June 11 | 13

Big Data in Business

Graduation Thesis

Author: Namrata Sakhrani

Student Details	
Family name , initials:	Sakhrani, N
Student number:	2126284
Project period: (from – till)	March 18, 2013 – July 1, 2013
Company Details	
Name company/institution:	Fontys Hogeschool/Andarr
Department:	
Address:	Rachelsmolen 1, Eindhoven
Company tutor	
Family name, initials:	Van Tol, Eric
Position:	Director
University tutor	
Family name , initials:	Hamers, Rien
Final report:	
Title:	Big Data in Business
Date:	June 11, 2013

Preface

Practical orientation in a professional IT/business environment is a substantial and characteristic element of the BIS curriculum. Therefore the student will participate in a graduation thesis during the 8th semester. At this time, the student will carry out the tasks required to complete the assignment

The student had to choose a company outside Fontys that will offer the opportunity to do a thesis in collaboration with the company for a period of 80-100 days.

BIS students must be aware of the international business and IT related aspect of their placement. The assignment must be oriented on the basis of IT Business management.

This Final Report was written by Namrata Sakhrani, 8th Semester Business Information Systems student at Fontys University of Applied Science in Eindhoven, The Netherlands. Namrata Sakhrani was born in Philipsburg, St. Maarten and lives in Rotterdam, The Netherlands.

This report is primarily meant for the BIS supervisor however, the company supervisors shall not be excluded as they have to verify certain elements.

The author would like to thank Eric van Tol, Director of Andarr and Rien Hamers, Fontys Internship Tutor for their guidance and support as company and univeristy supervisors throughout the period. Furthermore the author would like to thank, Mr.Sander van Kleef, BI Consultant at Ordina, Mr. Pieter Stel, BI Senior Manager at Bearing Point, and Mr. Gwellyn Daandels, Consultant at Cognizant, for their assistance as experts in the field during conduct of the thesis.

Executive Summary

Wikipedia defines Big Data as a “collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.” In other words, Big Data can be generalized as the consolidation of volumes of data sets from several different data sources to be processed by non-traditional methods. IBM has characterized Big Data according to four distinct V’s, these being, volume, velocity, variety, and veracity.

Volume: The mass amounts of data available in terabytes even petabytes.

Velocity: Streamlining the processing of all the data sets at faster speeds.

Variety: It is no longer only about structured data rather it is also about the unstructured and semi-structured forms of data available

Veracity: How reliable the data being processed is in regards to maintaining a high quality level of information. Therefore how trustworthy is the information that has been collected.

The report illustrates the concept of Big Data and its influence in the business world. The report is to aid in better understanding the revolution of data analytics into a multidimensional perspective. As data keeps growing and businesses are faced with the issue of an information overload, technology and techniques have been developed to make it easier to analyze and process this data even in its most raw form. The report has been divided into eight chapters, starting with a deeper overview of Big Data and concluding with recommendations on the adoption of Big Data in Business.

Chapter 1 will begin to introduce Big Data and the implications that come with understanding Big Data, followed by **Chapter 2** where several subsections which provide an in depth outlook on Big Data. These subsections discuss the value of Big Data, the underlying issues and benefits of Big Data, Big Data’s influence on certain industrial sectors and how should businesses approach adopting Big Data. Besides the literature study conducted, **Chapter 3** takes a different approach, where Big Data is explored through Field Research. Experts in the field of Business Intelligence and who are familiar with the concept of Big Data as well as companies that have taken on the Big Data endeavor were interviewed during the process. The findings of these interviews are described in this chapter. **Chapter 4** follows with an illustration to a Big Data framework which provides a conceptualization for businesses to consider when considering to taking Big Data practices. To conclude the Big Data research, **Chapter 5** includes recommendations on the overall outlook of Big Data in Business. To complement the research conducted, a Sample Business Case is provided to demonstrate the use of Big Data in Business in **Chapter 6**. Lastly, any references and additional research is included in **Chapter 7 and 8** respectively.

Table of Contents

Glossary

ELT: *Extract Load Transform*

ETL: *Extract Transform Load*

Data Warehouse: *A database used for reporting and data analysis. It is a central repository of data which is created by integrating data from one or more disparate sources.*

Data Mart: *Access layer of the data warehouse environment that is used to get data out to the users.*

MNE: *Multinational enterprise*

Metadata: *Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.*

OLTP: *Online transaction processing*

Key value stores: *A means of storing data without being concerned about its structure.*

Real-time: *Providing an almost instantaneous response to events as they occur.*

Structured data: *Data stored in a strict format so it can be easily manipulated and managed in a database*

Unstructured data: *Data with no set format, or with a loos format*

Chapter 1. Introduction

"Data have become a torrent flowing into every area of the global economy."[1]

Data has grown significantly over several years. Companies have created a cotton ball of data through several procurement methods. If one were to explore this sea of data, you would probably find information in its most raw form. Companies gather data on their employees, customers, suppliers etc. as well as from, what is better known as, the 'Internet of Things', where information is obtained from physical objects embedded with sensors with the ability to communicate. The Internet world has revolutionized the way the physical world connects. Therefore companies can easily keep track of the customers' purchasing records or employees' progress while at the same time the physical connection between objects archive a mass of raw data.

With an abundance of data, available at our fingertips, we are left to question what do we do now? All this data will only keep piling and due to its unstructured form, companies question its reliability and physical worth.

"A business running without accurate data is running blind."[2] Big Data is the product of the sheer volume of data flow across several digital mediums. 'Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. Generally any dataset over a few terabytes is considered to be Big Data. Big Data has been found to have the ability to influence economical decisions.' According to McKinsey Global Institute (MGI), a retailer is capable of increasing its operating margin if Big Data is used to the full potential. With the ever changing dynamic of the technological world, Big Data is believed to provide the potential to boost economic efficiency for both the public and private sector.

Big Data has been practiced by MNE's to gain competitive edge. Tesco, which is a UK based grocery company and is the world's 3rd largest retail chain, uses different strategic techniques, in terms of Big Data, to better understand their shoppers i.e. loyalty card program. As can be seen by this example, Big Data has successfully been applied by companies to provide beneficial prospects. However, this may come across to the intimidating as several businesses find it difficult to apply the analytics behind unstructured data. Businesses are encouraged to view this as a value adding mechanism through which they may be able to find the biggest and highest strategic opportunities. In regards to strategy, Big Data influences each process within a business besides just marketing and sales.

How does Big Data create value for companies? [1]

- Creating transparency
- Enabling experimentation to discover needs, expose variability, and improve performance
- Segmenting populations to customize actions
- Replacing supporting human decision making with automated algorithms
- Innovating business models, products, and services

---

[1] McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.
[2] Economist Intelligence Unit. 2012. *Lesson from the Leaders.* The Economist

Data readily made available or rather accessible to relevant stakeholders creates a more lucid organization. Here in, processes can improve, ensuring better response times and quality work. Big Data entails procuring more accurate and detailed results which could enable experimentation methods, thereby limiting variable performance. It creates a platform for identifying separate demographics deploying tailored products and services specific to needs. Big Data may also support human decision making in that it provides subjective analytical results. Lastly companies can manufacture new products and services according to Big Data results as well as improve business models.

Big Data requires a forward thinking mindset. In other words, Big Data can only be viewed as advantageous to a company if its leaders approach this concept with an objective and less than skeptical perspective. "Companies that interface with large numbers of consumers buying a wide range of products and services, companies that process millions of transactions and those that provide platforms for consumers digital experiences. These will be the big-data-advantaged businesses."[3] Besides the fact that Big Data seems like a very favorable option for businesses, the devil's advocate would include the several issues which influence the potential of applying Big Data. These issues have been identified in perspective to data policies set in motion implicating organizational boundaries i.e. privacy and security risks.

Due to the overwhelming volume of data, to capture, in essence, legitimate data, organization must consider adopting new technologies. Other factors come into play including the availability of organizational talent, uniform access of data amongst stakeholders which instigates a competitive atmosphere and finally, sectors lacking competitive intensity and performance transparency are most likely to not fully exploit the benefits the Big Data. Of course to every issue there has to be a defining line for compromise or rather the initiative to take significant measures to validate it. Through aid of the continuous transformation within the technological world and support by policy makers to enable economic growth will ultimately aid the ease of conformation to Big Data.

[3] McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.

Chapter 2. The Big Data Story

2.1. What is Big Data?

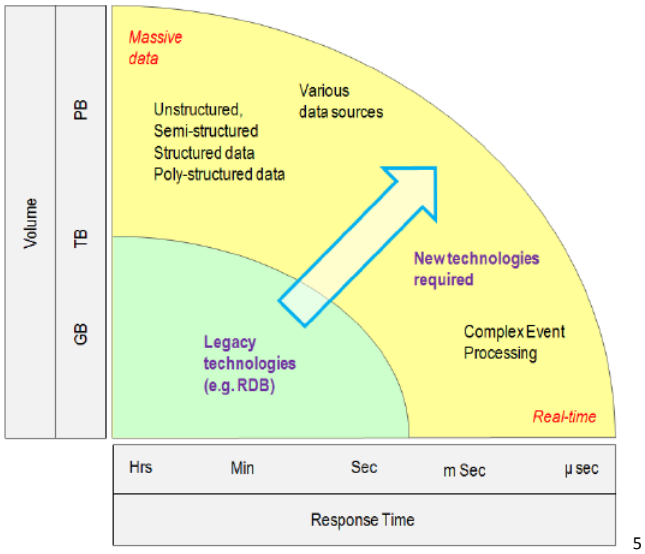
"Big Data is a relative term describing a situation where the volume, velocity, and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making."[4]

Big Data can be described in terms of four characteristics:

- Volume: with the increasing amount of data accumulated everyday, it is considered to be Big Data once over a few terabytes or even exabytes. At this point we are storing all kinds of data, i.e. environments, financial, medical, etc. and since the digital age. We record everything through the day by means of technology. Organizations also face implications of massive volumes of data as they store information on their employees, customers, shareholders, and other stake holders as well as departmental data.
- Variety: The amount of data accumulated is available in different data types and sources. The multiplicity of this variation means data will come in all forms including structured, non-structured/raw, and even semi-structured from several sources i.e. web pages, sensory data, e-mails, documents etc.
- Velocity: The rate at which data flows must be processed at an accelerating speed. This means the technology enabling this flow of data should be able to maintain the peed requirement for collecting, processing, and using data. Big Data has therefore reached a significant mass within the economy along with the intensity of the digital age, the development of data has rapidly escalated.
- Veracity: Otherwise known as data uncertainty, concerns the level of reliability and predictability associated with data types. The inherent imprecise nature of data causes speculation in being able to maintain high data quality despite the application of data cleaning methods. [2][3][4]

Big Data is more commonly interpreted through means of the 3 V's (sometimes even 4 V's) however the rise of this technology should also be interpreted in terms of the value it is capable of providing. It combines the characteristics of different data sets in order to form the bigger picture. There is a versatile array of data sources nowadays which are to be taken into consideration, for example, data available in the form of log contents, click streams, multimedia types i.e. video, photo, audio. With the evolution of the application of sensors embedded within devices, the amount of data collected creates several opportunities for an organization to be able to influence decision-making.

[4] SAS. *Big Data meets Big Data Analytics.*



[5]

The image above illustrates the ratio of volume to processing power. More clearly put, data that is usually collected through means of traditional BI methods i.e. relational database systems, are available in GB volumes however, if you take data analytics a step further you are faced with dimensions of data formats i.e. unstructured, semi-structured. This means new technologies are required to be able to handle this influx of data and in all the different forms it comes in. The objective is to make the unimaginable imaginable, wherein companies can process all this data in real-time at high speeds. Therefore Big Data opts to transform the concept of Business Intelligence Analytics onto a platform wherein the way we perceive information is multi-dimensional.

[5] Fujistu. 2013. *Solution Approaches for Big Data.*

2.2. Value of Big Data

The value of Big Data for any organization can be discovered at varied levels. Of course an organization must first determine the kind of objective it is aiming to achieve and the value of Big Data will be tested once it has been challenged. In other words an organization, if it invests in the right technologies, should expect to come across new insights.

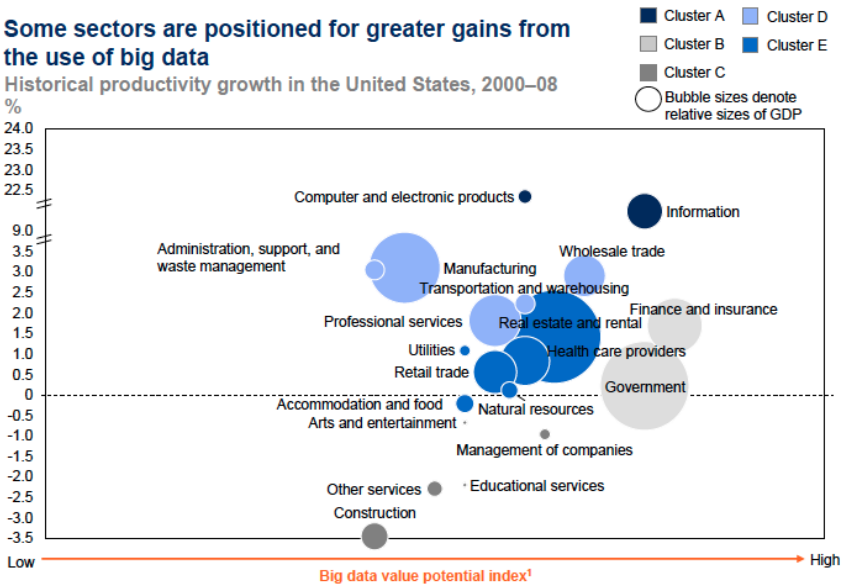
IBM outlines certain key Big Data principles a company should keep in mind when considering to take action in this perspective:

- Big Data solutions are ideal for analyzing not only raw structured data, but semi-structured and unstructured data from a wide variety.
- Big Data solutions are ideal when all, or, most of the data needs to be analyzed versus a sample of the data.
- Big Data solutions are ideals for iterative and exploratory analysis when business measures on data are not predetermined.
- Big Data solutions are beneficial as it preserves the fidelity of data and allows the company to gain access to mountains of information for exploration and discovery of business insights.
- Big Data is well sited for solving information challenges that don't natively fit within a traditional relational database approach for handling the problem at hand. [5]

Leveraging data within an organization can improve productivity growth in terms of producing higher quality products. Big Data provides organization with the upper hand in creating value for products and services. Therefore it can generate significant financial value across sector. According to MGI, Big Data will not only create way for productivity growth but as well as influence consumer surplus.[6] The large amount of data captured by consumers creates an economic surplus favorable to sectors that deploy Big Data i.e. a better match can be made between products and customer needs. Enhanced consumer surplus enables improved economic transparency in terms of pricing and revenue performance as well as public sector administration.

Big Data will influence the performance of several sectors at a time however affects them differently. Certain sectors are poised for greater gains du to transparency of barriers and data is readily accessible.

[6] Consumer surplus occurs when the consumer is willing to pay more for a given product than the current market price. [17]



7

The figure above illustrates sectors capable of greater gains from the use of Big Data.

Computer and electronic products and the information sectors benefit considerably from Big Data in their opportunity for standing productivity growth. Compared to other sectors, those experiencing substantial productivity growth are less affected by barriers which prevent them from experiencing a higher degree of data surplus. Sectors such as those of the public sector i.e. educations face the main issue of lack of data-driven mindset and available… therefore the value of Big Data varies according to every sector especially since its influence can only be determined by the exposure of available/captured data and the ability to effectively manipulate the data in favor of the respective industry sector.

“The value of Big Data that can be unlocked from analytics, also known, as, ‘data equity’, is increasing rapidly as technological innovations take hold.” [8]

There are several technological platforms through which data is captured both structured and unstructured. All the available information is only of real value if it is capable of providing strategically driven insights. Data equity capitalized on the company’s profitable range. The company gains further knowledge into logistical information, customer behaviors, and economic/market trends.

[7] McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.
[8] SAS. *Big Data meets Big Data Analytics.*

2.3. Underlying issues of Big Data

2.3.1. Data governance

As more data is recorded digitally means this could possibly result in attesting certain lawful or even speculative boundaries. Organizations face the issues concerning privacy, security, and legality. These issues conspire the validity of data. Privacy is a point of concern to customer especially when it comes to data pertaining to healthcare and welfare details. Even though the data acquired could potentially provide substantial information into identifying better medical and financial practices; the data is considered to be sensitive. With the abundance of data available, one must question how safe their details are in order to avoid it being easily accessed by a third party. There have been several instances during which security breaches have been experienced. Therefore this results in further speculating as to the credibility of data security.

For example, the latest security breaches have targeted top social media enterprises including Twitter and Facebook. Twitter faced the thefts of passwords since the account of the Associate Press news service was hacked and a false post was published claiming of explosions at the White House which resulted in a steep stock market decline specifically $136 billion loss in market value. [6]

With all this exposure to the data available, legitimizing ownership over the information presents the unwavering nature for legal action or more exclusively intellectual property rights. The governance of data presents several unfavorable issues which object to the concept of Big Data. To make Big Data seem more credible the implementation of data policies must be introduced.

