Big Data Analytics for Business Intelligence in Vertical Industries

Seminar Big Data WS 2015/16

Anton Okolnychyi, Andreas Koslowski

RWTH Aachen University, 52056 Aachen, Germany
{anton.okolnychyi, andreas.koslowski}@rwth-aachen.de

Abstract. In recent years, Business Analytics has taken advantage of management, computer and statistical science in order to introduce new revolutionary approaches of conducting business in various industries. Big Data becomes an integral part of huge number of analytic processes because amount of generated data increases in volume and conveys more and more crucial information to control decision-making, to perform analysis and guide the company. This paper should give an introduction to Big Data, how and why it is adopted in vertical industries, Business Intelligence and what kind of Big Data analytic tools are used in the different industries. It will describe and compare different strategies and technologies that are applied in vertical industries while adopting to Big Data.

1 Introduction

The business world is going through a revolution induced by the use of data to control decision-making and to perform analytics. A major reason for the business analytics revolution is the rapid proliferation of the amount of data available to be analyzed [1]. Tasks in modern analytics require a huge computing power, storage capacity, appropriate information technologies which are needed to gather, analyze and retrieve an asset from data. This becomes possible due to the constant evolution of the corresponding software and hardware techniques. The availability of computing power and data storage capacity have expanded at an exponential pace, and this trend seems unlikely to abate any time soon [2].

Big companies understand the value of their data and possible outcomes of usage of data analysis in adjusting and defining their business strategy. Some companies try to make use of data which contains opinions expressed by customers in order to improve the overall customer experience. Others analyze data to find patterns in the customer's behavior which can be used later to predict future needs and purchases. Data is a strategic asset in making recommendations. Often, data is analyzed in order to fine-tune the enterprise itself, with analytical insights used to refine internal processes, promote safety, and pinpoint operational issues the resolution of which can drive up efficiency, profitability, and competitive positioning [3]. This is especially true in the vertical industries,

where it is critical to adapt to the needs of the customers and predict the trends of each specific market. By analyzing the data the industries gather, it is much easier to adapt to current trends and the firms are able to offer the customers complete business solutions.

This paper provides an overview of how Big Data is used in the industry and clarifies how Big Data and its use will change the Business Intelligence, particularly in the vertical markets. It will compare how the different vertical industries are dealing with the topic of Big Data Analytics and what strategies are adopted while doing so.

2 History and Area of Application of Big Data

Big Data has a long history and no defined beginning. For ages people studied data to gain more knowledge and the amount of available data was continuously growing. In 1944 Fremont Rider published a book named "The Scholar and the Future of the Research Library" in which he estimated that the American university libraries were doubling in size every sixteen years [4]. In November 1967 B. A. Marron and P. A. D. Maine published the article "Automated Data Compression" in the "Communications of the ACM" [5] describing a fully automatic data compression to help keeping storage requirements for all information to a minimum. A key factor for the growth of information is the dramatically reduced cost of storage. In 1981 digital storage had a cost of $700/MB. This reduced to only $0.0002/MB in 2006 [6]. A major milestone was reached in 1996 where digital storage became more cost-effective than paper [7]. Even though there was a lot of information the term "Big Data" was used for one of the first times in 1999 in the article "Visually exploring gigabyte data sets in real time" by Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes [4]. One of the most important inventions for Big Data was the internet, which opened up many possibilities for new data to be generated. Social media, clickstreams, online transactions, and data streams from sensors just to name a few. With this constant stream of data, companies were able to analyze data of larger volume and even more value than before to improve their strategies. Nowadays the term "Big Data" is used to describe the process of capturing, processing, analyzing and visualizing of potentially large datasets in a reasonable timeframe not accessible to standard IT technologies [8].

Big Data is used in many different industries for a variety of applications. The main interest in Big Data is to help a firm being as cost efficient as possible trough Cost Reduction, Time Reduction and satisfying the customers. UPS for example, who started using Big Data in the 1980s, is now tracking 16.3 million packages per day and storing over 16 petabyte of data. Through sensors in over 46000 vehicles UPS is able to track speed, direction and breaking of those vehicles. They are using this information to plan routes and reduce time and fuel spent. With this system UPS saved 8.4 million gallons of fuel in 2011 [9]. But Big Data is not only used to help a company to spent less or make more money. It can also be used to analyze situations and solve problems that cannot be engaged using

standard methods. At the investigation in the "UN Oil-for-Food Program fraud" forensic analysis was applied on Big Data to analyze thousands of documents about communications, events and transactions [10].

3 Big Data Concepts and Architecture

The main attributes of Big Data are wildly known as the 4 Vs, namely volume, velocity, variety and veracity. Those attributes best describe what Big Data is and what it can be.

Volume The volume of data is the quantity and therefore the most visible and the primary attribute of Big Data. It is the fact that in order to work with Big Data a large amount of data is needed. In 2012 there were about 2.5 exabytes of data created each day. This amount is expected to double every 40 months [11] and it is estimated that by 2020 there will be 35 zettabytes, which are 10^{21} bytes, of digital data [12].

Variety The second attribute of Big Data is the existing variety of data. Data can be sorted into three categories. Structured, semi-structured and unstructured. For many years many large companies have been focused on structured data which is standardized in a data model. By having a model that describes exactly how this data will be stored, processed and accessed, the data can be easily entered, stored and analyzed. An example of structured data is a Key/Value system where each specific key has a specific value. Such data can be obtained by sensors that generate a value to a corresponding timestamp. In the past it was only possible to analyze structured data due to high cost and limitations of storage and processing power. But with new forms of databases it is possible to store and analyze unstructured and semi-structured data. Unstructured data can be data in forms of images, presentations, audio files like voice recordings or text files with natural language and is difficult to organize. Semi-structured data is a composition of both structured and unstructured data where tags or markers can be used to identify certain elements within the data. An example of a semi-structured data are emails which consist of a sender, recipient, date, time, and other fixed elements along with an unstructured text.By using tags and markers it is possible to organize semi-structured data [13].

