of complexity as briefly discussed in Table 2 The table also indicates maturity[49] and adoption by business[50] of each of the analytics application classes.

| Analytics application | Description |
|---|---|
| **Descriptive analytics**<br>"What happened?"<br>70% adoption | Descriptive analytics involves standard aggregate functions, compute descriptive statistics by simple arithmetic to help group or filter the data. Descriptive analytics basically summarizes what happened. |
| **Diagnostic analytics**<br>"Why did it happen?"<br>30% adoption | Diagnostic analytics tell the user why something happened. Especially device know-how modeled through semantic technologies and reasoning enriches the data model such that answers to why something happened can be extracted. Data mining also offers powerful tools to extract correlations, which then become interpretable through the semantically enriched data model. |

[49] Gartner, "Gartner's 2013 Hype Cycle for Emerging Technologies Maps Out Evolving Relationship Between Humans and Machines", 19 August 2013. http://www.gartner.com/newsroom/id/2575515
[50] Kart, op. cit. , 2013.

| Analytics application | Description |
|---|---|
| **Predictive analytics**<br>"What will happen?"<br><br>16% adoption<br>Plateau of productivity | Predictive analytics allows spanning a trajectory of what will happen. For this purpose, also an enriched data model is required. Data mining and machine learning offer methodologies and algorithms, which enhance the extraction of knowledge. Statistical models help summarizing the huge amounts of data; but most importantly that model can also be used to extrapolate. In case of temporal data, predictive analytics yield forecasts on the future development. In other cases where data is too expensive to measure or not available in the scale required, models are also be used to predict the data, i.e. outcome, that we don't know otherwise. Beyond this purely statistical view of data modeling, analytical system models, neural networks, or reasoning enhanced with domain and device know-how enable improved prediction of the behavior of a complex system. |
| **Prescriptive analytics**<br>"What shall I do?"<br><br>3% adoption<br>Innovation trigger | With prescriptive analytics the simple predictive model is enhanced with possible actions and their outcomes, as well as an evaluation of these outcomes. In this manner, prescriptive analytics not only explains what might happen, but also suggests optimal set of actions. Simulations and optimization are analytics tools that support prescriptive analytics. In cyber-physical systems, which are enhanced with real-time monitoring, model- and data-driven analytics is the essence of efficient prescriptive analytics, since the correlations that are discovered in the data can additionally be explained or verified and acted upon according to the system models. |

**Table 2: Data analytics application classes (with additional information on maturity and adoption from Gartner)**

Depending on the analytics application, which the business or engineering user needs, the requirements for data modeling and analytics are formulated. Behind these high-level classes of analytics applications there are many different types and techniques of analytics:

Data presentation may be the last or the first step of analytics: In a traditional sense, the results of the different analytics forms can be presented to the business user, as a last step, in the form of traditional or innovative **reports and dashboards**. This type of descriptive analytics is mainly passive, but with technological advancements, especially in web and visualization, descriptive analytics also becomes more interactive.

As a first step of analytics, data presentation, through **visual analytics**, enables data scientists and business users to faster gain a common understanding of the quality or value of data and formulate appropriate business and engineering questions to operationalize that value. Visual analytics also ensures the validation of assumptions about the quality of data, which in turn improves the data integration phase. Adapting visual analytics to the masses of Big Data is a challenge. At the same time, visual analytics can be a very powerful step in the process of knowledge discovery from mass data. The following depiction Figure 4 shows how visual data exploration and automated data analysis intertwine. In the Big Data domain this combined process needs to be automated and can hardly remain a manual process as it is today.

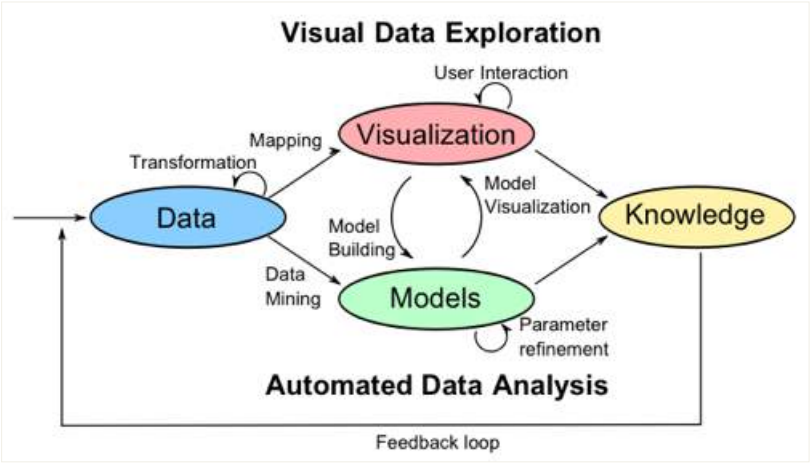


**Figure 4: The combined process of visual data exploration and automated data analysis[51] needs to be streamlined and automated to scale over Big Data**

