Open University of Cyprus

Faculty of Economics and Management

Postgraduate Programme of Enterprise Risk Management

Master's Dissertation

Big data analytics, fraud and early risk detection

Glykeria Economidou

Supervisor

Dr. Pandelis Ipsilandis

DECEMBER 2021

Open University of Cyprus

Faculty of Economics and Management

Postgraduate Programme of Enterprise Risk Management

Master Thesis

Big data analytics, fraud and early risk detection

Glykeria Economidou

Supervisor
Dr. Pandelis Ipsilandis

This Master's Dissertation was submitted in partial fulfillment of the requirements for
the award of the postgraduate title
On December 2021
by the Faculty of Economics and Management
of the Open University of Cyprus.

DECEMBER 2021

Summary

Introduction – Big data analytics provides enterprises productivity and innovation, giving them the opportunity to exploit full potential of data and respond quickly to changes. Big data are helping enterprises improve their services and products, making them introduce new techniques that manage more efficiently and effectively their processes.

Purpose – The research concentrated on how the businesses in Cyprus use the big data analytics, how it affects them and in which sectors are implemented. How big data techniques and algorithms contribute to risk management. Moreover, it demonstrates what risks the businesses that do not use data analytics have and how they are affected.

Methodology – Theoretical analysis of the topic is utilized, as well as a mixture of quantitative and qualitative techniques. The data have been collected by using questionnaires and interviews with companies and people who are using the big data analytics, asking them to respond on the importance of the use of the big data analytics.

Results – the participants that worked in very small companies, as well as medium and large companies, consider very important the collection and analysis of data relating to their organization operations whereas most participants that worked in small companies consider it moderately important as also most of the participants that were from 18-29 years old, said that they have moderate general understanding of big data, most of the participants that were from 30-49 years old, have full or good general understanding of big data general understanding of big data, whereas most participants that were from 50 years old and up, said that they don't a general understanding of big data at all.

Conclusion - The constant evolution of technologies combined with the science of Big Data has a huge impact on today's business, helping companies, large and small, to have a powerful weapon for improved decision making as well as better real-time information for all parts of the business. It is certain that Big Data has a wide range of possibilities with great prospects for the future, helping businesses to constantly evolve and grow, in order to achieve the maximum possible business benefit.

Περίληψη

Εισαγωγή – Η ανάλυση μεγάλων δεδομένων παρέχει στις επιχειρήσεις παραγωγικότητα και καινοτομία, δίνοντάς τους την ευκαιρία να εκμεταλλευτούν πλήρως τις δυνατότητες των δεδομένων και να ανταποκριθούν γρήγορα στις αλλαγές. Τα μεγάλα δεδομένα βοηθούν τις επιχειρήσεις να βελτιώσουν τις υπηρεσίες και τα προϊόντα τους, με αποτέλεσμα να εισάγουν νέες τεχνικές που διαχειρίζονται πιο αποτελεσματικά και αποτελεσματικά τις διαδικασίες τους.

Σκοπός – Η έρευνα επικεντρώθηκε στο πώς οι επιχειρήσεις στην Κύπρο χρησιμοποιούν τα big data analytics, πώς τα επηρεάζουν και σε ποιους τομείς εφαρμόζονται. Πώς οι τεχνικές και οι αλγόριθμοι μεγάλων δεδομένων συμβάλλουν στη διαχείριση κινδύνου. Επιπλέον, δείχνει ποιους κινδύνους έχουν οι επιχειρήσεις που δεν χρησιμοποιούν αναλυτικά στοιχεία δεδομένων και πώς επηρεάζονται.

Μεθοδολογία – Χρησιμοποιείται θεωρητική ανάλυση του θέματος, καθώς και συνδυασμός ποσοτικών και ποιοτικών τεχνικών. Τα δεδομένα έχουν συλλεχθεί με τη χρήση ερωτηματολογίων και συνεντεύξεων με εταιρείες και άτομα που χρησιμοποιούν την ανάλυση μεγάλων δεδομένων, ζητώντας τους να απαντήσουν σχετικά με τη σημασία της χρήσης των αναλυτικών στοιχείων μεγάλων δεδομένων.

Αποτελέσματα – Οι συμμετέχοντες που εργάστηκαν σε πολύ μικρές εταιρείες, καθώς και οι μεσαίες και μεγάλες εταιρείες, θεωρούν πολύ σημαντική τη συλλογή και ανάλυση δεδομένων που σχετίζονται με τις λειτουργίες του οργανισμού τους, ενώ οι περισσότεροι συμμετέχοντες που εργάστηκαν σε μικρές εταιρείες τη θεωρούν μέτρια σημαντική όπως και οι περισσότεροι οι συμμετέχοντες που ήταν από 18-29 ετών, είπαν ότι έχουν μέτρια γενική κατανόηση των μεγάλων δεδομένων, οι περισσότεροι από τους συμμετέχοντες ηλικίας 30-49 ετών, έχουν πλήρη ή καλή γενική κατανόηση των μεγάλων δεδομένων γενική κατανόηση των μεγάλων δεδομένων, ενώ Οι περισσότεροι συμμετέχοντες που ήταν από 50 ετών και άνω, είπαν ότι δεν κατανοούν καθόλου τα μεγάλα δεδομένα.

Συμπέρασμα - Η συνεχής εξέλιξη των τεχνολογιών σε συνδυασμό με την επιστήμη των Big Data έχει τεράστιο αντίκτυπο στη σημερινή επιχείρηση, βοηθώντας τις εταιρείες, μεγάλες και μικρές, να έχουν ένα ισχυρό όπλο για βελτιωμένη λήψη

αποφάσεων καθώς και καλύτερη ενημέρωση σε πραγματικό χρόνο για όλα τα μέρη του η επιχείρηση. Είναι βέβαιο ότι τα Big Data έχουν ένα ευρύ φάσμα δυνατοτήτων με μεγάλες προοπτικές για το μέλλον, βοηθώντας τις επιχειρήσεις να εξελίσσονται και να αναπτύσσονται συνεχώς, ώστε να επιτυγχάνουν το μέγιστο δυνατό επιχειρηματικό όφελος.

Contents

CHAPTER 1

Introduction

1.1. Background Analysis

Data is the most valuable asset in today's world. No one can deny that technology and consequently the Internet have changed education, our way of living, governmental operations and how businesses are functioning. "Big Data" has changed from a buzzword to a real value creator in the past years, boosting the performance of operations, businesses and economy.

Nowadays, data is generating at a very fast pace and will continue to grow with every passing second giving the opportunity to businesses and organizations exploiting such a big amount of data for their own benefit. Social media platforms, mobile phones, cameras, etc. generate data and data scientists get value from this data by analyzing it and try to find valuable patterns to enhance business decisions (Ahsaan & Mourya, 2019). Big Data can give analysts exclusive information into market trends, buying patterns, maintenance cycles and many other business issues, enabling more targeted decision making.

The increase of data is becoming a trend worldwide and this tremendous growth of data has created a new era, where big data are used in every sector and every economy. Moving to a new era of big data analytics, big data has become a buzzword in the last years in business fields, highlighting the importance and the significance it represents. The ability to collect, manage and analyze data is an important and valuable asset to any organization, promoting innovative activities and taking accurate decisions in crucial times for the survival of an enterprise.

The "Internet of Things" and data mining are integrating massive amounts of real-time data, and governments and companies are storing and analyzing these data to learn what are your habits, your preferences, your buying behaviors, your social and cultural dynamics, to find patterns and capturing people’s thoughts that lead to better service and profit, however, it also raises some privacy concerns (Watson, 2014).

Big data analytics provides enterprises productivity and innovation, giving them the opportunity to exploit full potential of data and respond quickly to changes. Big data are helping enterprises improve their services and products, making them introduce new techniques that manage more efficiently and effectively their processes. Big data analytics also contribute to security management and monitoring as one of their most common use is fraud detection and enterprises enhance bid data applications to eliminate and mitigate risk through data-driven decision-making technologies (Vassakis, Petrakis, & Kopanakis, 2018). We developed an understanding on how data mining techniques can be applied in fraud detection and investigated the benefits that big data analytics provide in fighting and preventing fraud.

The application of Big Data Analytics results in more strategic and operational risk decision making leading to more reliable businesses and shows that by using big data analytics within an industry results to process safety and effective risk management (Goel, Datta, Mannan, O’Connor, & McFerrin, 2017). Risk management and fraud detection involve in their structure monitoring of the behavior of users and require the researchers to collect and analyze data to further examine the problems in order predict and prevent improper behavior of the population.