Velocity Another important aspect is the speed with which the data is created or delivered. For many applications the speed is even more important than the volume of the specific data itself. For example the sensors in a car have to produce a constant stream of data at a high rate to keep the main processor up to date in case something unusual like a crash happens. Other uses of datastreams are the so called clickstreams that many firms collect from websites to make purchase recommendations and to display specified ads to the users [14]

Veracity Veracity is the varying quality relevance or value of data [12]. It is highly subjective and is more of a factor when one has limiting conditions like space or processing power. If one is dealing with those conditions it can be important to prioritize some data over other to get more value in return.

4 Big Data Analytics

According to T. Davenport and J. Harris, the Big Data analytics is the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions [15]. Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include different types such as structured/unstructured and streaming/batch, and different sizes from terabytes to zettabytes [16]. The main aim of big data analytics is to assist organizations in making smart business decisions by enabling data scientists and other analytics professionals to analyze large volumes of data that was previously inaccessible, unusable or just untapped. With a help of advanced methods and tools, such as text analytics, machine learning, data mining, statistics, and natural language processing, companies are able to perform analysis of previously untapped data sources to retrieve new insights which lead to better and more informed decisions. Just having a lot of data does not give anything. Only using the correct analytics one can extrapolate powerful insights. Amir Gandomi, Murtaza Haider say in [17], that Big data is worthless in a vacuum. Its potential value is unlocked only when leveraged to drive decision making. To enable such evidence-based decision making, organizations need efficient processes to turn high volumes of fast-moving and diverse data into meaningful insights. Charter Global in [18] defines 3 types of analytics: Descriptive, Predictive and Prescriptive.

Descriptive Analytics is the simplest kind of analytics. It allows you to retrieve smaller and more useful bits of information from huge data sets. Alternatively, it may provide a summary of what happened. Social media analytics is one of examples of Descriptive Analytics. It refers to the analysis of structured and unstructured data from social media channels. User-generated content (e.g., sentiments, images, videos, and bookmarks) and the relationships and interactions between the network entities (e.g., people, organizations, and products) are the two sources of information in social media [17].

Predictive analytics is the next step up in data reduction. It utilizes a variety of statistical, modeling, data mining and machine learning techniques to study recent and historical data, thereby allowing analysts to make predictions about the future[18]. Predictive Analytics is the use of historical data to forecast on consumer behaviour and trends [19]. This type of analysis makes use of the statistical models and machine learning algorithms to identify patterns and learn from historical data [20]. At its core, predictive analytics seek to uncover patterns and capture relationships in data. Predictive analytics techniques are subdivided into two groups. Some techniques, such as moving averages, attempt to discover the historical patterns in the outcome variable(s) and extrapolate them to the

future. Others, such as linear regression, aim to capture the interdependencies between outcome variable(s) and explanatory variables, and exploit them to make predictions [17].

Prescriptive Analytics does not predict only one future, it rather predicts "multiple futures" based on the decision-makers potential actions. Each decision is associated with a likely outcome. This can be used to choose the better action. T. H. Davenport and J. Dyche in [9] say, that there is an increased emphasis on prescriptive analytics. It provides a high-level of operation benefits for organizations and it places a premium on high-quality planning and execution.

5 Business Intelligence

The term "Business Intelligence" was used in scientific context for the first time in an article by H. P. Luhn, an IBM researcher, in 1958. In his article [21], Luhn described an automatic method to provide current awareness services to scientists and engineers who required assistance to handle an increase of scientific and technical literature. However, this term became widely used only in the 1990s due to H. Dresner. In 1989, he (later a Gartner Group analyst) proposed "business intelligence" as an umbrella term to describe "concepts and methods to improve business decision making by using fact-based support systems" [22].

Business Intelligence is one emergent area of the Decision Support Systems (DSS) discipline and can be defined as the a process that transforms data into information and then into knowledge [23]. Being rooted in the DSS discipline, BI has suffered a considerable evolution over the last years and is, nowadays, an area of DSS that attracts a great deal of interest from both the industry and researchers [24].

It is important to distinguish Business Intelligence from Business Analytics since there is a confusion in the literature. Business Intelligence is needed to run the business while Business Analytics are needed to change the business [25]. BI is focused on creating operational efficiency through access to real time data enabling individuals to most effectively perform their job functions. BI also includes analysis of historical data from multiple sources enabling informed decision making as well as problem identification and resolution. Business Analytics relates to the exploration of historical data from many source systems through statistical analysis, quantitative analysis, data mining, predictive modelling and other technologies and techniques to identify trends and understand the information that can drive business change and support sustained successful business practices. M. van Rijmenam says that Business Intelligence is looking in the rearview mirror and using historical data from one minute ago to many years ago. Business Analytics is looking in front of you to see what is going to happen. This will help you anticipate in whats coming, while BI will tell you what happened. This is a very important distinction as both will provide you with different, not less, insights. BI is important to improve your decision-making based on past results, while business analytics will help you move forward and understand what might be going to happen [25].

The concept of BI, although not new, is experiencing a renaissance which positions the concept in its infancy phase again [24]. Therefore a lot of new and ambiguous definitions have emerged. After some time, BI became a multi-dimensional concept. It can be described as a process, as a product, as a set of technologies or even as a combination of these. Shariat and Hightower in their article [26] characterize BI as a composition of processes, technologies and products: processes for collecting and analyzing business information; technologies used in those processes; and the product as "the information (knowledge) obtained from these processes". Most BI definitions in the literature may be divided into 2 groups: the first group defines BI as a process, while the second one - as a set of technologies. For instance, J. Dekkers and J. Versendaal and R. Batenburg in [27] define BI as the continuous activity of gathering, processing and analyzing data - supported by a BI system. On the other hand, S. Pemmaraju in [28] says that BI encompasses all of the software applications and technologies that a company uses to gather, provide access to, and analyze data and information about its operations.

Business Intelligence systems represent the natural evolution of decision support systems and put a strong emphasis on data-driven decision making, based on the integration of multiple data resources that reflect different aspects of organizational activity [29].

Business Intelligence tools aim at improving the quality and accuracy of information used in decision making processes by simplifying the storage, identification, and analysis of information [30]. BI systems let users at all organizational levels access data, interact with it, and analyze it toward improving business performance, discovering new opportunities, and increasing efficiency. Well-designed BI systems offer a global view of the entire organization, permit analysis of business activities from multiple perspectives, and enable rapid reactions to changes in the business environment.