Diagnostic analytics applications require the interpretability of data coming from a variety of sources and deep analytics thereof. **Semantic technologies** can enrich the data model such that reasoning can extract answers as to why something happened. For instance, ontologies can capture domain and system know-how, which can be modeled semantically. Ontologies improve the interpretability of data for humans and machines alike. **Automatic reasoning** can then be employed on the semantically enriched data, in order to reliably deduce root causes of a disturbance in complex systems. One of the benefits of automated reasoning is that it improves the statistical accuracy through additional classification techniques. Other techniques reach from rule engines to complex event processing engines. These engines in combination with business process management or embedded in cyber-physical systems enable enhanced diagnostics and decision support.

With the increasing amounts of data, especially due to the dynamicity of the underlying system, statistical modeling becomes more complex. **Data mining** aims at automating the modeling process. With respect to the vast amounts of data, data mining can be useful as the first step of **machine learning**. Tools and techniques reach from unsupervised learning, e.g. through clustering, or supervised learning such as with neural networks. Graphical models, such as Bayesian networks, enable the description of high-dimensional data. Which of these models are most suitable, needs to be tested through data exploration. The management of training data sets requires both domain and analytics expert know-how, but is indispensible in improving the accuracy of the models.

Sometimes also called “deep learning,” these techniques enable the realization of predictive analytics. Depending on the data and business or engineering question at hand the most cost-effective techniques can be chosen. There will always be a tradeoff between efficiently handling the data modeling and analytics complexity and the degree of actionability of the knowledge generated from raw data.

Especially in cyber-physical systems, reinforcement learning represents an opportunity for the learning-based optimization of autonomous agents. **Optimization**, in general, is about finding the “best” solution to a problem given the constraints of the system. As such, optimization is one of the driver technologies behind prescriptive analytics. At the same time, classical optimization problems are not designed to scale to the masses of data that industrial

[51] Keim, D. A., F. Mansmann, J. Schneidewind and H. Ziegler, “Challenges in visual data analysis”, *In Proceedings of the Tenth International Conference on Information Visualization*, 2006, pp. 9-16.

digitalization will bring along. Novel algorithm design and adaptation with domain know-how as well as adaptation for execution in distributed environments such as cloud and massively parallel processing will be key differentiators.

Prescriptive analytics can also provide decision options and show the impact of each decision option so that operations managers can proactively take appropriate actions on time. Prescriptive analytics combined with real-time data from field devices and to enhance decision automation in general or the control and automation capabilities of field devices, such as in industrial, building, and energy automation.

Whilst real-time prescriptive analytics is a promising research area, the myriad of engineering and business processes today still involve manual steps, in which field crew, technicians, or sales professionals generate textual data. The integration of such unstructured data requires **natural language processing and search** techniques, such as classification, indexing, or information retrieval and extraction. Natural language processing and search are technologies that allow the findability, interpretability, and traceability of unstructured data. Question answering is just one of the many subfields, which are relevant for businesses in the digital transformation.

## 4.3 EXAMPLES FROM DIFFERENT SECTORS

In some industrial sectors Big Data potential has been around as long as in the native online data-driven businesses. The use of functional neuroimaging (fMRI) in healthcare resulted in increased possibilities to improve diagnosis, however, also in the data volumes to be acquired, analyzed, and stored. One cubic millimeter of mouse brain represents a thousand terabytes (a petabyte) of image data[52]. Other sectors are still at the verge of digitization: The complete roll-out of smart meters, which can deliver energy usage data every 15 minute, will result in a 3,000-fold increase of data to be acquired, managed, and analyzed in the utilities business. However, there are vast difficulties in moving from analog meters and manual reads to digital meters[53].