2.3.2. Technological advancements

Technology is ever-changing. Therefore enterprises are compelled to adopt according to the dynamic nature of technological advancement. Data will always continue to grow in mass. Organizations must deploy new technologies to capture the data. Depending on the kind of data and the company's experience with Big Data, the appropriate technology to be adopting will vary. To successfully integrate, analyze, visualize, and consume the growing torrent of Big Data, innovation of technologies and techniques will aid individuals and organizations.

2.3.3. Organizational characteristics

Big Data has been used by many multinational companies to gain a competitive advantage. In several industries, organizations aren't well educated in Big Data. Therefore these companies lack the necessary knowledge required to gain from the full potential of Big Data. It thereby also poses a difficult situation for new entrants into the market which attempt at applying Big Data techniques as MNE's already have advanced significantly in this concept. Organizations aren't aware of how to structure workflows and processes subjected to Big Data in order to attain credible data insights. The right technology, talent, and awareness are required for an organization to derive any optimal action as a part of the Big Data act.

In addition, according to the specific sector/industry companies lack of competitive pressure, limits urgency for performance enhancement i.e. public sector industries generally don't feel obligated by economic pressure. Therefore these organizations do not acknowledge the benefits of Big Data. Big

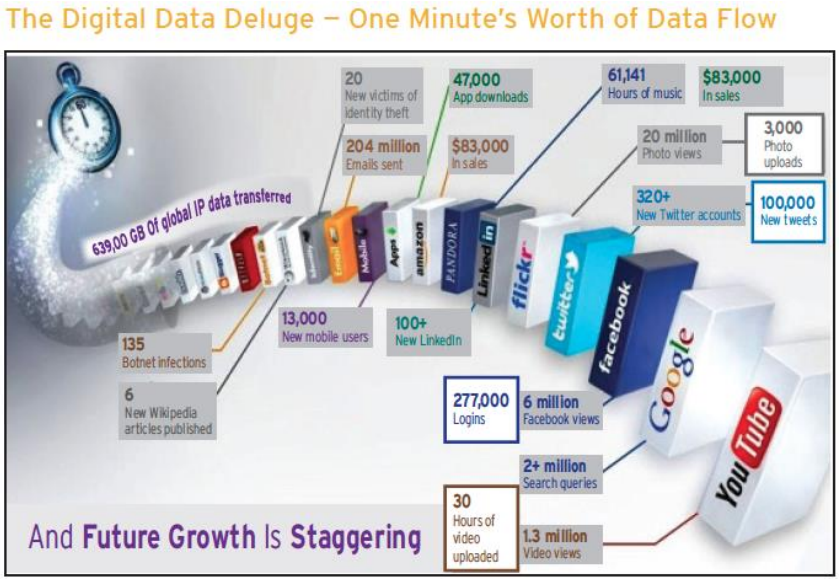
Data can in fact enable decision making criterion especially for example, the health public sector which van benefit from Big Data can determine better medical treatments for patients.

Separate parties view Big Data differently however, the big picture behind this innovation is the idea of improved service by individuals and organization simultaneously. Big Data is not intended to present a controversial future.

2.4. How BIG is Big Data?

Exponential is one word that could describe the growth rate of data. Data is consistently accumulated across several mediums. The degree to which it is recorded can tally up to petabytes of raw data. The siloed data along with new incoming data do not only serve an economical purpose rather may map for greater value creation. In other words, Big Data stands to hold the potential that can benefit all societal stakeholders.

The deluge of data is happening at an incredible pace. To clearly illustrate, from approximately 5 EB of data online in 2002 grew to 750 EB in 2009 and by the year 2021 it has been projected to reach over 35 ZB. Statistically speaking 90% of all data was created within the past couple of years compared to the last 4 decades. The world's leading commercial information providers deal with long standing stacks of business records. They are faced with over 200 million records of information of which are accumulated from different sources. Their databases are updated every 4 to 5 seconds. The challenge to deal with this influx of data will only magnify. [7]



[9]

According to the research issued by the information management company EMC, their analysis showed that the amount of data generated is increasing at an alarming rate, faster than the world's storage capacity is being amplified. The volume of data is calculated to grow by 44 times to 2020 which suggests an annual growth rate of 40 percent.[10] The ability to generate and process data has risen over the last few decades. The global storage and computing capacity from 1986 to 2007 was estimated to have grown by 23% annually. Through the years we have gradually digitized our means of storing data. The rise of digitization as a result is accountable for the 94% of data stored through means other than analog forms. The increasing volume and detail of information captured by enterprises, together with the rise in multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future. [1]

---

[9] Cognizant. 2012. *Big Data's Impact on the Data Supply Chain*

[10] McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.

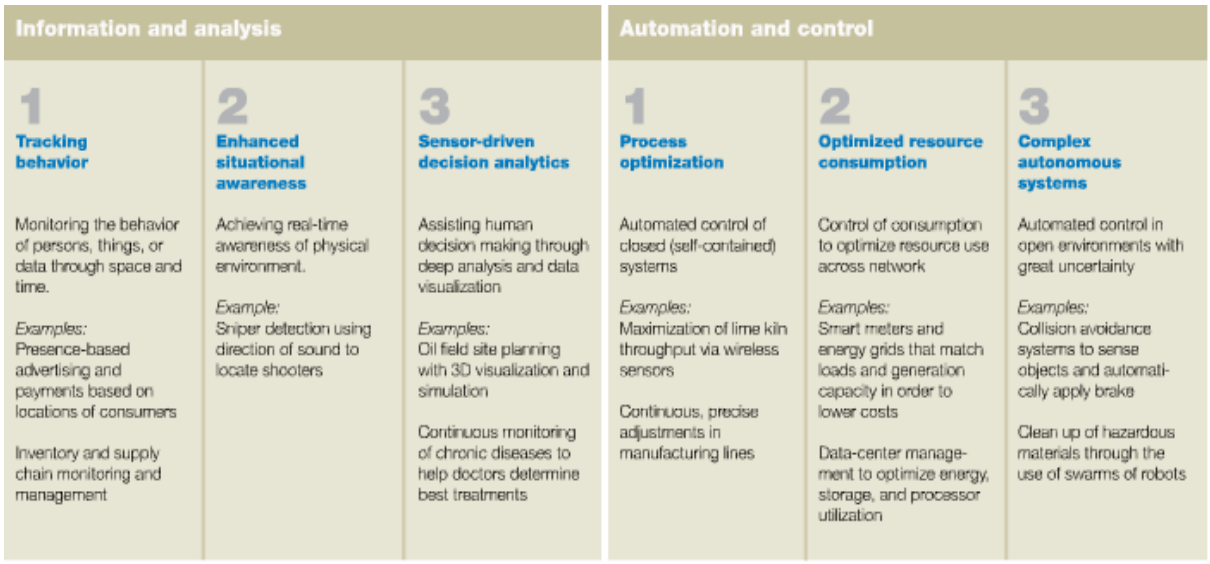
Information originates from several sources including proprietary databases, public means and from the Internet as well. “The physical world itself is becoming a type of information system.”[11] The Internet of Things implies a revolution to attaining information through physical information systems i.e. sensors, actuators found in physical objects which are linked to IP networks connected to the Internet. The Internet of Things provides companies with the advantage to gain a competitive advantage over their competitors. As the abundance of data continues to pile up, the advancement of technologies which enable the adoption of Big Data practices provide companies with new opportunities to compute and process large sets of information.

To conclude, let’s take another look into the amount of data collected on a daily basis. Twitter generates more than 250 million tweets per day, nearly 50 hours of video are uploaded every minute on YouTube, and over 200 million photos are uploaded per day on Facebook calculating to over 90 billion photos. This is simply the online multimedia platform, where the users across the globe are responsible for the endless supply of data. “With new electronic devices, technology and people churning out massive amounts of content by the fraction, data is exploding not just in volume but also in diversity, structure, and degree of authority.”

[11] Chui, Michael, Markus Loffler, and Roger Roberts. "The Internet of Things." *McKinsey & Company*. Mar. 2010.<http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_things>

2.5. Internet of Things

As mentioned in the previous section, the Internet of Things is basically reforming the schema of business models. Companies can adopt ways in which they may adhere to changing environment of the Internet realm. The Internet of Things can be defined as "the network of physical objects that contain embedded technology to communicate and sensor or interact with their internal states or the external environment."[12] This concept, or rather innovation, of thinking provides an organization with the ability to track behavior, enhance control and business automation processes, and optimize resourcefulness.



McKinsey has identified ways in which technologies along the basis of the Internet of Things compliment forms of application. As seen in the image above, six different applications serving separate purposes of Information and Analysis and Automation and control illustrate patterns companies face in the decision making process.

2.5.1. Information and Analysis

Once companies come across technologies that base their data to produce certain outcomes in terms of product development, internal and external operations etc., are dependent on the factor of information and analysis application.

1. Tracking behavior

Embedding sensors to certain products gives way for companies to track processes and movements along the way which provides the opportunity to institutionalize business models in a more efficient way. In other words this tracking behavior provides behavioral data. This type of data has provided companies across different industries to enhance their business processes serving both internal and external purposes. For example, insurance companies offer a service to install location sensors in cars to base the price on policies depending on how the customer drives the car and where it travels. Another example would be using sensors to track RFID (radio-frequency-identification) tags.

[12] "Internet of Things." *Gartner IT Glossary*. Web. <http://www.gartner.com/it-glossary/internet-of-things>.

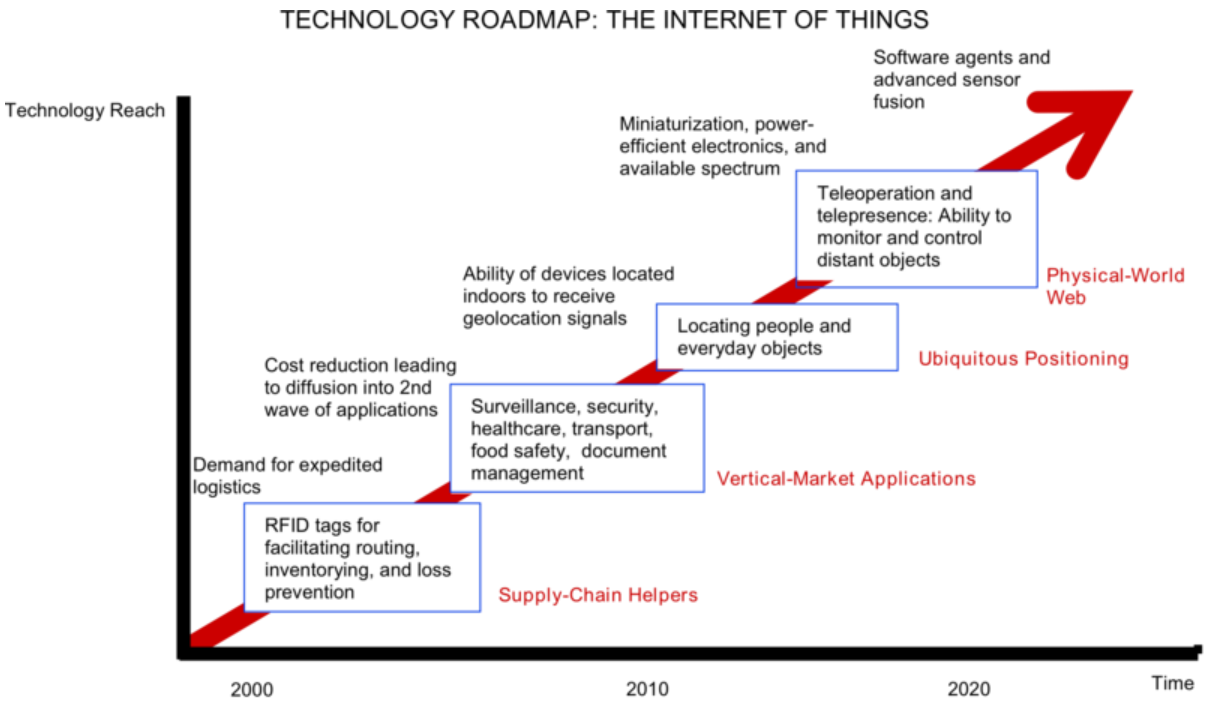
The logistics aspect within a business can greatly be improved through this method wherein products which move along the supply chain. This can lead to a reduction in production and other logistics costs. Tracking data has enabled several industries even within the aviation sector has adopted this technology, to be able to enhance business models.

2. Enhanced situational awareness

Sensors deployed amongst environmental applications i.e. infrastructure, meteorological conditions, which provides data to decision makers with information based on real-time events.

3. Sensor-driven decision analytics

The idea of embedding sensors for simply improving business processes or analyzing environmental conditions serve communal purposes however, this technology can be up-scaled. In other words, advanced software technologies and large storage capacities, which will allow for a more overall and complex decision making process. In the retail sector for example, companies can gather information on a shopper's adventure through stores to optimize store layouts and in turn increase revenues. Another important example would be the healthcare sector where in patients can be better treated based on real-time information collected when monitoring their behavior and symptoms. Doctors can diagnose patients with treatments better fitted to their conditions.



[13]

The image above illustrates the roadmap of the Internet of Things up to 2020 and further into time.

[13]"Internet of Things." *Wikipedia*. Wikimedia Foundation, 06 Oct. 2013. Web. <http://en.wikipedia.org/wiki/Internet_of_Things>.

2.5.2. Automation & Control

What good is data if it only provided the ability to analyze processes rather it should enforce a controlled environment through feedback automated during the analysis of processes. This ultimately modifies a process in an autonomous environment.

1. Process optimization

As per the aspect of process optimization, sensors feed information to computers that analyze this data and send signals to the system to adjust processes. For example, in the chemical industry this mechanism is adopted in order to improve scale of granularity through modification of temperature, mixtures or even during assembly lines. This in turn reduces the amount of waste and energy costs as well as prevents the need for much human intervention.

2. Optimized resource consumption

Utility services are deploying sensor ridden systems to provide customers with visuals displaying the energy usage and real-time costs that come with it. These sensors provide automated feedback which influence usage patterns enabling pricing differentiation. Residential customers can knowingly reduce consumption of household utilities based on time-of-use pricing i.e. shut down air conditioners or delay running dishwashers during peak times. It can also help commercial customers in altering periods of high intensity energy production during lower-priced off peak hours.

3. Complex autonomous systems

The Internet of Things enables the process of personifying a machine to make decisions as a human would. This calls for, ‘real-time sensing of unpredictable conditions and instantaneous responses guided by automated systems.’ Certain industries have adopted this concept to better the level of performance. The automobile industry has developed their systems to be able to take action in the case of likely collisions. Scientists have the exploring the word of robotics in terms of maintaining facilities and coordination of heavy machinery. [8]

2.6. Benefits of Big Data

2.6.1. Business Efficiency Gains

Big Data provides several benefits regarding the impact on firm revenues and costs, these gains are influenced by several factors including customer intelligence, supply chain management, PQR management and fraud detection.

Customer intelligence entails efficient segmentation and profiting of customers. In this way, production and sales capacity can grow significantly if customers' preferences are identified appropriately. In turn results in customer satisfaction and maximum output are induced through high performance analytics. The social media network aids in enabling businesses to keep track of customer behavior through the online platform. Therefore their attitude and views towards brands can facilitate businesses in product development and direct marketing. Ultimately, the pricing and production function can be regulated on the basis of market trends.

The supply chain process is influenced by demand-driven supply however with the implementation of Big Data; manufacturing functions can infer methods such as JIT and lean delivery processes. Companies can use the information obtained from Big Data analytics to forecast demands more accurately which can entail optimal inventory levels and reduction in expenditure for storage capacity. Businesses can determine supplier performance and make informed decisions to minimize any delays and prevent process interruptions while at the same time improve quality and affect price competitiveness.

Big Data can impact **the performance, quality, and risk management** within a business, as variability in opportunities is sublime. The quality of products can be improved as a result minimizes performance variability which means reducing the time consumed during manufacturing and marketing operations. With minimal disruptions in the production process provides a business with the ability to save significant capital expenditures on machinery and labor expenditure. With improved quality, managers can make swifter decisions in addressing customer matters thereby advocating brand equity of the business. Big Data offers a tactic to businesses is better mitigating risk profits through integration of siloed data and real-time analysis. This enables optimizing financial services in terms of determining investment opportunities which can lower instances of unanticipated losses. Performance management entails operating the business along the concept of maintaining efficient processes. Therefore monitoring performance will manage the degree of transparency and expenditure control within the organization. Integration of the PQR factor through the influence of Big Data results in significantly altering the business' prowess. Big Data provides a skillset that once tapped is capable of providing a company with the task to a resourceful and game-changing schema.

No one enjoys being extorted however from the average Joe to multinational companies even governmental bodies have experienced the unpleasant nature of fraudulent behavior. Big Data can assist in detecting **fraud patterns** in order to adopt new ways to combating these types of fraud. For example, customer intelligence insights can be modeled as 'normal' customer behavior in order to identify inconsistent occurrences which may signify suspicious activity.

2.6.2. Benefits from product innovation

Big Data impacts the research and development activities in regards to increasing operational efficiency as well as the innovation of new products. Increased new product development proposes for uplifted revenue and efficient capital expenditure.