6 Big Data in Vertical Industries

6.1 Objectives for Big Data

Like many new information technologies, Big Data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings. Like traditional analytics, it can also support internal business decisions. The technologies and concepts behind Big Data allow organizations to achieve a variety of objectives, but most organizations are focused only on several [9].

1. Cost Reduction from Big Data Technologies

   A lot of companies stick to the opinion that the cheapest way to store structured data is by means of such Big Data techniques like Hadoop clusters. Technical report [31] made by NewVantage Partnes says that it is needed $37,000 for traditional database, $5,000 for a database appliance, and only

$2,000 for Hadoop cluster. Organizations that were focused on cost reduction made the decision to adopt Big Data tools primarily within the IT organizations on largely technical and economical criteria [9]. Cost reduction can be the main goal as well as a secondary objective after others have been achieved. The organization's first goal might be to innovate new products and services from big data. After accomplishing that objective, it may want also to investigate how to do it with minimum expenses.

2. Time Reduction from Big Data

   The second common objective of Big Data technologies and solutions is time reduction. Macy's merchandise pricing optimization is a great example of time reduction by means of Big Data. This company managed to reduce the time to optimize pricing of its 73 million items for sale from over 27 hours to just over 1 hour. This possibility obviously allows Macy to re-price items much more frequently to adapt to changing conditions in the retail marketplace[9]. Another big advantage of time reduction is ability to interact with customers in real time, using analytics and data derived from the customer experience.

3. Developing New Big Data-Based Offerings

   One of the most ambitious things an organization can do with Big Data is to employ it in developing new product and service offering based on data. Many of the companies that employ this approach are online firms, which have an obvious need to employ data-based products and services [9]. Facebook can be one of the best examples. It was able to implement a huge number of product offerings and features, including People You May Know, Groups You May Like and several others. It is worth to mention the intelligent way of dealing with adverts at Facebook. It offers an advertising service (or sponsored posts) based on very specific criteria: interests or features of user profiles.
   T. H. Davenport in [9] emphasizes the importance of Google in developing products and services based on Big Data. This company, of course, uses Big Data to refine its core search and ad-serving algorithms. There are such products as Gmail, Google Plus Google Apps which are making use of Big Data techniques.
   J. Lucker states in [32] that customers are increasingly frustrated by the generic offers they are bombarded with from marketers, and most marketers continue their "spray and pray" mass marketing techniques of the past, seeing little reason to change. However, customers nowadays would like to have only relevant and personalized interactions, which are based on a customer's situation and preferences. According to J. Lucker, companies that fail to provide relevant offers will be left behind. He describes the concept of highly customized "next best offers" (NBOs) that reflect each customer's preferences, purchase history, and purchasing context in [32]. Organizations can create

NBOs be means of using data and sophisticated analytics in order to acquire customer loyalty. When done right, NBOs can increase cross sell/upsell opportunities, boost profits, and create competitive advantage.
Despite the revolution caused by the use of data, there are a lot of companies that treat customers in a generic way. J. Lucker believes that tailoring the communications and offers based on consumers' specific behavior and preferences is a significant opportunity to lead in a particularly industry for those organizations that know consumers better. Usage of sophisticated analytics to craft highly customized, relevant offers can allow companies to achieve a greater loyalty from consumers, greater profitability, and competitive advantage [32]. Most commonly, NBOs are intended to drive loyalty, contribute to making a purchase, or both. An NBO can be aimed to provide offers on products, services, information.
A lot of companies still do not have or cannot afford the infrastructure and applications which are required to integrate NBOs into the overall customer experience strategy. Organizations have to have complex and comprehensive modern Big Data and analytics infrastructure, where analysts can aggregate, organize and derive valuable information from internal and external data of different kinds. Lack of this makes accurate understanding and predictions of customer behavior impossible.

4. Supporting Internal Business Decisions

   When it comes to making business decisions, it is difficult to exaggerate the value of managers experience and intuition, especially when hard data is not at hand. Today, however, when petabytes of information are freely available, it would be foolhardy to make a decision without attempting to draw some meaningful inferences from the data [33]. In the same report, P. Nannetti declares that nine in ten of the company executives polled feel that the decisions theyve made in the past three years would have been better if theyd had all the relevant data to hand.
   For those uses of big data that do involve internal decisions, new management approaches are still necessary, but not yet fully resolved in practice. This is because big data just keeps on flowing. In traditional decision support situations, an analyst takes a pool of data, sets it aside for analysis, comes up with a model, and advises the decision maker on the results. However, with big data, the data resembles not so much a pool as an ongoing, fast-flowing stream. Therefore, a more continuous approach to sampling, analyzing and acting on data is necessary [34]. The most advanced firms are monitoring not only their own suppliers, but their suppliers' suppliers [9].

6.2 Relevance of Big Data

V. Gopalkrishnan, D. Steier, H. Lewis and J. Guszcza say in [1] that most companies which are looking at potential applications of Big Data and decide whether to adopt to Big Data solutions ask themselves three main questions:

1. What is the business goal, or organizational goal?

   The goal defines a context in making decisions on whether to apply Big Data techniques or not. The volume and variety of data should justify the application of automated analytic techniques and make them reasonable. There is an example of an insurance company in [1] which neglected usage of Big Data for certain scenarios due to the context of a task. Company decided to stick to the decade-old approach since information derived from Big Data application was not so crucial and could not give a big benefit.
2. Given the goal, is the available data suitable?
   The data suitability question should be asked in all circumstances. For instance, companies can make use only of the data which is representative of the people whose behavior company is trying to predict. A famous example from US political history applies here: in the 1930's, a poll predicted that Franklin Roosevelt would lose by a landslide to his Republican opponent. But he didn't. The poll data was biased because it was collected over the telephone, which was a luxury item disproportionately used by Republicans during the great depression.

3. What is the return of investment on Big Data?

   The Netflix Prize attracted worldwide interest and thousands of entries, but Netflix decided not to implement the winning algorithm [35] due to a number of practical reasons. Company should be sure that acquired benefits worth the complexity of the required implementation. Organizations should take into account all considerations, including costs of collecting or licensing data of sufficient quality, the costs of transferring, cleaning, integrating data across multiple sources, effort and skill sets needed to develop and maintain models, timing of data and result availability, cost of required infrastructure and marginal business value of the results [1].