Data analyses have also been used heavily for decades in industrial businesses. Using data analytics to increase operational efficiency has always been a business lever. However, in the following subsections many of the cited use cases with the much higher amounts of data and analytical capabilities than business-as-usual are still pilots. Digitization and automation clearly have not yet reached the tipping point. Nonetheless, the rise of Big Data (volume, variety, velocity) with increased digitization, enables data analysts to draw conclusions based on more data – given the capabilities to reduce time for analytics from days and months to hours and minutes.

Big Data computing will be the accelerator in many businesses in the traditional sectors, especially due to the open source enabled spillover of Big Data technologies from online data businesses. In the following, some examples are listed from some of the sectors reaping considerable value from Big Data and Big Data technologies. Some applications have been around for decades but can now be delivered more efficiently. Some applications are newly enabled through Big Data and their value potential can only be evaluated qualitatively yet.

[52] Shaw, Jonathan, "Why "Big Data" Is a Big Deal", *Harvard Magazine*, March-April 2014. http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal

[53] Probert, Tim, "Making dumb European rollouts SMART - Three golden rules of consumer engagement", *Intelligent Utility*, January/February 2013. http://www.intelligentutility.com/magazine/article/300331/making-dumb-european-rollouts-smart

Most of the examples show what value has been generated or what challenges have been overcome by utilizing big data computing. Naturally, it is difficult to find more details on how these industrial businesses have been using the mentioned technologies. Hence, these motivated examples are rather a point of reference.

### 4.3.1 Oil & Gas Industry

Oil & gas is a vertically integrated segment of the energy sector covering all activities from exploration, development, production, transport (pipelines, tankers), refinery to retail, all of which have many potentials regarding horizontal and vertical scale of big data opportunities.

- A fully optimized digital oil field can yield 8 percent higher production rates and 6 percent higher overall recovery according to industry estimates[54].
- Example Use Case Shell[55]: optical fibre attached downhole sensors generate massive amounts of data that is stored at a private isolated section of the Amazon Web Services. They have collected 46 petabytes of data and the first test they did in one oil well resulted in 1 petabyte of information. Knowing that they want to deploy those sensors to approximately 10,000 oil wells, we are talking about 10 Exabyte's of data, or 10 days of all data being created on the internet. Because of these huge datasets, Shell started piloting with Hadoop in the Amazon Virtual Private Cloud
- Others use cases[56]: Chevron proof-of-concept using Hadoop for seismic data processing; Cloudera Seismic Hadoop project combining Seismic Unix with Apache Hadoop; PointCross Seismic Data Server and Drilling Data Server using Hadoop and NoSQL; University of Stavanger data acquisition performance study using Hadoop.

### 4.3.2 Electricity Industry

The electricity industry is by far the most vivid sector for future big data applications, as it is at the verge of digitization and liberalization, both of which are very transformative for a sector (e.g. in comparison to the transformation of the telecommunications sector): with liberalized roles in power generation and retail as well as metering services, and regulated infrastructure operators of transmission and distribution networks; with exciting new market segments through decentralized and renewable energy resources and direct marketing, as well as demand response and energy efficiency markets with new players in the roles of energy data-driven service providers .

- Example use case EDF[57]: Currently, most utilities do a standard meter read once a month. With smart meters, utilities have to process data at 15-minute intervals. This is about a 3,000-fold increase in daily data processing for a utility, and it's just the first wave of the data deluge. Data: individual load curves, weather data, contractual

[54] Leber, Jessica, "Big Oil Goes Mining for Big Data - As petroleum production gets trickier, digital innovation becomes more crucial.", *MIT Technology Review*, 8 May 2012. http://m.technologyreview.com/computing/40382/

[55] Mearian, Lucas, "Shell Oil Targets Hybrid Cloud as Fix for Energy-Saving, Agile IT", *Computer World Magazine*, 4 April 2012. http://www.computerworld.com/s/article/9225827/Shell_Oil_targets_hybrid_cloud_as_fix_for_energy_saving_agile_IT

[56] Nicholson, Rick, "Big Data in the Oil & Gas Industry", *IDC Energy Insights*, 2012. https://www-950.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/RICK%20-%20IDC_Calgary_Big_Data_Oil_and-Gas/$file/RICK%20-%20IDC_Calgary_Big_Data_Oil_and-Gas.pdf

[57] Picard, Marie-Luce, "A Smart Elephant for A Smart-Grid: (Electrical) Time-Series Storage And Analytics Within Hadoop", 26 June 2013. http://www.teratec.eu/library/pdf/forum/2013/Pr%C3%A9sentations/A3_03_Marie_Luce_Picard_EDF_FT2013.pdf

information, network data 1 measure every 10 min for a target of 35 million customers (currently only 300,000)- Annual data volume: 1800 billion records, 120 TB of raw data. The second wave will include granular data from smart appliances, electric vehicles and other metering points throughout the grid. That will exponentially increase the amount of data being generated.