It is concluded that the application of big data in risk management and fraud detection can provide valuable insights on risk decision making companies leading to a safer industry and with data mining in real time, risk can definitely be reduced. Big data analytics focuses on developing new insights and understanding of business performance based on data and statistical methods. Data analytics combine skills, technologies, applications and practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. An empirical study using qualitative and quantitative statistical techniques, algorithms and tools for analyzing data in order to evaluate what promote risks and have significant impact in decision making.

1.2. Necessity and Importance of Research

This research offers a deeper understanding of the big data analytics, how they are used in several businesses and which risks arise. Therefore, it is discussed the identification of opportunities and threats by using the big data and how this benefit while using this kind of analysis.

CHAPTER 2

Literature Review

2.1. Definition and Characteristics of Big Data

Over the last years there is a huge growth in data. This trend is at a life-changing stage, where the amount of information gathered is at a very high level that surpasses the capability of existing data storage techniques. These "*datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze*" as McKinsey Global Institute defined in 2011 and moreover, many academicians define big data as an enormous size of data produced by high-performance applications that range from social network to scientific computer applications (Bhadani & Jothimani, 2016).

Big data phenomenon involves data collection, storage, management and analysis of massive amounts of complex data. Although over the years many scholars have tried to identify and interpret the meaning of big data, however, big data has no specific definition since there are a lot parameters yet remain undiscovered, and this tremendous growth in complex data, both structured and unstructured, promotes the need of discovering new advancements to better understand the usage of big data and gather relevant information for effective decision making.

2.2. Different Types of the Big Data

Big data are coming in a huge volume and in an inconceivable speed. These data come from various sources and can be classified in three types (analytics, 2020):

✓ Structured Data
✓ Unstructured Data
✓ Semi-Structured Data

In order to achieve a better understanding into what big data is, it is important to discover one of each category and understand what insights it can produce and all three can be applied at any level of big data analytics.

Structured Data**,** is the most common type of data since the majority of data managed and analyzed falls under the category of structured data. It is the quantitative data that are the easiest type of data to handle such as age, numbers, tables, etc. and analytical tools can go straight to collecting and processing the data to extract the valuable information. Structured data require little to no preparation before processing without the need to diversify or convert to the type needed before using to gain the results. Moreover, structured data are stored in data warehouses that are highly structured for a specific purpose and integrated by relational databases like SQL (Structured Query Language).

Unstructured Data**,** is defined as the unorganized data such as text that express human language. Unstructured data requires a level of understanding due to its complexity and need some time and effort to process in order to be readable and interpretable. In addition, unstructured data demand a well-structured application involving complex algorithms and experienced data scientist to understand what kind of information is extracting. The challenge of unstructured data is to be able to merge internal and external data into meaningful outcomes. An example including unstructured data is *when an application requests you to prove you are not a robot.*

Semi-Structured Data can be found in between structured and unstructured data. Mostly, data that are collected in any type, like picture, time, location or email address can be translated to unstructured data and integrated accordingly. For example, when a user sends an email, data collected (email addresses, location, time, the IP addresses, etc.) are grouped according to certain characteristics. However, semi-structured data can be proven a huge asset for an organization. It can identify patterns in datasets and organize them for gaining deeper insights.

The new era, the era of Big Data, has been characterized over the years in multiple dimensions, though, the majority of authors emphasize that the seven most important "Vs" that define Big Data are presented below (Mikalef, Pappas, Krogstie, & Giannakos, 2017), (Vassakis, Petrakis, & Kopanakis, 2018):

- ✓ **Volume** refers to the high quantity and unmanageable large size of generated datasets that are captured every second. Data is dependable to its volume in order to characterized as big data or not.
- ✓ **Variety** refers to the different types of data, structured, semi-structured and unstructured that are being generated and the increasing diversity of data formats. Data such as video, text, audio, web sources and social media actions, financial transactions, sensor data, GPS data, etc. furthermore, data that are connected with human behavior through their activities click-streams and actions in formats that is difficult to classify are characterized by variety
- ✓ **Velocity** refers to the frequency and the speed data are being generated by various sources. It is crucial to capture and analyze the data in real-time to take actions and remain agile in order to take informed decisions and gain advantages against your rivals. Data can be collected in real time or near-real time through sensors, social media posts and social trends and businesses have the ability to analyze them and take action at the time of the event.
- ✓ **Veracity** refers to data accuracy and unreliability provided by the different types of data. This uncertainty questions the quality of data and data must be differentiated according to trustworthiness according to the source that the data is collected from.
- ✓ **Variability** refers to the complexity and meaning of collected data along with the different ways the same data can be interpreted. For example, some datasets like texts can have a different meaning according to who is writing the text, for what purpose, etc. and the algorithms must take into account the factors affecting the meaning.
- ✓ **Visualization** refers to the visual representation of data and qualitative or quantitative information presented in a graphical layout in ways that makes it easy to interpret and understand comparable to usual forms of data.
- ✓ **Value** refers to the very high value the data have once are extracted and analyzed and the impact big data have on real world applications and problem solving. Moreover, collecting and analyzing big data is important to enterprises, shifting the value at an increasingly high level that helps organizations in decision making.

The data-driven enterprises that exploit data effectively for collection, storage, management and analysis of data assessing the condition financially and operationally gaining better understanding on the causes of the problems and taking informed decisions in improving their performance.

Many big data professionals believe that due to the increasing number of internet and social networking users, the amount of generated data is growing exponentially. IDC (International Data Corporation) predicts that by 2025 the worldwide datasphere will reach 175 zettabytes (1 zettabyte is equal to 10^{21} bytes) (Khvoynitskaya, 2020). This growth is shown in Figure 1 and provide a clear understanding of how rapidly data are growing and also the big challenge data scientists have in order to put every effort they can to handle such voluminous data.

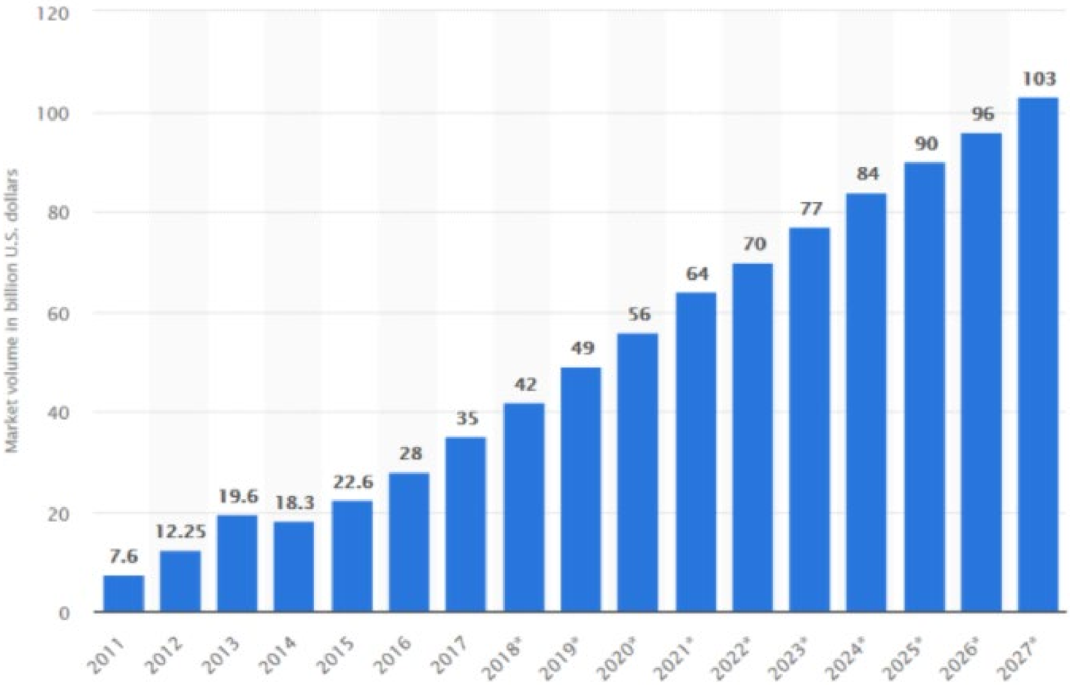


Figure 1: the exponential growth of big data

Source: Statista Research Department

2.3. The Features and Characteristics of Big Data

Most definitions of "Big Data" focus on the volume of data. However, in addition to the size of the data, there are other equally important features in "Big Data", among which is the variety of data and the speed of the data. The combination of Vs (Volume, Variety, Velocity) is a complete definition of the concept of "Big Data", in which there are no restrictions. In addition, each of these three Vs has its own peculiarities in terms of their analysis. (Figure 2)