2.6.3. Benefits from business creation

The gains acquired as a result of the usage of Big Data can support the growth of employment amongst small and medium sized businesses. New entrants into the market can experience greater business opportunities through a more refined means of market intelligence. As the knowledge is required to manage Big Data technologies, the employment of data analysts and/or data scientists is required as a result. [9]

2.7. Hello Big Data, Goodbye Traditional BI?

Through the span of history, Big Data has existed or at least the concept behind Big Data enabled organizations to find ways in shaping decision-making processes. With all the data being stored on paper or digital form, experts found a way to organize all this information in order to make the most of what could be retained. With the invention of different database systems, organizations have been able to record information on all aspects which has provided them with the ability to view their strategies with new prospects. However, they can only use the data that they need on a daily basis and as a result record specific datasets. Big Data does not only include those specific datasets, it generalizes all the data collected in every form which may have been stored through several accessible information sources. Big Data can improve the analysis process of data and organizations can start thinking in terms of continuous improvement.

Traditional database systems enhance the productivity and efficiency of processes within an organization to an extent. Management can attempt at correcting any abnormalities that may occur during data processing in order to avoid future occurrences. Whereas, Big Data provides an organization with the ability to react quickly to changing outcomes. Taking for example credit card companies, the marketing department is used to creating models that portray the most likely customer prospects from the information stored in a database warehouse. The processing of data can last for a while. Enabling Big Data practices can improve the way in which the marketing department monitors customer activities. In this way they can keep track of online and offline activities in a faster way wherein they can meet customer requirements and optimize offers.

Relational database management systems are not designed for mass data with a very large number of rows. A row-wise data organization is well suited for online transaction processing (OLTP) but when it comes to analytical tools, such irrelevant data will be read. Pre-processing of data is required before data can be stored. It means high amount of metadata, high storage capacities and slow access to data. The server is required to be powerful enough to maintain the totality of the increasing data. The database servers must be interconnected allowing access to all the data, which means parallel connectivity. This in turn results in decreasing server efficiency which becomes time consuming and expensive. Ultimately, having to maintain loads of data by means of a traditional database warehousing system would create disconnected data silos and analytics which will provide incomprehensible insights.

Below is a table outlining differences between traditional data warehousing analytics and Big Data analytics. [10]

Traditional data warehousing analytics	Big Data analytics
Analysis of data is done readily as information is well understood and in line with business metadata. Most of the data warehouse function upon ETL processes and database constraints which refine the information analyzed.	Most information is unstructured compared to the traditional refinement of any incoming data. More time consuming to scope for the right data however can provide more insight.
Relationships between concerned/similar data sets exist which define the purpose of the system, therefore analysis is focused as per the purpose.	Relationship amongst all the information is not defined however all the data can be linked through means of different data formats.
Row based databases	Columnar databases
Batch oriented therefore time consuming, waiting on other jobs to complete	Real time process and computing. Meant to support decision making at any point in time.
Parallelism is costly	Achievable through commodity hardware and new analytical software i.e. Hadoop, Hives.

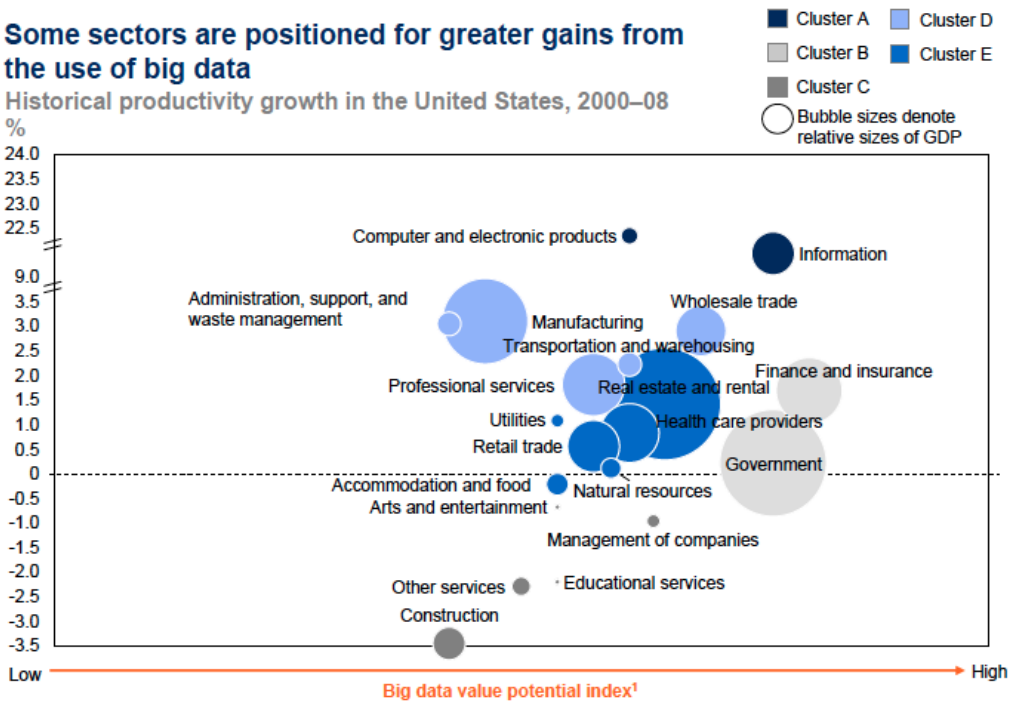
2.8. Industries that benefit from Big Data

Big Data has created value for companies across different industrial sectors. It presents them with opportunity of creating more transparency, effectively segmenting and targeting customers, improve performance levels and limit variation of business processes as well as support human decision making with automated algorithms.

Equens is a payments processor that uses a Big Data approach to prevent fraudulent cards transactions. Every transaction to be processed is compared in real-time to one million previous transactions in order to avoid any suspicious activities with the transactions and associated cards.

Car markers, Toyota, Fiat, and Nissan have cut down the time on product development by 50%. Toyota has been able to eliminate 80 %of the chances of defects occurring prior to building the first physical prototype.

Consumer driven models include companies such as Amazon that use customer data to determine the kind of recommendations to be made to customers. This is based on predictive modeling technique known as collaborative filtering. Another example is Tesco's loyalty program wherein it generates large amounts of data on customer activities which the company analyzes to make informed decisions for promotions and strategic segmentation of customers. [11]



14

The image above illustrates which sectors are set to gain more from the use of Big Data. (earlier referenced in section 2.4)

[14] McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.

As seen in the image above, the technologically driven sector, computer and electronic products and information sectors (Cluster A), experience greater momentum within the economy and benefit more from Big Data. The finance and public sector (insurance and government) can benefit greatly from Big Data granted issues of data governance do not pose at barriers and can be overcome. Compared to other sectors such as construction or education (Cluster C), where Big Data value potential is considerably less, however, sectors such as Retail, Manufacturing etc. (Cluster D & E) are likely to gain fairly from Big Data. [1]

All sectors are capable of adopting Big Data practices however, it depends on if they are able to overcome certain barriers as for some sectors these can create problems. Taking the public sector into consideration particularly education, there is a lack of data-driven potential and available data while with the healthcare sector, there is data available, however the investment in IT is relatively low. Therefore the way in which each sector is able to identify the potential of pursuing Big Data can vary due to different outlooks and ease of data accessibility.

2.9. How should businesses approach Big Data?

"Every organization wants to make the best informed decisions it can, as quickly as it can."[15] It is important a business implicates the right drives to street it along the lines of success. A business should approach the concept of Big Data with the intention to bring greater value. Organizations face the challenge of implementing Big Data practices. For a business to recognize the advantages of Big Data, they must first question whether there is data that can be used to serve certain purposes. Therefore the information sources form which all this data is recorded should be readily available for the business. The organization must be able to apply these sources in such a way that ultimately benefits it as well as all the stakeholders involved in manipulating the data i.e. collecting, handling, integrating, analyzing, acting. Hereafter, the business is to identify ta purpose for which the data may possible provide insight for. It may lead to new opportunities or even new ways to approaching old ideas or projects. The following stage would include simply running a simulation in order to determine any valuable insights.

A business must keep the following in mind when adopting Big Data practices.

- Link data in order to avoid creating new silos of information therefore feeding information from a variety of sources.
- It is usually advisable for a business to determine the problem or purpose for adopting Big Data however a business may also consider simply understanding the relationship it has with data at the point in time.
- With a purpose in mind, it is important to make certain the extent of data life cycles. In other words, the duration for which the information may be retained as new data will continue to increase and old data loses its credibility over a period of time. Besides this, with this retained data a company will have to determine whether it should continue to be stored in its present format i.e. unstructured and degree of accessibility. Essentially data is created, maintained and eventually defected at a certain time. Therefore the data is managed to the point of serving its purpose for the right length of time and then no long used.
- To carry out Big Data analysis, a company is to consider the right tools to be used, as in the software/technology needed to aid in processing the data, moving data, data integrity and analysis. In addition, the talent is required to maintain the data. Data scientists are namely proficient in their manner. With these tools in place, a company can over time gain from any identifiable relevant patterns in the data. [12]

[15] Fujitsu. *The White Book of Big Data: The in business analytics.*

2.10. Building a Big Data Platform

When it comes to designing the platform upon which an organization is to perform data analysis, it is important to keep in mind the kind of technologies and tools needed to deliver value for the business. A well-planned approach can provide a business with the ability to successfully leverage its data. For an organization to consider adopting Big Data practices, the kind of infrastructure requirements needed to enable the integration of unstructured data to enterprise/structured data must be selected according to certain criterion. Of course for every organization, the infrastructure will differ. The requirements in a Big Data infrastructure include 3 phases including data acquisition, organization, and analysis.

2.10.1. Data Acquisition

With the aspect of high velocity and variety of data, it is important to consider that the infrastructure supports a low, predictable latency during the process of capture and execution of queries. Therefore the infrastructure must be compatible enough to handle high transaction volumes as well as support flexible, dynamic data structures. To achieve this, normally NoSQL databases are used to acquire and store Big Data as they support dynamic data structures and are highly scalable. NoSQL systems simply capture all the data without first categorizing and parsing[16] data. Whereas SQL systems entail well-defined structures and impose metadata on the data captured to ensure consistency and validate data types.

Compared to having to design a schema outlining relationships, the system functions using a key to identify the data point, and the content containing the relevant data. With such a simple structure, it is relatively easy to make changes in the storage layer without additional costs involved. NoSQL databases can be interpreted as OLTP databases in reference to traditional BI methodology, as they provide very fast data capture and simple query patterns.

2.10.2. Data Organization

Assuming there is a high volume of data to be processes, organizing the data in its original storage location is more feasible in terms of time saving and costs. This prevents having to move the large sum of data. Therefore the infrastructure is required to be able to manipulate and process the data in its original storage location simultaneously supporting a high throughput, enabling batch processing, and handle a variety of data formats.

Apache Hadoop is a Big Data driven technology that operates along a similar basis wherein it allows large volumes of data to be organized and processed on the original data storage cluster.

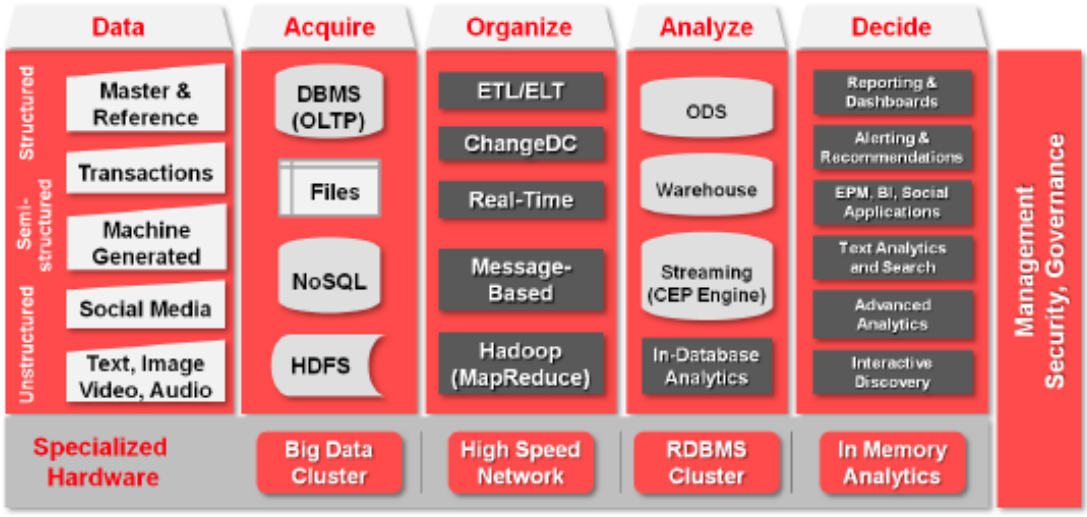
Hadoop Distributed File System (HDFS) and MapReduce programs support the storage and distribution of data across different nodes to generate aggregated results on the same cluster. These aggregated results are then loaded into a relational database management system.

[16] Parsing is the process of analyzing a string of symbols, either in natural language or in computer languages, according to the rules of a formal grammar. A parser is one of the components in an interpreter or compiler that takes input text and builds a data structure giving a structural representation of the input, checking for correct syntax in the process.

2.10.3. Data Analysis

In order to carry out the analysis of data effectively, the infrastructure must support the integration of analytical tools which support statistical analysis and data mining of a large variation of data types. Therefore new insight must be obtainable with the addition of being scalable to the data volume and delivering results with a fast response time. To create a more in depth and 360 degree view on all the data, integrating Big Data and traditional enterprise data can also provide a new outlook on old matters.

"Big Data strategy is centered on the idea of evolving the current enterprise data architecture to incorporate Big Data and deliver business value, leveraging the proven reliability, flexibility, and performance."[17] [3]



18

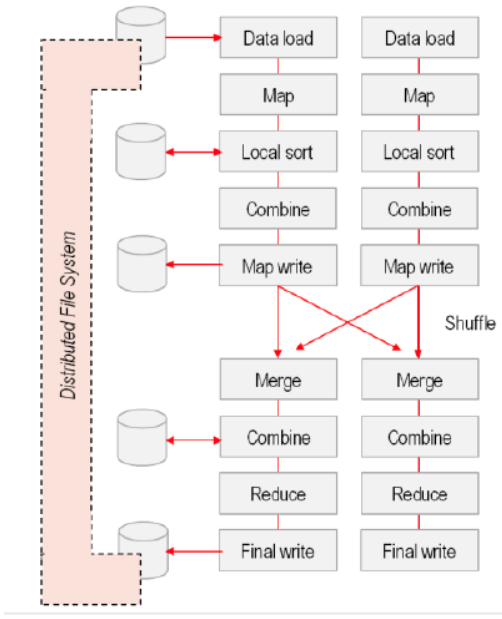
The image above is a depiction of the integration between traditional database system and Big Data database system.

[17] Oracle. Jan 2012. *Big Data for the Enterprise.*
[18] Oracle. August 2012. *Oracle information Architecture: An architect's Guide to Big Data*

2.11. Big Data process flow

When we approach the idea of data analytics, the traditional BI methodology is the process we consider when attempting to meet our objectives. This method of operating includes data collection through internal and historical information accumulated from predefined sources. After which the data is structured according to regulations which are identified within the RDMS system. Therefore data is more or less meant to provide static output, which means that all this information is acquired on the basis on a specifically formatted business model. Insights procured can be helpful to a certain extent however, eventually these become obsolete and companies may be late in their reactions to the market. On the other hand, Big Data analytics methodology functions on the border of proactive and real-time decisions. There is clearly an explosion of data however, if you only make use of 1% of that data available which seems enough in terms of economic feasibility and performance enhancement, the potential behind the remaining 99% is never explored and could mean great financial returns and valuable insights for organizations.

A Big Data platform is based on the concept of parallelization which means the architecture is based on a 'shared nothing' platform in order for there to be a smooth running system across servers. To illustrate this clearly, consider the electric circuit of Christmas lights, if one light bulb stops works, the other lights bulbs are not affected by the break in the circuit, rather they continue to remain lit up. The same concept can be applied in this perspective. Parallelization makes it possible for several servers to remain functioning despite any failures in another server granted the respected data has been replicated across the servers. It also allows for several actions to be executed simultaneously which increases performance level and delivers results immediately.



[19]

For parallelization to be possible, middleware is required which means a certain framework is to be designed in order to support the distribution of data.

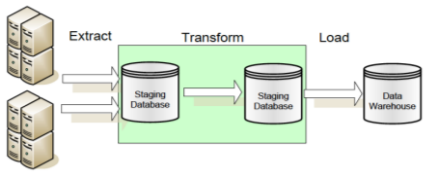
[19] Fujistu. 2013. *Solution Approaches for Big Data.*

First of a distribution system spread across the local storage of a cluster is needed in order to fragment the data divided amongst several nodes. Therefore whenever there is a job request, the system's coordinator partitions the necessary data and distributes it to the server nodes based on the defined rules. Each server is in charge of a certain amount of data and every piece of data is replicated and stored on more than one server. The job at hand is divided into tasks which are distributed to server nodes close to the data that needs to be processed. Taking for instance, the Map Reduce process which 'maps' the tasks to the server nodes and after intermediate output is produced which is tasked with creating the final output through aggregation, also known as 'reduce' tasks. Between the intermediate and final output, the coordinator can sort out the tasks to be assigned to the server nodes. The tasks are parallel to each other therefore functions independently creating a linear environment. (As compute jobs are sent to the data and not data to the compute jobs, I/O traffic is reduced by a great deal.)