- Example use case Power Grid Corporation of India[58]: Unified Real-time Dynamic State Measurement with a target integration of around 2,000 PMUs (starting with 9 PMUs). The PMUs provide time-stamped measurements of local grid frequency, voltage, and current at 25 samples per second (up-to 100 Hz possible). This amounts to about 1 TB of disk space per day. Traditional SCADA systems poll data from RTUs every 2-4 seconds. For comparison: A traditional SCADA system for wide area power system management has 50 times more data points than a system that locally processes high-resolution data received from PMUs. Yet, the SCADA system has to process less than 1% of the data volume of the entire PMU data accumulating in the same area[59].

### 4.3.3 Manufacturing Industry

Manufacturing industry also has access to increasingly more data, especially from operations as well as from other parts of the supply chain. The growing complexity of interdependent processes, when tamed through advanced analytics, can catapult a manufacturer's competitive advantage such as efficiency, flexibility, or agility.

Following is an example in healthcare products from a research by McKinsey including many other manufacturing examples[60]:

- Biopharmaceuticals (e.g. vaccines, hormones, and blood components) maker needs to monitor more than 200 variables within the production flow to ensure the purity of the ingredients, i.e. genetically engineered cells. Two batches with the identical process can have a difference in yield by 50 percent resulting in issues with production output. The maker could increase its vaccine yield by more than 50 percent—worth between $5 million and $10 million in yearly savings for a single substance through advanced analytics utilizing vast available data.

Another example[61] covering car manufacturing is described in the following:

Opportunity:

- Only a few experts have a complete overview over all available data for a car – from design to production to after-sales service.
- Providing and analyzing such data can improve quality and allow early error recognition.

Data & Analytics:

[58] Power Grid Corporation India, Ltd., "Unified Real Time Dynamic State Measurement (URTDSM)", *Technical Report*, February 2012.
http://www.cea.nic.in/reports/powersystems/sppa/scm/allindia/agenda_note/1st.pdf retrieved on 2013-08-09
[59] NERC., "Real-time Application of Synchrophasors for Improving Reliability", 18 October 2010.
http://www.nerc.com/docs/oc/rapirtf/RAPIR%20final%20101710.pdf
[60] Auschitzky, Eric, Markus Hammer, and Agesan Rajagopaul, "How big data can improve manufacturing", *McKinsey & Company Insights & Publications*, July 2014.
http://www.mckinsey.com/insights/operations/how_big_data_can_improve_manufacturing
[61] BITKOM, "Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte", 2012, p.77.
http://www.bitkom.org/files/documents/BITKOM_LF_big_data_2012_online(1).pdf

- Approx. 10 TB of data through new, more complex electronic car components
- Creation of a common API for the over 2000 internal users with different functionalities and analyses
- Collection of all data sources in a central data warehouse ensures data consistency and clear responsibilities
- E.g. text mining and reporting speed up error detection and correction for cars
- Optimization of analytics support, standardisation of analytics and reporting in after-sales and technologies.

Results:

- Decision making based on all quality related data for cars; synergies in after-sales and technologies
- Increase in customer satisfaction
- Increase in profits through early warning systems
- Increase in quality through identification of errors

### 4.3.4 Logistics Industry

Logistics industry has high potential in efficiency increase and interesting big data monetizing options, which are also highly driven by e-commerce.

- Trucking company Xpress is reported to have saved over $ 6 million a year by combining 900 different data elements (such as sensors for petrol usages, tyres, brakes, engine operations, geo-spatial data and driver comments across a fleet of 8.000 tractors and 22.000 trailers) from 10.000s trucking systems into Hadoop and analyze it all[62].
- DHL[63] employs more than 300,000 people worldwide with a fleet of 60,000 vehicles and 250 daily cargo flights. Developed a single point of collection for all costing data for every single package delivered with a complex pricing model - such as discounting shipments to fill up planes which would have otherwise flown half-empty. Value: able to save 7,000 man days per year on unnecessary costing exercises plus more accurate pricing – even dynamic pricing became possible.
- On the other side, only 9 percent of logistics professionals have complete visibility into their supply chains because of the lack of access to quality data and a reliance on manual processes, according to a 2013 KPMG report[64].