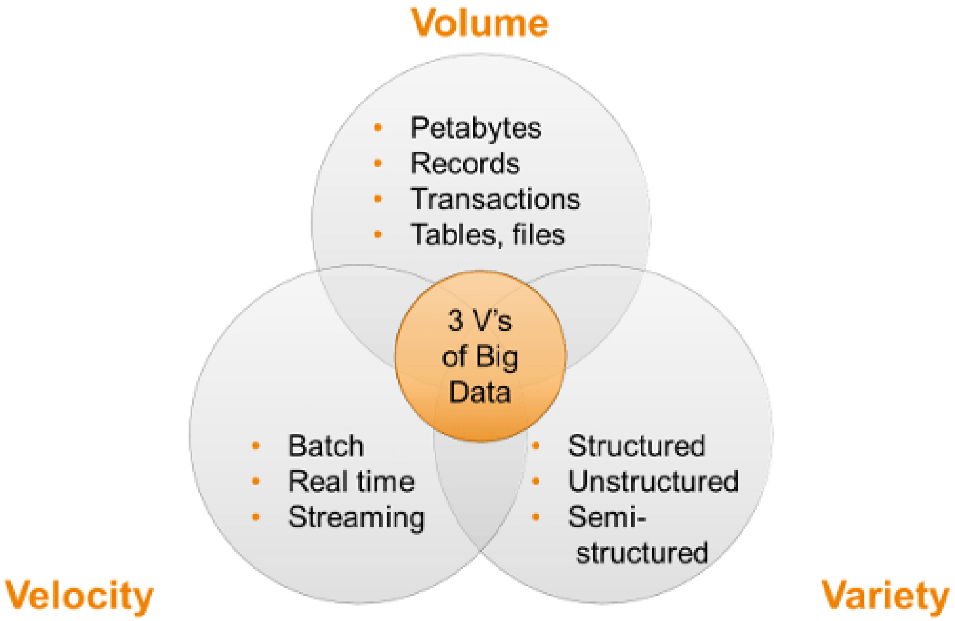


Figure 2: The 3V's of Big Data

Source: Pedro César Tebaldi Gomes (2014)

2.3.1. The Three Vs of Big Data

✓ Volume: Refers to large amounts of all types of data from any data source, including business transactions, social networks, and information from data transmitted between digital devices. The benefits of collecting and analyzing these quantities create a number of challenges in gaining valuable knowledge for people and companies.

✓ Velocity: Refers to the speed of data transfer. With the ever-increasing sources of data mining, the flow of data, as well as their content, is changing at high speeds, which makes them difficult to manage. For this reason, new algorithms and ways are needed for their smoother processing and analysis.

- ✓ Variety: Refers to a variety of different types of data collected through sensors, smartphones, social media and websites, displayed in video, text, audio, etc. Also, this data consists mainly of structured, numerical data (in relational databases) to unstructured files.

However, over the years, additional features have been added around the "Big Data", some of which are ( ):

- ✓ Veracity: The accuracy of the data can save organizations from many problems. Because data comes from many different sources, ensuring data validity by companies is important for their analysis and more specifically for their automated decision making.
- ✓ Variability: Refers to the variability of data. Because data streams are extremely unpredictable and constantly changing dramatically, it is difficult for companies to understand the meaning and significance of data. Although difficult, this can be achieved by using the appropriate algorithms
- ✓ Visualization: Data visualization is a special piece, as it is important for converting large volumes of data into a format comprehensible to the reader. Depicting them, although not technically difficult, is the most difficult part of "big data", because behind each graph is a complex story that is often difficult to capture, but also extremely important.

2.4. Different Types of Big Data

Data can be created by machine or by man. Human-generated data refers to data generated as a result of human-machine interactions. Emails, documents, Facebook posts are some of the man-made data. Machine-generated data refers to data generated by computer applications or hardware devices without human intervention. Data from sensors, disaster warning systems, weather forecasting systems and satellite data are some of the data generated by the engine. Figure 3 represents the data generated by a human on various social media, the emails sent, and the images taken, and the machine data generated by the satellite.

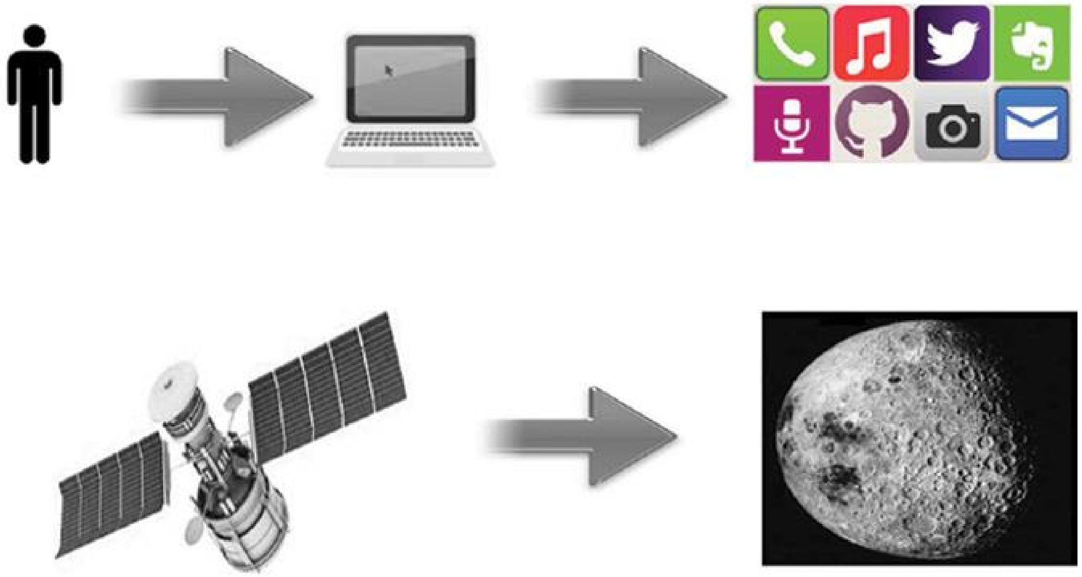


Figure 3: Human- and Machine-generated data Source: Machine-generated and human-generated data can be represented by the following primitive big data types
Source: Pedro César Tebaldi Gomes (2014)

<u>Structured data</u>

Data that can be stored in a relational database in the form of a table with rows and columns is called structured data. Structured data, often generated by businesses, is highly organized and can be easily processed using data mining tools. Examples of structured data are employee data and financial transactions.

<u>Unstructured data</u>

Data that is raw, unorganized, and incompatible with relational database systems is called unstructured data. Almost 80% of the data generated is not structured. Examples of unstructured data are vinesaudio, images, emails, text files, and social media posts. Unstructured data is usually found in either text files or binaries. Data in binaries has no recognizable internal structure, such as audio, video, and images. The data in text files is e-mail, social media posts, PDF files and word processing documents.

Semi-structured data

Semi-structured data are those that are structured but do not fit into the relational database. Semi-structured data is organized, which makes it easier to analyze than unstructured data. JSON and XML are examples of semi-structured data.

2.4.1. Big Data Lifecycle

Big Data brings great benefits, ranging from innovative business ideas to unconventional ways of tackling disease, thus overcoming many challenges. Challenges arise because much of the data is being collected by technology today. Big Data technologies are able to capture and analyze them effectively. The Big Data infrastructure includes new computer models capable of processing both distributed and parallel computations with scalable storage and performance. Some of the Big Data components include Hadoop (framework), HDFS (storage) and MapReduce (editing).

