# The Influence of Big Data on Retail Activities and Retail Performance

**Brkanić, Lovro**

**Master's thesis / Diplomski rad**

**2020**

*Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj:* **University of Zagreb, Faculty of Economics and Business / Sveučilište u Zagrebu, Ekonomski fakultet**

*Permanent link / Trajna poveznica:* https://urn.nsk.hr/urn:nbn:hr:148:088433

*Rights / Prava:* In copyright

*Download date / Datum preuzimanja:* **2022-06-28**

*Repository / Repozitorij:*

REPEFZG - Digital Repository - Faculty of Economcs & Business Zagreb


**University of Zagreb**

**Faculty of Economics and Business**

**Master's Degree in Trade and International Business**


# THE INFLUENCE OF BIG DATA ON RETAIL ACTIVITIES AND RETAIL PERFORMANCE

**Master thesis**

**Lovro Brkanić**

**Zagreb, June 2020**

**University of Zagreb**

**University of Zagreb**

**Faculty of Economics and Business**

**Master's Degree in Trade and International Business**

# THE INFLUENCE OF BIG DATA ON RETAIL ACTIVITIES AND RETAIL PERFORMANCE

**Master thesis**

**Lovro Brkanić, 0067504358**

**Academic year: 2019/2020**

**Mentor: Doc. dr. sc. Kristina Petljak**

**Subject: Trade Management**

**Zagreb, June 2020**

## Summary

The following master thesis *"The Influence of Big Data on Retail Activities and Retail Performance*" is aimed to define big data and big data analysis, and to evaluate its importance in a new data driven world. Furthermore, it will be investigated how big data can be used in retail activities as it is proved that the use of big data can improve retail performance. Therefore, this thesis will also present case studies of successful usage of big data analytics in retail companies where it will be proved that big data analysis can lead to an increase in sales and customer loyalty, improvement of store layout to fit the customers' needs and many more. To see the relevance of big data in retail across the globe, a qualitative research will be made. This research will be conducted by doing a case study analysis. The goal of this research is to find out how retail companies use big data to achieve competitive advantage.

The first part of the thesis will introduce the topic of big data and big data analysis along with objective of the study, research design and methodologies. The second part of the thesis deals with defining, describing, and evaluating big data and big data analysis. Furthermore, the third part of this paper will deal with the potential of using and implementing big data analytics in the retail sector as well as presenting the benefits that companies received after successful implementation of big data analysis. The fourth part of the thesis will show the empirical research along with research method, research results, limitations, and future potential of big data analysis. The final part of this study ends with the conclusion on the topic of big data analysis and the usage of big data analysis in the retail sector. Finally, list of references will be shown at the end the thesis.

Lovro Brkanić

Name and family name of student

## STATEMENT ON ACADEMIC INTEGRITY

I hereby declare and confirm with my signature that the master thesis

(type of the paper)

is exclusively the result of my own autonomous work based on my research and literature published, which is seen in the notes and bibliography used.

I also declare that no part of the paper submitted has been made in an inappropriate way, whether by plagiarizing or infringing on any third person's copyright.

Finally, I declare that no part of the paper submitted has been used for any other paper in another higher education institution, research institution or educational institution.

In Zagreb, 23.06.2020

(date)

Student:

L. Brkanić

(signature)



4

# TABLE OF CONTENTS

# 1. INTRODUCTION

## *1.1. Research background and objectives*

In today's world where companies strive for greatness, the competitiveness has never been bigger. The pace to which everything operates has never been faster, and the prices of many goods have never been lower. Whether a company focuses on cost leadership or differentiation, accomplishing sustainable business performance was never harder. Also, achieving competitive advantage over a competitor requires creative and innovative solutions that appear only so often.

Not so long ago, companies wanted something that would give them the edge over their competitor, something that would improve the service given to their customers. It was at that time when companies started looking more into their sales records to get to know their customers more. The more those sales records were investigated, the more the companies discovered. Information such as which products sell best at what time, which products are frequently bought together, which product is mostly purchased by people aged from 18 to 28. All these data (also known as big data) serve like gold to retailing companies, which is why the practice of analysing company data became so important.

Big data has played a big role for many retailing companies over the last decade. The field of big data has been developing fast, with great success, which is why the demand for the job of data scientists has never been bigger. But most importantly, it was discovered that through big data, many companies have the opportunity to achieve a competitive advantage. For this reason, the aim of this paper is to investigate the importance of big data. Also, it will be discovered in what ways do retailers use big data to increase company performance.

## *1.2. Research design and methodologies*

As the aim of this paper is to investigate the usage of big data in the retail sector, the best way to cover the topic is to conduct a qualitative research. Therefore, a qualitative research will be conducted by using secondary data, mostly collected from the internet. By using books, academic articles, case studies and websites, the field of big data will be analysed and presented.

### *1.3. Structure of the dissertation*

This paper is structured so it remains consistent from start to end. It starts with an introductory part where the aim of this paper, design, methodologies, and structure are presented. The paper then starts to deal with the main topic and gives an overview of the theoretical background of big data. Contents such as big data definition, 5 Vs of big data, big data lifecycle, big data platforms and many more. In the following part, the paper starts to deal with a more practical background of big data and in this part, some actual applications of big data will be presented. Furthermore, an empirical research will be conducted and presented. In this part, a comparison of big data usage between retailers from different countries will be made so it can be seen how big data usage differs in different parts of the world. Finally, a conclusion will be given which will include final thoughts on the field of big data.

# 2. A REVIEW OF BIG DATA

## *2.1. Definition of big data*

Big data is a term used for the massive quantity of information created by people and information technology. It can also be seen as a feedback of a certain operation, a piece of data that answers the questions of what, where and how many. The term big data dates all the way back to the creation of Internet. People were mentioning big data quite early, but the opportunities of analysing big data came not long ago. Big data became an exceedingly popular term as soon as companies found a way to use it to gain important insights about potential opportunities and challenges.

Akter and Wamba (2016, p. 178) gave a precise definition and they claim that big data can be seen as a holistic process that involves collection, analysis, use and interpretation of data for various functional divisions and therefore gaining actionable insights creating business value, and establishing competitive advantage . It has been a popular topic for years now which is why companies are beginning to invest in technologies that collect, distribute and store big data. Furthermore, companies are getting committed to analysing big data more and more, analysts and statisticians are getting more involved in the big data analysis process and their main goal is to extract a valuable piece of data in the enormous sea of big data.

It is said that the global data volumes grow by 40 per cent annually (Manyika et al., 2011). This is an enormous number considering the amount of data that is already available for use. Majority of these data are unstructured which makes it hard to use it to our advantage. Some projects based on big data evaluations will fail and it can happen sometimes. However, another project will most likely bring a competitive advantage to the company.

## *2.2. Reasons for big data usage*

Big data on its own is a vague term and to find something usable and valuable in the enormous pile of data is hard, but there are many reasons why investing time and resources in big data provides many benefits. As soon as a problem appears, a solution must be found, and it is always more beneficial if a data-based decision is made. This could be an inventory level problem, or the need for real-time analytics or even estimating future sales. Furthermore, a company could attempt to provide new benefits to customers or strengthen customer loyalty or participate in any other customer-centric analytics.

The need for big data analysis comes at different times for different industries. The dedication to exploit the beneficial use of big data will appear in different areas of business. A company that offers services will mostly concentrate on using big data to get to know their customers so they could provide new, exciting features to boost sales or customer loyalty. British Airways is an excellent example of using big data to improve customer satisfaction. Their "Know Me" program was an attempt to use data available about their customers and pictures from Google images so the staff of British Airways can personally greet their customers, mostly high-profile flyers (Henschen, 2013).

Manufacturers on the other hand claim that the best use of big data can be found in detection of product defects and supply planning *(TATA Consultancy Services, 2014)*. Every new product manufactured must go through product testing phase and each test will provide a feedback to the system. According to the results of the feedback, the data will signal either that the product is wholesome, or that it has a defect or some kind. In case of a defect, the system will try to identify the quantity of defects, the place where the defect occurred, and whether the defect occurred on every third, fourth etc. product. The manufacturer can therefore reconfigure the assembly line or use other materials in productions or something else. When it comes to supply planning, every manufacturer uses lots of different sorts of materials and it is crucial to keep track of quantities so that the production doesn't stop. By tracking those materials, setting of alarms or signals on when to order which is very important if the manufacturer operates on just in time model, production will be more efficient, and storage won't suffer from over capacity or below capacity.

Apple can be named as a manufacturer that used all above stated reasons for using big data. Apart from using big data to detect flaws in their products and keep track of their inventory, Apple also used the information they had on their customers to provide benefits which no other offered. As Steve Jobs said once, the customers don't know what they want until you show them, it is finding the consumer need that isn't already fulfilled.

### *2.3. Handling big data*

The process of collection of data is common and not overly complicated. By using programs and information technology and other resources, data can be easily stored in large quantities. Furthermore, data can be evaluated and divided into three categories. The first category contains data available and easily obtainable within existing data sources. The second category contains data that exists but cannot be available for numerous reasons. Finally, the

third category contains data currently unavailable, but can be generated with known technology (Olsson, Bull-Berg, 2015).

There are multiple ways of how data is generated. The most common source of data are commercial activities, meaning payment services and consumption patterns. Another source of data is internet traffic which includes social media activities and data collected from search engines. Furthermore, it is worth noting that a large quantity of data comes from movement-related data which includes the GPS and tracking activities based on certain locations. Finally, data can be collected through different types of sensors.

When it comes to storage of big data, there are a few ways a company can manage their data. Nowadays, most companies store data on cloud-based big data solutions because of its' effectiveness and the ease to expand the storage size. Storage of big data can also be outsourced; however, this is not the best solution to every type of company. Yes, outsourcing storage and data management is a smart choice if the company doesn't need every day real-time data-based decisions. It also gives the benefit of smaller investments and dedication towards data management.

On the other hand, in case a company deals with sensitive information, or needs fast access to its data, it is better that the company uses in-house cloud-based or hybrid solution regarding data storage and management. In this case, data scientists and other employees of the company can access any type of information they want. Also, in-house storage enables employees to control and customize data as well as take care of security (Mousannif et al., 2016).

However, as much as this option offers companies more benefits, the company must figure out if the need for in-house data storage is inevitable because in this situation, with lots of benefits comes lots of costs and obligations. Therefore, the company has to take into consideration investing into big data technology system, as well with hiring and training big data experts. This is a complicated process that requires lots of time and resources, but after the implementation and successful use, the company will most likely create a competitive advantage. Companies that use, or should use, in-house marketing are mostly companies that have loyalty programs such as retailers. Another example are companies that offer personalized offers, advertisement, or recommendations such as Netflix or Amazon.

Another option of data storage for companies is the public cloud. This option attracts mostly small and medium size businesses with limited information technology infrastructure. Also,

some large companies opt for this option as well because it is cheaper than in-house data storage. The public cloud also offers fast distribution which helps companies concentrate more on the value of data instead of data management (Intel IT Centre, 2013).

As cheap as it is, the public cloud has some disadvantages to consider. The first disadvantage is the inability to customize and control data. Also, the costs of maintaining public cloud in the long run may bring increasing costs which can exceed the costs of maintaining in-house data storage (Mousannif et al., 2016). Finally, one of the biggest threats with choosing public cloud is the security issue. Therefore, before choosing public cloud because of its low costs, every company needs to consider multiple factors before relying on the public cloud option. Firstly, the company needs to acknowledge to which extent will the control over data be provided. Furthermore, the company has to carefully examine the security the public cloud provider offers. It is important to check the measures of security and the process of accessing data, confidentiality, and accountability. Finally, it is important to check the legal framework that is responsible for data transfer and disclosure.

### *2.4. 5 Vs of big data*

In the large pool of data that a company receives, it is hard to find the right data to support the subject of analysis. The data collection process happens every second which means that the amount of collected data cannot be analysed with conventional methods (Kunz et al., 2017). It is even harder to split and distinguish important data from unimportant. Therefore, analysts started to characterize and describe Big data with the 3 Vs, volume, velocity, and variety. After a while, analysts added two more Vs, veracity, and value. Seeing big data through the 5 Vs makes it even easier to get fully familiar with the big data concept in general (Wedel, Kannan, 2016). And even though some argue that the only important Vs are volume, velocity and variety, veracity and value will be covered in this paper as well.