Companies can decide whether it is relevant or irrelevant to apply Big Data in their use-cases based on aforementioned questions .

6.3 Overview of Big Data Tools

The four main goals defined before encourage companies to apply Big Data techniques. There is a huge variety of innovative data tools to support engineers in achieving those goals which simplify the process of building the correct Big Data architecture. P. Warden in [36] explains the emergence of new Big Data tools by the fact that techniques originally developed by website developers coping with scaling issues are increasingly being applied to other domains. In the same book, he states that research techniques from computer science can be effective at solving problems and creating value in many real-world situations. This draws the attention from commercial organizations and facilitates the investments made in research area as well as in developing new tools. It worth to mention relatively

cheap hardware which is available nowadays. More companies can afford to do large-scale data processing. However, it is impossible to distinguish tools based on objectives for Big Data since most of tools can be used in achieving different goals. Each company creates its own Big Data architecture which significantly depends on the context of the task and many other reasons. All of that means that each tool may be involved into processes which are aimed to get completely different benefits for company.

However, it is possible to outline common areas in which all companies are solving main problems in order to build the correct Big Data architecture. Those areas are storage, servers, processing, machine learning and visualization. There are many tools which can be used in each area and they will be briefly described in this section.

1. NoSQL

   Developers want to have a lot of flexibility and control over the database. In some cases, they do not want to spend too much time designing a schema for a database since not all requirements might be known and the overall picture can be still unclear. That's why they would like to have an easy way to create new and experimental systems to try new solutions.
   This widespread demand for solutions, and the comparative ease of developing new systems, has led to a flowering of new databases [36]. The main common property is that most of them do not support the traditional SQL interface. This fact contributed to the name which we hear a lot nowadays - NoSQL. All NoSQL tools are designed to trade the reliability and ease-of-use of traditional databases for the flexibility and performance required by new problems developers are encountering.

   MongoDB

   MongoDB is a database aimed at developers with fairly large data sets, but who want something thats low maintenance and easy to work with. Its a document-oriented system, with records that look similar to JSON objects with the ability to store and query on nested attributes [36]. It supports sharding and is well suited for MapReduce operations. You can query it by means of Javascript queries.

   CouchDB

   CouchDB is also a document-oriented database. It has a lot in common with MongoDB. For example, it provides a JavaScript interface. However, there are some crucial differences. CouchDB supports querying, scaling, and versioning. P Warden claims in [36] that CoucheDB uses a multiversion concurrency control approach, which helps with problems that require access to the state of data at various times, but it does involve more work on the client side to handle clashes on writes, and periodic garbage collection cycles

have to be run to remove old data. It doesnt have a good built-in method for horizontal scalability, but there are various external solutions like BigCouch.

Cassandra

Cassandra is a distributed key/value system. It was initially introduced by Facebook as an internal product, but then it was open sourced. It is quite complex and requires additional learning in order to apply it effectively. On the other hand, it provides a lot of power and flexibility. It is very similar to the Google's BigTable.

Redis

Two features make Redis stand out: it keeps the entire database in RAM, and its values can be complex data structures. Though the entire dataset is kept in memory, its also backed up on disk periodically, so you can use it as a persistent database. The processing speed slows down if data expands beyond available memory and the operating system starts paging virtual memory to handle accesses.

2. MapReduce

MapReduce is a highly-popular paradigm for high-performance computing over large data sets in large-scale platforms [37]. The main goal of this technology, pioneered by Google, is to enable distributing work. There is a whole set of tools to actually enable that.

Hadoop

Originally developed by Yahoo! as a clone of Googles MapReduce infrastructure, but subsequently open sourced, Hadoop takes care of running your code across a cluster of machines. Its responsibilities include chunking up the input data, sending it to each machine, running your code on each chunk, checking that the code ran, passing any results either on to further processing stages or to the final output location [36].

Hive

With Hive, you can program Hadoop jobs using SQL. Its a great interface for anyone coming from the relational database world, though the details of the underlying implementation arent completely hidden.

3. Storage

Big Data processing operations deals with data in a way that traditional file systems are not designed for. Data is usually written and read in large

batches at once. Efficiency is a higher priority than features like directories that help organize information in a user-friendly way. The data also must be persisted within multiple machines in a distributed way. Therefore, new specialized technologies appeared.

S3

Amazons S3 service lets you store large chunks of data on an online service, with an interface that makes it easy to retrieve the data over the standard web protocol, HTTP. It is missing some features like appending, rewriting or renaming files, and true directory trees. But it is a key/value database available as a web service and optimized for storing large amounts of data in each value.

Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is designed to support applications like MapReduce jobs that read and write large amounts of data in batches, rather than more randomly accessing lots of small files. Unlike S3, it does support renaming and moving files, along with true directories.

4. Visualization

Since amount of collected and analyzed data is huge, decision makers are making use of data visualization software that allows to see analytical results presented visually. Such tools help them to find patterns and even predict the future.

Gephi

Gephi is an open source Java application that creates network visualizations from raw edge and node graph data. Its very useful for understanding social network information; one of the projects founders was hired by LinkedIn, and Gephi is now used for LinkedIn visualizations [36].

Processing

Initially best known as a graphics programming language that was accessible to designers, Processing has become a popular general-purpose tool for creating interactive web visualizations.

Protovis

Protovis is a JavaScript framework packed full of ready-to-use visualization components like bar and line graphs, force-directed layouts of networks, and other common building blocks.

6.4 Big Data Use Cases in Vertical Industries

The goal of this section is to provide concrete use cases and scenarios in which Big Data is applied by different companies in various industries.

Healthcare As a result of evolution in the IT industry, new techniques were developed in healthcare domain which provide all stakeholders new sources of knowledge. Pharmaceutical-industry experts, payors, and providers are now beginning to analyze big data to obtain insights. Information provided by all origins is extremely complex, diverse and timeliness. Companies analyze the data in order to improve the overall quality of healthcare and reduce the amount of money they have to spend. Many innovative companies in the private sector - both established players and new entrants - are building applications and analytical tools that help patients, physicians, and other healthcare stakeholders identify value and opportunities [38].