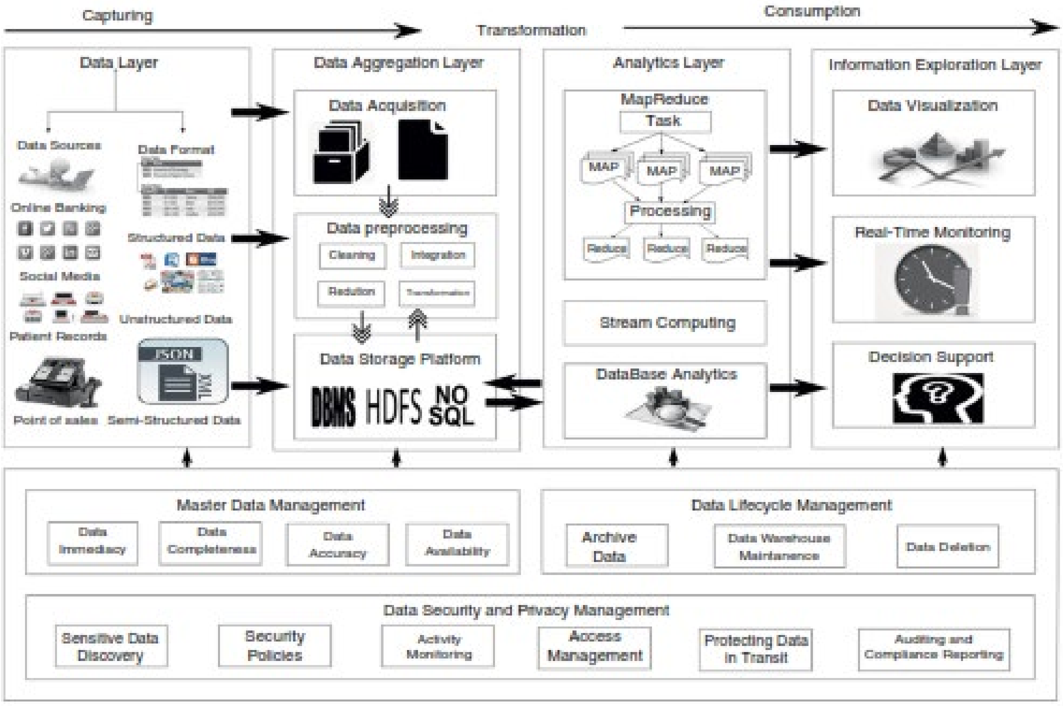


Figure 3 illustration of the life cycle of Big Data.
Source: Pedro César Tebaldi Gomes (2014)

Data that reaches high speeds from many sources with different data formats are recorded. The recorded data is stored on a platform such as HDFS and NoSQL and then pre-processed to be suitable for analysis. The pre-processed data stored on the storage platform is then transferred to the analytics level, where the data is processed using Big Data tools such as MapReduce and YARN and then the processed data is analyzed to reveal hidden knowledge from them.

Analytics and Machine Learning are important concepts in the Big Data life cycle. Text analysis is a type of analysis performed on unstructured text data. With the development of social media and e-mail transactions, the importance of text analysis has increased. Predictive analysis of consumer behavior and analysis of consumer interests are performed on text data extracted from various online sources, such as social media, retail websites and more. Machine learning made text analysis possible. The analyzed data is visualized by visualization tools such as Tableau so that it is easily understood by the end user to make decisions.

1. <u>Data creation</u>: The first phase of the Big Data life cycle is data generation. The range of data produced by differentiated sources is gradually expanding.

2. <u>Data collection</u>: The Big Data lifecycle data collection phase includes the collection of primary data, the transmission of data to the storage platform and their preprocessing. Gaining data in the Big Data world means gaining high-volume data that is growing at an ever-increasing rate. The raw data collected in this way is transmitted to an appropriate storage infrastructure to support processing and various analytical applications. Preprocessing includes data cleansing, data integration, data transformation, and data reduction to make data reliable, error-free, consistent, and accurate. The data collected may have redundancies, which take up storage space and increase storage costs, and can be addressed by pre-processing the data. Also, much of the data collected may not be relevant to the target of the analysis and should therefore be compressed during preprocessing. Therefore, efficient data preprocessing is necessary for it to be cost

effective as well as for efficient data storage. The pre processed data is then transmitted for various purposes, such as data modeling and data analysis.

3. Pre-processing of data: Data preprocessing is an important process performed on raw data to convert it to comprehensible form and thus provides accessin consistent and accurate data. Data generated from multiple sources is inaccurate, incomplete and inconsistent due to their bulk and heterogeneous sources, and it makes no sense to store junk and dirty data. In addition, some analytical applications are a prerequisite for quality data. Therefore, for effective, efficient and accurate data analysis, systematic data processing is necessary. The quality of the source data is affected by various factors ( ).

For example, data may have errors such as a negative value field that results from transmission errors or typographical errors, or intentional incorrect data entry by users who do not wish to disclose their personal information. These errors mean that the field lacks features of interest, which may result from non-applicable field or software errors. Data inconsistency refers to discrepancies in the data, such as date of birth and age. Inconsistencies in the data occur when the data collected come from different sources, due to inconsistencies in the naming of contracts between different countries and inconsistencies in the form of import. Data sources often have redundant data in different formats, and therefore duplicate data must also be removed during data preprocessing to make the data substantial and error-free. There are several steps involved in data preprocessing ( ):

1. Data consolidation
2. Data cleaning
3. Data reduction
4. Data transformation

Data consolidation

Data consolidation involves combining data from different sources to give end users a unified view of data. There are many challenges in data consolidation. For example, when extracting data from an individual's profile, the first name and last name may be interchanged in a particular culture, in which case unification may occur incorrectly. Data redundancies often occur when consolidating data from multiple sources. Sources such as organizations, smartphones, personal computers, satellites and sensors generate different data such as e-mail, employee data, WhatsApp chat messages, social media posts, online transactions, satellite imagery and sensor data. These different types of structured, unstructured and semi-structured data must be consolidated and presented as unified data for data cleaning, data modeling, data storage and then for data extraction, conversion and loading (ETL).

Data cleaning

The data cleansing process replenishes missing values, corrects errors and inconsistencies, and removes data redundancy to improve data quality. The greater the heterogeneity of the data sources, the higher the degree of impurity. As a result, more cleaning steps may be included. Data cleansing involves several steps, such as detecting or recognizing the error, correcting the error, or deleting the incorrect data, and documentation of the error type. A detailed analysis of the data is required to detect the type of error and inconsistency in the data. Data redundancy is the duplication of data, which increases storage costs and transmission costs and reduces the accuracy and reliability of the data. The various techniques involved in handling data redundancy are surplus detection and data compression. Missing prices can be filled in manually, but it is tedious, time consuming and not suitable for the huge amount of data. A universal constant can be used to fill in all missing values, but this method creates problems when integrating data.

Data reduction

Processing data into huge volumes of data can take a long time, making data analysis either impossible or impractical. Data reduction is the concept of reducing the volume of data or reducing the size of the data, ie the number of features. Data reduction techniques are adopted for the analysis of data in a reduced form without losing the integrity of the real data and at the same time giving quality results. Data reduction techniques include data compression, dimension reduction and number reduction. Data compression techniques are applied to obtain the compressed or reduced representation of real data. Dimension reduction is the reduction of a number of featuresthat is, a technique that removes irrelevant or unnecessary features. Numerology reduction is a technique adopted to reduce volume by selecting smaller alternative data.

Data transformation

The data transformation refers to the transformation or consolidation of data into an appropriate form and their conversion into logical and important information for data management and analysis. The real challenge in data transformation comes into the picture when the fields in one system do not match the fields in another system. Before the data is transformed, the data is cleaned and manipulated. Organizations are collecting a huge amount of data and so the volume of data is growing rapidly. The collected data is transformed using ETL tools.

2.5. The Advantages of Big Data

Data volume is growing rapidly and companies are storing large amounts of data and analyzes to better make decisions and optimize their strategy. Big Data technologies help in this storage and analysis of data faster. For this reason, companies from all sectors are increasingly attracted to the way Big Data is managed and analyzed. An important criterion here is that companies need to respond faster to change, not only to current but also to future challenges. Thus, Big Data is becoming more and more necessary for decision makers to make better decisions.

Big Data through the analysis of data from many sources, such as scanners, mobile phones, the internet and social media platforms offer significant benefits to businesses. Therefore, according to Frankel and Reid, (2008) one of the most important benefits of Big Data is their ability to display the most important and useful information for businesses. Big Data and the analysis of these data help managers and business executives to evaluate the information they have collected and use it to their advantage, thus reducing costs and time for better decision making as well as creating new ones. products and services that target the needs of their customers.