In regards to data integration, Big Data analytics processes data through the approach of Extract Load Transform (ELT). This means data is extracted from several sources and refined (integrity and business rules can be applied), hereafter it is loaded into the data warehouse environment almost immediately. Within the warehouse, the extraction and loading process are isolated from the transformation process. Data is transformed into the kind of specific output format indicated. By making the loading process independent of the transformation process, data can be optimized whenever a new job request is made known. Also separating the processes enables the project to be divided into small chunks which entails better predictability and manageability with reduced risks and costs.

Compared to the traditional method of extracting data from data sources and then transforming it to a format which has already been stated to provide the kind of output needed, the transformed data is loaded into the data warehouse ready for presentation. Processing data in this way works basically from the end backwards. The kind of output required is pre-determined therefore the data is specifically extracted and is transformed according to the rules designed in order to produce the desired outcome.

In summary, Big Data follows a process which allows data to be continuously extracted and loaded into the data warehouse without having to go through the process of being transformed immediately for only a specific output. Data is generated through different data sources. Then data is extracted from the data sources and cleaned. If the data structure needed for analytics is known it can then be transformed into a more usable form before loading into the data store. Data can then be analyzed by submitting relevant queries for further visualization. [13]



[20]

[20] Data Academy. 2008. *ETL vs. ELT.*

2.12. Big Data Technologies & Techniques

2.12.1 Tools & Technologies

Currently there are several Big Data products on the market which have been identified as providing prominent development in the aggregation, manipulation, management and analysis of Big Datasets. These technologies serve different purposes in relation to making data integration and analysis possible. Below is a table illustrating these technologies. [1][14]

Name	Type	Description	Distribution
Big Table	Columnar DB	Distributed database built on the GFS	Proprietary
Google File System	Distributed File System	Google's core data storage system	Proprietary
Cassandra	Columnar DB	Based on Bigtable and Dynamo	Open source
Dynamo	Key value DB	Amazon's core data storage system	Proprietary
Hbase	Columnar DB	Based on Big Table. Non relational Database	Open Source
MapReduce	Computation framework	Programming framework for processing huge datasets on a distributed system	Open Source
S3	Key value DB	Simple Storage System	Closed source
MongoDB	Document DB	Document oriented, scalable and fast, written in C++	Open source
CouchDB	Document DB	Document oriented database	Open Source
Hadoop	Framework	Software framework for processing huge datasets on certain kinds of problems on a distributed system	Open Source
Pentaho	BI suite	Integrated reporting, dashboard, data mining, workflow and ETL capabilities	Open Source
Neo4J	Graph DB	Stores data in graphs	Open source/proprietary
Lucene	Library	Indexing and search library used by NoSQL database system	Open source
R		Software environment for statistical computing and graphics	Open source

2.12.2 Techniques

When pursuing Big Data, there are several ways in which an organization may look to deriving value, particularly through a relatively flexible approach. By leveraging Big Data in perspective to the organization's context, strategy and capability, these techniques can harness significant output. Of course, these techniques are not only applicable to large data sets as they can also be applied to smaller volumes of data. Below is a list of techniques which may be particularly helpful through common interest of several different organizations. [1]

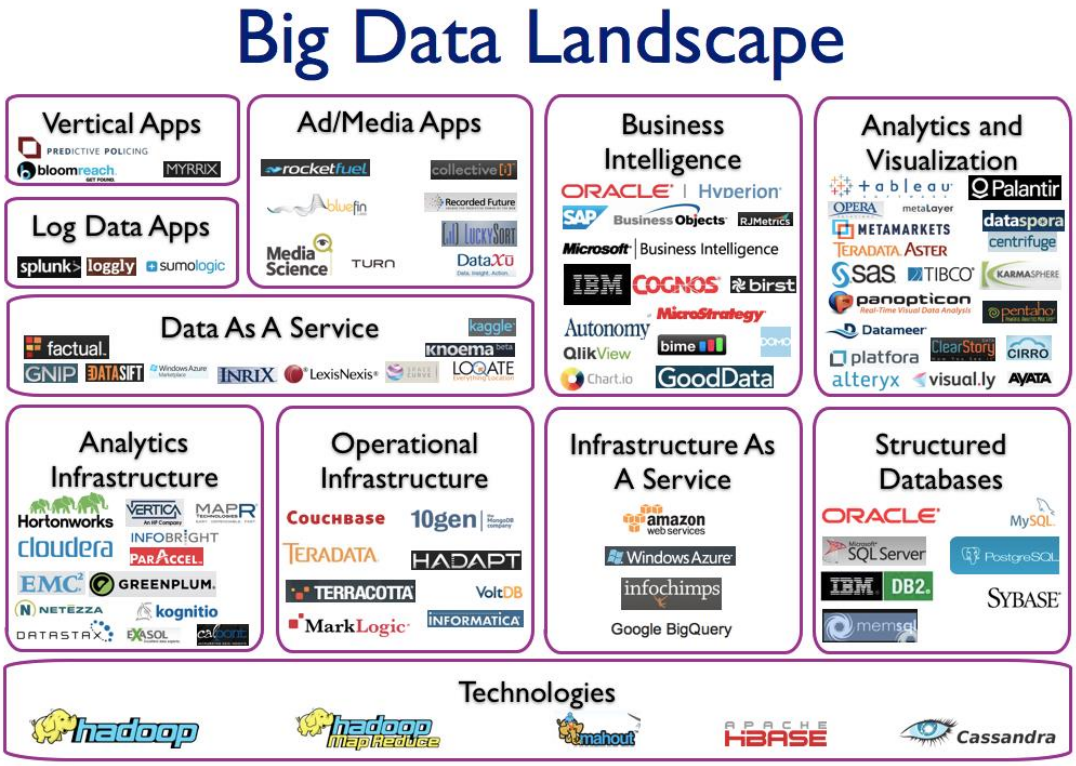
- A/B testing: A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate.
- Cluster analysis: A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance. An example of cluster analysis is segmenting consumers into self-similar groups for targeted marketing.
- Crowdsourcing: A technique for collecting data submitted by a large group of people or community (i.e., the "crowd") through an open call, usually through networked media such as the Web.
- Data fusion and data integration: A set of techniques that integrate and analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data.
- Data mining: A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management.
- Machine learning: A subspecialty of computer science (within a field historically called "artificial intelligence") concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data.
- Network analysis: A set of techniques used to characterize relationships among discrete nodes in a graph or a network.
- Optimization: A portfolio of numerical techniques used to redesign complex systems and processes to improve their performance according to one or more objective measures.
- Pattern recognition: A set of machine learning techniques that assign some sort of output value (or *label*) to a given input value (or *instance*) according to a specific algorithm.
- Predictive modeling: A set of techniques in which a mathematical model is created or chosen to best predict the probability of an outcome.
- Regression: A set of statistical techniques to determine how the value of the dependent variable changes when one or more independent variables is modified.
- Spatial analysis: A set of techniques, some applied from statistics, which analyze the topological, geometric, or geographic properties encoded in a data set.
- Statistics: The science of the collection, organization, and interpretation of data, including the design of surveys and experiments. Statistical techniques are often used to make judgments about what relationships between variables could have occurred by chance (the

"null hypothesis"), and what relationships between variables likely result from some kind of underlying causal relationship.

- Supervised learning: The set of machine learning techniques that infer a function or relationship from a set of training data.
- Simulation. Modeling the behavior of complex systems, often used for forecasting, predicting and scenario planning.
- Time series analysis: Set of techniques from both statistics and signal processing for analyzing sequences of data points, representing values at successive times, to extract meaningful characteristics from the data.
- Unsupervised learning: A set of machine learning techniques that finds hidden structure in unlabeled data.
- Visualization: Techniques used for creating images, diagrams, or animations to communicate, understand, and improve the results of Big Data analyses. For example, Tag cloud which helps the reader quickly perceive the most salient concepts in a large body of text. Clustergram is used for cluster analysis displaying how individual members of a dataset are assigned to clusters as the number of clusters increases. History flow charts the evolution of a document as it is edited by multiple contributing authors. Another visualization technique is one that depicts spatial information flows.[21]

[21] These techniques have been chosen according to how useful they could be in regards to Big Data processing. Each technique has been defined as mentioned in the McKinsey & Company MGI White paper.

2.12.3 Overview Big Data Landscape



22

The image above displays an overview of the virtual landscape of software applications regarding Big Data implementation. The technologies concerned with the extracting and loading of data include Hadoop, an open source data processing framework, and MapReduce, a programming framework supporting the distribution of data tasks, and HDFS, Hadoop Distributed File System. In collaboration with the basic low level infrastructure, additional BI tools can be added to the system to create a more concise and constructive database system.

[22] "Big Data Vendor Landscape - Rose Business Technologies." *Rose Business Technologies*. 27 June 2012. Web. <http://www.rosebt.com/1/post/2012/06/big-data-vendor-landscape.html>.

2.14. Big Data Use Cases

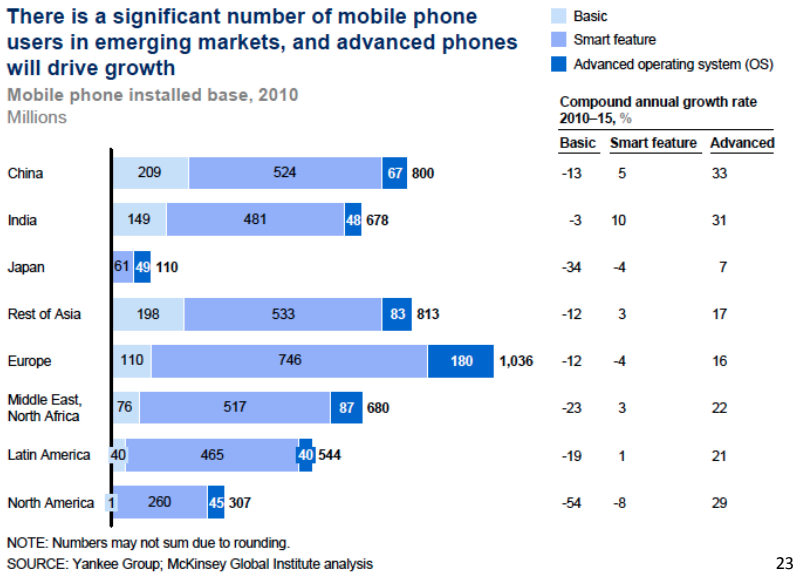
Use Case Personal data collection

When walking down the street, probably one in every 5 persons, is using their phones with the likely choice it is a smartphone. The age of technology has created a median through which both the concerned organization and the end user can acquire significant value. There is an explosion of data happening of which some of this data is collected as only a few bytes of a person's location details. Technologies such as GPS has made is more convenient for us to locate a device that could only be a few meters away. With the ease of accessibility to the volume of personal location data, it is not only beneficial or prominent to a single sector, rather it can provide value to several different sectors including, telecom, media and retail. Value creation is unavoidable. According to MGI research more than $100 billion in revenue can be generated by service providers and as much as $700 billion in value to the consumer and business end users.

Locating someone on a grid map has become more or less convenient. Earlier an individual's credit / debit card payments based on POS terminals typically provided as personal location identifiers sources. However with the increasing number of smartphones being used, triangulating a person's whereabouts through the use of cell tower signals has triggered a moment in which several services have emerged in leveraging the data for public use. For example, providing users with the ability to find friends or locate shopping stores in the vicinity.

Smartphones are equipped with GPS capabilities which triangulate the location within about 15 meters using a constellation of orbiting satellites. In addition Wi-Fi networking capabilities also act as a source for determining locations. Besides these smartphones technologies in play, companies such as UK Path Intelligence based in the UK, monitor signals sent by individual mobile phone to track foot traffic within malls and amusement parks.

The global pool of generated personal locations data estimated to nearly 1 PB in the year 2009. It is believed to grow by 20% annually. Currently Asia generates the most amount of personal location data due to excess use of mobile devices. "Growth in the use of mobile telephones is set to grow rapidly in developing markets."



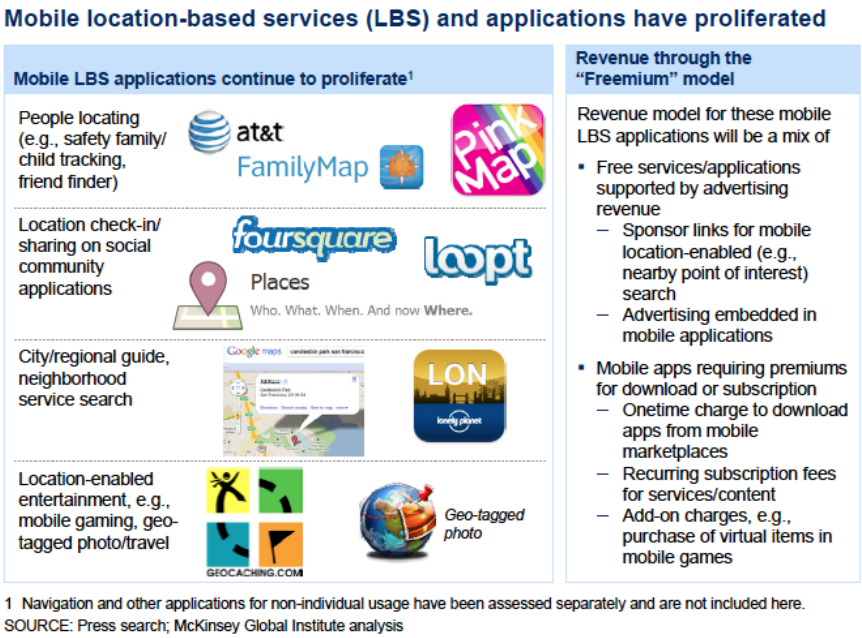
As can be seen, the rise of personal location data is increasing significantly as there is a rise of devices enabling navigational technology. McKinsey & Company has identified 3 main categories of applications of personal location data. These are as follows:

- Location based application services for individuals
- Organizational use of individual personal location data
- Macro level use of aggregate location data

Location based application services for individuals

1. Smart routing:
   Based on real- time traffic info. Provide end users with up to date info on points of interest and weather conditions. Provide drivers with suggestions of routes to take based on data congestion activity. Penetration of smartphones and other navigation devices with GPS capabilities will increase the use of smart routing. Digital map data must be kept up to date for smart routing to be effective (which is difficult in emerging markets)
2. Automotive telematics:
   GPS and telematics enable a source of services concerning personal safety and monitoring.
   For example, GM's Onstar can provide the driver with information / alerts as to when they need repairs and can located vehicles during emergencies by collecting real-time vehicle location and diagnostics info a central monitoring site.
3. Mobile phone location based services:
   Provide services to end users on their mobile devices, including safety related apps or for finding points of interest.
   Value generated in this way could accrue to $80 billion to mobile location based service providers by 2010.

[23] McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.



Organizational use of individual personal location data

1. Geo targeted advertising:
   Create value in providing segmented only to consumers i.e. Personalized ads for a few shop compared to traditional forms of advertising. This provides higher relevance to consumer of the time of making a purchase decision.
2. Electronic toll collection:
   GPS enabled mobile phones provide user with the ability to locate toll booths and pay the toll at once compared to having to use separate transponder devices.
3. Insurance Pricing:
   Combining personal location devices and vehicle telematics offers insurers with the ability to determine price risk based on the individual behaviors.
   Provide incentive for individuals to be more careful with the way they drive.
   Insurers can offer real-time alerts on traffic and weather conditions, high risk parking areas, and changing speed limits.
4. Emergency response:
   Enable emergency service dispatch to quickly identify the location of a person during an emergency.

Macro level use of aggregation location data

1. Urban planning:
   Help makes decision on road and mass-transit constructions, mitigation of traffic, congestion, and any high- density development.

[24] McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.

Agencies can monitor information on high and low peak traffic hotspots, patterns in transit use, and shopping trends which may in time lead to the cutting down on congestion and emission of pollution

2. Retail business intelligence:
   Provide retailers with the ability e to analyze shopping patterns through aggregation of foot traffic density and speed. This identifies where shopper slow down and speed up in response to retail campaigns. Leads to making business decisions on store layout, product development and merchandising.

Personal location devices have the potential to influence the value generated through the aggregation of information attained by means of mobile technologies. It is estimated to create created value of $100 billion for service providers only. Of course there are barriers which could challenge the growth of value creation. There are the issues that business and policy makers face in terms of privacy and security concerns and well as technological obstacles in terms of error detecting etc. Despite this, there is more of an opportunity available. "Creativity and innovation will shift the value potential upward, and a long tail of specialized apps will combine to offer substantial total additional value." [25]

Use Case Public Sector

Within the past few years, governments across the globe have experience the impact of the recession. Many governments have been trying to ameliorate the impact it has had on the public deficit. According to MGI research, governments can recuperate better by influencing their productivity levels. It is speculative whether Big Data is the answer to achieving a higher level of productivity. Europe's public sector can potentially reduce costs in administrative activities upto 20% and create new value though subsequent gains.