*Table 1.* **The 5 Vs of big data**

| The 5 Vs of big data | |
|---|---|
| **Volume** | The total amount of data within an organisation |
| **Velocity** | The speed of processing data |
| **Variety** | The differences in the type of data (structured and unstructured) |
| **Veracity** | The quality of data, represents the relevance/importance to a specific project |
| **Value** | The benefit it brings to the organisation |

Source: Kunz et al. (2017)

***The Volume***

The volume of big data is plain and simple, it is the amount of data collected in a certain period (Anuradha, 2015). The volume of big data collected in each period varies depending on the size of the company, number and type of business operations, the number of social media platforms a company uses and many more. However, even if the company is small, the amount of data collected will be big enough to make the data analysis complicated and time consuming.

Sales for example, produces millions of data every day for a company like Wal-Mart. Even for a small retailer we are talking about thousands of sales records every day. The difference is enormous, but the amount of data still requires a proper analysis and resources, mostly time, money, and the help of big data expertise. That being said, Wal-Mart as the biggest retailer in the world had to invest lots of resources towards the big data collection, storage and analysis. A smaller retailer does deal with a smaller amount of data than Wal-Mart, but that doesn't mean that the amount of data generated by a smaller retailer doesn't require resources or expertise. The retailer needs to do the same thing as Wal-Mart, only in a smaller scale.

Furthermore, when we look at sales, it can be said that there is much more to it than the pure sale of products or services. Sales, in terms of big data represent a much bigger importance in the analysis because it can be looked at from different perspectives.

Sales can tell a retailer how many customers passed through his stores. It can tell the number of products a certain customer buys during one purchase or if there are goods that customers buy in pair, such as milk and cereal. That way, the retailer will have a better clue on how the store should be laid out and which goods should be close to each other in the store. Sales can also give the information such as at what time did most of the sales occur and which products sold at a specific time of the day or month.

After analysing just one component of a retailers' business operations, it is evident how much data can be generated daily, what these data represent and how it can be used to acquire useful knowledge.

***The velocity***

Velocity represents the speed of the data processing (Ylijoki, Porras, 2019). It can also be described as the frequency of changes in data and the demand for real-time analysis and

decision making (Anuradha, 2015). This is of much importance because it is better to have a limited amount of data in real-time than lots of data with delay. That way, a retailer for example can see the changes in inventory in real-time and is able to react accordingly. An example of huge velocity of big data are social media networks like Facebook, Youtube, Twitter etc. where every second thousands of pictures, videos and tweets are being posted.

***The Variety***

The variety of big data are also known as different formats of data that are available (Kunz et al., 2017). Those data can be structured and unstructured. Structured data are data that are clearly defined which makes them easy to search and analyse (Lycett, 2013). The pattern of the structured data itself makes it easily searchable as well with the fact of the place it can be found which is relational database (RDBMS). Structured data can be generated both by humans and machines, but it remains structured as long as it is created within the relational database. The most common example of a collection of structured data is an Excel spreadsheet. Also, similar to creation of structured data, the process of finding structured data can also be performed both by human generated queries and technology. Data found in an Excel spreadsheet are displayed in rows and columns, usually generated by humans and the program itself enables us to easily find, filter, sort, and analyse anything we want.

Methods of finding and analysing structured data goes way back to the 1970s, as opposed to unstructured data which remains hard to find and analyse in a time and cost-effective manner. SQL, also known as structured query language, was developed in the early 1970s by IBM and is the first program that enabled highly efficient structured data management. SQL gave the opportunity to its users to find certain patterns and gain useful insights that can boost the company performance.

Unstructured data is called this way for a reason, and that is because this type of data is stored without any organizational method and in many different forms. Furthermore, unstructured data can also be created by humans or machines. As opposed to structured data, unstructured data does not contain words and texts only. Essentially, unstructured data can be an E-mail, photo, video, social media websites or any other website. Take Facebook for example, every day people upload different content and the daily uploads go far beyond millions. Photos, videos, comments, status updates, all different forms of data and those are usually unstructured. Therefore, when it comes to unstructured data, no relational database is included.

The difference between structured and unstructured data is obvious then. The biggest difference outside the fact that structured data can be found in relational databases and unstructured cannot, is the difficulty of conducting an analysis. In the vast sea of unstructured data, it is hard to conduct a valuable analysis because the type of content unstructured data contains are hard to search, sort and filter out (Comuzzi, Patel, 2016). Also, the size of structured data is much smaller than of unstructured data and the percentage of storage space goes much in favour of unstructured data. However, this doesn't mean that the analysis of unstructured data shouldn't be neglected. Analysts are constantly finding new ways to investigate unstructured data to extract value.

It is known that social media websites are the biggest generator of unstructured data. The size of content grows in large exponential measures and to analyse a specific matter is hard. Companies not so long ago began using social media websites for many purposes, promotion, engagement with customers and even sales of products. Also, companies found out that their customers became very keen of expressing their satisfaction or unsatisfaction with a specific product. However, it was complicated to filter out and find these specific posts that are concerned with their product. Chris Messina, the first person to start using hashtags on social media platforms gave an excellent solution to that problem. The essential idea of hashtags was grouping photos, messages, videos, posts, and everything else, therefore making it easier for everyone to search a specific hashtag and investigate the results. Also, by doing so, companies quickly found the potential in the hashtag as it was one of the solutions of analysing unstructured data from social media platforms. Therefore, companies supported and continued the trend and started pushing their customers to use hashtags and the name of their products so they could easily search and see what customers think about their products or do with their product. It can be seen again that up until the mass usage of hashtags, there were just an enormous amount of data that had no value. Companies had no use of social media websites apart from promotional activities and now, all these unstructured data yield valuable insights.

***The Veracity***

Veracity represents the data quality, context, accuracy and the large amount of sources for data, meaning that it is difficult to understand where it comes from, who the originator is, whether it is accurate/correct and finally, what the meaning of data is (Kunz et al., 2017). As

it was already mentioned, it is hard to distinguish relevant data from irrelevant, accurate from inaccurate.

A common example can prove this point. Say you went on a trip to a city you have never been before, and you want to eat something. The application TripAdvisor will come in handy in the selection of restaurants because of all the reviews and ratings that can be found. However, the problem comes when hundreds of reviews are displayed, and you don't have the slightest idea whose review should be considered when choosing the restaurant. Some reviews are positive, some negative and a problem can arrive from both sides.

A positive review could be a comment from the owners' friend. And whether the food is good or not, there is still the conflict of interest which makes the comment highly irrelevant. On the other hand, a negative review can come from the competitors' side which will also make the comment irrelevant, whether the food is good or not. An accurate and relevant review can only come from a person that matches your gastronomic taste and is willing to spend a close enough amount of money on food.

***The value***

The value of data itself represents the potential to improve the business performance within the context of data analytics and business intelligence (Chen et al., 2012). When it comes to big data however, for a data to be valuable, value must be assigned to that piece of data. What this means is that this valuable piece of data needs to be found, analysed and only then it can be seen if the data is valuable. After the right set or piece of data has been found, the company can get useful insights about their customers or products. Then, a company can use these data to offer the right promotions at the right time or improve their products or service accordingly.

Another term that is quite important regarding the value is the return on information (ROI). The return on information represents the outcome of using data and shaping the business process accordingly (Kunz et al., 2017). A low return on information points to the insignificant benefit of the business process that was shaped by that piece of data considered to be valuable. In the case of low return on information, it can be concluded that either the data wasn't valuable enough to improve the business process, or that the data was used in the wrong field of business.

## *2.5. Big data platforms*

### 2.5.1. History of big data platforms

The need for big data platforms dates a long way back. Since the very beginning of tracking records such as tax reports or population records, the process of organizing large collections of data was always time consuming and the demand for a more sophisticated method was large. And even though the term big data didn't exist, it doesn't mean the size of data people handled back then wasn't big.

Big data as a term was not coined up until the mid-20$^{th}$ century. At that time, people were well aware of the ever-growing size of data created by human actions and computers. By that time, methods of organizing large sizes of data were created. However, it was in the late 19$^{th}$ century when the first sophisticated method of processing and organizing data was invented. Herman Hollerith, an American statistician, and a pioneer of data processing created a computing machine for punched cards. The computer was able to read and summarize data (data such as age, gender, marital status etc.) based on the holes that were punched on the card. To this day, Herman Hollerith is responsible for the development of the multinational company IBM as well with the development of data processing methods.

Nowadays, there are plenty of big data platforms available. Modern, sophisticated platforms being modified and updated for more than ten years to fit the needs of every company, no matter the size or industry. Some are free, some are cloud based, some fit every company and some fit only large, multinational companies. The variety of big data platforms is already becoming overwhelming making it harder for companies to decide which one suits their needs perfectly. Nevertheless, a large offer of big data platforms increases the competition which can only lead to benefits for each company that uses, or wants to use, big data platforms.

### *2.5.2. Big data platforms today – Apache Hadoop*

Over the last 30 years, growing importance of big data has led to a large demand for a way to process large amounts of data. The emergence of Internet and the commercial use of it started creating amounts of data only supercomputers were able to process. Furthermore, companies such as Yahoo, eBay and Amazon started analysing the customer behaviour. At start, the analysis of customer behaviour was conducted by looking up to data such as click-rates, the

customers location and what they searched for on the website. Those data brought value to these companies which led them to dig deeper into the field of big data.

The result became the creation of Apache Hadoop in 2006. A revolutionary program with an ideal solution to data processing, having an incredibly strong processing power with the ability to store large amounts of data. Apache Hadoop is an open source program which means it is available for everyone to use and modify according to their needs (cwiki.apache.org, last accessed June 2020). At the time Hadoop was released, not many companies thought of using it. After all, the ways of using big data to gain benefits were not exactly known as it is today. Therefore, only large, multinational companies (mostly e-commerce companies such as Amazon and eBay) started using it and experimenting.

As Apache Hadoop was being used, many companies and their data scientists started making modifications with the intent of improvement, adapting it more to the company's big data needs. These modifications led to creations of new big data software solutions. However, the modules Hadoop is based on often remained but with some alternations.

Apache Hadoop has 4 modules which carry out most of its main functions, one of which is the MapReduce (Marr, 2016). The MapReduce is a parallel processing software framework, meaning it performs two tasks at the same time. The words map and reduce both have their meaning and each carries out its own task. The Map carries out the first step, which is reading the data from the database, splitting the data into smaller parts making it suitable for analysis, and finally distributing to the nodes (computers in charge of processing and storing data). The Reduce part enables the user to search and extract data from the database by performing mathematical operations.

However, as much as effective MapReduce is, using it isn't easy and to find someone who adequately knows how to use it can be tricky. On the other hand, SQL, which is another big data solution, doesn't have as complicated procedures and it is easier to find someone to operate in it. Although, SQL isn't a solution for every company as it can store less data than Hadoop and it works only with structured data.

The second module of Apache Hadoop is the Distributed File System. Alongside MapReduce, the Distributed File System carries out one of the most important functions. It is in charge of storing the data across multiple storage devices. Third module of Apache Hadoop is Hadoop Common. In simple terms, Hadoop Common is responsible for enabling the users' computer system to read the data stored on Hadoop. By providing the tools in Java,

Hadoop Common makes it possible for the user to read data, whether the user has Windows or Linux. Finally, the fourth module of Apache Hadoop is YARN. YARN stands for "Yet Another Resource Negotiator" and it is in charge of managing resources of the system for the purposes storing data, running the analysis and Hadoop itself.

No big data platform is easy to use, it requires knowledge and expertise to set up the platform, make it available for use and finally use it. But when it comes to Apache Hadoop in its raw state, it does come up as complicated to use, even for IT professionals. For this reason, other big data platforms were created, such as Cloudera which basically simplifies the task of installing and running a Hadoop system. However, it is still the most used big data solution for a couple of reasons. The software itself is free (although maintaining it is not), it is suitable for most companies, offers powerful processing power and many more. For these reasons, more than half of the Forbes top 500 companies claim they use Apache Hadoop (Marr, 2016).

### *2.6. Big data lifecycle*

Big data lifecycle can be defined as stages data goes through, from the beginning when the data is created and stored, to the end when the data is archived or deleted. It is important to understand the data life cycle because it gives the complete picture of how data should be managed for it to become valuable. The general idea behind data lifecycle is not only to present the stages data itself goes through, but to present the human processes that are involved as well.

The life cycle that will be presented, also called smart data lifecycle (Smart DLC), is divided into thirteen categories, each being equally important, but not easy to carry out. Those stages are: *Planning, Management, Collection, Integration, Filtering, Enrichment, Analysis, Visualization, Access, Storage, Destruction, Archiving and Security* (El Arass et al., 2016).

**Planning**

As in any other business venture, data also requires a great deal of planning. A good plan makes every process more likely to result in success and the same is with big data. Essentially, whatever the data management team decided in this stage will be imbued throughout the whole data lifecycle. In the planning process, the project team will go through each stage and create instructions for themselves on how they would like to carry out the project in the long run. Also, the team usually sets up a specific time frame within the project

needs to be carried out, as well with human and material resources that are needed to complete the project. Usually, the goal of the project is known even before the data lifecycle starts, whether the company wants to conduct this project for purposes of innovation, improvement or exploring. However, according to the main goal that the team wants to carry out, it is specified in this stage which data they want to gain access to, collect and finally manage and analyse.