2.6. Big Data Challenges

Opportunities often coexist with challenges. On the one hand, Big Data offers opportunities in many areas. On the other hand, the security and privacy challenges posed by Big Data also catch the eye of the people. These challenges last a lifetime in Big Data, which can be described as data collections, as well as data storage and management, transmission, analysis and destruction. Therefore, regardless of data collection and data corruption, Big data challenges can be summarized as follows ( ):

1. Privacy and security challenges:

It is the most important issue with Big Data, as it is a very sensitive issue with technical and legal significance. A big issue is a person's personal information when combined with large external data sets, leading to new facts about that person and how likely it is that these kinds of facts about the person are secret and unwanted by the data holder or any person to learn about him. At the same time, user information is collected and used to add value to the organization's business. This is done by creating ideas in their life that they do not know. Another important consequence is the social stratification where an illiterate person would benefit from Big Data predictive analysis and on the other hand, the disadvantaged would be easily identified and treated worse.

The Big Data used by law enforcement will increase the chances of some people suffering from adverse consequences without the ability to resist or even be aware of discrimination.

2. Challenges to data access and information sharing:

If data is to be used to make accurate decisions in a timely manner, it becomes essential that it be made available in an accurate, complete and timely manner. This makes the data management and governance process a bit more complicated, adding to the need to make data open and available to government agencies in a standardized way with standard APIs, metadata and formats, leading to better decision-making, business intelligence and improvements. in productivity. The anticipation of data exchange between companies is strange due to the need for companies to excel over others. Sharing data about customers and their activities threatens the culture of confidentiality and competitiveness.

3. Storage and processing challenges:

The available storage space is not enough to store the large amount of data generated by almost everything: Social media sites are the same big contributors along with sensor devices, etc. Due to the strict requirements of Big Data on networks, storage media and servers, outsourcing data to the cloud may seem like an option. However, uploading this large amount of data to the cloud does not solve the problem. This is because Big Data information requires the collection of all data and then their connection to some means of extracting important information. Terabytes of data require a long time forto upload to the cloud and in addition this data change so quickly that it will be difficult to upload this data in real time.

Data transfer from the storage point to the processing point can be avoided in two ways. One is processing only the storage and the results can be transferred, or only that data that is important can be transferred. But both of these methods require maintaining the integrity and origin of the data. Processing such a large volume of data also takes a long time. In order to find suitable data, the whole data set must be scanned, which is somewhat impossible. Thus, indexing right at the beginning while collecting and storing data is a good practice and significantly reduces processing time.

4. Challenges of analysis:

Big Data brings with it some huge analytical challenges. The type of analysis that needs to be done on this huge amount of data that can be unstructured, semi-structured or structured requires a large number of skills in advance. In addition, the type of analysis to be done on the data depends to a large extent on the results to be taken, ie the decision-making. This can be done using one of two techniques: either by incorporating huge volumes of data into the analysis or by determining in advance what data is relevant.

5. Skills Challenges:

As Big Data is in its infancy and is an emerging technology, it must attract organizations and new people with different new skill sets. These skills should not be limited to techniques but also extend to research, analytical, interpretive and creative. These skills need to be developed individually and therefore require training programs to be developed by organizations. In addition, universities must introduce a Big Data curriculum to produce specialized staff in this technology.

2.7. The Big Data Software Tools

In the near future, all organizations will adopt the big data in every section of their organization. This process of data acquisition will provide a most accurate result on problem solving. Nevertheless, this challenge requires new technologies and more advanced algorithms and tools for processing and analyzing data. Many researches show that most companies are collecting data but only a few of them have the knowledge of tools and data platforms to effectively analyze them.

With the improvements in Internet of Things and computing technologies, it is much easier to store data in clouds and extract information with the help of software and tools such as Artificial Intelligence and Machine Learning Applications. Several software tools and techniques are being created in order to process the data which are being collected by numerous digital devices and applications that generate massive data in various forms (unstructured data). Big data tools can be categorized according to the services they provide based on the problem they would like to explore.

Big data tools as presented in Figure 2 are classified as: Big Data Analysis Platforms and Tools, Databases/Data Warehouses, Business Intelligence, Data Mining, File Systems, Programming Languages, etc. (Grover & Kar, Big Data Analytics: A Review on Theoretical Contributions, 2017). We are about to discuss and analyze some of the most popular of each section.

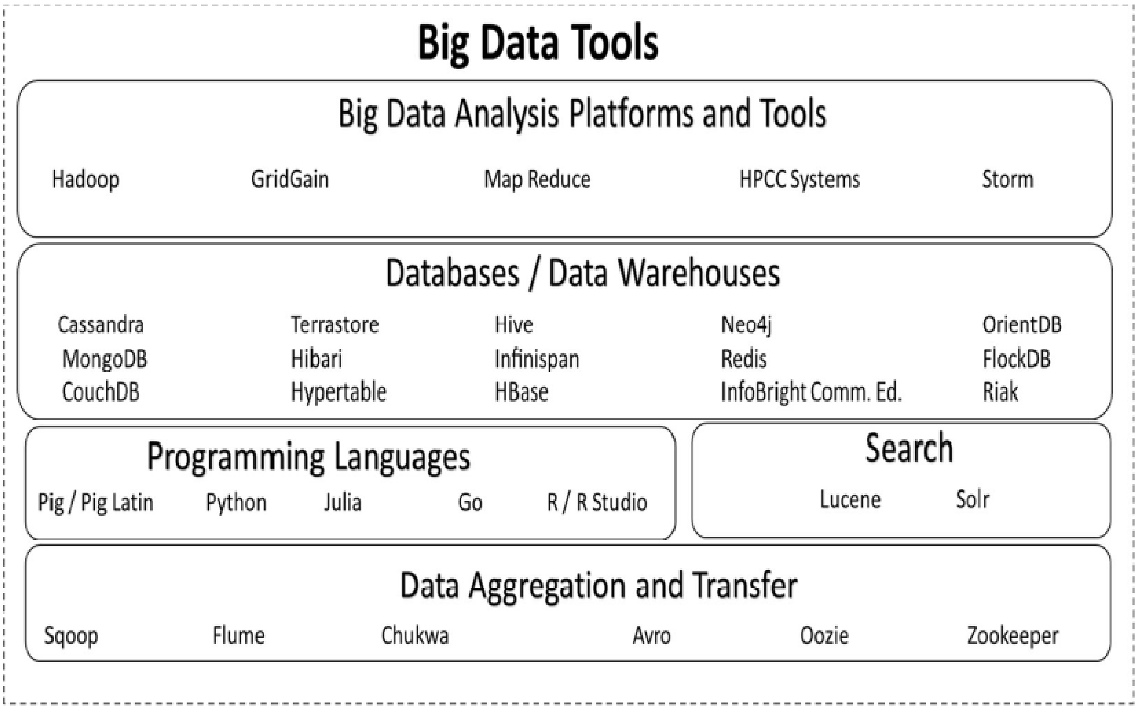


Figure 2: An overview of various big data tools (Grover & Kar, Big Data Analytics: A Review on Theoretical Contributions and Tools used in Literature, 2017)

2.8. Big Data Analysis Platforms and Tools

Hadoop is an open-source apache framework used for storing structured and unstructured data and running applications on clusters of computers using simple programming models (Apache Hadoop, 2021). There is no doubt that this is the most widely known tool around the world. Some of the companies that use Hadoop programming model are Amazon, IBM, Intel, Microsoft, Facebook (Software Testing Help, 2021).

Map Reduce is a technique and a programming model used for processing large data sets and contains two functions, Map and Reduce. Map is used to filter and split the sets of data into maps and Reduce is used to shuffle and reduce the data, returning output value. MapReduce programs are written in various languages such as: C++, Java, Python, Ruby, R, etc. (Grover & Kar, Big Data Analytics: A Review on Theoretical Contributions, 2017)

Apache Storm is a free and open-source distributed computing system. Apache Storm is extremely resilient, extensible, reliable and can be used with any programming language with the ability to be processing over a million tuples processed per second per node (Apache Storm, 2019).

2.8.1. Databases/Data Warehouses

Cassandra is a database used for storing huge amounts of structured data from different sources which is suitable for online web and mobile applications and can satisfy each customer's requirements. "*Apache Cassandra database is the right choice when you need scalability and high availability without compromising performance*" (Apache CASSANDRA, 2016). Cassandra is in use at Apple, Ebay, Netflix, Microsoft, McDonalds, Netflix, New York Times, and many other companies that have massive amounts of active datasets.