The data generation by the public sector is usually in textual and numerical form. 90% of this data is generated in digital form by means of e-government initiatives. While this may be, public sector agencies limit data accessibility across the public domain promoting inconsistencies and lack of transparency. However, there are certain Big Data contributing factors which can administer the productivity of the public sector. These factors include:

- Creating transparency
- Exposure of variation in performance of processes
- Customization through segmentation
- Replacing human decision making with automated algorithms
- Innovation of new business models

Creating transparency

Efficiency can be improved through the ease of accessibility of large public sector databases by the government for external and internal stakeholders. Government agencies can regulate the

[25] McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.

collection of large volumes of data across organization silos by avoiding the re-entry of information by separate public entities. This can reduce errors and speed up the processing time. Governments have adopted the principles of 'open data' which entails making raw government databases available to the public i.e. data.gove.uk.

Exposure of variation in performance of processes

Productivity across government agencies external and internal can be improved with less variation in total performance. Performance dashboards that display operational and financial data allow the agency to measure variability of performance amongst separate departments. This can lead to reevaluation of approaches to improve productivity. For example, tax agencies use integrated monthly scorecards that measure revenue collected, track customer satisfaction scores, staff engagement scores, and feedback from the public and agency heads. This method may also act as a benchmark technique among foreign branches and internal agencies. Providing consumers with the access to information service may create incentive for providers to improve their performance.

Customization through segmentation

"Segmentation and tailoring government services to individuals and population cohorts can increase effectiveness, efficiency, and citizen satisfaction." Tax agencies can segment individual and business taxpayers in order to monitor collection activities. Effectively segmenting population can narrow the gap between actual and potential productivity.

Replacing human decision making with automated algorithms

Applications that use automated algorithms to analyze data help find any anomalies in order to prevent errors or cases of fraud. In terms of the public sector, inconsistencies in payments of tax collections may flag any further auditing required.

Innovation of new business models

Through collaboration between the public and private sectors could lead to new value creation which may entail improving services, management of practices, and develop new & existing programs. With Big Data available at the fingertips of the government, performance levels can be improved.

From the five factors or as referred to as levers by MGI, the overall concept behind applying Big Data within the public sector will potential lead to boosting overall productivity. According to MGI, these Big Data levels can provide benefits in terms of monetary valued gains including improved allocation of funding into programs, higher quality services.

In conclusion, by providing services targeting to citizens and businesses can reduce time consumption, improve public trust and accountability in the public sector. Increased transparency, implementing measurement tools i.e. dashboards, and taking the initiative to improve accuracy of databases to avoid any fraudulent information can offer the opportunity to making better decisions which may affect productivity positively within the public sector. [1]

2.13. Summary

Big Data is believed to bring forth new opportunities to enterprises across different industrial sectors. It is no longer only about the SAP or Microsoft CRM traditional ways of collecting data, rather it is all the information collected from more nontraditional ways i.e. social media, sensors, video, email etc. If we know it or not, the fact of the matter is that the data accumulated in this way actually provides a multidimensional view into several undiscovered queries. It may seem as it may not be worthwhile to explore this untapped array of data, on the contrary taking advantage of the data available can enable organizations to make more in depth decisions. There is no need to get bid adieu to the traditional forms of data aggregation, in fact combining different forms of data to provide a bigger picture can provide substantial knowledge about the company's customers, internal operations, and other external factors.

McKinsey Global Institute has estimated that data is set to grow 40% per year between 2009 and 2020. At this rate, the nature of data could be characterized as ubiquitous. Such an explosion incurs the opportunity to improve business processes, consumer targeting, new service and product development etc., we are only faced with the hesitation to move forward. Big Data is easily misconceived to be appropriate for businesses that acquire a vast amount of data through means of different sources however, the "big" can also be perceived as indefinite, which means that as long as the data is there regardless of its volume, there is the potential to gain insights from. Technologically based companies i.e. web companies, are early adopters of Big Data as they come across a large volume of data on a daily basis. Therefore they analyze this data which includes visitors' activities, to better cater to their customers' needs and target their interests. However, Big Data is not only meant for them, it is even suitable for businesses within the telecommunications, retail, financial institutions, even the healthcare industrial sectors. As the Internet of Things has become inherent with installment of sensors within several devices, the healthcare sector can benefit from the data collected from such sources to customize patient treatments or even possibly discover new ailments to diseases. While this is all still very speculative in the eyes of many, there are businesses that have taken the initiative to pursue Big Data. The one challenge these businesses may face is the adoption of Big Data technology to their framework.

Like the traditional method to data analytics, Big Data entail the collection, refinement, organization, and analysis process. To enable this process, new technologies have been developed to aid in managing the vast quantities of the different types of data in a more feasible and cost-effective manner. Certain technologies which are popular within the market currently are Hadoop which is an open source platform for the consolidation and transformation of large volumes of data as well as MapReduce, a programming framework supporting the processing of large data sets on distributed nodes to generate aggregated results. Key enablers for analyzing Big Data are tools for statistical and advanced analysis. These tools must be able to work with distributed data to perform analysis regardless of where the data resides, to scale as Big Data volumes grow, to deliver response times driven by changes in behavior, and to automate decisions based on analytical models.

Imagining the bigger picture always seems far-fetched compared to what it really entails. Big Data offers companies with the opportunity to explore the dynamics of their business in turn improving processes and providing a competitive advantage. The way companies leverage the data available to them will determine the kinds of insights which may become apparent. Big Data does not

necessarily include taking on a large project to see the effect, rather businesses should pilot projects to test the waters in order to get familiar with the techniques involved. Therefore businesses can approach Big Data through different alternatives including combining traditional forms of data analytics to acquire both financial and operation gains.

Chapter 3. Field Research

"Today it is a business priority, given its ability to profoundly affect commerce in the globally integrated economy." Big Data has been sought by companies with the incentive to better understand the nature of data generation. From social media analytics to real-time data generation, organizations are exploring ways into acquiring new value creation, many organizations view the concept of Big Data simply by the volume of data. This may be the case in some instances however; Big Data characterizes the ability to efficiently extracting insights in its most raw form. So how do organizations define Big Data? A survey conducted by IBM showed that the participants viewed Big Data as an opportunity to a greater scope of information compared to the minority that viewed it as the latest buzzword. It continues to mention that nearly 50% of the respondents have been planning Big Data initiatives and 28% are already in the nascent stages of implementing Big Data activities. Such efforts do show the potential behind Big Data analytics especially when providing businesses with the ability to target strategies with a wider perspective.

Many companies intend to address customer-centric objectives where they are committed to improving the overall end user experience. Therefore with the vast array of data sources available i.e. transactions, social media, loyalty cards, companies are capable of effectively targeting specific demographics. Besides focusing on consumer targeting, companies also look to adopt Big Data practices with the incentive to optimize operational functions within the organization.

The way in which organizations approach Big Data practices vary as the kind of information available to each is significantly different, especially across industries. Most may explore the potential of internal data extracted from customer transactions, events, emails and structured data. For an organization to fully realize the potential of Big Data, they must develop a process for exploring the dimensions, engaging in the implementation of Big Data practices followed by the project execution.

Big Data is no longer a theorized hypothesis rather it has transitioned into a dogmatic approach for data analytics application. To understand the impact of Big Data within the business, further research was conducted into the field of data analytics. Experts familiar with world of Big Data were interviewed to gain a better understanding of the credibility Big Data has within the business domain. As well as companies that have tackled Big Data and implemented its practices within their current database infrastructure, were interviewed to understand how they approached the concept and their reasons for doing so.

Experts from Ordina, Cognizant, Andarr and Bearing Point were interviewed during this process, and Managing Directors from Vektis and Bol.com were interviewed as part of the company research. Interviews were conducted according to a set of questions formulated regarding certain issues and points of interest on Big Data in Business.

Topics Discussed

Interviews were conducted on the basis of literature research on Big Data. Issues such as the impact Big Data has on the economy, ownership of data, the approach businesses should adopt, the tools and technologies, the transition from traditional means to real time analysis, each expert's opinion on the future and concept behind Big Data, the effect on the data supply chain, advantages of Big Data practices. Regarding the company interviews, the objective was to gain perspective on the

approaches taken by each company as a Big Data practice. The questions asked during the interviews can be found in Appendix A. As per the company interviews, each company was asked the same set of questions these particularly being:

- Reason(s) for considering a Big Data approach
- The way in which implementation of Big Data was approached
- Concerns when considering Big Data as a solution
- Drivers enabling Big Data practice
- The Big Data Infrastructure
- Outlook on the future of Big Data

Chapter 3.1. Expert Findings

Ordina

Interviewee: Sander van Kleef

Position: Business Intelligence Consultant

Sander believes Big Data is simply a part of the hype as the word has caught on this new technological phenomenon. However, the power of data is undeniable regardless of the sum of data accumulated. During consulting sessions with businesses, Big Data is not exclusively discussed. It may not seem obvious at the moment however, Big Data will eventually show more promise down the road. There will be a need for the concept behind Big Data. Many businesses question the fact of where and how can they use Big Data in a way that is suitable for their purposes. The more data the more time consuming it may become which could make it seem unappealing. Consulting a business entails investigating their business needs; Big Data is not the primary solution to the needs outlined but rather traditional tactics to Business Intelligence are applied. Compared to 50 years ago, the digital age has progressed the way in which we process information. It is more organized and with the intention to improve internal processes through BI tools including internal database management for conversion rate purposes. The main advantages of Big Data include the ability to maintain webshops efficiently in perspective of intelligent decision making. Storing information on customer activities will help the business to identify the customer's preferences and cater effectively in order to create a relationship.

Businesses may be hesitant to adopt Big Data practices due to the lack of talent in the field and especially monetary issues that may be faced. For a business to be able to gain value by means of Big Data analytics is through determining the purpose/goal they want to achieve. Big Data can mean starting small and eventually setting up for a higher parameter. There are several tools which enable processing data in an effective way in order to understand any form of data. Big Data can be more definitively described from a technical perspective. Big Data is ready as part of a technical solution, basically the relational database system has been improved gradually as datawarehouses require more space to maintain the growth of data. From a business perspective, value is the most important, the opportunities must be clear if the technology is applied. The concept behind Big Data should be applied alongside a traditional database system instead of only focusing on the idea of the 3V's (volume, velocity, variety). Simply because there is a large volume data should not necessarily imply the need for Big Data practices rather considering traditional BI initially to solve the problem should be considered before going ahead with Big Data as the solution.

Cognizant

Gwellyn Daandels

Position: Director Enterprise Information Management Benelux

Big Data from the beginning is more valuable to IT however IT facilitates a business model. A business does not give much thought to Big Data due to budgeting issues. In terms of costs, data governance becomes a matter of focus. In order for a business to add value, a few questions have to be addressed about Big Data. What is happening, what should be being done, how is it happening,

why is it happening? Big Data can be identified as a buzzword however mustn't assume that all the data is stored simply in one location rather it is available by different means, internal, machine, etc. data. Big Data has been able to build interest through economic outlook. As there is more revenue pressure, consumers buy less and are no longer brand loyal rather price driven. Therefore businesses are more purpose leveled in order to keep consumers interested. The action distance analogy comes into play in that they must be able to react to customers better than competitors. For examples, Netflix improved their analytical algorithm and agility in order to make decisions effectively and ensure implementation. Big Data is the frontier to competition. Big Data presents the opportunity of gaining momentum ahead of competitors however, in the EU Big Data is given less priority compared to the companies in the US. Europe could harmonize data privacy laws, enable education systems with knowledge to new technological advancements as there is the need for innovators. Therefore when inquiring how influential Big Data is in the Netherlands, it is safe to say that there is a lack of creativity amongst businesses.

Companies have been making the general move into Big Data at an agile pace. They are equipped with the means which allows them to experiment more often, analyze data effectively in order to find a pattern. Since ERP, certain companies have become inflexible in their ways as in leveraging their system becomes a guessing game. Therefore viewing traditional methods to information management as inadequate brings about much skepticism to viewing Big Data as the new method in data analytics.

Big Data brings about the concept of next generation data supply chain wherein, costs of maintenance and technology are relatively cheaper, processes are optimized in order to improve products and services for customers. Driving Big Data analytics is a task which requires companies to make a difference for customers rather than on just making a difference to the products. As currently the public sector has been spending less and more revenue pressure, companies have to realize that consumers have become price sensitive and less brand-loyal, which means delaying purchase decisions. To rectify this, adopting Big Data analytics and agile movement will lead to the revolution in data analysis. Brainstorming ways in which the data available can be used could improve output for information service providers for example. If their infrastructure is not scalable and flexible enough to handle the data then this becomes a sort of bottleneck. Wherein if it is inexpensive and doesn't work could influence optimization efforts and any further expenditure. Information service providers have to address the concept of data governance, which means deciding on the manner in which the data could be used without any objection by the public.

Big Data in the future can be characterized as more of a commodity therefore with time, it will be used more evidently and in a user friendly manner. For Big Data to get to this stage it may take approximately 3-4 years.

Andarr

Eric van Tol

Position: Director at Andarr

At the moment there is so much data that mainly only the structured form of data is primarily implemented compared to the other forms of available data. If we consider different sources of data types it could change the way we work. Raw data plus the technology capable of handling this data is the tipping point to the Big Data epidemic. All the data available may not necessarily be used however we should view it as an opportunity through which trends could be set. MIT created an algorithm which predicted 95% of the time what the trending topic would be in real time through analysis of patterns in tweets. Big Data has become more of a social aspect compared to being a more technical aspect. The availability of this data and the technology behind its implementation is cheaper and readily available. Despite this aspect, businesses are hesitant in moving forward with this acquisition whereas larger companies with sufficient budget use it as part of tactical measure for a competitive edge. The main advantage to Big Data analytics is the aspect of decision making which entails understanding the story being told by the data. As everything is recorded nowadays, analyzing the data can assist a business in retrieving valuable insights. With the continuous array of data growing by the second, the concept of data analytics will be around for a long time and as a result people will become more in tuned and clever in using the data. Businesses within the US and Europe differ in their approaches to data analytics or rather their pace differs. However, surprisingly enough according to Google trends companies within the Benelux region have shown significant interest into the matter of Big Data, which shows the potential data analytics has on business decisions.

Experimentation with different Big Data technologies should be an endeavor businesses should look more into. The framework to determining the kind of Big Data technologies to pursue can be categorized into 3 parts these being the hardware, data management, and analysis and visualization. For every business it is important to keep in mind the purpose they hope to achieve thereby determine a certain framework which may guide them in choosing the right tools.

In conclusion, Big Data presents a platform for a deeper analysis into the dimensions of the decision process. Having the want to make use of this data in a certain way and the reason to using the data followed by an in depth analysis will lead to making better decisions.

Bearing Point

Pieter Stel

Position: Senior Manager of Digital/Customer Management

Big Data has been around for years, we have always had large amount of data stored in all forms before the digital age. The concept of Big Data has only become more evident right now however even if we weren't aware of it, data analysis has been a part of history. Big Data as defined by IBM and other Big Data advocates is clearly identified through the 4 V's (volume, velocity, variety, veracity). Through the eyes of the public, Big Data is recognized through the volume of data available. The real challenge is connecting all the sources of information in an intelligent and speedy

way to provide the opportunity with the immediate need for any necessary information. Big Data is currently not a priority amongst businesses due the culture as not everyone is aware of its value and continue to being ignorant in addition to the current economic crisis which discourages investing and businesses becoming prudential with their finances. Accenture conducted research on the comparison of companies that do invest or apply data analytics to those that don't. It was seen that companies that did do analytics recovered faster from the economic crisis in 2008. The IT sector has hyped the benefit behind Big Data therefore the idea to have a lot of storage has been conceptualized. This idea should be reformed into something more appealing by which even smaller companies may find interest in investing and potentially benefit from.

A business can determine the value of their data through control testing wherein they may compare between 2 different states of data. At the same time it is important to include the creative side of thinking when determining the kind of information to use as well as a support and knowledge system. Big Data could be viewed as part of data driven business optimization initiative. Basically approaching data analytics includes 2 main parts including the decisions to be made and the execution of these decisions into actions. Based on strategy/objective, there comes a certain process of data driven optimization.

Data Collection -> Data Analytics -> Business Decisions -> Execution -> Performance and Evaluation. In order for businesses to retain profit from Big Data practices it is important for them to recognize the business value to be attained through means of revenue increase and possible cost reductions which may ultimately result in top line growth and lower cost respectively. There is a difference between strategic data and unidentifiable data therefore the value of this data is better determined through trial and error in order to become more effective in its purpose.

For every business it is initially difficult to identify the purpose or rather the objective to driving data analytics however there are a few factors to consider in this case, these being the personal belief of the possibilities that come with Big Data analytics, secondly the creativity elements comes into play with analytical power as the 3rd factor. These factors are important to consider when considering Big Data analytics as an opportunity.