**Management**

Management of data lifecycle is a part of the planning process; however, it is also something that is connected to all other operational phases from data testing to archiving. Furthermore, it is said that management of data involves any other sort of use or handling of data. Data lifecycle management is equally important as the planning stage because of how often it is needed throughout the whole project. The effective management of data will also increase the chances of the success of the whole project. However, to know how to manage data correctly, it is crucial that the management is performed according to what was decided in the planning process.

**Collection**

Data collection is the first stage where the data itself comes to light. After the generation and storage of data, whether that happened inside or outside (or both at the same time) of the company, authorization to collect the data is needed. It is crucial not to neglect the security aspect of this stage because in many cases, the data analysis team are dealing with information that company wants to protect at all costs. Therefore, the flow of data needs to be enabled to the team in charge of data management but in a secure manner.

As it was previously stated, there are structured and unstructured data and, in many cases, both are needed to conduct a proper analysis. However, when the data is collected, both structured and unstructured data usually demand modification and conversions to make them easier to organize and sort out. Also, it is important to collect as much quality data as possible within a reasonable time frame. The reason why this is mentioned is because when it comes to big data, even when the collection is over, there is still a large chance that valuable data are remaining somewhere on the storage space. Therefore, it is important to collect just enough data to recognize or create a pattern that will lead to valuable insights (El Arass et al., 2018).

**Integration**

When it comes to Big data lifecycle and integration this is the stage that will make a difference in terms of difficulty of conducting the analysis, especially when there were two or more sources from where the data was collected. Integration is also important because it allows the analysts to recognize the pattern within all data more easily. After the collection is complete, the team will find that the data collected is in various forms and formats. Therefore, separating data, organizing, and regrouping plays a large part in the data lifecycle.

Furthermore, organizing unstructured data can turn out to be time consuming and non-efficient which is why the project team needs to collect just the right amount of unstructured data so that it doesn't prolong the project beyond acceptable measures. On the other hand, when it comes to structured data, the team will find and collect this type of data in an organized structure and prepared for analysis (Chaki, 2015).

**Filtering**

In this stage of data lifecycle, the project team needs to ensure that data overflow is restricted. Furthermore, the project team must separate valuable, quality data from the so-called noisy data and errors (El Arass et al., 2018). In many cases, the data management team collects an over excessive amount of data in the collection phase knowing that through filtering, valuable data for that project will be left for further usage. The discarded data can be separated into completely meaningless set of data and meaningful data but for another project.

As in all parts of the big data lifecycle, it is important to carry out the filtering according to the plan. The goal determined in the planning phase will steer the project to choose a set of guidelines which the data management team will follow in order to perform the filtering successfully. Therefore, the team needs to ensure that after the filtering has been made, only data useful for the chosen project remains. However, the team needs to be careful because an excessive amount of filtering can present a problem as well. Therefore, the team needs to find the right balance between what was planned, and which set of data needs to stay for further usage in the project.

**Enrichment**

Enrichment of data in the data lifecycle represents the modification of data with the goal of increasing its value. A piece or a set of data can seem blank at first which is why modification can contribute to enrichment of data. As in stage of filtering, the enrichment of

data must be performed according to the plan for the contribution to be worthwhile. However, the enrichment cannot be performed in excessive amounts because the piece or set of data must keep its original meaning.

**Analysis**

The analysis of big data is the most important of all stages in the data lifecycle. After the previous stages and all the preparation that was performed, this is the stage where every effort that was put into the project must make it worthwhile. Not to discriminate the previous stages, but the efforts that the team made can become worthless if the analysis isn't performed at its best. It is of much importance that maximum value is extracted from the data that was brought into this stage (El Arass et al., 2018).

Before the analysis starts, the planning team needs to define the objectives of the data analysis. Those objectives are generally defined according to specific demands a certain department has. So, for example, the company is planning to release a marketing campaign that involves certain actions such as placing commercials, intensified social media presence and lot of other practices. The company invested a lot of money and resources into the campaign which is why the company wants to conduct big data analysis to make sure to target the right market, at the right time and in the right place. The marketing managers and the big data analysis project team will therefore come up with a list of expectations from the analysis. According to those expectations, the project team will have to select certain methods of data mining, data sampling and analysis that will be used in order to achieve those objectives.

Data mining, a term that is often used in big data terminology, is the process of examining large quantities of data with the goal of recognizing meaningful patterns. When examining large quantities of data, the data analysis team must always keep in mind the predetermined criteria because only then, the discovered patterns will be relevant and valid. Data sampling on the other hand, begins with the process of extracting a set of data from the original group of data that the data scientists collected. Afterwards, the data scientists are able to examine and analyse the extracted set of data within a smaller time frame while still producing accurate findings (IBM Software, 2013).

### Visualization

In the stage of visualization, the data scientists create a summary of the results of the analysis in a way that is quite easy and quick to read. The reason for implementing visualization in the data lifecycle is that the field of data analysis is not common to all managers or employees. Therefore, the original results of the analysis can look like a page filled with letters and numbers without any meaning whatsoever. This is completely understandable which makes this stage very necessary because after all, the managers are the ones that will bring the final decision backed by the results of the data analysis.

When it comes to creating the summary, the team needs to choose the best way of displaying the data analysis results because not every analysis is conducted in the same way, which is why each result demands its own type of summary. Whether the analysis report will be constituted by graphs, diagrams, pie charts, text, or any other sort, it needs to be easy to read and analyse (Khan, 2014).

As it has been said, the whole process of data lifecycle is quite long and can be time consuming. Therefore, each stage, including this one, must be conducted in the most time-efficient manner because when the data analysis report is sent to the management, the management team won't be happy if they lose too much time on analysing the report because the visualization stage isn't performed well enough.

### Access

At this stage, whatever data storage provider the company is using, the data consumer will go through authorization process so the database can be accessed. The data manager, usually a big data platform, will fulfil the needs of the data consumer. Allowing the access to the company's data while securing it is the task of the data manager. Furthermore, allowing the data consumer to find, analyse and receive data is the main objective of this stage (El Arass et al., 2018).

### Storage

As the word itself describes, the main task of the storage is to collect and save all generated data. Also, the storage is programmed to save data according to a specific method that fits the owner, such as saving data and sorting it by the time and date when the data is created. The storage of data comes at the beginning of the data lifecycle. Although, the need for storage will stretch throughout most cycles. Having the ability to save, group, sort, extract and delete

data are functions that every storage enables and each of these functions are frequently used when the analysis is being conducted.

It is said that any storage must fulfil four characteristics for it to be satisfying for any company (El Arass et al., 2018). Also, it was previously mentioned how companies that wish to use the services of a storage provider needs to consider what the best option is, in-house storage or public cloud. These concerns are more based on the costs that the storage service will create and the need for real-time analytics. Another concern companies have when it comes to storage is the security which will be the first characteristic that any storage must fulfil. The level of security that every storage provider offers will differ and depending on the type of information the company deals with; it needs to find the right provider that will offer the company with the right level of security. However, every storage, whether it is a cloud or a hard drive, must fulfil the minimum requirements of security, especially if we talk about a cloud storage.

The second characteristic of storage is the size of the storage itself. Depending on the quantity of data a company generates, the company decides on the size of the storage. The storage must be able to store the massive amount of data that comes every day. It also must be able to keep storing for at least a couple of years before the storage requires either deleting data or increasing the size of the storage, depending on the value the company sees in the data. Some storages can save Terabytes, some can save Petabytes, Exabytes, Zettabytes. Depending on the speed of data generation, hence the time required for the data to fill up the storage space, the company must decide upon the size of the storage space. However, data generation in most cases can be overwhelming which is why larger storage space is usually required.

Third characteristic that a storage must fulfil is flexibility. The flexibility of the storage is defined by the opportunities it gives to its user starting from accessing the storage to the intelligent methods of data positioning. The ease of analysing data can depend on this certain characteristic because flexibility will offer the data analysts to view the data in many shapes and forms. Finally, the collection of data from the storage also comes into this category because to a certain extent, it depends on the flexibility.

Fourth and final characteristic of storage is reliability. The word itself is self-explanatory and to describe why anything must be reliable is quite unnecessary. However, when we think of a reliable storage, this will mean a few things. First, the storage will collect and save all the

data generated. Second, the storage will never lose or corrupt any data that is already saved. Third, the storage will perform automatic data backup often enough to secure no data is lost even if any problem occurs. Finally, the storage must provide the restore capability option which will enable the company to restore any lost or corrupted data. These prerequisites will ensure the reliability of the storage.

**Destruction**

Destruction is a stage in data lifecycle where the usability, value and significance of data comes to an end. After the data has been used in the analysis and brought value to the company, it is necessary to delete the selected data from the storage. There are multiple reasons why it is better to delete the data rather than save it. Primary reason is cleaning the storage and freeing the space for new, more useful data. It is also important to do that because the costs of maintaining the storage could potentially rise if the storage gets over capacitated which will then present the need for more storage space (El Arass et al., 2018).

The second reason why it is better to delete the data used in the analysis is, so it doesn't get mixed up with new data as well with old data that wasn't used in the analysis. When it comes to destruction of data that was never used in any analysis, data importance potential should be distinguished. As it was previously mentioned, one set of data can be useful for one project, and another set can be useful for a different project. Therefore, deleting the data that was separated during the collection stage is questionable because the extraction of value can happen, just in another project.

Another reason why some companies and data management teams don't prefer to delete data is because they want to save the data for future comparisons of old and new data. Also, they believe that the need for using the data will present itself again even though the value was already extracted. However, it can be illogical keep the data on the storage when the data scientists already created analysis reports and summaries that are concerned with that particular set of data, thus, if ever comes the need to view the data again, the value can be found in those reports. But in the end, it all comes to what was defined in the planning stage of data lifecycle.

**Archiving**

Archiving data means long-term storage of data that can potentially be useful again in another lifecycle (El Arass et al., 2018). It was stated above that some companies like to keep the old

data for some future comparisons or for the simple reason of having the history of the data generated and collected. If the company decides to do so, archiving is a smart thing to do because it separates the old data from the new ones. And as it is known, when a company has such a large collection of data, it is crucial to keep the data organized. Therefore, the company can choose to archive the data, for example, in the order of when the data was created. Another possibility is to archive and organize the data according the last use.

After the company decided to archive the data, there are three steps that must be done. Beginning with the so called, long-distance storage. This means that the chosen data will be separated from the main storage into another storage. The benefits of doing so are freeing space in the main storage as well with organizing it in a manner the company wishes. The second step includes encryption techniques that are usually performed by IT experts. The encryption of data is performed for security reasons so that only authorized parties can access the data. Finally, data retrieval mechanism is implemented so that the authorized personnel is able to request and receive any set of data saved in the long-distance storage (Lin et al., 2014).

**Security**

The security in data life cycle is always present. It is also one of those stages that demands special attention because security of data is especially important, especially for some companies. There are multiple reasons why certain companies require top level big data security. Companies that have unique production methods will secure their data at all costs. Also, companies that have loyalty programs which require private information. Those data must always be kept safe and for this specific reason, data masking is usually performed to hide the original content from everyone, even data analysts. All big data storage providers are required to offer a solid security system that will protect, keep, and distribute data in the safest way possible.

So, when it comes to maintaining high security level, three important factors should be considered and those are access control, data integrity and privacy. Access control refers to the process of allowing access to authorized personnel to collect and use data for certain projects. For any project that a company assigns to data scientists, the project team will need special authorization to enter the storage and collect, modify, delete, archive or any other action.

Data integrity is a term that has a broad meaning, but two words specifically will describe it best, accuracy and consistency. The role of the security system is to keep data as is. This means that no change should occur to the original outlook of data. Also, it means that the data should remain unchanged from the moment it arrives to the database till the end of its life cycle.

Finally, privacy of data should be maintained throughout the whole data life cycle as well. As it was mentioned, some companies store highly sensitive data that no one apart from a few persons is allowed to see or know. Whether it is concerned with a secret production method or private information of the companies customers, those data should be either masked, well-hidden or locked under a unique password so that no one, including the data analysts, isn't able to access (El Arass et al., 2018).

### *2.7. Advantages and disadvantages of using big data*

Primarily when data generation started to skyrocket, people saw no use of it and didn't consider using big data to gain useful insights. Then in the early 2000s, companies such as Amazon, eBay and Netflix started using big data analysis to get to know their customers, what customers think about their company and many more. Every company in the world should know their customer base and target market because decisions are made easier when the company knows for whom the products/service are made for.