### 4.3.5 Retail Industry

In the era of personalization and always connected end users, the retail industry – the last link of the supply chain, buying goods directly from manufacturers and selling them to end users for profit – finds itself in the midst of the Big Data question of how to utilize all the clues to sentiment, behaviour, and preferences of end users now available in unstructured formats generated through user interaction with social and mobile services[65], to increase their retail profits.

[62] Big Data Startups, "Trucking Company US Xpress Drives Efficiency With Big Data", 2013. http://www.bigdata-startups.com/BigData-startup/trucking-company-xpress-drives-efficiency-big-data/
[63] Passingham, Michael, "DHL cuts shipping costs with big data analytics", *V3.co.uk*, 23 Oct 2013. http://www.v3.co.uk/v3-uk/news/2302384/dhl-cuts-shipping-costs-with-big-data-analytics
[64] Envirnomental Leaders, "Big Data Improves Shipping Supply Chain Sustainability, Visibility", 30 October 2013. http://www.environmentalleader.com/2013/10/30/big-data-improves-shipping-supply-chain-sustainability-visibility/
[65] Rometty, Ginni, "A New Era of Value", Recorded at Retail's BIG Show, 13 January 2014. http://www.onlineevent.com/NRF/annual13/A-New-Era-of-Value-A-Conversation-with-Ginni-Rometty/

The following is yet another example where big data already results from the operations of large retailers, and how big data computing can deliver operational efficiency by moving operations into more real-time and making business decisions more agile[66]:

Opportunity:

- Business lacking the ability to react to market conditions and new product launches

Data & Analytics:

- 8.9B sales line items, 1.4B SKUs, 1.8B rows of inventory, 3200 stores
- Entire solution moved from mainframe to Hadoop
- Calculating price elasticity over 12.6B parameters

Results:

- Price elasticity now measured weekly against all data instead of quarterly against a subset
- $600K annual savings; 6000 lines of batch code reduced to 400 lines of PIG

# 5 THE EMERGING BIG DATA TRENDS, TECHNOLOGIES, AND NEEDS

Big Data computing has proven its capability to cope with the volume and velocity aspects of Big Data as also evidenced by the sectorial applications referenced in the previous Section. Variety, however, seems to be the ongoing challenge. In IT and data analysis the most challenging task has always been data integration. Now with Big Data this challenge has only gotten bigger. Semantic technologies promise to bring order that can be automated and as such scaled to massive amounts of dispersed data sets originating from various sources.

This Section touches upon some of the emerging needs and technology uses such as Linked Data, Analytics Engine, or In-field Analytics in the following Subsections. These are only indicative of the current and future development of Big Data technologies.

## 5.1 INCREASING IMPORTANCE OF DATA INTERPRETABILITY

Volume and velocity dimensions of data encourage thinking in terms of scalable storage and fast in-stream computing. When handled cost-efficiently, the solution allows coping with massive amounts of fast incoming data. Most value lies in integrating data from various sources, cross-combining the different perspectives the data from various sources has captured on the same event. For example, precision medicine providers are gaining a more refined understanding of what works for whom by integrating molecular data with clinical, behavioral, electronic health records, and environmental data[67].

Variety has various dimensions itself: variety of representation and formats, variety in trustworthiness of the source, variety in terms of the underlying data models and concepts, temporal and spatial dependencies, etc. As such a recent survey found that the unresolved

[66] Laney, Doug, “Information Economics, Big Data and the Art of the Possible with Analytics”, 2012, p. 40. https://www-950.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/Gartner_Doug-%20Analytics/$file/Gartner_Doug-%20Analytics.pdf

[67] Miliard, Mike, “Data variety bigger hurdle than volume”, *Healthcare IT News*, 3 July 2014. http://www.healthcareitnews.com/news/data-variety-bigger-hurdle-volume

challenges the variety of data brings with it, forces data scientists to leave most of the data on the table[68].

Data integration, hence, plays an important role. With small data volumes and slow or even batch throughput of data, data integration could be handled with manual data curation or data format conversions. Integrating Big Data at each user's premises, however, is neither scalable nor efficient. Rather the Linked Data[69] concept promises to solve the integration problem by making data interpretable and linkable at the source of its creation. Users of Linked Data can more efficiently integrate Internet-scale data, because data becomes self-descriptive. Due to a general agreement on RDF[70] and OWL as basic data representation language for Linked Data, many syntactic challenges are resolved.