Big Data in the US and in Europe is being applied at different rates. According to research conducted by Gartner, Big Data is the main driver for job generation in the US. However, this could be speculative within Europe. There may be job creation however more in the data analytics field, which may result in the shift of jobs. Big Data in the future will stay around. It is a hype at the moment however, in the short term it will be overestimated while long term it will be underestimated.

Summary

The research conducted concludes several similarities in the way Big Data Analytics is perceived. The experts interviewed during the process were business professionals familiar with the concept of Business Intelligence. Each expert provided substantial information on their individual opinions on the revolution of data analytics. As each expert was approached with a variation of questions, it is suffice to say the general opinion of Big Data is that the theory behind the phenomenon has been floating around for several years through history. Before the age of technology and the rampant incline to the attention of data aggregation, we have been collecting information in all forms online and offline. The influx of data accumulated on a daily basis is capable of providing a bigger picture. While the common conception was that Big Data is the buzzword across industries, the technology enabling it can provide substantial gains for organizations. Every organization approaches data analysis with an objective in mind; however the way they go about executing this differs depending on the kind of data sources and the technologies suitable for further data processing.

The timeline for Big Data to become as described appropriately by Gwellyn Daandels, a commodity, was more or less within 5 years. Big Data should be able to be applied by several businesses in the process and possibly be able to aid in the reforming of the economy. As developed economies are currently experiencing declines in consumer expenditure and an increase in public deficits, businesses are hesitant to making the step into unfamiliar territory. However, it is important for organizations to realize the general adoption of Big Data does not necessarily incur high costs compared to traditional methods to data analysis. Technologies such as Hadoop have provided companies with open source framework which allows them to perform Big Data analysis more economically.

Usually an organization would identify the objective they want to achieve through traditional BI means after which they procure the required data from internal information sources to only be formatted in a structured form. Big Data on the other hand functions on the basis of collecting different types of data, unstructured, semi-structured, etc., which is then analyzed to provide further insights into possible business strategies. Big Data provides the opportunity for initiatives.

Below is a table summarizing the outlooks of each expert on Big Data including certain similarities and differences on the concept.

Factors	Ordina	Andarr	Bearing Point	Cognizant
Definition	3 V's	It has become a part of the social aspect, however has the potential to aid in decision-making situations.	3 V's	The frontier to competition. Data is not only available in one location anymore.
Conviction	Big Data has potential however, require more evidence to convince companies of its credibility as "Big Data" can be misinterpreted as Big Data can also mean small data.	Confident in the evolution of data analytics and believes in searching for the answering in the vast amount of data available	Assertive of the fact that Big Data is credible endeavor for businesses however, the process of data analytics will build up to the use of Big Data	Confident in the future of Big Data, views it as a commodity for the future of doing business.
The hype	The hype doesn't make it credible enough. It is only a buzzword for now.	It has been hyped by the big companies like IBM and Accenture, therefore there is a strength in the buzzword		
Future Outlook	Growth of data is undeniable and will gain credible value even if not labeled as Big Data	Will gain momentum as it has potential to aid businesses	It will gain momentum with time as over the long term it will underestimated	Will become a commodity

Chapter 3.2. In-company findings

Vektis

Vektis is a company that provides information based on healthcare transactions.

In other words it stores information obtained from several health institutions with details on individual healthcare history within the Netherlands. Their role influences the way in which insurance companies and relevant government bodies use certain information to make decisions. Because Vektis acquires as much data from different sources it is important for them to receive it in a structured format. As allowing any inconsistent data to be entered into the system will entail further time consumption for rectification purposed. Vektis deals with several different kinds of organization healthcare entities including practitioners, universities, ivms, insurance companies, and the government. They provide services at the lowest level. To provide credible information, it is important to understand the way in which the business functions and the link between processes. Therefore making connections from the information stored to be comprehensible.

Vektis has not pushed the boundary on the concept of Big Data as welcoming any unstructured data will provide an incomprehensive picture. A Big Data solution for Vektis would be to anonymize the data to a certain level so the information may be used for more means; however once again making the connection becomes a task.

Currently, Vektis has adopted Netezza, an IBM developed database software technology, which has enabled the company to improves its overall performance and its levels of data processing power. Instead of opting to expand their infrastructure which would entail much expenditure, Netezza provided them with three main advantages:

- More database administrators
- Indexes
- Smart programming

Before installing Netezza, the company performed several pre-pilot studies to measure the system's feasibility.

Vektis is capable of providing deeper insights to healthcare institutions from the information collected. The chairman of Vektis believes the Netherlands specifically will get to the point of acquisitioning Big Data analytics within a time span of five to ten years. However at this time for Vektis, if the company were responsible for data aggregation of different data types, may cause some chaos in turn resulting in possible consequences.

Bol.com

Bol.com first started off as a book reseller 15 years ago. It has now become a platform for resellers to sell different categories of products. It is also available on different technological platforms.

The company deals with data mainly concerning consumer transactions. Therefore they analyze the consumer's behavior during the time they visit the website. The company's current database architecture includes an oracle database system to deal with transactional data, a Hadoop file

system, and a MapReduce algorithm. Currently, the Oracle database system supports interactive queries for traditional data marts and Hadoop is not used for the same function, but rather for analytical processing, it is a batch-oriented solution. The company is looking to shift from depending on Oracle functionality by implementing possible tools/technologies to the architecture.

- Install HBase, which supports RAM and makes interactive queries possible on Hadoop.
- Install Hives, which would add an SQL and layers on the HDFS however; it does not provide interactive functionality.
- ETL tool wherein the performance would improve and log in with pricing would be reduced.

Bol.com adopted the practice of Hadoop to their system because they are now able opt effectively process over a 100 days' worth of log history compared to the capacity of handling one day of log history. By increasing the number of days of log history they are able to provide more products for consumers, which increase the conversion rate. By being able to process over 50 million clicks in a few hours has allowed them to expand the horizon of services to customers. Taking into consideration the concept behind Moore's Law, the speed of data being generated is happening at a much faster rate. The machine initially used in the system was not enough to handle as much data however, with the Hadoop cluster, adding more machines to the system increases processing powers, especially in parallel. This step has allowed the company to operate linearly and the approach is horizontally scalable. The infrastructure is relatively simple as their data sources are extensive. The infrastructure includes 10 machines within the cluster with the capacity of processing 700 days of log files. The company intends to add more machines to the cluster and expand processing capacity of data, which would include also tracking the customers' scroll activity on the webpage as well as statistics obtained from mobile devices. Bol.com's focus is on what customers' activities are on their own channel instead of engaging in external data sources i.e. social media. Compared to competing businesses i.e. Marktplaats, Bol.com has adopted similar data analytical initiatives.

Summary

The main difference between Vektis and Bol.com is that they both serve separate sides of the demographic spectrum. Their end-users' needs differ as their business functions aren't relatable within the same industrial sector. Therefore understanding the acceptance, realization, and utilization of Big Data occurs at different paces. Vektis caters to the healthcare industrial sector, which includes public service institutions, compared to Bol.com, whose focus on consumer products sold on an online platform therefore incurring a technological aspect. Each company has approached the concept of Big Data in a different manner.

Vektis requires standardization during aggregation of their data therefore explore possibilities in managing large amounts of data however opts for a more structured manner of organization. Bol.com on the other hand collects large amounts of data as well however they are required to process information on a real-time basis as consumer trends within the market are indefinite and the company must be able to respond to consumer needs with a faster response time. By adopting the practice of Big Data, Bol.com has been able to make the long tail[26] representation a compatible business prospect. As Vektis provides the data collected to healthcare institutions, with the current database architecture it has in place, it is able to bring about a great deal of insight into certain healthcare issues which may not necessarily be of interest by the public sector. If Vektis were to take a deeper plunge into the Big Data whirlpool, it would entail analyzing further information into unstructured data collected by, for example, sensors, which provide substantial data on healthcare beneficiaries and in turn become an important source of information for healthcare providers. The degree of acquisition of Big Data between each company will differ due to the kind of end user each provides services to. Different industrial sectors are reluctant to take steps into adopting technological advancements immediately however there are some where the advantages are quite clear. Data will keep growing in every form, however how affective it is in providing an organization with the ability to improve its processes is controversial.

[26] Long tail of some distributions of numbers is the portion of the distribution having a large number of occurrences far from the "head" or central part of the distribution.

Chapter 4. Big Data Framework

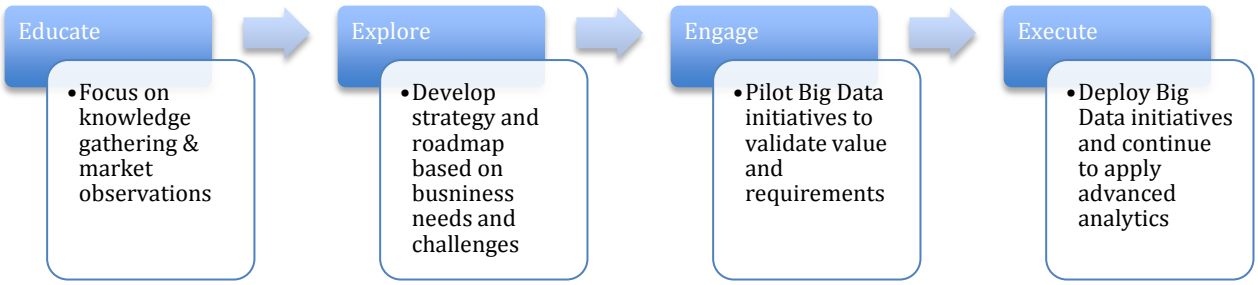
From the research conducted, both desk and field, a framework has been formulated to outline aspects to which businesses may need to consider when approaching the concept of a Big Data practice. First off, the framework will identify a criterion, which shows stages within the process of Big Data acquisitioning.

Decision Support & Automation

At this stage, a business must be able to engage in a collaborative understanding of the implications of Big Data on the business. Therefore before taking an even bigger step, the company should educate themselves on the knowledge aspects of Big Data including research on observations made in the market similar to their business, the concept behind Big Data. The company must be able to identify whether Big Data practices can complement their current business strategies.

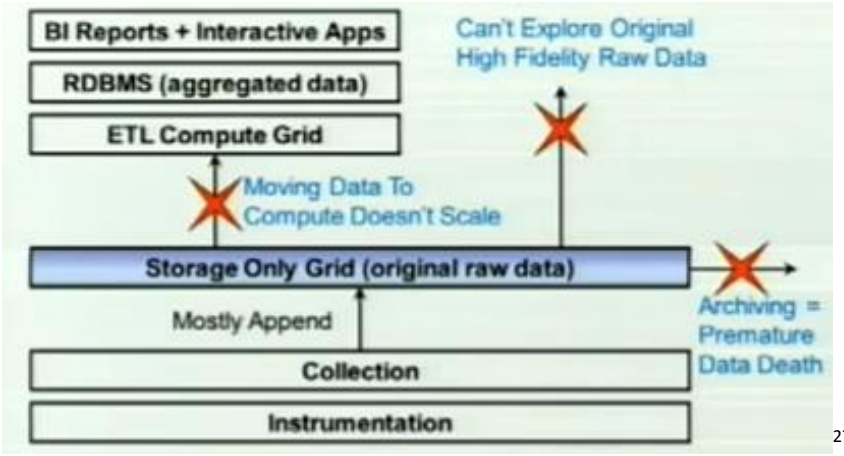
The company is then to explore the possibilities of Big Data. Therefore describe the challenges that Big Data will tackle and the possible rate of achieving this. During this process, it is advisable that businesses outline a roadmap which will display the strategies (short and long term goals), challenges, current database infrastructure within the company (including the kind of data it deals with i.e. structured/unstructured), alignment of Big Data with business processes, solutions Big Data will bring, potential future outlook and lastly the costs involved.

According to roadmap, the company is decide whether to deploy Big Data practices or if another methodology is more suitable i.e. traditional BI. Assuming Big Data is the likely solution, the company can test the waters in that they may pilot a few projects in order to analyze the outcomes and validate the value and requirements before taking on more scalable Big Data initiatives, which is followed up in the execution phase.



The image above describes the Decision & Execution Approach discussed in stages of Educate, Explore, Engage, and Execute. [4]

To better determine whether the Big Data approach is identifiable with the company's business needs, Cloudera formulated a brief framework portraying the feasibility aspects between Relational Database Management System (RDBMS) and Big Data Technology.



[27]

The image above provides an illustration which shows the process of data analysis in the traditional way using a RDBMS system. The data is initially collected and stored in Storage Only Grid where it is then sent to an ETL Grid, where it must then go through a compute grid for refinement of data to an acceptable form. All the data cannot be processed as moving data to compute doesn't scale. Any data that was stored in the Storage Only Grid no longer can be referred therefore exploring any fidelity raw data is not possible and once data has been archived, it isn't economical to retrieve that information for future use. These are certain drawbacks of an RDBMS system compared to the capabilities employed by Big Data. In the Big Data world, Hadoop is the primary framework for processing large volumes of data.

When to use the right tool for the right job

RDBMS

- Interactive OLAP Analytics (<1sec)
- Multistep ACID Transactions
- 100% SQL Compliance

Hadoop:

- Process unstructured and structured data
- Scalability of storage/Compute
- Complex Data Processing
- Can perform OLAP Analytics however not as fast[15]

[27] University, Standford. "Introducing Apache Hadoop: The Modern Data Operating System." *Youtube.com*. YouTube, 04 Sept. 2012. Web. <http://www.youtube.com/watch?v=d2xeNpfzsYI>.

Analytics & Discovery

The next stage in the framework includes the aspect of Analytics & Discovery. At this stage, the company will apply certain analytical tools which will aid in data mining. In other words it will be used to detect patterns in the data which may provide insight into the value of the data processed.

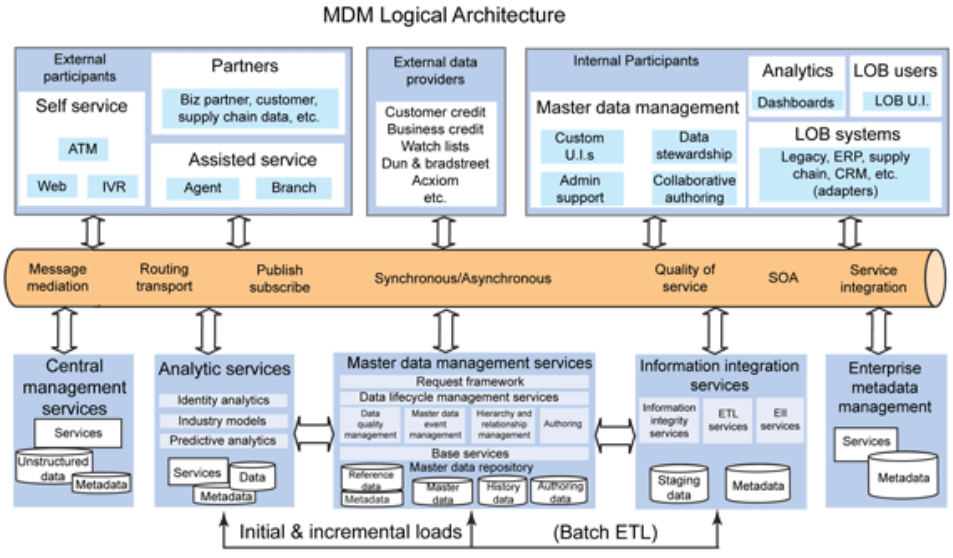
Data Organization & Management

To maintain integrity and governance of data, the concept of Master data can be incorporated within the data architecture. Master data is intended to provide credible and verifiable data at the disposal of the enterprise. Master data management (MDM) in a nutshell, comprises of the tools necessary for governing and leveraging data.

The objective behind the MDM approach is to provide value to Big Data by giving it a more plausible and reassuring structure and meaning. MDM tools are tasked with matching products with traits identified from web comments as well as engage in the data provided by mobile devices including information on the time and location consumers purchased certain products.