#### *2.7.1. Advantages of using big data*

**Easier decision-making process**

The first advantage of using big data is easier and smarter decision-making. When a company decides to start another project, big data associated with the project can be evaluated and analysed in order to know how to execute the project. This does not mean that every project or business venture calls for big data analysis, especially those smaller, less risky ones. Big data analysis, as it will be explained, is time consuming and loosing time and money on big data analysis before each project isn't efficient. However, when it comes to important business ventures, big data analysis can highly influence on the success of the project (Satyanarayana, 2015).

**Reduce costs**

Gaining insights from big data can also help in the reduction of costs (Marr, 2016). By using big data analytics, business optimization can be performed which can lead to cost reductions. Initially, big data tools offer data storage at a much cheaper price than the traditional databases. When speaking from a technological point of view, storage of big data alternatives such as horizontal scalability or scale-out offers the opportunity to upgrade the storage capacity when needed. By adding nodes, storage capacity and computing power as well are increased in an efficient manner. However, through big data analysis, it can be found in certain areas that the company was overcommitted financially as the importance of that area wasn't as big as it was thought.

**Gives valuable insights on target customers**

Discovering shopping habits of different demographic characteristics has been one of the most valuable advantages of big data. The analysis of big data has brought many companies better understanding of what their customers like. Furthermore, targeting customers becomes much easier by using big data analytics (Satyanarayana, 2015). Having the information such as customers likes/dislikes is crucial because it massively increases the chance to succeed when releasing new products for example. Each demographic group has certain needs that must be fulfilled and because of big data analytics, fulfilling those needs was never easier.

**Competitive advantage**

As it was previously mentioned, many companies accomplished competitive advantages by using and analysing big data. Big data always offers valuable information, it only needs to be found. Competitive advantage can be achieved through cost leadership and differentiation. Cost reduction was already stated as one of the advantages of using big data and this can lead a company to having a competitive advantage.

Large volumes of data can be stored in a cheap manner which enables companies to analyse these data quickly, and this could finally lead to the increase in product offers for example. A smartly created product offer will increase sales and return on investment and in the end the company will have created a competitive advantage. Furthermore, big data analysis can create many benefits in many different areas. Through big data analytics, a company can come up with new products and services that suits the customers' needs. Also, companies have created new business models after realizing its potential through big data analysis.

Gaining new customer behaviour insights is important as well because the information can always be found in big data. By using those customer behaviour insights, the company can easily increase customer satisfaction which will in turn increase customer loyalty (Satyanarayana, 2015). Sign-ups and personalization of the customer experience are mostly made according to the findings a company has on its customers.

### *2.7.2. Disadvantages of using big data*

When it comes to big data there are challenges that can be hard to overcome. Just the term itself can seem very vague because data without meaning and value is just data. Also, the need for real-time data is huge and inevitable, but sometimes the system doesn't process the data fast enough so it can immediately be used. When a problem arises and must be solved quickly, such as a stock-out situation, data needs to be available immediately to that department so decisions could be made. Getting most of the value from data sometimes means extracting the knowledge from data as fast as possible.

Using big data and the technology it requires can be a risk as it is vulnerable to cyber-attacks and stealing information (Satyanarayana, 2015). There are many companies in this world that generate data which can be useful to others. Google for example collects data from people concerned with everything known on this world. Even though Google uses the most secure database, a hacker skilled enough can pass through the security and enter the so-called cloud. Other companies do not use the best security when it comes to protection of data and are much more vulnerable to these attacks. Cyber-attack happens because of multiple reasons, primarily because of the lack of authentication mechanisms. This is because of the need of real-time data; the availability of data needs to exist in every department of a company needs to be easily distributed and accessed. Furthermore, problem arises because of the lack of use of secure channels for accessing information. In case of someone passes through security and gains access to the data, the company can suffer from stolen data, removing/deleting data from the system, adding false data to the database or finally modifying the existing data.

Another challenge that comes along with big data are people. As stated earlier, data without meaning and the appropriate analysis is just data. Therefore, companies are obliged to dedicate towards building a successful team of data scientists. Finding skilled data scientists can be a significant problem as the demand for the job is high, but the number of skilful data scientists doesn't meet the demand. Data scientists have many roles when it comes to big

data analysis, such as data engineering, statistician, scientific method, advanced computing and many more (Almeida, Fernando, 2017). Also, it is mentioned that data statistician jobs are in large demand nowadays and has the potential of growth in the information technology sector in this century (Tharwat, 2017).

# 3. BIG DATA IN RETAIL

## *3.1. Importance of big data in retail*

The retail business can be characterized in many ways. It is a complex, enormous industry that is a part of peoples' everyday lives. But mainly it is a business which is constantly changing. Any business that goes through many changes throughout the year requires special attention to details because external factors such as seasons, weather, holidays, consumer habits are the ones that provoke organized preparation. These factors are inevitable, and in the retail world, they mean everything. Of course, there are factors that cannot be predicted which is why retailers aren't able to be prepared for everything at any time. Just a single rain that wasn't anticipated can boost the sales of umbrellas. A human error that occurs every now and then can lead to missed sales. Therefore, retailers are forced to anticipate events that are likely to happen and prepare themselves in the right way.

This is where big data comes in handy. Using big data gives retailers a better chance to anticipate these unexpected events. It can serve as a guideline on almost every part of the retailing business. A big data analysis can lead to smaller chance for stockouts, or even excessive stock. It is known that inventory management is one of the most important segments of retailing, and by using big data, it can be performed at a much higher, effective level. Also, using big data can lead to an increase in sales. Retailers found out they can boost sales by creating personalized offers based on the data considering the targeted customer. This way, retailers are making sales that wouldn't have been made if the personalized offer hadn't been presented to the customer.

These are just a few examples of big data usage and it already seems enough to convince retailers to start using big data. Therefore, most retailers that have enough resources to finance and manage big data departments started storing, keeping track, and analysing big data. Primary data retailers look at are the data generated within their own organization. This includes data pulled from the point of sale (POS) device, website activities, store activities, e-mails and many more. Also, retailers pull information from various external sources. Those sources are mostly meteorological data, economic data, telecoms data, social media, gas prices etc (Marr, 2016).

However, as much as useful big data analysis is for the retail industry, it cannot be seen as an answer to every question. It should rather be seen as a tool that enhances business operations.

Or a tool that will decrease chances of failing, rather than giving you a guarantee of success. Relying solely on big data will never be a solution to every issue that comes up. Big data in retailing serves as a hint towards an existing problem Furthermore, it can serve as a hint towards something that can be done better or more efficient. Of course, there are exceptions to this matter, and this can include some straightforward issues which happen from time to time. But when it comes to more complex issues, decisions that are backed up by data have a higher chance of succeeding.

Walmart is the best example when it comes to using big data to enhance retailing activities and stimulate growth. They are the pioneers and most certainly inventors of big data usage activities in the retail world. Walmart has strongly dedicated themselves to finding new ways of using big data to anticipate opportunities in the future. For them to succeed in these activities, Walmart was forced to find the right personnel for the job. And big data being quite a new field whose potential isn't fulfilled to its maximum potential; Walmart encountered a challenge.

Finding adequate, skilled, and educated employees in the field of big data wasn't an easy task. And Walmart being a revolutionary in using big data in retailing, those employees had to be outstanding. To overcome this step, Walmart posted a challenge online. The challenge was to anticipate the influence of promotional and seasonal events, such as stock-clearance sales, on the sales of other, different products. Those who participated in the challenge, had to create a model closest to actual, real-life data gathered by Walmart and those who were the closest, had the opportunity to work for Walmart's data science team (Marr 2016).

### *3.2. Applications of big data in retail*

For years now retailers have been constantly developing new ways on how to use data. Also, retailers have been learning new ways of extracting value and knowledge from data. But most importantly, retailers wanted to apply that knowledge to many different fields of their business so they could create better offers and more suited in-store environment for its customers (Marr, 2016). Lots of resources have always been directed towards achieving customer satisfaction, especially during the last decade. The reason was simple, satisfied customers equals loyal customers. And having loyal customers leads to a long-term stream of money coming in the business from that single customer.

And at the end of the day, increasing sales is what most businesses care about. After pointing the obvious, it is logical that retailers, as many other industries, invest more time, resources and money towards big data applicability.

Personalization for example, is an excellent application of big data to connect with the customer on a more unique and intimate manner. It creates a feeling of familiarity which often leads to sales that otherwise wouldn't be made. Retailers found the opportunity and took advantage of the information received about its customers likes and dislikes. Therefore, retailers often know what their customer wants even before the customer itself. Then, when the offer is presented, the customer gets an instant feeling of rush and satisfaction which causes a substantially bigger chance of selling the product. Especially if the offer is time limited.

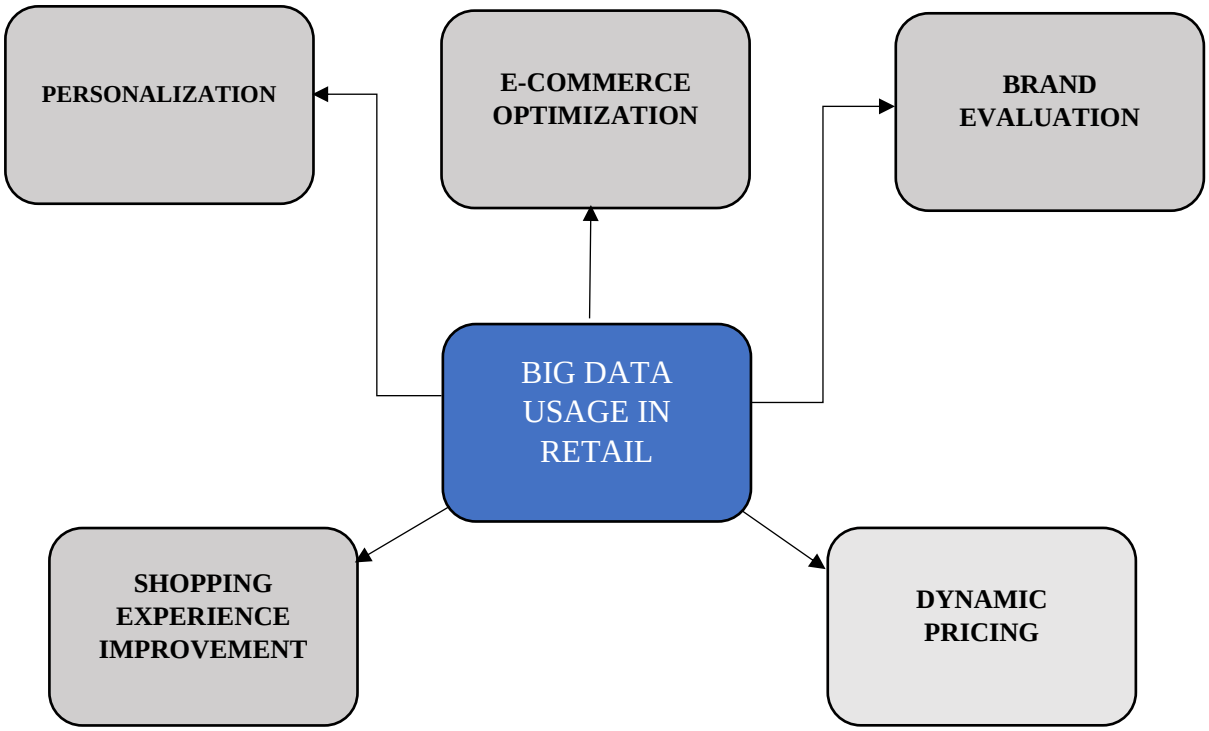
This could spark a moral debate on whether it is right or wrong for retailers to do so. The reason is people who are more prone to impulsive buying habits are more likely to continuously indulge in these types of purchases. In many cases, these offered products do seem necessary, but the reality is, the customer didn't realise that specific need until the offer was presented to them. This could be referred to as creating or pushing artificial needs and demands towards customers in order to sell a product that otherwise wouldn't have been sold. Or maybe would have been sold but in a month, year or two, when the need actually presented itself. In many cases, products bought out of those artificial needs are being used a limited time until a new product comes along that satisfies some other needs.

Additionally, there is the aspect of disrupting the privacy of customers where the customers don't like the feeling of them being analysed to get these personalized offers. Sometimes, customers get the feeling that the retailor knows too much about their purchasing habit along with their personal information such as personal preferences. As it has been said, this could be a debate in the future. Retailers currently have little obligation to ask subtly for the customers permission to gather and use data about them. And even though the main advantage of using the data is to create a better experience for the customer itself, it is possible that customers will start feeling like retailers are starting to cross a line in terms of privacy.