P. Patil and R. Raul and R. Shroff and M. Maurya define in [39] the following sources of helathcare data:

1. Health-information exchanges (HIEs).

   Since a lot of healthcare companies, hospitals, doctors have adopted new technologies, they became an integral part of health-information exchange programs. Such exchange programs allow all stakeholders to access the broader range of information.

2. Activity and Cost Data.

   Medical insurance companies provide information about which services were provided and how they were reimbursed.

3. Pharmaceutical R&D Data.

   The pharmaceutical R&D departments describe the side effects and other harmful actions caused by specific drug treatments.

United Healthcare is an organization that pursuing Big Data in healthcare industry. This company was focused on structured data analysis for many years, but now it is primarily focused on unstructured data. The level of customer satisfaction is one of the most important thing for each healthcare company. Not only because service levels are being checked and verified by government, but also because customers are free to choose different health plans or change the company at all. T. H. Davenport describes in [9] one use case of Big Data at United Healthcare which is related to processing phone calls. United Healthcare converts the voice data into text, and then analyze it with a set of tools for "natural language processing". The company is capable of identifying which

customers are dissatisfied by means of complex analysis. Healthcare companies always had the task to figure out which customers are complaining, but the tools for solving this challenge are different nowadays. United Healthcare puts the data initially into a "data lake" using Hadoop and NoSQL storage, so the data does not have to be normalized. The natural language processing - primarily a "singular value decomposition", or modified word count - takes place on a database appliance.

Map Reduce is heavily used for large-scale medical image analysis. Some of the use cases are described in [39]. For instance, Map Reduce is involved into parameter optimization for lung texture classification using support vector machines (SVM). People are also making use of Map Reduce for indexing content-based medical images. Two approaches for content-based image indexing were compared and implemented in the MapReduce framework: component-based versus monolithic indexing. The former resulted in an unexpectedly long runtime due to requirement to write to a very large CSV(Comma-Separated Values) file of approximately 100 Gb for 100,000 images. The result showed that MapReduce was not performing well with intensive input-output tasks. On the other hand, the monolithic strategy showed to be well-suited for MapReduce, which allowed indexing 100,000 images in about one hour using 24 concurrent tasks.

There are also challenges and problems that each company is facing during implementation of Big Data in healthcare industry. P. Patil emphasizes in [39] the following ones:

1. Awareness.

   Companies must understand the complexity and necessity of data services.

2. Exploration of business application model.

   Most software companies are not familiar with the medical business. Therefore, they are unable to discover key business intelligence with a help of Big Data tools and techniques.

Logistics and Transportation There is also a huge number of new data sources that generate relevant data for logistics and transportation. For instance, companies can make use of smartphones, tablets, log files, road and vehicle sensors, GPS and many other devices. In addition, there are customer's applications and websites that participate in the process of generating relevant information. Organizations have to adopt Big Data solutions to refine the quality of services due to the amount of available data and its characteristics.

One of the companies in India is monitoring real-time vehicles to improve operational efficiency. The authors collect data from vehicles about fuel, speed, acceleration, GPS location coordinates using sensors and GPS devices with other data such date, time, driver's id and then send this information to clustered servers running Hadoop. All data is stored to a Hadoop Distributed File System(HDFS). An analysis is made then weekly or monthly over these terabytes

of data using Hadoop analytics system in order to improve the transportation company productivity and help reduce the costs. The analysis of data allows also monitoring driving behavior and answering to some questions like: which vehicles are wasting fuels? which drivers have the highest risks? [40]

Another project was created as a result of collaboration between IBM and Dublin City Council which took place in Ireland from 2010 to 2013. IBM helped the city of Dublin to improve its public bus transportation network and reduce the increasing traffic congestion problem [41]. There was an additional requirement not to make any major modifications for the city historic infrastructure. The system handled GPS, speed, stop, fare data from bus systems, traffic flow data, road whether conditions, road works and maintenance data, Dublin event data with a help of clustered servers running IBM Big Data analytics to build a real-time digital map of Dublin city transportation network. Advanced analytics on the collected data helped to identify traffic problems and answer questions such as the optimal time to start bus lanes, the best place to add more bus lanes, etc. The project helped the city to better monitor and manage the traffic in real time, which accelerate decision making and improved traffic flow and mobility in the city [40].

There is a Big Data analytics project that was started in 2012 by Opera Solutions company. The main goal of the project is to improve the quality of British Airways services for its customers. The objective is to understand clients' needs better than any competitive airline company [41]. The project collected different types of information about 20 million customers via websites, smart phones and tablets applications, blog sties rating, likes on social medias, conversations with call centers, etc. After that, a big data analytics system stores and analysis all the structured and unstructured data to identify customers' preferences, characteristics, problems, and to provide them high quality services [40].

The exponential growth in the production and storage of mobility-related data has been accompanied by rising concerns relating to the adequacy of regulations ensuring privacy. These concerns have been fuelled by the personally identifiable nature of much of the data being collected and the fact that it is often collected without the full knowledge and informed consent of the data object. Location-based data is particularly vulnerable to breaches in privacy. Despite concerns over privacy, location-based data enhances services available to individuals and may contribute to significant improvements in safety, traffic operations and transport planning [40].

Education The emerging techniques and tools in Big Data allow to derive meaningful information and assets from education data that can improve and refine the educational process, help students and teachers. There has been growing interest in the education community to take advantage of Big Data to improve learning performance of students, enhance working effectiveness of teachers and reducing administrative workload [42]. Learning is a way to discover new things that we do not know about to increase our knowledge and expand our understanding of the world, which is a lifelong activity [43].

A huge amount and variety of educational data contributes to diverse kinds of information have been increasingly gathered from e-learning platform. This data can be further augmented with information collected from other sources such as twitter, facebook, student surveys, blogs. Techniques from statistics, computer science and machine learning are heavily used in learning analytics, education data mining to benefit from Big Data in educational sector.