MDM can greatly aid in regulating data governance. A way in which Big Data could be governed in collaboration with Master Data is as follows:

- Matching Big Data with what can reliably be assigned to Master Data
- Classifying unstructured data to be as accurate as possible
- Using common metadata between Big Data and MDM initiatives
- Deciding on the levels of quality and veracity in line with the organization, and then focusing on the Big Data that meets those levels
- Applying data privacy and security policies to the Big Data in correlation to Master Data. [16]

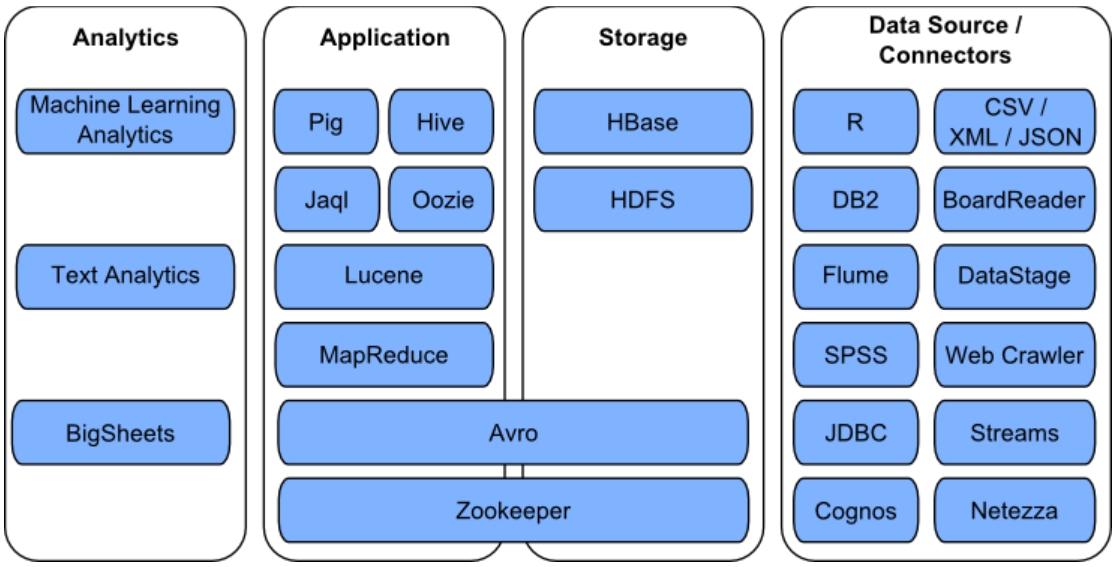


28

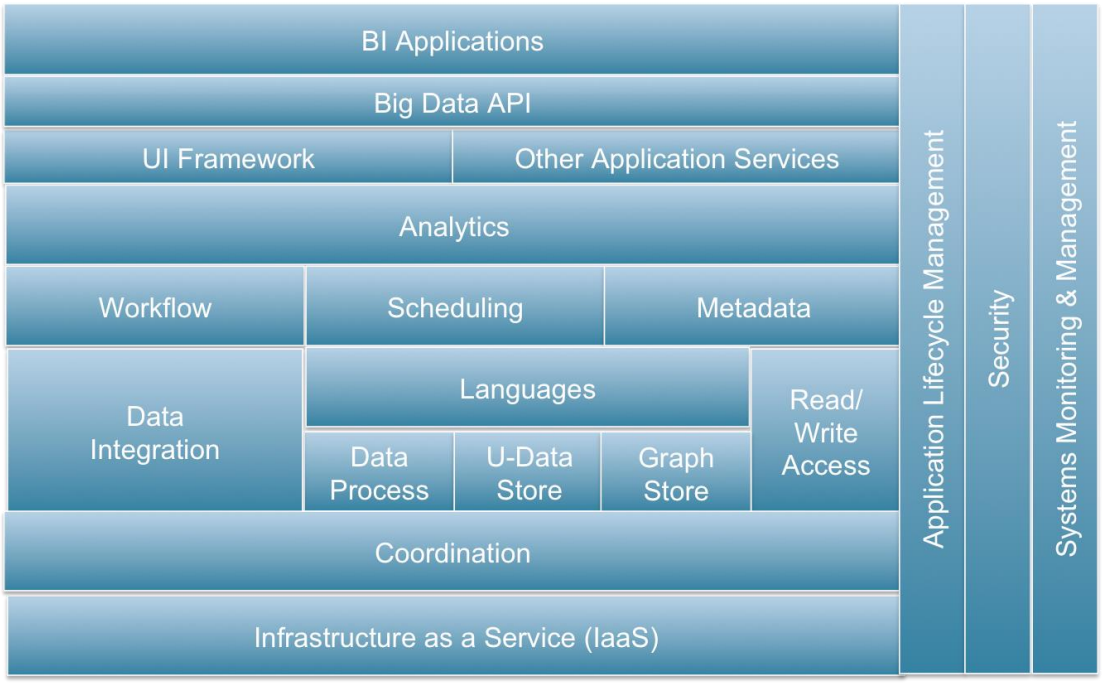
[28] "Monitoring and Tuning InfoSphere Master Data Management Server." *IBM*. 10 Oct. 2010. Web. <http://www.ibm.com/developerworks/data/library/techarticle/dm-1010mdmservertuning1/>.

The image above simply provides an overview into the components common in a typical MDM architecture.

Infrastructure



The infrastructure above presents a possible Hadoop technology stack which includes the different tools involved. Further explanation into each type of tool identified above is mentioned in Appendix B. The infrastructure above can be translated into a more mapped illustration displayed below.



Moving from bottom to top, the data is extracted from different sources and loaded into the Hadoop Distribution File System where it is then distributed by MapReduce algorithm across the nodes to later be analyzed in order to provide the required output.

Conclusion

The framework provides an overview into the stages involved in the process of application of Big Data practices within a typical business, therefore its influence within different industrial sectors will vary depending on specific needs. The framework is to act as a guide to understanding the process involved in approaching Big Data. Each business is faced with different kinds of data, large and small, therefore implementing a Big Data practice in accordance to the business' processes is imperative. If approached in an incorrect manner may provide inconsistent results and cost more than intended initially. As observed during the field research, each company within their own industrial sector viewed the concept of Big Data along the same lines however, approached it in a way so that their main business processes were optimized.

Chapter 5. Conclusion & Recommendations

Big Data has gained momentum across the business world since its discovery. Organizations have found comfort in adopting Big Data practices as they have realized the positive outcomes to be gained from the technology. Big Data taps into parts of the business to create more value. It is therefore imperative for organizations to begin with reevaluating its strategic objectives and discover the extended potential the business has left untapped. While some organizations find it difficult to engage in Big Data initiatives, there is a significant number of companies across many industries adopting this concept. To these companies, the data they collect on a daily basis is capable of illustrating a bigger picture in which it may aid in improving their business processes, customer service, and encourage new product development. Companies within the sector of consumer products, come across large datasets of customer transactions on a traditional platform however, these companies are also engaged with the social media spectrum therefore collect large amounts of data published by their customers concerning reviews, blogs, videos, photos etc. on their products. These kinds of data types are categorized to be unstructured and traditional methods aren't equipped to be able to deal with this data format. Therefore, Big Data technologies such as Hadoop and MapReduce have been developed for the convenience of processing such data sets in a reasonable time span.

An organization should set up a roadmap when pursuing Big Data practices. For example, it may consider drawing up an enterprise-wide Big Data blueprint. This means they must consider the vision, strategy, and requirements for Big Data and its alignment with internal business processes and users. In turn, it will create a common understanding of what the organization hopes to achieve with Big Data with a pragmatic approach in mind. The blueprint should define the scope of Big Data as whole, including the business challenges being faced currently, the requirements to its application, and the architecture of the Big Data infrastructure i.e. data, technologies and tools.

Moving on to the next stage, for cost purposes, it is advisable that companies first evaluate the kind of the data that currently exists within the organization. As a result, this may assist in making possible short term goals achievable thereby gaining gradual experience and eventually exploiting the scalability of handling larger volumes of data and tackling more complex initiatives. In the process of analyzing data projects, for a company to gain credible insights, it is a prerequisite for businesses to consider acquiring certain analytical tools such as R, Netezza etc.

Let's consider the concept of Big Data in the eyes of the business world with an analogy. Imagine being on a boat in the middle of the ocean, during this time the fisherman is faced with dealing several obstacles along the way, dangerous sea creatures, storms, etc. Just like in the business world, companies are faced with several obstacles along the way. In the process, a fisherman's purpose is to go out to sea to explore and look for fish in order to bring back to sell at the local market, in the same way a company analyzes the business market to explore for new opportunities to attract customers. Therefore if the fisherman were to only go out to sea and only fish for one kind of fish despite being surrounded by an abundance of available sea creatures, the fisherman will have more of a difficult time in gaining a competitive advantage over the other fisherman at the local market. In a similar manner, a company comes across volumes of data, which provide the company with new opportunities to exploit in the market and gain a competitive edge however, if businesses only focus on one type of data form, they may just be left behind and rendered opportunistic. It may

seem easier said than done, however, if organizations do not make use of the massive surge of data being made readily available to them, the business world will keep sailing in a single direction constantly facing obstacles but with less to show for than what it possible.

Adopting Big Data can provide businesses with a learning curve In that they come across certain aspects they never gave much thought too. Leveraging the data available will provide companies with insights which could potentially increase productivity and reduce working capital. While this may all seem hunky dory, there is the issue of data privacy and security especially since in today's age, any information put on the online platform is inevitably stolen or even used by governing bodies to analyze for any harming evidence of terrorism which creates an uncomfortable feeling for the user. From the time a person wakes up to the time of bed, details about the day have probably been stored somewhere in some server, with or without the person's consent. While many protest this form of data accumulation, the fact of the matter is that this data can be used to produce positive outcomes beneficial for the public particularly. While it may seem unethical to the public, it is a company's pursuit to portray this use of data as beneficial for the customer.

For an organization to consider adopting Big Data practices must consider their strategic objectives before moving forward and determining whether the technology to be implemented are in line with the business processes. As in the field research conducted of the companies engaging in Big Data, their business purposes differ due to the industrial sector they are engaged in and therefore cater to different demographics. Each company tackled ways in which they could address large amounts of data while at the same time being able to manage its business processes in a more efficient and cost effective manner. At the moment, Big Data may seem as if it is not for everyone as there is some hesitation by companies to take this step however, with time Big Data will become a part of the way a business processes information or even embedded within a company's businesses processes.

Recommendations

1. Big Data isn't just a tech-based initiative, rather it is a business-led project which means a wide range of skills is required when formulating a Big Data strategy. This will enable businesses to identify which problems it faces and determine whether it has the right data needed to carry out a Big Data practice. After determining the problem at hand, shaping a possible solution Big Data is to resolve should be outlined along with the value creation factor. To carry out these strategies, the next step should be established, therefore, what should the business do next to make it happen.
2. Companies that do not have as many funds compared to larger businesses, experimenting with Big Data practices in an inexpensive manner is a possibility as they can implement technologies such as Hadoop which is an open source framework for data processing. Any business will be hesitant to take the initiative to adopt a new approach, however, first piloting small projects to test the viability of Big Data within the business would help the business gain an understanding as to whether it is appropriate for the respective company.
3. As Big Data is still gaining momentum within the business world, there is a lack of solid knowledgeable experts in the field, investing in the education in one or two data analysts to better understanding the ins and outs of Big Data could give a business a stronger IT foundation.
4. Big Data currently seems to be hyped up within the business world however, while this may seem true, the hype has helped businesses to realize the value of the data available to them. As many experts mentioned that, Big Data will catch on within the next five years, businesses can right now benchmark other organizations within the same industrial sector, which have pursued Big Data to mirror their actions in order determine whether Big Data is feasible within the respective company. Therefore this prevents businesses from leveraging their data as if it were a guessing game.
5. Companies that are to deal with large volumes of transactional data, it is advisable they use a traditional form of data warehousing however, depending on the discrepancy of data, businesses could adopt certain tools that more cost effective and perform in SQL such as Hive or implement ELT tools. At this time however, combining a traditional data warehouse and non-traditional warehousing system, can provide a business with a 360 degree outlook into the business' potential, with the integration of the different data types.

Chapter 6. Sample Business Case

The airline industry is a part of the industry that is currently facing tough economical times due to the decline in consumer confidence. Since incidents such as 9/11 and those following have caused travellers to re-evaluate their travel frequencies and the increased securities have made travelling experiences less comfortable. In addition, due to the events occurring within the global economy, consumers are also hesitant to spend extra on travel costs and are therefore looking for cheaper alternatives. As result, this creates a competitively driven market for airlines to engage in. Airlines are finding it difficult to commoditize travelling.

Every airline at this time is probably attempting to reach their customers on a more personalized platform therefore trying to improve the customer experience in order to gain loyalty. On the customer experience lifecycle, the airline has a change to either drive or sink their chances for brand loyalty.

Of course, airlines have explored the world of mobile applications, which includes providing customers with ease of access to booking, cancelling, making payments for flights, obtaining schedule details, and being able to check in online. Therefore this data is mainly transactional or structured. However, what airlines have allowed missed opportunities to 'fly' by them where other third parties have taken the initiative to aggregate all kinds of information catering to creating a fulfilling experience for travellers.

The Mobile App World

According to Edtech Magazine, the number of mobile-connected devices is projected to exceed the world's total population within 2013 itself. The use of mobile devices grew by 70% globally with an 81% rise in usage of smartphones. Travellers are always on their mobile devices therefore collecting data from their online activities can provide insight into the way they personally endured the travel experience. Making use of this technology, airliners can record selections made by customers and the search results during the process of booking flights. They may be able to determine the kind of customer visiting the webpage and be able to generate suggestions which provide the customer with an all rounded experience pre-travels and as well as post-travels.

Big Data and the Mobile App World

Let's consider the framework earlier discussed in Chapter 4. The framework illustrates 4 stages with which a company should consider when adopting Big Data practices.

Decision Support & Automation

For an airline it is important to consider the experience it provides for the traveller from the moment a customer enters the airline's webpage till the point of leaving airport terminals. Therefore the airliner should consider the approach of educate, explore, engage and execute. This would entail

Learning the aspects of Big Data capabilities and the way in which it can bring value for the airliner. The company can create a roadmap outlining the strategies (short and long term goals), challenges to be addressed by Big Data, current database infrastructure within the company (including the kind of data it deals with i.e. structured/unstructured), alignment of Big Data with business processes, solutions Big Data will bring, potential future outlook and lastly the costs involved. An additional tool which may aid in determining value creation, is a value diagram which illustrates forces that drive shareholder value, operational levels that enable them and the metrics associated with each lever.

Taking the example of an airline's ticketing service. Many travellers prefer to have a mobile boarding pass instead of having to print it, therefore certain levers such as time, convenience and costs can drive a value chain across the stakeholders involved in the process. For the customer, it would save check-in time, and increase the convenience of checking-in in more than one place, and possibly lower ticket prices which ultimately could result in the airlines being able to increase revenue, reduce overhead costs and working capital. Therefore as can be seen, an airliner can identify the roadmap which could be taken to achieving short and long term goals ultimately with the use of Big Data by analyzing user's responses to the online platform.

What is important to keep in mind is the kind of data processing technology whether RDBMS or open source technologies such as Hadoop are suitable for the airliner's needs. As airlines store large amounts of data, mostly transactional which include sensitive data about the customer's personal details, adopting an open source platform may not seem viable however, creating an environment where in a traditional data warehouse system collaborates with a Big Data processing system such as Hadoop. Hadoop can process information such as ratings, reviews, frequent searches, click rates, etc, and the data warehouse system can aggregate transactional data. As result this will allow the airliner with the opportunity to cater to their customers more efficiently.

In the *Analytics and Discovery*, several tools can be applied to analyze the information collected in order to provide valuable insights for the airliner for statistical needs. Followed by the *Data organization and Management* phase which basically includes governing the credibility of the data. Lastly, the *Infrastructure* of the complete data distribution system will include the different tools involved in creating a Big Data technology stack, depending on the requirements identified by the airliner in earlier stages.

In conclusion, an airliner may consider the concept of Big Data to provide positive outcomes which may enable to better understand their customers' travel preferences. With Big Data tools, an airline company can collect several forms of data collected through mobile apps in particular to monitor a traveller's journey on an online platform.

Chapter 7. Bibliography

1. McKinsey Global Institute. McKinsey & Company 2011. *Big Data: The next frontier for innovation, competition and productivity*.
2. Economist Intelligence Unit. 2012. *Lesson from the Leaders.* The Economist
3. Oracle. Jan 2012. *Big Data for Enterprise.*
4. IBM. 2012. *The real world use of Big Data*.
5. Christ Eaton, Tom Deutsch, Dirk Deroos, George Lapis, and Paul Zikopoulos. IBM. *Understanding Big Data.* McGraw Hill.
6. Womack, Brian, and Chris Strohm. "Twitter Said to Bolster Security After AP Account Hacked." *Bloomberg*. N.p., 25 Apr. 2013. Web. <http://www.bloomberg.com/news/2013-04-24/ap-twitter-account-hacking-exposes-social-media-weakness.html>
7. Cognizant. 2012. *Big Data's Impact on the Data Supply Chain.*
8. Chui, Michael, Markus Loffler, and Roger Roberts. "The Internet of Things." *McKinsey & Company*. N.p., Mar. 2010. Web. <http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_thing>
9. Cebr. April 2012. *Data Equity: Unlocking the value of Big Data.*
10. "Traditional vs Big Data Analytics." *Traditional vs Big Data Analytics*. SYS-CON Media, 2008. Web. <http://srinivasansundararajan.sys-con.com/node/1968472/mobile>.
11. Logica. June 2012. *Excellent Business Decisions Powered by Big Data.*
12. Fujitsu. *The White Book of Big Data: The definitive guide to the revolution in business analytics.*
13. Fujistu. 2013. *Solution Approaches for Big Data.*
14. Essent, May 2012. *Big Data & Essent*
15. University, Standford. "Introducing Apache Hadoop: The Modern Data Operating System." *Youtube.com*. YouTube, 04 Sept. 2012. Web. <http://www.youtube.com/watch?v=d2xeNpfzsYI>.
16. IBM. Nov 2012. *Master Data management: The key to leveraging Big Data.*
17. "Consumer Surplus." *Definition*. Web. <http://www.investopedia.com/terms/c/consumer_surplus.asp>.
18. Oracle. *Integrate for Insight.*
19. Data Academy. 2008. *ETL vs ELT.*
20. Mark Albala. Cognizant. 2011. *Making Sense of Big Data in the Petabyte Age.*
21. SAS. *Big Data meets Big Data Analytics.*
22. Accenture. 2012. *Building the Foundation for Big Data*.
23. Oracle. August 2012. *Oracle information Architecture: An architect's Guide to Big Data*
24. Oracle. Jan 2012. *Big Data for the Enterprise.*

Chapter 8. Appendix

Appendix A

Expert Interview Questions

Questions proposed to Ordina, Mr. Sander van Kleef

1. How do you make sense from so much data?
2. How would you define big data?
3. How important is big data in today's age?
4. Do you think big data will be around for the next few years/decade?
5. What is the main advantage of big data?
6. If businesses adopt big data practices what would be a few drawbacks/inconveniences that they could experience?
7. The kind of businesses you consult, what do you they usually look for when it comes to big data? Do they believe or can you convince them easily enough of big data' worth to their business?
8. How popular is big data in the Netherlands?
9. What do you normally propose as a big data solution to a customer?
10. There is the issue of data security and privacy, how do you solve this issue, especially when people speculate big data's credibility in this perspective?
11. Who own all this data?
12. Which industrial sector benefits most from big data?
13. How can big data help business in the Netherlands particularly?
14. In terms of economic purposes, can big data influence the way company carry out their business in NL to an extent?
15. There's too much data and at this point data is expected to grow at alarming speeds, what is the point of recording everything when it will only consumer more time?
16. What are current big data technologies?
17. What is your opinion on big data?