### *3.3. Big data usage in retail*

In the following figure, possible big data usages in retail can be seen.

*Figure 1.* **Big data usage in retail**




Source: QBurst (2015)

**Personalization**

Personalization can be described as shaping a unique experience by offering customized offers to consumers based on demographic characteristics, behavioural targeting, psychographic segmentation and purchasing pattern analysis. It is a method of reaching out to consumers with the intention of selling the right products to the right customer, at the right place and at the right time. Personalization has proven to be a successful method of boosting sales and increasing profits for all businesses, but retailers especially. A study has shown that 22% of retailers use big data to create personalized offers, while 35% are planning on the implementation of personalization (QBurst, 2015).

It is needless to say that personalization exists because of big data. Without the information retailers gather on their customers, personalized offers would remain on the level of offering seasonal products. Before there was personalization, retailers were able to target mass audiences by offering seasonal products such as sun cream during the summer, or hats during the winter. At that time, retailers didn't collect information that much, didn't have the

sophisticated databases and algorithms to collect, analyse and generate personalized offers efficiently to a single customer.

However, since retailers have the access to the demographics of their customers, the likes and dislikes based on the customers purchasing habits and many other characteristics, retailers are able to produce a much more advanced personalized offer. A person that has consistently bought black jeans instead of blue ones is most likely going to receive an offer of a special discount on a new pair of black jeans, rather than blue ones. These offers are being generated even on a more complex level. The database tracks the time and date as well with the products bought on that particular day. Therefore, if a single customer consistently bought a toothpaste on the 15$^{th}$ of every month, it is highly likely that the customer will receive a special offer or a discount on toothpaste around that time of the month.

Now someone would wonder how retailers are able to connect the purchase to the customer. When it comes to online shopping, people are usually required to create an account. Each purchase the customer makes is stored in the database under the name of the customer's account. On the other hand, in-store data collection is performed through loyalty programs. Loyalty programs have existed for a long time and as opposed to the modern loyalty programs we have now, its main goal was customer retention. A customer would receive a form of token which could be exchanged later on for a discount, free product, or something in that manner. Nowadays, the goal of loyalty programs is not only customer retention, but also the collection of data on their customer so these personalized offers could be created. And as it has been mentioned, personalized offers boost sales which is a large advantage for retailers.

**E-commerce Optimization**

A standard brick and mortar retailer will always be more focused on the in-store sales performance rather than web sales. People are still more used to buying groceries in stores and even though that may never change, retailers still had to implement the omnichannel marketing in their business. As it has been mentioned, things can change very quickly nowadays, and it is important to keep up with modern trends. E-commerce optimization refers to creating and customizing a web store in such a way that it is relevant, engaging, easy to use and nice to look at. Also, e-commerce optimization as a term means constant upgrades and updating your web store until it perfectly fits your consumers' needs (QBurst, 2015).

When it comes to evaluating the performance of online stores, retailers analyse big data. By analysing heatmap studies, retailers are able to see which parts of the web shop are most

frequently used (Khomtchouk et al., 2017). Furthermore, retailers are able to see on which page customers spend the most time, which products are frequently clicked etc. The analysis of heatmap studies should be performed every now and then, especially when new assortment of products is added to the web store or when the outlook of the web store is changed. As some of these factors change, so does the consumers behaviour on the web store. Therefore, optimizing the web store according to the feedback presented by big data and these heatmap studies is of large importance.

Furthermore, website analytics is an important part of e-commerce optimization as well. For those that prefer online shopping, the retailer needs to perform a search engine optimization for its online store. After conducting search engine optimization, the web shop will start being more relevant. This means that by searching certain key words connected to the products that the retailer offers, or the retailer itself, the web shop will pop up on the screen of the person conducting the search. Also, analysing which factors boost traffic on the web shop is of large importance because it will let the retailer know which factors are more important, and which are less important. For example, a new, exciting product is released which created large interest and demand from the consumers. The retailer has that exact product in their assortment and started advertising the availability of the product through Facebook. Because a large number of people searched about the product, the advertisement created by the retailer started popping on the Facebook of the people who searched about the product. If the offer from the retailer is good enough, many people will open the advertisement, and some may even buy the product. This is an example of a factor boosting traffic on the web shop, hence why the retailer should continue investing in that Facebook advertisement.

Conversion rates are also especially important when it comes to analysing the performance of a web store. It is the best indicator of either success or failure. The conversion rate can be explained as a percentage of website visitors that completed a desired goal (Saleem, 2019). For a retailer, the conversion rate would then be explained as the percentage of website visitors that completed a purchase in regard to the total number of website visitors. In order to have a satisfying conversion rate, the web store needs to be created with lots of consideration for it to be fully optimized. The retailers must carefully consider the web design. The design of the web store needs to be appealing to the targeted customers and it has to be easy to use. When the web store design fits the targeted customers' eye so to say, the visitor will more likely continue using the site and search through the product assortment. Product assortment

is the most important factor and is the one that will decide whether the conversion rate will be satisfying or not. If the product assortment reflects the visitors searching needs, the visitor will become a customer. The product assortment needs to be relevant, nicely displayed, and consistent.

Finally, the process of purchasing a product needs to be as straightforward and simple as possible. This is crucial because the website visitors can easily be demotivated if the purchasing process turns out to be even slightly complicated. Also, the web store needs to give the sense of reliability because you do not want the customers to start questioning the safety of purchasing the product from your website.

The process of e-commerce optimization is never ending. Data will always require attention and keeping track of the success of the web store will solely rely on the feedback of big data analysis. The conversion rate can always be improved; however, the improvement of the conversion rate will be limited in some cases. A fashion retailer that solely operates online as a web store will have much bigger opportunities when it comes to improving the conversion rate, rather than a brick and mortar retailer that sells fast moving consumer goods (FMCG). A brick and mortar retailer would have to work on changing the purchasing habits of their customers and redirect them to online purchasing. This is not an easy task as people usually prefer to buy FMCG in stores rather than online. Therefore, the web store of a FMCG retailer will usually have a low conversion rate as people visit the web store for the purposes of collecting information rather than purchasing a product. However, if that particular website visitor made an in-store purchase afterwards, this could also be called a successful conversion rate.

**Brand Evaluation**

The goal of every company is to maintain customer satisfaction at the highest level. How customers perceive a company and their product/service is important for the company to acknowledge. Ever since people started using social media platforms, it became easier for companies to acquire knowledge about the customers perception. For this reason, data analysts started gathering information from social media platforms (QBurst, 2015).

Before there were social media platforms, companies couldn’t gather these information as easy as they can nowadays. People are now willing to discuss publicly about satisfaction/dissatisfaction regarding a product/service from a certain company. Furthermore, companies realized that the opinion of certain people such as celebrities and influencers have

a great impact on how people perceive a certain product/service. If a negative opinion of a certain product reaches a lot of social media users, it could have a large, negative impact on sales. In these situations, companies are forced to react quickly, whether that means reaching out to customers with an explanation, or even modifying the product according to the consumers' need. For this reason, data analysts are very dedicated to keeping track of the feedback and analysing the data collected from social media platforms.

As opposed to standard big data analysis (business operations analysis), analysis of big data from social media platforms is usually easier to conduct. Social media websites such as Facebook have created data analytic tools in order to help companies, especially small and medium sized companies that don't have a data analysis team nor enough resources to focus on analysing data. These data analytic tools will measure the success of a page, post or a campaign that a company ran in a specific time frame. It offers a lot of advantages as these analytic tools always function without error, it saves a lot of time and money and the results of the data analysis are always presented in a clear way (analytics.facebook.com, last accessed June 2020).

**Shopping Experience Improvement Using Big Data**

In-store shopping experience is one of the most important parts of the retailing business. It will determine whether the customer will come back to the store or not. The shopping experience is determined by a lot of factors and by combining them, the retailer creates a unique experience to the customer (QBurst, 2015). Furthermore, each customer has their own preferences and needs, and according to those, the customer will choose its favourite retailer.

The store design, external and internal, is the first thing the customer will notice. Depending on the store design, the customer will be able to establish expectations about the store he is entering. A nice and fancy store design will catch the eye of most customers, but it will also give the impression that the products inside the store may be exclusive or overpriced. On the other hand, a warehouse alike design will not be very appealing to a customer, but the customer will most likely get the impression of affordability. Also, retailers will differ in other aspects such as product assortment (product width and range), size of the store, location of the store and the services they offer inside the store.

However, regardless of the store design, product assortment and all other features, most large retailing companies will analyse the data generated and collected inside their stores. The analysis of in-store data plays a particularly important role as it can shape the way a store

looks, the way it is organized, the number of employees working on certain days and many more. Retailers found multiple ways to monitor the behaviour of their customers and collect data to detect potential spots for improvement. Room for improvement can often be found in certain aspects such as shelf layout or product arrangement.

Video analytics is one of the more effective methods of collecting data. It plays a large role in analysing the consumer behaviour inside the store. By analysing the video footage collected from the stores' cameras, the retailer can identify the so-called hotspots (QBurst, 2015). Hotspots is a term used to describe a place that is frequently visited within a store. Furthermore, a place within a store where people spend most of their time can also be referred as a hotspot. As soon as hotspots within the store are identified, retailers will use the opportunity to find ways of using them to their advantage. Increased promotional activities on the hotspot is one way of exploiting the benefits the spot offers.

Presenting new products, deals and offers is always effective on the hotspot. The retailer could also push products that did not sell, or products with close expiry date could be sold at a discounted price.

Another method of how retailers collect data is at the point-of-sale (POS), the spot where every customer concludes its shopping. Point-of-sale records are by far the most extensive collection of big data a retailer could have inside their store. It saves all information regarding every product that an individual customer purchased. And when it comes to analysing the POS records, there is a lot of possibilities for the retailer. The retailer can easily measure the success of a promotional activity by analysing POS records by comparing the number of people that bought the particular product opposed to those who did not. Furthermore, it is also useful for the retailer to find out which part of the day, week or month of the year is the one that when most sales are made.

Furthermore, getting to know the customer that walks in the store is crucial for the retailer. And as it was mentioned before, loyalty programs play an important role because then the retailers are able to connect the shopping cart to the customer. Grouping customers in terms of age and comparing their shopping habits can prove useful. A promotional activity that is targeted towards a certain group of people will prove to be more successful if the retailer gets familiar with the groups' shopping habits. Another information retailers like to analyse is the number of discounted products customers bought during one of their shopping. That way, retailers measure both attractiveness and effectiveness of selling discounted products.

Furthermore, it will tell the retailer in what measure their customers care for discounted products.

**Dynamic pricing**

Dynamic pricing is a method frequently used by retailers to achieve different selling prices for the same product. It is an intelligent way of maintaining a healthy inventory level yet earning extra revenue. Brick and mortar retailers aren't exactly able to carry out dynamic pricing in the same way e-tailers can. Offering discounts on certain products will reach every customer that walks inside the store, opposed to e-tailers' method of dynamic pricing where customers receive different offers.

Dynamic pricing demands constant data analysis for it to be successful. Furthermore, following market trends (supply and demand) and inventory level will also provide dynamic pricing with relevant information for it to decide whether discounts should be given or not (QBurst, 2015).

Also, the retailer can rely on the dynamic pricing system without worrying that a product will be sold for less than it should be. For each product, the retailer sets a minimum and maximum asking price. The minimum price of a product usually holds a tiny mark-up which will allow the retailer to cover the costs of listing the product and delivering the product to the customer with little or no profit left.

There are several reasons why the dynamic pricing system gradually decreases the selling price of a product to the minimum. Firstly, it could be the case that the product was both unattractive and overpriced. This case easily leads to a low demand, or no demand at all which will eventually lead the dynamic pricing system to offer discounts or gradually decrease the price until a customer becomes satisfied with the offer or price. Usually when this case occurs, it is the retailers' fault and there could be multiple reasons for that. One of the most common mistakes regarding this case is offering a product that does not fit the retailers' customer base. Another reason may simply be that the retailer wanted to try something new and had higher hopes for the product.

These mistakes have no correlation with any of the external factors that can impact the sales of a product. Economic factors and changes in market trends, social factors, legal factors, technological, all these external factors can influence poorly on sales. The retailer in this case cannot be blamed, especially if these events were unpredictable. In this case, whether the

product is usually in high or low demand is partially irrelevant because eventually, price drops will occur as well with offering discounts. Finally, a product often reaches its minimum selling price because of an excessive inventory level, and an excessive inventory level of any product is or could be the outcome of all the reasons stated above.