Linked Open Data (LOD) is the ongoing, but relatively young, effort to provide such self-descriptive data on the Web. However, recent statistics show that the underlying principles are still hard to realize: Over 60 % of all LOD sources use a proprietary vocabulary, which means it is again the data consumer's responsibility to normalize the vocabularies[71].This is typically referred to as semantic heterogeneity – stemming from the variety of data description originators: When multiple parties are involved in describing a body of data from the same domain, there will always be different versions of the same[72]. Current research projects such as Optique[73] concentrate on reducing these semantic challenges, to give end users scalable semantic access to Big Data. The challenge of variety in Big Data is also the driving business and research question behind the project. As stated on the projects web site: in traditional businesses such as Oil & Gas, 30–70% of engineers' time is spent looking for and assessing the quality of data.

## 5.2 INCREASING IMPORTANCE OF LEGAL AND SECURITY ASPECTS

Data sovereignty started as a Cloud-specific discussion[74]: Through the global setup of Cloud providers, and their inherent backup strategy to store redundant copies of data in different data centers across the globe, the questions arouse as to which legal frameworks applied to the data, the nation's in which it originated, or the nation in which the data center was located. What about copies of data residing in different data centers in different countries? Many countries' legislation now requires that customer data needs to be kept within the same country as the customer resides[75].

With Big Data, the focus enlarged – and the term is being referred to as "personal data sovereignty:" The ability of an individual to have control over all of their personal data. Through the waves of Big Data, the user-generated content data through interactive web sites, social networks, and always-on consumer electronics such as smart phones has been playing

[68] Paradigm 4, "Leaving Data on the Table: New Survey Shows Variety, Not Volume, is the Bigger Challenge of Analyzing Big Data", *Survey*, 1 July 2014. http://www.paradigm4.com/wp-content/uploads/2014/06/P4PR07012014.pdf
[69] For further information see http://linkeddata.org/
[70] For further information see http://www.w3.org/RDF/
[71] Christophides, V., "Web Data Management: A Short Introduction to Data Science", *Lecture Notes*, Spring 2013, p. 15, http://www.csd.uoc.gr/~hy561/Lectures13/CS561Intro13.pdf
[72] Halevy, Alon Y, "Why Your Data Won't Mix: Semantic Heterogeneity", no date. https://homes.cs.washington.edu/~alon/files/acmq.pdf
[73] For further information see http://www.optique-project.eu
[74] Vaile, David, Kevin Kalinich, Patrick Fair and Adrian Lawrence, "Data Sovereignty and the Cloud: A Board and Executive Officer's Guide - Technical, legal and risk governance issues around data hosting and jurisdiction", *Cyber Lay and Policy Community*, Version 1.0, 2 July 2013. http://www.cyberlawcentre.org/data_sovereignty/
[75] WhatIs.com, "Data Sovereignty",no date. http://whatis.techtarget.com/definition/data-sovereignty

an important role. One of the core principle of the Big Data business application "Customer Experience" is the understanding if the customer, through collecting and acquiring as many behavioral or usage data as possible from these various sources. Personal data sovereignty seems to be the only feasibly solution to privacy and confidentiality protection. However, without technological enhancements keeping track of all personal and company data that has been shared over time and different service providers is an unattainable task[76]. On the other hand, the current situation of no knowledge at all leaves individuals and organizations very vulnerable. Again Linked Data could provide a technical solution by making all the usage links back-traceable to the original data item. Nonetheless, sophisticated algorithms and applications would be required to check whether personal or company data is used according to preferences or company restrictions.

In any case, not only individuals but also companies are at risk of conveying too much information or not being able to retrieve information that has been shared before. Energy service providers offering energy efficiency programs collect detailed energy consumption data, which when combined with other data sources and analyzed can reveal a lot if not all about a company's operations.

The current 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 or Linked Data (see 5.1), have the potential to trigger a paradigm shift towards an open and sharing economy also within the business-to-business industries, so-called Data Ecosystems. Such a paradigm shift would break down traditional organizational and geopolitical barriers more than ever before. In response, many countries will have regulated new compliance requirements by amending current laws or enacting new legislation that takes into account the changing realities through digitization.