Questions proposed to Andarr, Mr. Eric van Tol

1. How would you define big data?
2. How important is big data in today's age?
3. Do you think big data will be around for the next few years/decade?
4. What is the main advantage of big data?
5. If businesses adopt big data practices what would be a few drawbacks/inconveniences that they could experience?
6. Gartner defines big data differently compared to other sources. "big data is high volume, -velocity, and –variety information assets that demand cost effective innovative forms of information processing for enhanced insight and decision making" would you describe big data along the same lines or do you view big data along the lines of the 3V's?

7. How popular is big data in the Netherlands?
8. There is the issue of data security and privacy, how do you solve this issue, especially when people speculate big data's credibility in this perspective?
9. Who own all this data?
10. Which industrial sector benefits most from big data?
11. How can big data help business in the Netherlands particularly?
12. In terms of economic purposes, can big data influence the way company carry out their business in NL to an extent?
13. There's too much data and at this point data is expected to grow at alarming speeds, what is the point of recording everything when it will only consumer more time?
14. What are current big data technologies?
15. What is your opinion on big data?
16. What would you rename big data to be?

Questions proposed to Bearing Point

1. How would you define big data?
2. How important is big data in today's age?
3. What is your outlook on the future of big data?
4. In order to start investing in big data, what questions do you believe a business should ask themselves before taking this initiative?
5. What would you rename big data to be?
6. Why is big data not a priority within businesses?
7. Why is there such hype on the topic of big data presently?
8. How could big data influence business decisions to the point of it being profitable?
9. Is big data the frontier to competition?
10. What are a few factors that drive big data analytics?
11. Why isn't big data use more prominently in Europe compared to the US?

Questions proposed to Cognizant

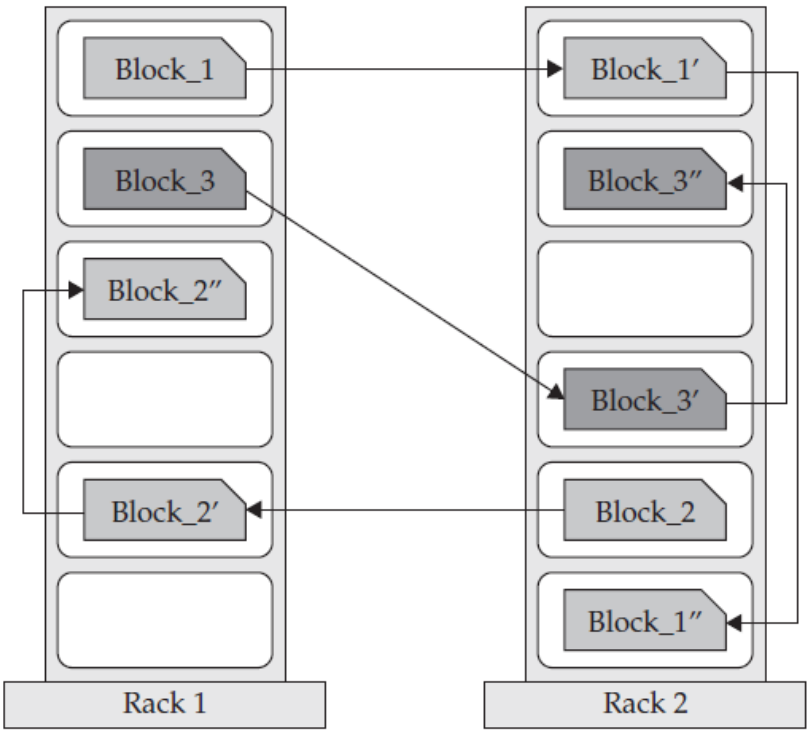
1. How would you define big data?
2. How important is big data in today's age?
3. What is your outlook on the future of big data?
4. Why is big data not a priority within businesses?
5. Why is there such hype on the topic of big data presently?
6. Is big data the frontier to competition?
7. What are a few factors that drive big data analytics?
8. Why isn't big data use more prominently in Europe compared to the US?
9. Why are traditional methods to information management inadequate for big data segmentation?
10. What does the next generation data supply chain mean? How does this influence integrity, scalability, and security of data?
11. What are a few big data technologies?

Appendix B

Hadoop Technology Stack

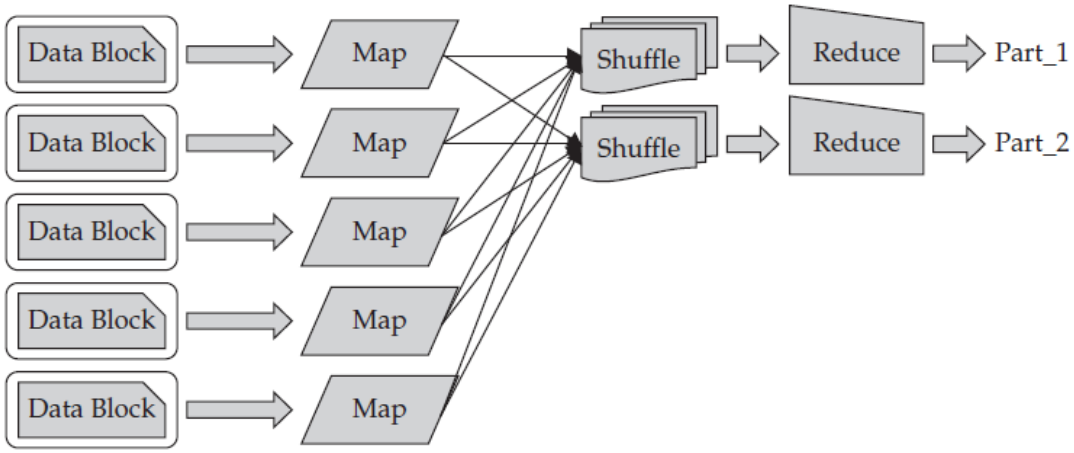
HDFS

Data is broken down into smaller pieces so that the functions being executed can be performed on smaller subsets, providing scalability. Hadoop's goal is to use servers with inexpensive internal disks in large numbers with performance increased by MapReduce. With such a large size the system is much more likely to fail however, because of its larger size overall and data being stored into small blocks, built-in fail tolerances are implemented in the form of data being copied multiple times across the system. This works by copying all the data into two additional servers by default, additionally allowing for higher availability and the ability to run these jobs in each place where stored , allowing for better scalability.



MapReduce

MapReduce is the heart of Hadoop, allowing for massive scalability across thousands of serves in a Hadoop cluster. MapReduce refers to two separate tasks, one where it takes a set of data and converts it into another set of data where individual elements are broken down, and another where it takes this output and reintroduces it with other outputs and repeats the process. This continued operation allows for the breaking down of billions of rows and columns into the few needed.



Application Development in Hadoop

Pig and PigLatin
Developed at Yahoo! Pig is designed to focus more on analyzing data versus writing to mapping and reducing programs. It is designed to handle any type of data and is made up of two components; the language, PigLatin, and the runtime environment where they're executed.

Hive
Unlike Pig it is based off of SQL, meaning for many it is based off a language that is already known. This gives it an immediate advantage however it is more limited in the commands it understands. The statements used, HQL, are broken down into MapReduce jobs and then executed across the Hadoop cluster.

Jaql
Allows for the processing of both structured and nontraditional data and was donated by IBM to the open source community. It specifically allows you to select, join, group, and filter data stored in HDFS, a combination of Pig and Hive. It is designed to process large data sets and can rewrite high-level queries into low-level queries consisting of MapReduce jobs.
Getting Your Data into Hadoop

Flume
Flume was created to allow you to flow data from a source into your Hadoop environment. In flume you work with sources (any data source), decorators (an operation on the data that can transform the data), and sinks (the target of the specific operation). Flume is especially useful in scenarios where data flows from many sources, manipulate it, and then drop it into the Hadoop environment.

ZooKeeper
ZooKeeper is an open source Apache project that provides a centralized infrastructure. It maintains common objects needed in large cluster environments, such as configuration data, hierarchical naming, etc. This centralized infrastructure allows for better organization of name, group, and

synchronization services across hundreds of servers. It can also assure that tasks across these servers are synchronized by maintaining the status type information.

HBase

HBase is a column-oriented database management system that runs on top of HDFS, well suited for sparse data sets. HBase are written in Java and comprises a set of tables where each table contains rows and columns, much like traditional databases. HBase allows for attributes to be grouped together into column families

Oozie

Oozie is a project that simplifies workflow and coordination between jobs, providing users with the ability to define actions and dependencies between actions. Oozie will then schedule these actions to execute when required dependencies have been met. A workflow can be scheduled to begin on a given time or based on the arrival of specific data.

Lucene

Lucene is an extremely popular open source Apache project for text search. Predating Hadoop, it offers full test indexing and searching with Java applications. Lucene will break down documents into text fields and index according to these fields. This indexing is what gives Lucene its speed and power.

Avro

Avro provides data serialization services. When writing Avro data to a file, the schema that defines the data is written to the file. This makes it easy for any application to read the data at a later time, since the schema is stored within it. Because of this data can also be versioned since the schema of an older data remains stored within the file.

Enterprise Integration

Netezza

The Netezza Adaptor lets you leverage the simplicity and flexibility of Jaql in your database interactions. It supports splitting tables and partitioning the table, this allows for SQL statements to be processed in parallel.

R Statistical Analysis Applications

R is a language and environment for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques and is highly extensible. One of its strengths is its ability to produce high quality plots including mathematical symbols and formulae where needed.

Data Visualization

BigSheets

Unlike the core Apache Hadoop components, BigInsights includes tooling for visualization and performing analytics on large sets of data. This hides the complexity of MapReduce, which allows the users to focus on the results and not how they were achieved. BigSheets is a browser based visualization tool that utilizes the familiar spreadsheet interface and requires no programming. BigSheets works in three simple steps; collect data, extract and analyze data, and explore and visualize data. Common outputs are the Tag Cloud, Pie Charts, Maps, Heat Maps, and Bar Charts.

Text Analytics

Text Analytics features a text engine that gain insights via looking through text data, insights such as customer web patterns, finding fraud indicators, and assessing customer sentiment from social media messages. The goal is to read unstructured text and distill insights from them and organizing this into databases.

Machine Learning Analytics

Machine Learning Analytics provides a platform where high-performance statistical and predictive analysis on data within a Hadoop cluster. It includes a high level machine learning language that is similar to R and includes a number of precanned data mining algorithms. It was developed by IBM whose primary goal was high performance and ease of use for analysts needing higher performance results from Hadoop context data.

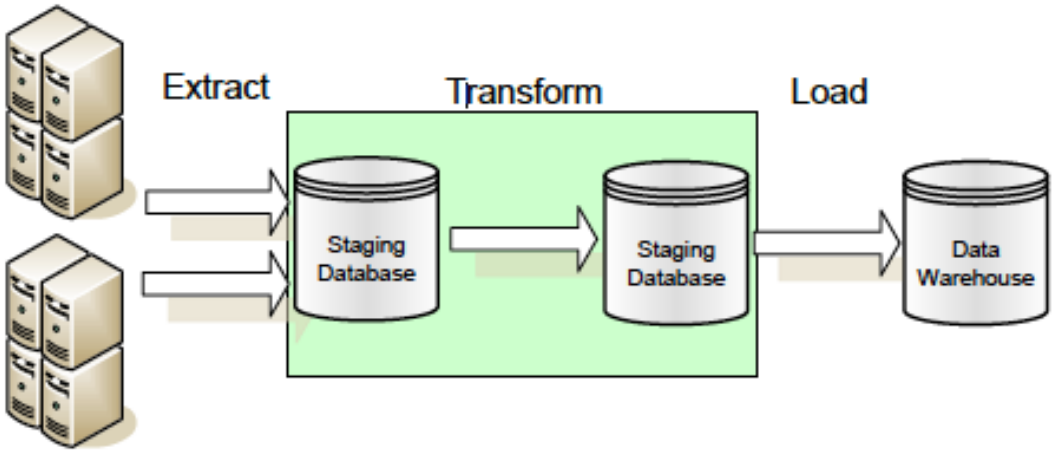
Large-Scale Indexing

An indexing component is also made for building large-scale indexing and search solutions. BigIndex can optimize, merge, and replicate indexes as well as indexing over Hadoop. The goal was for searches through hundreds of terabytes to return results within a second. This is completed by implementing multiple types of indexing: Partitioned index (an index partitioned into separate indices), Disturbuted index(and index distributed into shards where a collection of shards create one logical index), and Real-time index (real-time sources add data to the index).

Appendix C

ETL VS ELT

ETL

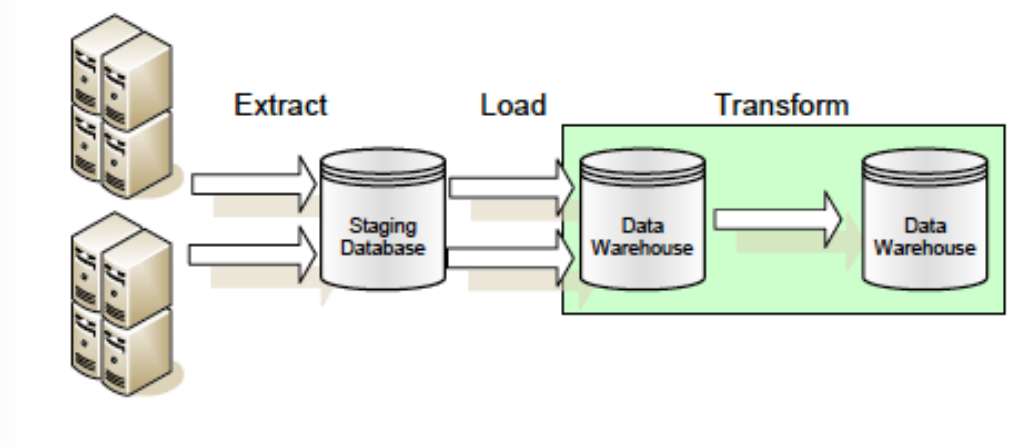


The figure above illustrates a traditional approach to data warehouse development, otherwise known as Extract, Transform, and Load (ETL). Data is extracted from relevant data sources which it is then transformed in line with business rules that dictate to produce the required data format of the output. Data is refined during the transformation process to maintain quality and integrity of data. This is done in the staging area of the process. Once the data has been refined to the targeted outcome, it is loaded in the data warehouse for further use.

The process of ETL is designed what seems to be backward processing, as the required output is pre-determined. The process of data aggregation follow certain business rules which define each stage from extraction to loading to the data warehouse. This form of data processing is specifically intended for processing structured information.

Strengths	Weakness
Extraction and processing overhead is reduced due to less time required during the development phase	Not very flexible as it targets the relevant data, any other data type will have to be added to ETL design
Contains only the data that is relevant	If the size of information grows, hardware may not be scalable or it is expensive to invest in additional hardware
Availability of tools is high as many can even be customizable	

ELT



The figure above illustrates a non-traditional method to data processing, otherwise known as Extract, Load, and Transform (ELT). Data is extracted from different data sources during which time it is checked with integrity and business rules. It is then loaded into the data warehouse, where it can be then transformed into the targeted output format.

Isolating the extract and load process from the transformation process means that any data can be stored for future use as well as projects can be broken down into smaller parts making the system more management.

Strengths	Weakness
Flexibility within the warehouse structure.	Not a custom within the current business world in regards to data warehousing
Scalability of hardware is possible.	
Managing projects is easier as project can be broken into smaller chunks.	