On the other hand, for a product to reach its maximum selling price, a couple of factors often needs to be satisfied. First, there is the case of a brand-new product that was largely expected by many people, hence creating a large demand where a certain percentage of people are willing to pay the maximum selling price. After some time, depending on the demand for the product, the dynamic pricing system will start decreasing the price to attract new buyers that weren't able/willing to pay the maximum selling price. Furthermore, in the case of more exclusive products where the retailer isn't able to order a large amount, the price of the product will often remain at the maximum selling price or close to it. External factors can also weigh in and increase the prices as in the case of reaching the minimum selling price.

The dynamic pricing system is an extremely useful tool. Retailers that can implement it have a guarantee that the price of their products will always be maintained at the realistic maximum selling price depending on all the circumstances mentioned above. It is extremely dependent upon constant data analysis and collecting data from various sources. In terms of big data utility, this is a highly advanced mechanism. Since every e-tailer is open 24/7, the dynamic pricing engine works constantly, collecting new data while analysing internal and external factors. It tracks the time when the web store is the busiest in order to increase prices when needed or decrease prices when the web store traffic is low. Also, it tracks shopping habits and activity of each user on the web store, and it offers discount more frequently to loyal customers. Each minute, the dynamic pricing system examines millions of variables without fault.

However, dynamic pricing does have a couple of disadvantages. Each customer of a company that uses dynamic pricing can be a subject of price discrimination. A customer, especially a loyal one, can easily be irritated if he found out he paid for a product more than his friend. Therefore, the customer can feel that he was cheated. Furthermore, this situation could jeopardize the loyalty of the customer and losing a loyal customer is a huge loss to any company.

Finally, a careful customer that is prone to exploring its choices and analysing prices of the same product on different web stores will find out which retailers are prone to overpricing its

products and which retailers have realistic prices. Once a potential customer finds out where a certain product was more expensive, it is most likely that the customer won't consider the web store next time. These disadvantages can be controlled by carefully selecting the minimum and maximum selling price. Also, comparing prices to the prices of the competitor is extremely important in order to adjust prices if necessary. Finally, loyal customers should be rewarded with more frequent discounts. By doing so consistently, the retailer will most likely avoid any inconveniences.

### *3.4. Case studies of successful usage of big data in retail*

**Walmart**

It is widely known that Walmart is the largest, most successful brick and mortar retail store. By offering prices lower than their competitors while providing services others did not, Walmart achieved global success earning more than any other retailer in the world. In less than 60 years, Walmart expanded rapidly to 27 countries, having around 11,500 stores, and employing over 2.2 million associates worldwide (corporate.walmart.com, last accessed June 2020).

Each week, Walmart serves around 260 million customers. A truly astonishing fact, especially when taken into consideration that Walmart records more customers than the worlds' population within a year. Having said so, it comes to no surprise that Walmart earned 524 billion of dollars in the fiscal year of 2020 worldwide (corporate.walmart.com, last accessed June 2020).

As a leader of the retail industry, Walmart has proven to be one of the first retailers to notice the value data can bring to a company. Analysing data as a whole, separating it into different business segments and making decisions based on the data analysis results has proven to lead to increases in sales, store traffic, growth, and many other positive outcomes.

However, having millions of products sold to millions of customers each day creates excessive amounts of data even for a company as large as Walmart. Therefore, managing, processing, and analysing all the data that comes into the company is no easy task. In fact, the rate at which the amount of data that was generated within a day was beginning to be so excessive that Walmart saw the need to commit a large proportion of resources towards data management. It was in 2011 when Walmart established Walmart Labs led by the "Fast Big Data Team" (Marr, 2016).

**Walmart Labs**

Walmart Labs, also referred to as the Data Café, is located in Bentonville, Arkansas. Originally, Walmart Labs was called Kosmix, formerly established and owned by Anand Rajaraman and Venky Harinarayan somewhere near the beginnings of year 2000 (walmartlabs.blogspot.com, last accessed June 2020). Rajaraman and Harinarayan built a rather large and successful company that offered a powerful search engine. It was able to present the most relatable data to the keyword that was searched across various, yet relevant platforms.

In 2011, after fully realising the potential and the positive impact data led decisions have, Walmart decided to acquire Kosmix. Furthermore, in 2013, Walmart decided to acquire a start-up called Inkiru, a company that dealt with predictive intelligence (Rao 2013, last accessed June 2020). Therefore, by merging Kosmix and Inkiru, Walmart created a powerful data management team with strong analytic capabilities. In 2015, Walmart announced its intention to build the world's largest private data cloud, able to process 2.5 petabytes of information every hour. If Walmart succeeds in doing so, it will enable them to handle data storage all by themselves, without the help of other data storage providers.

All these initiatives, the acquisitions and forming of a strongly capable data management team covered the aspect of business that happened within the company, meaning the data stored within the server was generated by actions such as sales, inventory management, supply chain and logistics management. Covering all these were the primary objective as it should have been. However, Walmart recognized the importance of data generated from outside the company as well. Data that concerned Walmart, maybe not in an obvious manner, but when filtered and grouped together, it carried a lot of value, especially when it comes to predictive intelligence.

This is when Walmart decided to create the Walmart's Social Genome Project (Marr, 2016). This project enabled Walmart to keep track of all the posts from various, most popular social media websites. Data scientist then have the ability to quickly analyse posts where Walmart was mentioned in order to predict their customers purchasing needs. Furthermore, monitoring these posts helps Walmart solve customer dissatisfactions, find out opinions on certain products, store experiences etc. Walmart even developed a search engine which allows their data scientists to analyse search inputs on their official website.

When we look upon all these initiatives Walmart took in the last ten years, it becomes clear how much belief Walmart has in the field of big data. As the world's leading retail company, investing into the infrastructure big data demands, acquiring big data companies, hiring highly skilled data scientists wasn't easy and surely not financially insignificant. Today, Walmart Labs employs around 6,000 employees, an enormous number which only proves the massive amount of data Walmart generates. Even the cost of maintaining and updating Walmart labs surely brings significant financial weights.

However, the effort Walmart took towards big data analysis only confirms its importance. It proves the value it can bring to a company despite all the costs. And by looking at the example of Walmart, one shouldn't think that the use of big data only applies to large companies. The costs of setting up the infrastructure as well with the costs of maintenance, employment etc varies upon scalability and the size of the company. But the value big data can bring to a company, no matter the size, varies only upon the quality of big data usage.

### *3.5. Using big data to improve retail performance - Walmart*

"If you can't get insights until you've analysed your sales for a week or a month, then you've lost sales within that time." By saying that, Walmart's senior statistical analyst Naveen Peddamail, wanted to point the important correlation between big data and timely analysis. Keeping track of data and conducting analyses helps Walmart prevent losing sales that otherwise would have been lost. Therefore, because of Walmart Labs, Walmart managed to reduce the time it takes from a problem being spotted, to a solution being proposed from approximately two, three weeks to just 20 minutes (Marr, 2016).

A perfect example was given to prove the inevitable correlation of big data and timely analysis. There were multiple times when the sales of a certain product were declining. As soon as this decline was noticed, the issue was sent to Walmart Labs where analysts acted quickly to resolve that issue. Within a day, results of the analysis were presented saying a pricing error occurred. The error was simply corrected; hence the price went back to normal and sales recovered within a few days (Marr, 2016).

Another example worth mentioning happened during one Halloween. Walmart's data scientists spotted that sales of Halloween themed cookies were selling well in most locations, whereas there were a couple of stores that weren't selling those cookies at all. A message was sent to those stores, informing them about the issue. The issue was quickly resolved after

store employees found out cookies weren't even put on the shelves. Following these examples, time truly is money (Marr, 2016).

At Walmart Labs, a lot of time is devoted towards predicting certain events, traffic intensity and many other figures. In general, predictive analytics plays a big part in maintaining efficiency of business operations. Predictive analytics is concerned with certain aspects of business such as sales predictions, customer traffic and inventory levels. Tracking changes in weather forecasts have large impacts on sales of certain products as well as inventory level of certain product categories. A rainy weather forecast will most likely boost the sales of umbrellas, raincoats and other products that keeps you from getting wet. Therefore, every Walmart in the areas affected by rain needs to ensure the right quantities of these products in order to avoid stock-outs.

It has been said by Walmart Labs that they pull data from 200 various sources. Sources such as meteorological data, economic data, telecoms data, social media data, gas prices, even a database that tracks whether new events near a Walmart store appeared. This fact alone shows the complexity of making any type of prediction. Combining internal data and external data is no easy task, especially when Walmart is the case. Furthermore, Walmart Labs is dedicated mostly to analysing data generated in the past two to three weeks. And in the two, three weeks' time, Walmart generates 200 billion rows of transactional data. When this database is combined with external sources, an immense size of data is created. Therefore, an extraordinary set of skills and capabilities is required to handle such predictive analyses.

One-way Walmart tries to anticipate traffic is concerned with Walmart Pharmacies. By combining data from POS devices and health care database, Walmart wants to predict the number of prescriptions that will be handed during a day and to see which time of the week/month is the busiest (corporate.walmart.com, last accessed June 2020). Doing so, allows Walmart to plan the number of employees working on certain days, working hours and staff schedule in general, depending on predicted traffic. A correct prediction in this case can lead to many benefits both for the company and the customer. From the perspective of the Walmart, extra employees are only paid when they are truly needed which means money is saved. Also, it increases the efficiency of the pharmacy store which leads to an increase in turnover and more sales conducted. On the other hand, the customers will be satisfied because of the fast, quality service.

The same method is applied to improving the speed and efficiency of store checkout. By analysing past experiences and conducting predictive analysis, Walmart is able to determine which days and times of the day will be the busiest (corporate.walmart.com, last accessed June 2020). A precise forecast will allow Walmart to have the ideal number of employees at the store checkout, at all times. By doing so, shoppers get to finish their shopping faster and each Walmart store will have the ideal number of employees. Furthermore, this method is also used to decide the best forms of checkout. Taking into consideration the number of customers per day, average age of customers, type of goods that are sold within the store, size of the store and many more, Walmart decides the number of self-checkout cash registers and facilitated checkout cash registers. This method is used by a lot of retailers as it is known that it can decrease personnel costs.

Another common, yet valuable usage of big data is getting to know the customers and finding out about their shopping patterns and preferences (corporate.walmart.com, last accessed June 2020). A lot of time is devoted to applying this method mostly because it demands constant updates and feedbacks on changes in customer behaviour. Considering the fast paced world we live in, where we have the constant emergence of new products and innovations, disruptive businesses etc. And when combined with new trends and lifestyles that, in fact, change very frequently nowadays, it is no surprise that the behaviour of customer, as well with their preferences change on a yearly basis. Also, many of these changes in trends are actually highly dependent upon new findings in health, science and technology. Therefore, Walmart Labs keeps track of these new findings as they can easily change and decide upon new shopping patterns customers will absorb.

When the customer behaviour analysis is conducted, Walmart is ready to make decisions based on the analysis. It plays a big role in how the shelves look like, how they are stocked, and the way products are displayed. In store marketing is also based upon the customer behaviour analysis. Furthermore, new products are chosen based on the analysis as well. Finally, based on the popularity of certain private brands, Walmart decides which brands will receive a good, visible spot on the shelf and which brands won’t (corporate.walmart.com, last accessed June 2020).

Walmart found new ways of using big data to improve every business segment, and the same was applied to supply chain and logistics. By using data and simulations, Walmart found new ways to optimize transporting routes (corporate.walmart.com, last accessed June 2020). Also,

by tracking routes truck transporters take, Walmart was able to save time and money through route optimization. Truck drivers are also given a work schedule according to availability and distribution of goods. Furthermore, the steps worker takes during loading, unloading, and transporting goods are also counted in order to analyse where improvement can be made.

# 4. EMPIRICAL RESEARCH ON BIG DATA USAGE IN RETAIL

## *4.1. Research method*

The research method chosen for this paper is the qualitative case study analysis method. The main reason why case study analysis method is chosen is because it will appropriately conclude the topic on the big data applicability. Analysing and presenting actual cases of big data usage in retail will certainly determine both impact and importance of big data in retail. Having said so, it will be seen in what measure do retailers assign value to big data. Also, it will be seen in what measure does big data bring value to the retail industry.

Furthermore, to comprehend the actual usage of big data in the retail sector, multiple case studies were selected. By using the selected case studies, this research will demonstrate how does big data usage help different retail companies across the world achieve competitive advantage. Furthermore, it will be interesting to cover the possible differences between the actual usage of big data, or the extent to which retail companies from different countries rely on big data. Also, it will be interesting to see the level of commitment (in terms of resources, data scientist's employment etc.) retail companies have across the globe.