Hive is an open-source data warehouse of Apache used for querying and managing basically large structured datasets through SQL language (APACHE HIVE TM, 2014). HBase is an open-source data model of Apache to get quick access to data, for handling and analyzing structured and unstructured data. HBase's goal is managing large tables while providing a fault-tolerant way of storing big data (IBM, 2021)

2.8.2. Programming Languages

Python is one of the simplest and most user-friendly programming languages. It is an open-source language supported by Windows, Linux and Mac platforms. Python due to its high code readability is easy-to-learn and easy-to-use programming language for use in scientific computing, mathematics, engineering and implementing machine learning techniques for quick analysis and also supports classification, regression, clustering, etc. (Bhadani & Jothimani, 2016). Moreover, Python helps you write a code with various functionalities in less extent and less time, having support with near-endless libraries that enable associations with online applications (Grover & Kar, Big Data Analytics: A Review on Theoretical Contributions, 2017).

R is an open-source statistical computing language which provides many statistical (linear and nonlinear modeling, statistical tests, classification, clustering, etc.) and graphical techniques for analyzing massive datasets. As Python, R is also available on Windows, Linux and Mac platforms and has all the features of a standard programming language for handling big data (Bhadani & Jothimani, 2016), (The R Foundation, 2021). Furthermore, R due to its efficient strong oriented nature, it provides excellent extensibility for data analysis. R is one of the most sophisticated coding languages for data manipulation through process to visualization.

Structured Query Language (SQL) was initially developed by IBM and is the most common used database query language for managing structured data amongst the data science field. SQL is mostly used to manage and organize large data sets in relational database management systems (RDBMS) (L, 2020), (WIKIPEDIA, 2021). On the contrary, NoSQL stores and manages unstructured data and NoSQL databases are increasingly used in big data and real-time web applications (WIKIPEDIA, 2021).

2.9. Categories of Big Data Analytics

The term big data analytics is concerned with data science, business intelligence and analytics. Data Science is a collection of principles used for deriving information and knowledge from data. Data Scientists must be highly trained and experienced professionals in order to be able to discover patterns and relationships to create value for the organization and solve difficult problems in critical situations. There are four types of analytics and all are equally important, descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics. These types are shown below:

Descriptive analytics, describe what had happened using business intelligence techniques trying to get insights and find correlations by identifying patterns using descriptive statistics such as mean, median, standard deviation, etc. The main purpose of this category is to summarize the findings in order to better understand each situation and address the strengths and weaknesses that can help in decision making (Mehta, 2017). By using this type of analytics, organizations can analyze consumer behavior and financial performance through data visualization, reports and graphs etc. (Vassakis, Petrakis, & Kopanakis, 2018).

Diagnostic analytics focus to determine what had happened and find the causes on past performance in order to provide to businesses the option to take better decisions in the future and avoid negative results (Vassakis, Petrakis, & Kopanakis, 2018). Diagnostic analytics identify anomalies in data and relevant statistical techniques are used to find correlations that can interpret such anomalies.

Predictive analytics used to predict future results and forecasts what might happen providing an estimation of a future outcome. It utilizes various techniques including data mining, data modelling and machine learning to predict future trends (Chong & Shi, 2015) and the probabilities of the occurrence of an event. Using all the available data, predictive analytics can uncover patterns and identify correlations in data that can be used for prediction of customer behavior, preventing risk and detect fraud (Vassakis, Petrakis, & Kopanakis, 2018). Predictive analytics are usually used by well-trained data scientists and can support complex forecasts and provide accurate predictions. Therefore, organizations that take advantage of predictive analytics can determine future trends and patterns in order to improve their business models by introducing innovative services and/or products (Mehta, 2017).

Prescriptive analytics is actually the next step of predictive analytics. It usually used to provide optimization to accomplish the best outcomes and helps businesses understand patterns and uncertainties to make smarter decisions with lower risk. Prescriptive analytics focuses on recommendation and providing assistance and advise to researchers in decision making which increases the performance of the big data system and the organization (Chong & Shi, 2015). Moreover, prescriptive analytics using big data, can create solutions for the problems, that lead to realizations and observations which maximizing business returns (ProjectPro, 2021).

2.10. Big Data Applications

Businesses are heavily relying in big data analytics for business intelligence and data analysis in domains of customer analytics, decision making, technological innovations, etc. The applications of big data are broad and extensible. Big data analytics is important because big data provide efficiency and improve the performance of the organization enable them to succeed against their competitors.

Big Data Industry is rapidly growing due to the broad usage in various areas such as forecasting, sentiment analysis, healthcare and privacy, auditing and finance, risk and fraud detection, telecommunication, insurance and privacy etc. Most of these analytical applications were not possible in the past, since big data technologies were unable to store such a huge volume of datasets and uncapable to process in a timely manner. It is worth noting that all the below applications that will be analyzed are associated somehow with **risk and fraud detection**. The big data applications are discussed as follows:

Healthcare

In the healthcare field there are many studies describing the money invested in order to access on the clinical insights. Data analytics obtain information from various sources in the healthcare sector, such as clinical data and scans, symptoms, pharmacy and laboratories, doctors notes, patient's monitors and generated sensor data. Financial Times (Neville, 2020) indicate that investors are giving billions of dollars to big pharmaceuticals to access "*clinical insights contained in vast troves of anonymized patient records*" and thus, obtaining predictive insights and advancing "digital health". These clinical data can provide to physicians an image of the patient and develop a treatment methodology that can be cost and time reducing while at the same time giving the best available treatment. This can benefit both the patients and the doctors/hospitals by reducing the hospitalizations.

For example, in the pandemic of COVID we are going through, by obtaining significant information from sources like social media, can help in assessing the exposure to the risk of getting the flu and in early detection of the pandemic and take precautionary measures. Moreover, another application concerns the healthcare sector is the monitoring of medical complaints and complications that might arise, in which the real-time analysis of data can minimize fraud and reduce medical insurance claims. Many studies had shown that big data analytics are crucial for predicting the impact the disease has and preventing medical malpractice, reducing costs (Chong & Shi, 2015).

Telecommunication

In the telecommunication industry data driven companies can use targeted marketing in order to analyze a number of factors affecting the customer preferences such as demographic data (age, gender, education, language preferences), usage preferences (internet usage, value-added services, working, entertainment) to create a model of customer preferences and offer a customized service that satisfies each client and increase their revenue. The huge advantage big data provide is the ability to help any company to innovate and redevelop its products according to consumer's needs. Additionally, telecommunication companies are facing telecom frauds very often, and thus, big data researchers must find new ways to overcome this and minimize exposure to frauds. In some studies, it is showing that data scientists are working with telecom analytics solution teams to provide a fraud detection system using predictive analytics and artificial intelligence (Bhadani & Jothimani, 2016).

Finance and Banking

Financial companies are using advanced big data technology to store and manage huge volumes of data. Big data can help in preparation of regulations and anti-money laundering manuals, fraud mitigation, etc. Predictive analytics is important to detect trends and identify trades in investment banking and capital markets, where real time data flow need extraction and further analysis.

The use of big data in fraud detection by identifying patterns and trends, can determine fraudulent activities faster and more accurately for credit card fraud and money laundering. Big data technologies applied in business models of financial institutions and banking can help in risk controlling in order to be ready to react faster in new risks, optimizing its portfolio analysis. Banks can improve their processes by testing risk systems based on big data techniques, software recognizing fraud patterns, and by applying risk assessments. The volume and velocity of collected data can also address challenges in credit card fraud, having a significant impact on the review that has to be done before authorizing a card for a transaction.

Morgan Stanley *invests in cybersecurity and fraud prevention technologies* (Morgan Stanley, 2020), for example they are using big data in identifying voice prints. Their model includes identification of unauthorized intrusion, fraud monitoring, generating alerts and notifications, like updating passwords and enhancing strong encryption protocols to protect their client's information.

2.11. Big Data Analytics and Risk Management

In the risk management area, big data are providing the ability to assess and identify potential threats and reduce risks, enabling every organization to detect and analyze fraudulent activities with more effective ways improving their performance and efficiency. Big data analytics has contributed to the development of risk management. The tools available allow a firm to forecast a potential risk and quantify the danger, enabling the business to implement smarter risk mitigation strategies. Data analytics systems are in position to detect weaknesses and potential risks, ensuring that the organization will facilitate a robust strategic plan and risk analysis ensuring that the business will remain profitable (Kopanakis, n.d.).