Beyond regulatory innovation, technology innovation is still required. The ability to do big data analytics over encrypted data, for example, would solve many of the issues. However, so-called homomorphic encryptions have not yet been implemented successfully[77]. Current security development moves into the same direction as Linked Data, namely, to readapt the existing concepts at the most atomic level: data[78].

[76] Obar, Jonathan A., "Phantom Data Sovereigns: Walter Lippmann, Big Data and the Fallacy of Personal Data Sovereignty", 25 March 2013. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2239188

[77] Yergulalp, Serdar, "IBM's homomorphic encryption could revolutionize security", *InfoWorld*, 2 January 2014. http://www.infoworld.com/t/encryption/ibms-homomorphic-encryption-could-revolutionize-security-233323

[78] Escaravage Jason, and Peter Guerra, "Enabling Cloud Analytics with Data Level Security – Tapping the full Potential of Big Data and Cloud", 2013.
http://www.boozallen.com/media/file/Enabling_Cloud_Analytics_with_Data-Level_Security.pdf

## 5.3 Evolution from Query Engine to Analytics Engine

If there is one consent in the entire Big Data discussion, then this: It's not about the numbers, but the value. As the Big Data technology evolution also shows (see 3.1) we are currently moving beyond data management towards integrated data analytics. In order to be cost-efficient and scalable in the face of Big Data, analytics algorithms and data access need to be better aligned. We are moving from "in-database analysis" to "in-dataspace analysis." The data may reside anywhere – for the sake of cost-effectivity, the users do not want to replicate data at their premises just to do analytics, the users want to extract insights and use this insight to generate business or research value. However, it is not that trivial to "move algorithms to data."

Especially in industrial settings there is a lot of heavy lifting involved, e.g. simulations and forecasts, optimization etc. These techniques 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. At the same time, as a recent survey found[79]: although 49 percent of the respondent data scientist could not fit their data into relational databases anymore, only 48 percent have had used Hadoop or Spark – and of those 76 percent said they could not work effectively due to platform issues.

The serving layer of the lambda architecture (3.4.2) will need to provide a flexible and extensible abstraction, which we call more specifically the *Analytics Engine* (Figure 5). The analytics engine also needs to serve other analytical tasks such as manual data discovery, or automated pattern matching and as such satisfy the different timeliness requirements of the different analytics classes and techniques (see 4.2). Additionally, it must be noted that the data acquisition, which is only implicit in the lambda architecture, is an essential part in industrial settings and cyber-physical systems as will be discussed in the next Subsection. The Analytics Engine must be able to facilitate the deployment and execution of analytics algorithms at or near the source(s) of the data and hide the complexities of distributed data analytics from the user.

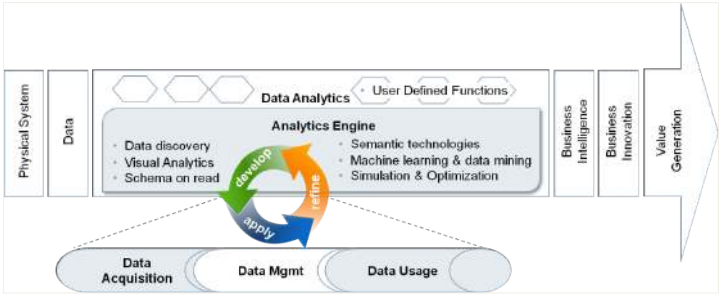


**Figure 5: Analytics Engine allows for the development, application, and refinement of knowledge and discovered knowledge in a semi-automated manner**

The Analytics Engine, as depicted in Figure 5, 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.

[79] Paradigm 4, "Leaving Data on the Table: New Survey Shows Variety, Not Volume, is the Bigger Challenge of Analyzing Big Data", *Survey*, 1 July 2014. http://www.paradigm4.com/wp-content/uploads/2014/06/P4PR07012014.pdf

## 5.4 EVOLUTION OF SMART DATA THROUGH IN-FIELD ANALYTICS

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 in Subsection 4.2, 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 can be in the sub-second range, meaning that the speed layer of the lambda architecture is not sufficiently fast: 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 level. 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. 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, in-field data, which is available in cyber-physical systems, has no presentation in online data businesses yet and hence in the original lambda architecture.