L. Cen and D. Ruta and J. Ng in [44] describe the following use cases of Big Data application in education:

1. Performance prediction.

   This use case is focused on predicting how well a certain learning task will be performed by a student or a group of students. This information might be used to assist those students who will most likely have problems. Each student has his own performance which depends on a lot of factors and education process is aimed to teach all students with different abilities. So, unique approaches for each special group of people can be crucial in achieving the overall success of a course. An excellent and precise prediction framework may be involved into creating an optimal curriculum, reducing the abnormal workload at a certain point in time. All of these actions can positively influence the educational experience. Moreover, universities can make use of these predictions in admission process in order to pick up the corresponding students.
   Accurate and reliable prediction can be achieved based on student profiles, historical performance, demographics data of current and former students by considering all important factors that could affect learning performance both before the learning task and during the course. Big data technologies increasingly allow collection, storage and processing of huge amounts of data spanning long periods of time and carrying a large variety of data types and samples. These aspects make them ideally suited to measure and store various characteristics of the students engaged in a learning task both individually and in a group [44].

2. Intelligent course recommendation.

   Students are usually offered a wide range of courses and modules which they can study. It is a hard task for students that have not decided about their future yet or only started their educational path. A good recommendation system should look at available learning resources, avoid conflicts in time allocation, satisfy school formal requirements, match the interest of a student, maximize learning efficiency and fulfill student's educational goals.
   L. Cen in [44] states that the input information to such recommender system should include students profiles, demographics information, historical student performance, the field of interest, students career preferences, and course related information like attendance records, student evaluation, class schedule, etc. A ranking score for each of the candidate courses would then

be delivered inline with the expected performance prediction obtained in response to the information retrieved from educational data.

3. Personalized learning.

   All people have different learning abilities, kinds of personality, models of thinking, background and knowledge in various areas. Therefore, their learning effectiveness and efficiency can be quite different even with exactly the same learning conditions and environment. Personalized learning is focused on adopting the general educational process to specific needs and properties of each different group of students to maximize their learning potential, and hence to fulfill the main education goal.
   For a particular student or group, personalized learning covers: recommendation of specific modules, courses, knowledge items and support materials that are identified as the best fit to maximize the particular student or group potential; optimization of knowledge acquisition process to match patterns of effective learning of individual students; and customization of knowledge composition that are best retained by students [44].

Other industries There are a lot other industries that are taking advantage of Big Data in their businesses. This section provides a less detailed overview of some such industries.

1. Retail industry.

   Retail industry was a pioneer in adopting Big Data solutions and nowadays all big companies in this industry heavily use Big Data analytics for a wide range of tasks. For instance, Macys.com is focused on customer-oriented analytical applications involving personalization, ad and email targeting, and search engine optimization. Macys.com utilizes a variety of leading-edge technologies for Big Data, most of which are not used else where within the company. They include open-source tools like Hadoop, R, and Impala, as well as purchased software such as SAS, IBM DB2, Vertica, and Tableau. Analytical initiatives are increasingly a blend of traditional data management and analytics technologies, and emerging Big Data tools. The analytic group employs a combination of machine learning approaches and traditional hypothesis-based statistics [9].
   Sears is also a company that is making use of Big Data in retail industry. They implemented an enterprise data warehouse in the 1980s while most retailers were still relying on manually-updated spreadsheets to examine their sales numbers. These days the company is using Big Data technologies to accelerate the integration of petabytes of customer, product, sales, and campaign data in roder to understand increase marketing returns and bring customers back to into the stores [9]. The company is now leveraging open source projects Apache Kafka and Storm to enable real-time processing.

2. Banking.

   Banks are struggling to profit from increasing volumes of data. Capgemin Consulting states in [] that there is a set of impediments which prevents banks from making use of Big Data. For instance, time taken to analyze large data sets, shortage of skilled people for data analysis, unstructured content is too difficult to interpret, the high cost of storing and analyzing large data sets. The use of customer data invariably raises privacy issues. By uncovering hidden connections between seemingly unrelated pieces of data, big data analytics could potentially reveal sensitive personal information[45]. Another big problem is a legacy systems. Deutsche Bank has been working on a big data implementation since the beginning of 2012 in an attempt to analyze all of its unstructured data. However, problems have arisen while attempting to unravel the traditional systems - mainframes and databases, and trying to make big data tools work with these systems. The bank has been collecting data from the front end (trading data), the middle (operations data) and the back-end (finance data). Petabytes of this data are stored across 46 data warehouses, where there is 90% overlap of data. It is difficult to unravel these data warehouses that have been built over the last two to three decades. The data integration challenge and the significant investments made by the bank in traditional IT infrastructure pose a key question for the banks senior executives - what do they do now with their traditional system? They believe that big, unstructured and raw data analysis will provide important insights, mainly unknown to the bank. But they need to extract this data, streamline it and build traceability and linkages from the traditional systems, which is an expensive proposition [46].

6.5 Comparative Study of Big Data Tools

As shown above there are many different Big Data Tools and various use cases for Big Data in vertical industries. But some Big Data Tools are more suitable for specific use cases than others. If a company picks a wrong tool for their project it could be more difficult or even impossible to meet the set goals.

Healthcare In Healthcare there are numerous different stakeholders who have different goals for Big Data analytics. For example Patients want their everyday technologies combined with medical care and want to be able to "comparison shop" for medical treatment [12]. Researchers want a high quality and quantity of data and new tools to improve their workflow. Pharmacy companies want to understand the causes of diseases to be able to provide treatments for those faster. One strategy adopted in healthcare to improve the general health is to support the patients by improving the understanding and to enhance the communication between the patients and the provider. Enhancing self-care is one way of doing this. With the rise of the mobile phones nearly everyone has a device which is capable of fast processing and motion sensing. Through tracking sleep and eating patterns apps like "MyFitnessPal" or Apple's "Health" are able to suggest a