In any of the aforementioned areas, big data can ensure that potential threats will be identified, compared and analyzed so that it will find ways to act quick and efficiently to shield the business against fraudulent activities. Early risk detection is more important to make key decisions and early warning systems are used to predict probable anomalies and reduce risks arise from sudden changes. Besides that, early warning systems give the chance to risk management teams to exploit the opportunity to avoid and mitigate potential risks by taking precautionary measures along with creating an emergency plan (Koyuncugil & Ozgulbas, 2012).

Fraud risk management can track and analyze user's behavior, identify suspicious actions, apply various levels of treatments based on business intelligence arbitration and create a safe and clean environment. Proper analysis of critical data is achieved by a relevant process includes building a model, training the model, evaluating its performance and implementing the appropriate model. Data mining methods associating clustering and sequential patterns, applying statistical and mathematical algorithms to discover symmetrical patterns and red flags indicating the existence of fraud.

In addition, enterprises must ensure that the data collected is from authentic sources in order for accurate forecasting to take place and maximize the success in the long-term and sometimes make changes and cut costs where is essential. With a combination of big data, proper strategy, data governance and appropriate technology that the enormous quantities of produced data will not be wasted but used to extract valuable information and take proper and timely risk decisions.

2.12. Challenges in Big Data Analysis and Risk Management in Business

In addition to the benefits that big data analytics can have, there are challenges that need to be addressed when implementing analytics in combination to the business' risk management. The biggest challenges of big data analytics are to move forward with technological advancement in a cost-effective way as well as to develop organizational decision-making processes for data use, information absorption, and intelligence conversion (Sanders 2014). In addition, there are frequent issues with data policies, data quality, and big data analysts.

Data policy issues become more important as data is increasingly being digitized. These issues include, for example, data privacy, security, copyright and liability (Mayika et al., 2011). Better access to information raises privacy issues that need to be assessed as opposed to ease of access to information in combination to the business' risk management (Brown et al. 2011). When outsourcing technological and analytical skills, the risks and dependencies associated with data security and copyright must be recognized. The addition of explicit data security requirements to contracts and the determination of liability for data breaches can be used to protect, for example, competitively sensitive data (Mayika et al., 2011).

The new Regulation on the protection of personal data of the European Union (EU) entered into force on 25 May 2018, with an impact on the protection of the privacy of EU citizens and the privacy approaches of organizations. The purpose of the regulation is to protect the privacy of all EU citizens and data breaches in an increasingly data-driven world. The biggest regulatory change will be the extension of the GDPR jurisdiction, which will apply to all companies that process EU members' personal data, regardless of company location. Organizations that do not comply with the GDPR may be fined up to 4% of their annual global turnover or € 20 million in the event of a higher amount. It is noteworthy that the regulations apply to both data controllers and processors, including cloud repositories (EU GDPR Portal, 2017).

Data quality consists of data consistency and data completeness. Big data comes from various sources, not all of which are verifiable in combination to the business' risk management (Hashem et al., 2015). Confidence in data quality is weakened when inconsistencies or data deficiencies are caused by, for example, data entry errors, faulty system design, or subjective judgment by the data carrier. The amount of increase in unstructured data creates challenges for the organization and the formation of quality data to create value (Sanders 2016). Data quality issues are included in pre-analysis processes, which consist of creating, structuring, and organizing data for further use or analysis (Manyika et al., 2011).

Talented data analysts are required to leverage big data, as many companies use resources to gain big data capabilities and continually improve their learning processes through the implementation of big data analytics. Therefore, as technological possibilities continue to develop, respectively, the skills required for the use of analytical tools increase (Sanders 2016).

Businesses, then, need to hire and retain such analysts and also create a culture that values the use of big data in decision making. However, the key is not just hiring analysts to extract and prepare data, conduct statistical work, create models, and program business applications. Equally important are analytical entrepreneurs who are willing, able and willing to use better information and analysis to enhance their work, in addition to working with analysts (Manyika et al., 2011).

2.13. Summary of Findings

Information Technology and Business Intelligence creates new opportunities and improvements that were impossible in the past. It is proven that big data is a promising topic and the growing role of big data in all kinds of industries is undisputable. Advances in business and data mining techniques have brought tremendous improvements to business operations and develop evolutionary ways for handling data. The variety of applications and risk management provide an exclusive analysis and decision-making processes.

It is undeniable that big data provide a dynamic competitive environment for any enterprise in any industry that implementing such technologies in their frameworks offering a unique experience to achieve results. As my analysis showed, building a risk management framework, tailored to the era of big data, is required in order to be able to monitor and detect frauds establishing risk control processes and functions.

However, some privacy issues arise when data mining processes are linked to surveillance. There is an increase in privacy concerns associated to big data analytics, since a huge portion of big data insights consists of information being collected through any data source that includes customer’s details. Big data organizations and scientists fail to understand the difference between privacy and convenience, and therefore they should put an effort and take actions in resolving these concerns.

Big data analytics is gaining more attention nowadays, but there are some research problems that need to be addressed. For example, the existing storage and processing systems might not be capable enough to efficiently store the increasing growth of data and also processing the oversupply of unstructured data as traditional databases lack scalability and expandability (Bhadani & Jothimani, 2016). Furthermore, big data researchers should pay attention to the value of data and the difference between big data and whole data and hence, proper data filtering techniques must be developed. Most applications of big data such as fraud detection require only the valuable data to extract important and relevant information, entailing real time or near real time data. It would definitely be good to do continuous research and studies in order to be always informed and prepared for any possibility.

CHAPTER 3

Research Methodology

3.1. The Purpose of the Research

The research concentrated on how the businesses in Cyprus use the big data analytics, how it affects them and in which sectors are implemented. How big data techniques and algorithms contribute to risk management. Moreover, it demonstrates what risks the businesses that do not use data analytics have and how they are affected.

3.2. Research Questions

The research questions are concerned to the following

- ✓ What are big data analytics and how they help/affect businesses?
- ✓ How businesses adopt big data analytics and use them in decision making??
- ✓ Which big data analytics techniques and implementation tools are used?
- ✓ How big data analytics are used in the economy of Cyprus and what are the risks of using big data analytics?
- ✓ How big data analytics help businesses in risk detection?

3.3. Research Methodology

Theoretical analysis of the topic is utilized, as well as a mixture of quantitative and qualitative techniques. The data have been collected by using questionnaires and interviews with companies and people who are using the big data analytics, asking them to respond on the importance of the use of the big data analytics. In advance, following a systematic literature review, the reserarcher has separated the review in stages. Firstly, the researcher developed a review protocol and performed an in-depth research for studies, followed by data extraction by using qualitative and quantitative data and analysis of past findings (McCombes, 2019).

Secondly, the researher have developed a questionnaire collecting and observing relevant data and information allowing me to better understand the usage of big data in risk and fraud detection. Lastly, the researcher has evaluated the research results according to the purpose and the scope of this thesis, discussed the research directions that should attract more attention in the future and also addressed the research gap in the literature.

3.4. The Sample of the Research

The 49,3% od the participants were female; another same percent were male and there was one participant that didn’t want to specify his/her gender. Most of the participants were from 30-39 years old, the 22,7% were from 18-29 years old, the 13,3% were from 40-49 years old and the rest 10,7% were 60 years old and up. Half of the participants had higher or the highest degree in Education (50,7%), the 45,4% has either College or bachelor’s degree and the rest 4% were High-school graduates.

In advance, most of the participants (46,7%) work in large organizations consisting of more than 150 employees, the 21,3% work for very small organizations consisting of less than 10 employees, the 18,7% work for small organization that have 10-49 employees and the rest 13,3% work for medium size organizations that have 50-149 employees. Most participants said that their organization offer Financial services (17,3%), the 12% are in Legal sector, the 10,7% in Education, an 9,3% in Technology, an 8% in Telecommunications and another same percent in Public sector.

CHAPTER 4

Results Analysis

4.1. Demographic Features

The 49,3% od the participants were female; another same percent were male and there was one participant that didn't want to specify his/her gender. Most of the participants were from 30-39 years old, the 22,7% were from 18-29 years old, the 13,3% were from 40-49 years old and the rest 10,7% were 60 years old and up. Half of the participants had higher or the highest degree in Education (50,7%), the 45,4% has either College or Bachelor degree and the rest 4% were High-school graduates.