After the research has been conducted, a review will be presented which will mostly consist of case study comparisons. This will mostly apply to searching for differences in the ways how big data is used around the world, if there will be any. Furthermore, it will be determined whether retail companies from different countries use big data for the same purposes. The purpose of finding these differences is mostly concerned with one question, do retail companies have the same knowledge/opinion when it comes to big data applicability.

### *4.2. Research results*

#### *4.2.1. The South Africa case study*

"The use of big data analytics in the retail industries in South Africa" is the first case study which was analysed (Ridge et al., 2015). As the name itself says, its main purpose was to investigate do retailers in South Africa use big data analytics, in which fields to they use big data and for what purpose. As a developed country in Africa, the authors were curious to which extent are retailers familiar with big data, as well with big data analytics.

The case consists of an introductory part where the term big data was explained as well with some other related terms. Terms such as the 4 Vs, or big data platforms were mentioned as they were important for the following part of the case. The authors then saw the purpose to explain many reasons why companies use big data. Reasons such as predictive analytics, price optimisation, inventory management, in-store behaviour analysis and many more, explaining each and every one with the main purpose of presenting a clear and concise overview of many advantages big data can offer to South African companies. Finally, the authors explained the purpose of the case study as well with the method of the research they conducted.

The research method authors decided to use is the interview method, which suited perfectly to the main purpose of the case, which is determining the use of big data in the retail industry in South Africa. Having said so, the authors eliminated the small sized retailers as they do not have the skills, resources, or purpose to use big data analytics. Therefore, the authors concluded only medium and large retail companies will be interviewed. Authors contacted eligible people within companies, mostly managers and IT professionals qualified to talk about the questions they had prepared.

***Big data knowledge***

After the interviews have been conducted, the authors concluded that big data analytics are not used to its full potential. Moreover, all the advantages big data offers are actually being used in a quite limited manner. However, most interviewees are aware of the value big data can offer.

When asked about the definition of big data, discrepancies could be found in the answers interviewees gave to the authors. The majority of people were informed in the field of big

data and did not express any negative connotations, some even referred to big data as business intelligence (BI). However, there were a couple of interviewees that did not show interest in the field of big data. Finally, there were a couple of interviewees that showed cynicism, saying the buzz that was created around big data was with the intention to push the sales of big data platforms to companies.

The interview proceeded to discuss about other big data related terms, one of which is the 4 Vs. First of which is the volume. Interviewees shared the same opinion, upon which the only thing that differed is the size they store on a daily basis. When it came to velocity, it was concluded that no retailer uses real time data monitoring, hence the data processing occurs a while after the data reaches the system.

When asked about variety, the large majority of retailers said that 90% of the data analysed is structured data, by which they mostly meant data such as transactional data (POS data). Unstructured data is being neglected, however, some interviewees said that they are starting to take it into consideration upon future analyses. On the other hand, those that use unstructured data said it was used in order to find out opinions of their customers on certain products/services. Those data are mostly collected from social media websites it was said.

The last one was value, where interviewees, again, mostly talked about how they to generate value, but only from structured data. But in terms of the value it brings to their company, most of the respondents said they struggled to find areas which could be improved by using big data analytics. One area upon which retailers agreed could be improved is targeted marketing, hence using big data analytics to get to know their customers better. Although, the main problem still was the correlation between the value that big data analytics can bring to the company, and the costs of implementing advanced big data department. Still, most of the respondents were aware of the value it can bring into the company (Ridge et al., 2015).

***Big data usage***

The knowledge interviewees had on big data seemed to be in a somewhat mature phase, where most respondents showed awareness upon the values big data analytics can bring. However, when it comes to big data usage, it could be seen that the South African retail companies are still in the early stage of big data usage implementation. Faced by many barriers, the bigger retail companies in South Africa showed limited big data usage.

Sentiment analysis is one of the few areas that South African retail companies use (Ridge et al., 2015). By doing so, companies find out what their customers think about certain products. However, only a few of the responding retail companies actually perform sentiment analysis. The required data for this type of analysis comes mostly from social media websites, e-mails and surveys which are all unstructured data. As it has been mentioned, the analysis of unstructured data is conducted by very few interviewees and most of it is done through external sources. The insights gained from this type of analysis can be very favourable for retailers as it can be used when deciding upon choosing new products, or simply if the retailer wants to find out which group of people share the same affections or interests towards a certain product.

Data mining has also been mentioned as a way of using big data to gain valuable insights. As it has been reported, it is used in various fields, such as monitoring refrigeration to decrease energy consumption, or improving marketing campaigns. Another benefit of data mining comes for the purposes of predicting the number of staff members required for a particular store, depending on the store traffic. Although these data mining techniques are useful, they cannot be seen as something advanced in terms of big data analytics. As reported by the interviewees, most of the retail companies use data mining regularly.

Another method of using big data which was mentioned was exception reporting. Exception reporting can basically be seen as an alarm system to the company, or an indicator which shows that a certain business process isn't working properly. It is used in many ways and if the system is well established, it can monitor lots of different areas such as the security system or inventory levels and many other things, which is why most of the interviewees reported they used it.

Finally, the interviewees mentioned using predictive analytics and statistical analyses. Most companies, especially large retail companies perform predictive analytics. When planning a large business venture which requires lots of resources, companies like to have a data team able to perform predictive analysis to minimize the risk of failure. Statistical analysis is performed more often than predictive analysis and the majority of interviewees confirmed performing it on a weekly/monthly basis.

### *4.2.2. The Italy case study*

"Big data for business management in the retail industry" is a case study with investigative intentions, trying to reveal the big data usage across retailing companies in Italy (Santoro et al., 2019). It's main purpose was to find out to what extent do Italian retail companies rely on big data analytics, data based decisions, in which fields are data based decisions implemented and much more. Italy being a developed country in Europe, the authors had no dilemma whether big data analytics was used. The question was mostly to what degree do retailers use big data.

The case study was introduced by a discussion about the increasing competition among many different industries. With the help of digitalization and the use of big data, many companies achieved competitive advantage by increasing firm efficiency and performance. Also, it was pointed the fact that the retail industry has become competitive over the last couple of years and will become even more competitive in the future with all the cost/price reductions retailers offer. Afterwards, the authors explained the structure of the case study, as well with the method of research that was chosen. Just as in the South Africa case study, the authors thought that the interview approach would suit best for the type of research they intended to conduct. Given their circumstances, the authors were able to contact five marketing managers from five different Italian retailing companies. One of the five marketing managers was eventually left out of the discussion as soon as they realised the retail company, he/she worked for didn't use big data analytics.

In order to paint a perfect picture, a theoretical background was given, where a summary of all the most relevant information concerning the field of big data were presented. Afterwards, the discussions they had with the marketing managers were shown, where the authors summarized the key parts, mostly concerned with big data usage. Finally, the interview results were analysed, and the authors gave their final opinion, as well with the conclusion.

***Big data knowledge***

When reading the case study, it soon became clear that the interviewees were highly informed in the field of big data. Although the interviewees were marketing managers with no or little IT background, their knowledge about big data proved the growing importance. Also, showcasing the need to adapt and learn new skillsets seemed to be crucial when managing large companies.

The discussion on the topic of big data usage was flowing smoothly as both sides were well informed and throughout all four interviews, no scepticism could be noticed. On the contrary, interviewees were aware of all the values big data could bring and showed an open mind towards new ways of big data usage. Therefore, the interviewees showed no discrepancies when the describing their views on what big data is.

When talked about the 4 Vs of big data, the interviewees, again, showcased a perfect example when it came to big data management. Each retail company deals with large volumes of big data. Furthermore, it was said that all four retail companies collect, manage and process data internally at a highly effective way. Having said so, employing skilful employees were no easy task, still, all four companies managed to assemble a highly skilled data science team.

The discussion about velocity, or the speed of processing data again, showed that the retail companies had advanced big data management, having both real-time analysis and batch processing analysis. As any modern company that exploits the values of big data, having real-time analysis gives many opportunities and most importantly, reduces the time to spot any problems significantly.

The variety of big data these Italian retail companies use show advanced big data processing as well, taking both structured and unstructured data into consideration. Even though structured data is usually the type companies are more focused on, unstructured data seemed as important as structured data. Finally, it was talked about value. The value extracted from big data is what makes everything worthwhile. And these companies extract value in many different ways to apply it in many different fields (Santoro et al., 2019).

***Big data usage***

As reported by the authors, all four Italian retail companies showed great big data knowledge, and even better examples of big data usage. Upon many areas that are improved by big data analytics, the retailers highlighted marketing and logistics as two key areas, having great potential to be even more improved by big data analytics in the future. However, implementing data-based decisions in other areas was said to be the new driver of achieving competitive advantage (Santoro et al., 2019).

All retailers agreed on one thing, big data plays a big role in the retail business. Their devotion towards big data analytics definitely proves it, as well with the fact that most of their business strategies and decisions are supported by data. Also, it was even mentioned

how data-based decisions led to cost reductions in many areas. When making important decisions, managers will take big data analysis results into consideration before making a final decision as it reduces the risk of failure. Furthermore, it was noted how experience and intuition most managers possess does not necessarily affect the final decision as much as the objectivity of data and the analysis. The attitude of supporting decisions based on past experiences has become less convenient as the reliability of data grew.

Segmentation of clients was also one of the key areas where the Italian retailers used big data. By analysing past purchases through loyalty programs, interviewees reported how easier it is for them to decide upon new products. Also, it was mentioned how big data analytics allows them to differentiate in order to reach out to new customers. Differentiation is never a safe way to reach to customers, however, with the clever use of big data, retailers have a bigger chance of successfully implementing it. Furthermore, customer targeting was proved to be more accurate. Having the ability to understand various customer profiles gives retailers a better opportunity to direct all sorts of promotions. Also, the interviewees mentioned it enabled them to adapt their offerings without many difficulties.

Effective management of commercial channels can also be achieved by clever big data usage. All interviewees said how they were able to improve the firms' position within the distribution chain, making them more competitive on the market and getting the opportunity to offer products at lower prices.

Furthermore, optimizing firms processes as well with logistics optimization was mentioned as areas that can be significantly improved by big data analytics (Santoro et al., 2019). By tracking routes their trucks take, retailers were able save time and money. Interviewees also mentioned how they were able to decrease operating costs by improving the efficiency of deliveries, inventory management and store efficiency. Another area where the Italian retailers managed to reduce cost is in the human resources department. It is a common practice in the retailing business to track store traffic on a daily basis so that unnecessary work force is avoided. When all these cost reduction methods are combined, the savings become largely significant, giving the retailer the competitive edge needed to outperform others.

Finally, the interviewees reported using predictive data analysis to improve and enhance planning processes. Opening new stores is an activity most retailers undergo on a yearly basis. Especially when a retailer is expanding on a different market. In those situations,

retailers try to predict information such as the approximated store traffic depending on many different demographic and economic factors. That way, retailers are able to create accurate budget forecasts according to the size of the store.

### *4.2.3. Research findings – Case study comparison*

The following table displays the differences in big data usage.

*Table 2.* **Differences in big data usage**

| DIFFERENCES IN BIG DATA USAGE | |
|---|---|
| **SOUTH AFRICA** | **ITALY** |
| Retailers do not use real-time analysis | Retailers use real-time analysis |
| Retailers do not analyse unstructured data | Retailers analyse unstructured data |
| Retailers do not use big data to optimize business processes | Retailers use big data to optimize business processes |
| Retailers outsource most of the big data analytics | Retailers perform big data analytics themselves |
| Segmentation of clients is not performed upon big data analysis findings | Segmentation of clients is performed upon the findings of big data analysis |

Source: Ridge et al. (2015), Santoro et al. (2019)

The two cases that were reviewed both presented different results. From the South Africa case, it could be seen that the knowledge they have upon big data is quite advanced, yet the usage of big data analytics is still in an early stage (Ridge et al., 2015). Faced by many barriers, the South African retailers are struggling to find both resources and reasons to take big data usage to the next level. On the other hand, the Italian retailers showed a great example of how big data should be used. Resources Italian retailers invested into big data management and analytics paid out of for them, leading to cost savings and performance improvement.

The responding South African retailers reported that the big data usage is still in a premature phase. Big data analytics conducted internally is still performed in a basic manner, without a data team dedicated towards advanced analytics that would give companies a competitive edge. Furthermore, the analysis of unstructured data is conducted externally, which also indicates on to low dedication towards data analytics. Although big data can offer retailers a huge advantage, it can also bring costs and barriers that can discourage companies from using it. This exact problem seems to be bothering the interviewees from South Africa as it was reported that they are still questioning the trade-off between investment costs and business performance improvement (Ridge et al., 2015).