better lifestyle to improve someone's health. The app "Health" works like a big data tool. It collects raw data from different sensors and other apps and can either show graphs of this data over different timeframes or let other apps analyze the data. Another approach is made by Humetrix's "iBlueButton" which is a mobile health information exchange app. It offers a secure way for patients and providers to communicate medical records via device-to-device communication [12]. "Ginger.io" is another application that enhances self-care. It is a cloud-based platform that collects data about the movement of the mobile phone of a patient with a chronic disease to analyze the patients behaviour between appointments [12]. But this is not the only strategy used by companies to provide healthcare application. Some companies try to improve the general health by increasing the awareness and knowledge of the general public about personal health. For instance "Asthmapolis" is an app that uses sensors on asthma inhalers to monitor the location and time of use to then tell the user where he used the inhaler the most. By doing so, the user can bring the asthma under control [12]. Most of these mobile apps gather data that has been collected from sensors or was manually entered by the users. The majority of those entries is structured data and therefore there is no need for a document-oriented database. The easiest way to deal with such data is with a NoSQL key-value database like the Google BigTable or Cassandra [36]. But Big Data in healthcare is not limited to be used to help the patients through apps on their mobile phones. Companies are also pooling data to build a better ecosystem. IBM developed a technology platform called "Watson" that uses natural language processing and large amounts of unstructured data to make faster and more precise diagnostics than a doctor could do [47]. Researchers in healthcare have to work with very large datasets of all types of data. For such purposes a normal NoSQL database is not efficient enough and a database that can process large amounts of unstructured data in a short amount of time is needed. The MapReduce algorithm splits the workload over a cluster of servers and is therefore able to analyze the dataset in a reasonable amount of time. Big Data Tools like Hadoop implement this algorithm and are widely used [36].

Logistic and Transportation When trying to improve logistic and transportation with Big Data, companies often try to improve their routing. This way they can reduce the fuel and the time spend driving [9]. This is done by collecting very large amounts of data from sensors to track speed, GPS data, breaking, and weather information amongst other things [9]. For uses like the Dublin public transit system it can be useful to additionally collect traffic flow data and to track events that take place in the city, including the amount of people expected to attend those events [40].

All this data is either structured like the data from velocity sensors or GPS data and semi-structured like the information of an event with the number of attendance plus the name and location of the event. Most of this data can be stored in a non document-oriented database since there is not a lot of unstructured data, but it should be possible to run real-time analysis. Though there

are other Big Data Tools that meet these conditions like the NoSQL databases Cassandra or Redis, Hadoop could be used to analyze the data collected by the sensors. Hadoop can do fast real-time analysis through the MapReduce algorithm and has useful functions like the visualization of the data. Since Hadoop is used quite often as a Big Data Tool, it can also be extended with other tools. Like in healthcare another use for Big Data Tool is to get a better understanding of the customer. Data about the customer is collected through the various websites with clickstreams, likes, ratings, comments, and through conversations with call centers. Since it contains a high amount of natural language, most of this data is semi-structured or unstructured. The analysis of this data does not have to be real time and can be done every month, week or day. But by making use of real time analysis, companies are able to provide a high quality customer support and they could even advertise new products or special offerings to unsatisfied customers. Because natural language analysis is much harder than the one of structured data, a more powerful tool is needed to maintain a good performance. Many companies utilize the tool Hadoop but most tools that implement the MapReduce algorithm and can work efficient with unstructured data can be reasonable choices.

7 Challenges in Application of Big Data Analytics in Vertical Industries

The adoption of Big Data Solutions can sometimes be difficult and complicated. There are numerous challenges that companies have to deal with to use Big Data efficient. These challenges can be technical or economical, but there can also be complications considering ethics.

Jeffrey Alan Johnson describes in [48] the ethical challenges when mining data from individuals rather than using sensors to collect non personal information. One of those challenges is to not violate a subjects privacy. The Arizona State University tries to identify students who intend to transfer, which is evidently an rather intimate information most students would like the university not to know about [48]. To avoid this ethical or even legal problem, call centers often ask for a permission to record the conversation for the use of converting it into text form and eventually analyzing it. Some of this data is collected anonymously, but if the data is not made anonymous it can be easy to target a specific subject and violate their privacy. Even when non anonymous private data has to be stored and analyzed to give the subject recommendations or feedback, the data has to be safe from all attacks that may occur and must not be made public without the agreement of the subject. If private information such as names, addresses, phone numbers or even credit card numbers or health records are publicly accessible, the person whom the information belongs to is vulnerable to manipulation, discrimination, harassment or identity theft. Another ethic concern especially in educational fields is the loss of individuality. Some programs in education are treating all subjects identically based on the assumption that they behave the same. Others represent the subject as a collection of attributes

rather than an individual. Therefore programs like an automated course recommendation systems can push a student to enroll in courses taken by other students with similar attributes ignoring other more important attributes or the opinion of the student [48].

But even when dealing with data not associated with a person there can be challenges in the application of Big Data. In healthcare large amounts of important data are generated by healthcare providers. Data like diagnoses, prescriptions, and conversations are all natural language and therefore unstructured. Unstructured data can be difficult to store and to analyze but an additional problem in healthcare is, that a majority of the data is handwritten and hard to read or virtually unreadable [12]. This lack of easy to analyze data possesses the problem that it can be very difficult to perform fast analysis or to store the given data efficiently. To solve this problem many companies extent their resources in order to analyze the data at a reasonable pace.

8 Conclusion

Big Data is a very useful way to improve a company's strategies. There are many different ways to use Big Data for Business Intelligence in Vertical Industries and there are various Big Data Tools available. For many years companies used structured data for Business Intelligence, but only in recent years especially through the invention of the internet new information sources were created. With this new data, companies were able to analyze it in big chunks. The knowledge derived from this data can then be used to get a direct or indirect cost reduction through time reduction. Some companies extended their offerings with new Big Data-Based Offerings while others build their company and all their products around Big Data and support internal business decisions with Big Data.

For the implementation of Big Data, companies can choose from a wide array of Big Data Tools. But some Big Data Tools have limitations on how fast they work or what type and quantity of data they can store. Many companies also have to deal with different challenges while adopting to Big Data solutions. Difficulties in storage or an excessive amount of unstructured data are some of these challenges. But one of the most challenging problems is the ethical question while making use of personal information. At the moment this problem is most apparent in relation to governments [49].