Prescriptive analytics must be applied on fresh and accurate data as close to the event as possible, both in terms of time and space, to be able to accurately identify and classify these events. It may still not be hard real-time, but within hundreds of milliseconds from the event. For achieving such timeliness, the in-field analytics is shown in Figure 6 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 in the field synchronizes patterns from the knowledge discovery of business analytics and monitoring in the backend.** 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.
- **Learning algorithms will predict the future communications requirements, so as to move data to the proper domains before the data is required**[80]. A recent

[80] Helzer, B., M. Clement, Q. Snell, “Latency tolerant algorithms for WAN based workstation clusters”, In *Proceedings of the Seventh Symposium on the Frontiers of Massively Parallel Computation*, 21-25 Feb 1999, pp. 52 – 59, doi: 10.1109/FMPC.1999.750584.

study found that more than 36 percent of data scientists say it takes too long to arrive at insights because the data is too big to move to their analytics software[81]. Additional research is also taking place in the optimization of Wide Area Network links to solve the remaining performance issues originating from old communication protocols[82].

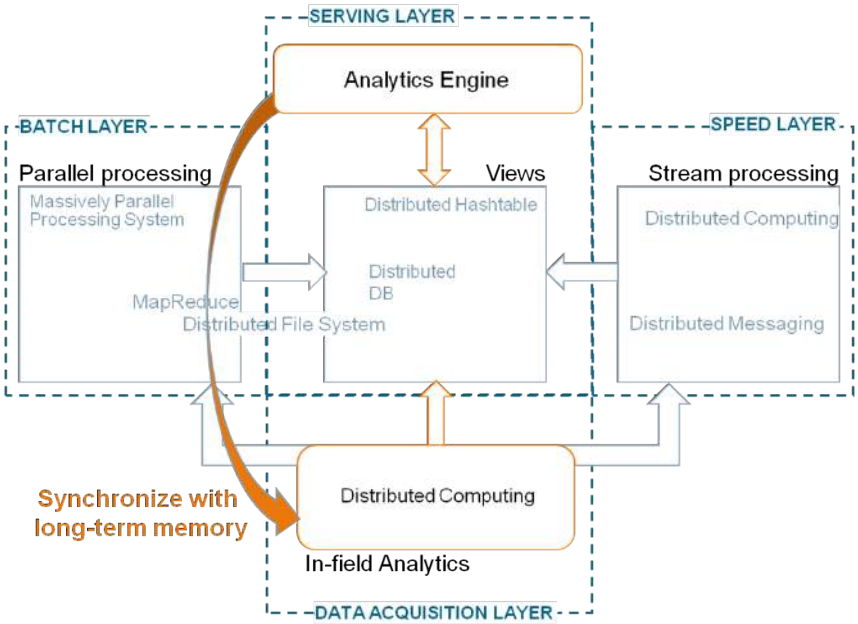


**Figure 6 Cyber-physical systems include computing resources in the data acquisition layer - In-field analytics can tap into this potential and deliver insights fast enough for actual prescriptive analytics**

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: from data acquisition, to management, to usage. For in-field analytics, as an example, the synchronization of discovered knowledge in the backend with the field level is essential and requires such an Analytics Engine as discussed in Section 5.3.

# 6 CONCLUSION

The information explosion has always been an issue of wide interest. With the beginning of the millennium, we reached a mark where for the first time the size of digitally stored information surpassed non-digital for the first time. Since then every three years, there has been a major advancement in handling the volume and velocity of data. However, variety of data at big scales, i.e. also coming from a large variety of sources, at very differing formats and quality still represents a challenge today; especially for businesses that have strict requirements towards veracity of data, such as in industrial automation. So, technology-wise the next frontier is how to cross-analyze various data in a cost-effective manner. Some research trends into semantic technologies enabling cross-domains analytics efficiently, whilst preserving privacy and confidentiality are already emerging.

Regarding Big Data applications, some businesses still have not reached the tipping point of digitization. Although there will be potentially massive amounts of data coming at high speed from a variety of sources – once smart devices are rolled out – it is hard to find actual

[81] Paradigm 4, op. cit., 2014.
[82] Khazan, Roger I., "A One-Round Algorithm for Virtually Synchronous Group Communication in Wide Area Networks", 22 May 2002. http://groups.csail.mit.edu/tds/papers/Khazan/khazan-phd.pdf

Big Data in all three dimensions of volume, velocity, and variety. Almost all references to Big Data applications in traditional industries researched for this deliverable concentrate either on the scalable management of Big Data or advanced analytics (however, not on Big Data in all three dimensions). Hence, application-wise the next frontier is how to apply the newest technological capability of real-time analytics, which certainly will require some time for maturing and adaptation in the particular business domains.