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. to be done

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 analysed [1]. Tasks in modern analytics require a huge computing power, storage capacity, appropriate information technologies which are needed to gather, analyse 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 of companies try to make use of opinions expressed by customers in order to improve the overall customer experience. Others analyse data to find patterns in customer's behaviour which can be used later to predict future needs and purchases. Data is a strategic asset in making recommendations. Often, data is analysed 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 analysing the data the industries gathers, it is much easier to adapt to current trends and they 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 it's use will change the Business Intelligence, particularly in the vertical markets. It'll compare how the different vertical industries are dealing with the topic of Big Data Analytics and what strategies are adopted while doing so.

2 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 [4]. 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 [5]. 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 analyse 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 [6], 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 [7] 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 [6].

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[7]. Predictive Analytics is the use of historical data to forecast on consumer behaviour and trends [8]. This type of analysis makes use of the statistical models and machine learning algorithms to identify patterns and learn from historical data [9]. 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 [6].

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.

Thomas H. Davenport and Jill Dyche in [10] 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.

3 Business Intelligence

The term "Business Intelligence" was used in scientific context for the first time in an article by Hans Peter Luhn, an IBM researcher, in 1958. In his article [11], 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 Howard 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" [12].

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 [13]. 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 [14].

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 [15]. 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. Mark 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 [15].

The concept of BI, although not new, is experiencing a renaissance which positions the concept in its infancy phase again [14]. Therefore a lot of new and ambiguous definitions have emerged. After some time, BI became a multidimensional 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 [16] 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 [17] define BI as the continuous activity of gathering, processing and analysing data - supported by a BI system. On the other hand, S. Pemmaraju in [18] 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 [19].

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 [20]. 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.

4 Context of Big Data in Vertical Industries

4.1 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 [10].

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 [21] 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 [10]. 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[10]. 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 [10]. 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 [10] 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.

Next Best Offers (NBOs)

J. Lucker states in [22] 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 [22]. 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 [22]. 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 [23]. 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 [24]. The most advanced firms are monitoring not only their own suppliers, but their suppliers' suppliers [10].

4.2 Big Data Tools

The four main goals defined in the previous section 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 [25] 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 [25]. 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 [25]. 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 [25] 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 [26]. 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 [25].

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 [25].

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.

4.3 Business Questions

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 [27] 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].

References

1. Gopalkrishnan, V., D.Steier, 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 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 stratss egic capability, Deloitte Review (2011)
4. Davenport, T.H., Harris, J.G.: Competing on Alalytics. The New Science of Winning. Harvard Business School Press (2006)

5. IBM: What is big data analytics? http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html (2010)
6. 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)
7. Duggan, S.: Big data analytics: Descriptive, predictive and prescriptive. http://www.charterglobal.com/big-data-analytics-descriptive-predictive-and-prescriptive (2015)
8. Mosavi, A., Vaezipour, A.: Developing effective tools for predictive analytics and informed decisions, University of Tallinn (2013)
9. Shmueli, G., Koppius, O.: Predictive analytics in information systems research. MIS Quarterly **35**(3) (2011) 553–572
10. Davenport, T.H., Dyche, J.: Big data in big companies, SAS (2013)
11. Luhn, H.P.: A business intelligence system. IBM Journal (1958) 314–319
12. Power, D.J.: A brief history of decision support systems. Technical report, DSS-Resources.COM (2007)
13. Golfarelli, M., Rizzi, S.: Data Warehouse Design: Modern Principles and Methodologies. The Tata McGraw-Hill (2009)
14. Shollo, A.: The Role of Business Intelligence in Organizational Decision-making. PhD thesis, LIMAC PhD School (2013)
15. Heinze, J.: Business intelligence vs. business analytics: Whats the difference? https://www.betterbuys.com/bi/business-intelligence-vs-business-analytics/ (2014)
16. Shariat, M., Hightower, R.: Conceptualizing business intelligence architecture. Marketing Management Journal (2007)
17. 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
18. Pemmaraju, S.: Converting hr data to business intelligence. Employment Relations Today **34** (2007)
19. Yogev, N., Even, A., Fink, L.: How business intelligence creates value: An empirical investigation. International Journal of Business Intelligence Research **4**(3) (2013)
20. Negash, S.: Business intelligence. Communications of the Association for Information Systems **13** (2004)
21. Davenport, T.H.: Big data executive surveythemes and trends. Technical report, NewVantage Partners (2012)
22. Lucker, J.: Know what customers wnat before they do, Harrvard Business Review (2013)
23. Nannetti, P.: The deciding factor: Big data and decision-making. Technical report, Economist Intelligence Unit (2012)
24. Davenport, T.H.: How strategists use big data to support internal business decisions, discovery and production. Strategy and Leadership **42** (2014) 45–50
25. Warden, P.: Big Data Glossary. OReilly Media, Inc (2011)
26. Cardosa, M., Wang, C., Nangia, A., Chandra, A., Weissman, J.: Beyond the hype: Big data concepts, methods, and analytics. In: Proceedings of the second international workshop on MapReduce and its applications, University of Minnesota (2007) 27–34
27. Amatriain, X., Basilico, J.: Netflix recommendations: Beyond the five stars. http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html (2012)