References

1. Gopalkrishnan, V., Steier, D., Lewis, H., Guszcza, J.: Big data, big business: Bridging the gap. In: Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Beijing, China, ACM, ACM Press (2012) 7–11
2. Times, N.Y.: Advanced offer path to further shrink computer chips. `http://www.nytimes.com/2010/08/31/science/31compute.html` (2010)
3. Guszcza, J., Lucker, J.: Beyond the numbers: Analytics as a strategic capability, Deloitte Review (2011)

4. Press, G.: A very short history of big data. http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/ (2013)
5. Marron, B.A., de Maine, P.A.D.: Automatic data compression. Commun. ACM **10**(11) (1967) 711–715
6. PC, Magazine: Storage: From highway robbery to runaway bargain (2007)
7. Morris, R.J., Truskowski, B.J.: The evolution of storage systems. IBM systems Journal **42**(2) (2003) 205–217
8. NESSI: Big data white paper: A new world of opportunities. http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf/ (2012)
9. Davenport, T.H., Dyche, J.: Big data in big companies, SAS (2013)
10. Hans, S., Swineheart, G.: Finding the needle - using forensic analytics to understand what happened - and what might happen. Deloitte: Boston, Masachusetts, USA (2011)
11. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D.: Big data. The management revolution. Harvard Bus Rev **90**(10) (2012) 61–67
12. Feldman, B., Martin, E.M., Skotnes, T.: Big data in healthcare hype and hope. October 2012. Dr. Bonnie **360** (2012)
13. Beal, V.: Structured data. (http://www.webopedia.com/TERM/S/structured_data.html)
14. Russom, P., et al.: Big data analytics. TDWI Best Practices Report, Fourth Quarter (2011)
15. Davenport, T.H., Harris, J.G.: Competing on Alalytics. The New Science of Winning. Harvard Business School Press (2006)
16. IBM: What is big data analytics? http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html (2010)
17. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics, Toronto, Ontario M5B 2K3, Canada, Ted Rogers School of Management, Ryerson University (2014)
18. Duggan, S.: Big data analytics: Descriptive, predictive and prescriptive. http://www.charterglobal.com/big-data-analytics-descriptive-predictive-and-prescriptive (2015)
19. Mosavi, A., Vaezipour, A.: Developing effective tools for predictive analytics and informed decisions, University of Tallinn (2013)
20. Shmueli, G., Koppius, O.: Predictive analytics in information systems research. MIS Quarterly **35**(3) (2011) 553–572
21. Luhn, H.P.: A business intelligence system. IBM Journal (1958) 314–319
22. Power, D.J.: A brief history of decision support systems. Technical report, DSS-Resources.COM (2007)
23. Golfarelli, M., Rizzi, S.: Data Warehouse Design: Modern Principles and Methodologies. The Tata McGraw-Hill (2009)
24. Shollo, A.: The Role of Business Intelligence in Organizational Decision-making. PhD thesis, LIMAC PhD School (2013)
25. Heinze, J.: Business intelligence vs. business analytics: Whats the difference? https://www.betterbuys.com/bi/business-intelligence-vs-business-analytics/ (2014)
26. Shariat, M., Hightower, R.: Conceptualizing business intelligence architecture. Marketing Management Journal (2007)
27. Dekkers, J., Versendaal, J., Batenburg, R.: Organising for business intelligence: A framework for aligning the use and development of information. In: BLED 2007 Proceedings, Deloitte (2007) 625–636

28. Pemmaraju, S.: Converting hr data to business intelligence. Employment Relations Today **34** (2007)
29. Yogev, N., Even, A., Fink, L.: How business intelligence creates value: An empirical investigation. International Journal of Business Intelligence Research **4**(3) (2013)
30. Negash, S.: Business intelligence. Communications of the Association for Information Systems **13** (2004)
31. Davenport, T.H.: Big data executive surveythemes and trends. Technical report, NewVantage Partners (2012)
32. Lucker, J.: Know what customers wnat before they do, Harrvard Business Review (2013)
33. Nannetti, P.: The deciding factor: Big data and decision-making. Technical report, Economist Intelligence Unit (2012)
34. Davenport, T.H.: How strategists use big data to support internal business decisions, discovery and production. Strategy and Leadership **42** (2014) 45–50
35. Amatriain, X., Basilico, J.: Netflix recommendations: Beyond the five stars. `http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html` (2012)
36. Warden, P.: Big Data Glossary. OReilly Media, Inc (2011)
37. Cardosa, M., Wang, C., Nangia, A., Chandra, A., Weissman, J.: Exploring mapreduce efficiency with highly-distributed data. In: Proceedings of the second international workshop on MapReduce and its applications, University of Minnesota (2007) 27–34
38. Groves, P., Kayyali, B., Knott, D., Kuiken, S.V.: The big data revolution in healthcare. `http://www.pharmatalents.es/assets/files/Big_Data_Revolution.pdf` (2013)
39. Patil, P., Raul, R., Shroff, R., Maurya, M.: Big data in healthcare. International Journal of Research in Information Technology (IJRIT) **2**(2) (2014) 202–208
40. Ayed, A.B., Halima, M.B., Alimi, A.M.: Big data analytics for logistics and transportations. In: Proceedings of 2015 4th IEEE International Conference on Advanced logistics and Transport (ICALT), Research Groups in Intelligent Machines, University of Sfax (2015) 311–316
41. OECD: Big data and transport. Technical report, International Transport Forum (2015)
42. Schmarzo, B.: What universities can learn from big data higher education analytics. `https://infocus.emc.com/william_schmarzo/what-universities-can-learn-from-big-data-higher-education-analytics` (2014)
43. King, I.: Big education in the era of big data. In: Proceedings of Federated Conference on Computer Science and Information Systems, The Chinese University of Hong Kong (2014)
44. Cen, L., Ruta, D., Ng, J.: Big education: Opportunities for big data analytics. In: Proceedings of 2015 IEEE International Conference on Digital Signal Processing (DSP), Etisalat British Telecom Innovation Centre, Khalifa University of Science, IEEE (2015) 502–506
45. Coumaros, J., Buvat, J., Auliard, O.: Big data alchemy: How can banks maximize the value of their customer data? Technical report, Capgemini Consulting (2012)
46. Preez, D.: Deutsche bank: Big data plans held back by legacy systems. `http://www.computerworlduk.com/news/data/deutsche-bank-big-data-plans-held-back-by-legacy-systems-3425725/` (2013)

47. IBM: What is watson? (http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html)
48. Johnson, J.A.: The ethics of big data in higher education. International Review of Information Ethics **7** (2014)
49. Beuth, P.: Millionen neue jobs! http://www.zeit.de/digital/internet/2016-01/hatespeech-csu-facebook-vorab-loeschen (2016)