First barrier South African retailers are facing is the substantial capital investment big data requires. Investments into big data platforms, employment of data scientists and then paying their wages is a long-term commitment South African retailers are not ready for. Second barrier faced is the lack of available analytical skills. Finding skilful employees, educated in the field of big data and retail seems to be a huge problem. Educating their employees is another option, but this requires time, money and most importantly, the will of their employees to take such a step. Outsourcing big data analytics seems to be a valid option in these circumstances, but the question regarding costs would still be present. Having to justify the costs of big data deployment, however, still seems to be the biggest issue.

When it comes to the Italian retailers that participated in the case study, it could be seen they have already developed an advanced big data management system. Used in many different areas, big data analytics increased company efficiency and has brought many costs down to a minimum (Santoro et al., 2019). Of course, at one point in time they also faced barriers that had to be overcome by investing a significant amount of time, money, and resources. It is interesting to note that the responding Italian retailers found an excellent solution towards finding skilful employees. It was reported that at the beginning of big data deployment, finding employees with the right skillset was hard. To overcome that, retailers started collaborating with universities, posting job offers with required skillsets for the job. Attending students would then see the offer, educate themselves knowing that a job will be waiting for them. This method could be used in South Africa as well and would help solve the problem of finding skilful employees.

Furthermore, Italy being a different, more competitive market with more demanding consumers, it was necessary to take the steps of big data deployment. Still, big data usage proved to be extremely beneficial for them. Also, by looking at this example, the trade-off between the benefits big data offers and the investment costs proved to be fair. Also, when comparing the Italian retail industry with the South African, it could be seen that the level of competition is much fiercer in Italy. The evidence to prove that comes merely from the fact that most large, Italian retailers had to adapt to big data usage as soon as possible to keep the competitive edge. Whereas in South Africa, no retailer indicated the possible implementation of internal big data analytics in the near future, as the barriers are hard to overcome (Ridge et al., 2015). Having said that, it is possible that as soon as one South African retailer implements a more advanced big data usage system, the others will follow.

### *4.3. Limitations and future research*

The comparison of big data usage in South Africa and Italy was conducted in a qualitative manner. Furthermore, the cases upon which the comparison was also made in a qualitative manner. Having said that, it would be especially useful to see this comparison being made in a quantitative manner. If, for example, a quantitative analysis was presented in the Italy case study, then companies from other countries would be more able to see whether implementing big data analytics would pay off, taking into consideration the financial situation of a company of course, as well with the costs of implementing big data analytics.

There is no denying that big data brings much value to any company, and as it was seen from the Italy case, implementing big data analytics does improve the business performance overall. However, a question did arise while the South Africa case was analysed, and that is, how far behind is the South African retail industry compared to the Italian? If so, would the Italian retail companies still recommend the South African retailers to implement big data analytics as soon as possible? Finally, to what extend should the financial performance have influence when deciding upon big data implementation? These questions would help solve the problem whether South African retailers, or from any other country in fact, are ready to implement big data analytics. When said ready, it is meant solely on the trade-off South African (and many other) retailers are questioning, the need for big data analytics compared to the costs it brings. Therefore, a quantitative research would definitely help solve the problem.

For further research, it would be extremely useful if a quantitative research were conducted. Presenting all the costs that big data implementation brought to a company would be helpful for others to see. Furthermore, it would be even more helpful if other companies were able to see to what extent does big data usage influence on cost reductions. Of course, the cost reductions would vary depending on the size of the company and their cost structure, however, if actual results were presented, other companies would have a better clue whether or not implementing big data analytics would pay off for them.

# 5. CONCLUSION

Big data is a term that was extremely popular in the last decade, however, after seeing its importance, it can be concluded that big data will be even more popular in the next decade. The number of jobs offered for data scientists is on the rise and many companies haven't even implemented big data into their everyday business, especially when talking about small and medium sized businesses. Having said that, it is clear that the development of the field of big data won't slow down. Furthermore, it can be said that the field of big data isn't being used to its full potential, yet. New methods of using, managing and analysing data are being made up every day. And not only that, there are many industries that still haven't found a way to use big data, yet I believe that at some point, they will. The impact it makes in many industries is too significant for it to be neglected.

One of many proofs how big data shaped a certain industry can be seen in retail. As it was seen from this paper, it is safe to say that big data improved many business processes, making it more efficient, time and money wise. When we talk about marketing and logistics especially, big data gave an immense amount of opportunities to retailers. Also, that added bonus in customer service which gave the competitive edge to companies using big data analytics.

However, when it comes to big data, there are lots of barriers to overcome. Many companies struggle to find the resources to invest into big data implementation. Implementing big data into business processes cost time and money, but most importantly, skills and education. Finding educated workers has been a problem mostly because the field of big data is so new. Also, it isn't easy to find a data scientists that are also an expert in the designated field he/she is working in. Nonetheless, any company that successfully implemented big data usage achieved a competitive advantage. Overcoming those barriers is something that will be seen more in the near future, which is why the dedication towards the development of big data will continue.

# 6. References


[1] Akter, S. and Wamba, S.F. (2016), "Big data analytics in E-commerce: a systematic review and agenda for future research", *Electronic Markets, Vol. 26 No. 2, pp. 173-194.*

[2] Almeida, Fernando. (2017). Benefits, Challenges and Tools of Big Data Management. *Journal of Systems Integration. 8. 12-20. 10.20470/jsi.v8i4.311.*

[3] Anuradha, J. (2015), "A brief introduction on Big Data 5Vs characteristics and Hadoop technology", *Procedia Computer Science, Vol. 48, pp. 319-324.*

[4] Chaki S., 'The Lifecycle of Enterprise Information Management', in *Enterprise Information Management in Practice, Springer, 2015, pp. 7–14*

[5] Chen, H., Chiang, R.H. and Storey, V.C. (2012), "Business intelligence and analytics: from Big Data to big impact", *MIS Quarterly, Vol. 36 No. 4, pp. 1165-1188.*

[6] Comuzzi, M. and Patel, A. (2016), "How organisations leverage Big Data: a maturity model", *Industrial Management & Data Systems, Vol. 116 No. 8, pp. 1468-1492.* https://doi.org/10.1108/IMDS-12-2015-0495 (last accessed June 2020)

[7] El Arass, Mohammed & Souissi, Nissrine. (2018). Data Lifecycle: From Big Data to Smart Data. Available at: https://www.researchgate.net/publication/328769944_Data_Lifecycle_From_Big_Data_to_Smart_Data (last accessed June 2020)

[8] Hencschen D. (2013), „Big Data Success: 3 companies share secrets", available at: www.informationweek.com/big-data/big-data-analytics/big-data-success-3-companies-share-secrets/d/d-id/1111815 (last accessed June 2020)

[9] IBM, 'Wrangling big data: Fundamentals of data lifecycle management', 2013, Available at: https://edisciplinas.usp.br/pluginfile.php/263618/mod_folder/content/0/Fundamentals%20of%20Data%20Lifecycle%20Management.PDF?forcedownload=1

[10] Intel IT Center (2013), Big Data in the Cloud: Converging Technologies Big Data in the Cloud: Converging Technologies, *Intel IT Center*

[11] Khan N. et al., 'Big data: survey, technologies, opportunities, and challenges', *Sci. World J., vol. 2014, 2014*.

[12] Khomtchouk BB, Hennessy JR, Wahlestedt C. shinyheatmap: Ultra fast low memory heatmap web interface for big data genomics. *PLoS One. 2017;12(5):e0176334. Published 2017 May 11. doi:10.1371/journal.pone.0176334*

[13] Kunz, W., Aksoy, L., Bart, Y., Heinonen, K., Kabadayi, S., Ordenes, F., Sigala, M., Diaz, D. and Theodoulidis, B. (2017), "Customer engagement in a Big Data world", *Journal of Services Marketing, Vol. 31 No. 2, pp. 161-171.* https://doi.org/10.1108/JSM-10-2016-0352 (last accessed June 2020)

[14] Lin L., T. Liu, J. Hu, and J. Zhang, 'A privacy-aware cloud service selection method toward data life-cycle', in *Parallel and Distributed Systems (ICPADS), 2014 20th IEEE International Conference on, 2014, pp. 752–759.*

[15] Lycett, M. (2013), "Datafication: making sense of (big) data in a complex world", *European Journal of Information Systems, Vol. 22 No. 4, pp. 381-386.*

[16] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011), Big Data: The Next Frontier for Innovation, Competition, and Productivity, *McKinsey Global Institute*

[17] Marr B. (2016), "Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results ", West Sussex, United Kingdom, John Wiley and Sons Ltd

[18] Mousannif, H., Sabah, H., Douiji, Y. and Oulad Sayad, Y. (2016), "Big data projects: just jump right in!", *International Journal of Pervasive Computing and Communications, Vol. 12 No. 2, pp. 260-288.* https://doi.org/10.1108/IJPCC-04-2016-0023 (last accessed June 2020)

[19] Olsson, N. and Bull-Berg, H. (2015), "Use of big data in project evaluations", *International Journal of Managing Projects in Business, Vol. 8 No. 3, pp. 491-512.* https://doi.org/10.1108/IJMPB-09-2014-0063 (last accessed June 2020)

[20] QBurst (2015), „Application of Big Data in Retail; 5 Ways Retailers Can Use Big Data Analytics", Available at: https://www.qburst.com/downloads/application-of-big-data-in-retail.pdf (last accessed June 2020)

[21] Ridge M., Johnston K.A., & O'Donovan B. (2015), „The use of big data analytics in the retail industries in South Africa", Available at: https://www.researchgate.net/profile/Kevin_Johnston4/publication/292945665_The_use_of_big_data_analytics_in_the_retail_industries_in_South_Africa/links/58ef34bf0f7e9b37ed16ebe0/The-use-of-big-data-analytics-in-the-retail-industries-in-South-Africa.pdf (last accessed June 2020)

[22] Saleem H., Uddin M.K., Rehman S.B. & Saleem S. (2019), “Strategic Data Driven Approach to Improve Conversion Rates and Sales Performance of E-Commerce Websites. *International Journal of Scientific and Engineering Research. 10. 588-593.*

[23] Santoro, G., Fiano, F., Bertoldi, B. and Ciampi, F. (2019), "Big data for business management in the retail industry", *Management Decision, Vol. 57 No. 8, pp. 1980-1992.* https://doi.org/10.1108/MD-07-2018-0829 (last accessed June 2020)

[24] Satyanarayana L.V. (2015), “A Survey on Challenges and Advantages in Big Data”, *IJCST Vol. 6, Issue 2*, https://pdfs.semanticscholar.org/582d/57611256a88843c4bcc90b3012a60dafb2d6.pdf

[25] TATA Consultancy Services (2014), “Manufacturing: big data benefits and challenges”, available at: http://sites.tcs.com/big-data-study/manufacturing-big-data-benefits-challenges/ (last accessed June 2020)

[26] Tharwat, M., 2017: Data scientist: 21st century sexiest job for free. John Snow Labs. Available at: http://www.johnsnowlabs.com/dataops-blog/data-scientist-21st-century-sexiest-job-for-free/ (last accessed June 2020)

[27] Wedel, M. and Kannan, P.K. (2016), “Marketing analytics for data-rich environments”, *Journal of Marketing*.

[28] Ylijoki, O. and Porras, J. (2019), "A recipe for big data value creation", *Business Process Management Journal, Vol. 25 No. 5, pp. 1085-1100*. https://doi.org/10.1108/BPMJ-03-2018-0082

[29] https://analytics.facebook.com/features (last accessed June 2020)

[30] https://corporate.walmart.com/our-story (last accessed June 2020)

[31] http://walmartlabs.blogspot.com/2011/05/goodbye-kosmix-hello-walmartlabs.html (last accessed June 2020)

[32] https://techcrunch.com/2013/06/10/walmart-labs-buys-data-analytics-and-predictive-intelligence-startup-inkiru/ (last accessed June 2020)

[33] https://corporate.walmart.com/newsroom/innovation/20170807/5-ways-walmart-uses-big-data-to-help-customers# (last accessed June 2020)

[34] https://cwiki.apache.org/confluence/display/HADOOP2/Home (last accessed June 2020)

## List of figures



## List of tables

# 7. Student CV

Lovro Brkanić, born 07.03.1994. in Zagreb, where I finished primary school and XIII. mathematical gymnasium.

In 2012 I enrolled in the Bachelor's Degree in Business program at the Faculty of Economics and Business and finished it in 2018. The same year, I enrolled in the Trade and International Business master's degree program.

In 2017, I attended and finished a primary course in accounting (RRIF accounting courses).

Besides Croatian, I speak English fluently and have excellent writing skills. Also, I have basic knowledge of German language.

My computer skills are advanced and have no problems using Microsoft Office.

I have a drivers' licence, B category.