Table1.Demographichs features of our sample

		N	%	Cumulative Percent
Gender	Female	37	49,3	49,3
	Male	37	49,3	98,7
	Other	1	1,3	100,0
Age	18-29	17	22,7	22,7
	30-39	34	45,3	68,0
	40-49	10	13,3	81,3
	50-59	8	10,7	92,0
	60+	6	8,0	100,0
Education	High School Diploma	3	4,0	4,0
	College	11	14,7	18,7
	Bachelor's Degree	23	30,7	49,3
	Master's Degree	32	42,7	92,0
	Ph.D. or higher	3	4,0	96,0
	Post graduate professional qualification	3	4,0	100,0

Information of the Organization they work for

Most of the participants (46,7%) work in large organizations consisting of more than 150 employees, the 21,3% work for very small organizations consisting of less than 10 employees, the 18,7% work for small organization that have 10-49 employees and the rest 13,3% work for medium size organizations that have 50-149 employees. Most participants said that their organization offer Financial services (17,3%), the 12% are in Legal sector, the 10,7% in Education, an 9,3% in Technology, an 8% in Telecommunications and another same percent in Public sector.

Table 2 Size and sector of their Organization

	N	%	Cumulative Percent
Very Small (1-9 employees)	16	21,3	21,3
Small (10-49)	14	18,7	40,0
Medium (50-149)	10	13,3	53,3
Large(150 or more)	35	46,7	100,0
Manufacturing	4	5,3	5,3
Entertainment and Media	2	2,7	8,0
Telecommunications	6	8,0	16,0
Energy and Utilities	2	2,7	18,7
Public sector	6	8,0	26,7
Research / Academia	4	5,3	32,0
Transport	1	1,3	33,3
Healthcare	4	5,3	38,7
Financial Services	13	17,3	56,0
Technology	7	9,3	65,3
Legal	9	12,0	77,3
Education	8	10,7	88,0
Insurance	3	4,0	92,0

Imports	1	1,3	93,3
Sales	5	6,7	100,0

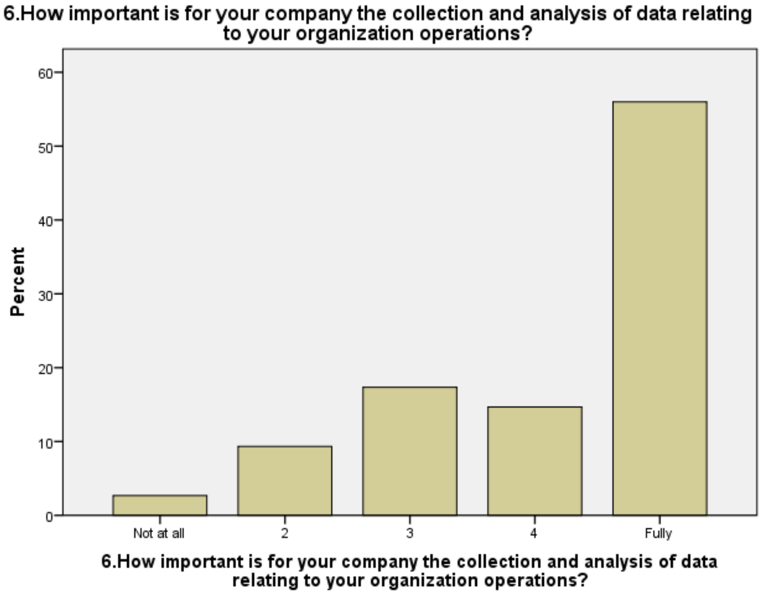
4.2. Main Questionnaire Analysis

According to table 3, the majority of the participants (56%) agreed with the opinion that the collection and analysis of data relating to their organization operations is highly important for their companies, a 14,7% consider them important, a 17,3% consider them neither important nor unimportant, whereas a 12% percent consider them of little importance or not important at all.

Table 3.How important is for your company the collection and analysis of data relating to your organization operations?(Q6)

	N	%	Cumulative Percent
Not at all	2	2,7	2,7
2	7	9,3	12,0
3	13	17,3	29,3
4	11	14,7	44,0
Fully	42	56,0	100,0
Total	75	100,0	

Chart 1

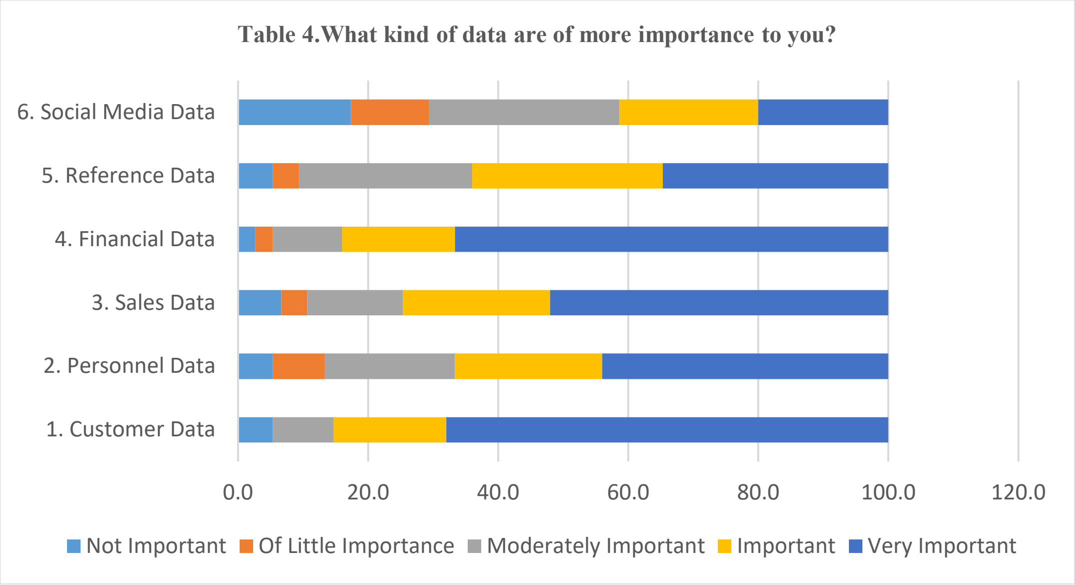


According to table 4, the majority of the participants consider almost all kinds of data very important. More specifically, the 68% of the participants consider most important data Customer data, then Financial Data (66,7%), Sales Data (52%), Personnel Data (44%) and Reference Data (34,7%). The 29,3% of the participants consider Social media data moderately important.

Table 4.What kind of data are of more importance to you?(Q7)

	Not Important	Of Little Importance	Moderately Important	Important	Very Important
1. Customer Data	5,3		9,3	17,3	68,0
2. Personnel Data	5,3	8,0	20,0	22,7	44,0
3. Sales Data	6,7	4,0	14,7	22,7	52,0
4. Financial Data	2,7	2,7	10,7	17,3	66,7
5. Reference Data	5,3	4,0	26,7	29,3	34,7
6. Social Media Data	17,3	12,0	29,3	21,3	20,0

Chart 2.



According to table 5, the majority of the participants (26,7%) fully agree with the opinion that they have a general understanding of big data, a 17,3% agree as well, a 28% consider them neither important nor unimportant, whereas a 28% percent consider them of little importance or not important at all.

Table 5. Do you have a general understanding of big data?(Q8)

	N	%	Cumulative Percent
Not at all	15	20,0	20,0
2	6	8,0	28,0
3	21	28,0	56,0
4	20	26,7	82,7
Fully	13	17,3	100,0
Total	75	100,0	

According to table 6, the majority of the participants (25,3%) fully agree with the opinion that their organization has experience with big data, a 16% agree as well, a 22,7% said that their organization is neither experienced nor unexperienced with big data, a 21,3% said that their organization has little experience and the rest 14,7% has no experience at all.

Table 6. Does your organization have experience with big data?(Q9)

	N	%	Cumulative Percent
Not at all	11	14,7	14,7
2	16	21,3	36,0
3	17	22,7	58,7
4	12	16,0	74,7
Fully	19	25,3	100,0
Total	75	100,0	

According to table 7, the majority of the participants (45,3%) currently analyze big data in their company, a 30,7% don't currently analyze big data and the rest 24% don't know whether their company analyze big data at the moment or not.

Table 7.Do you currently analyze any of the data collected in your company? (Q10)

	N	%	Cumulative Percent
Yes	34	45,3	45,3
No	23	30,7	76,0
Don't know	18	24,0	100,0
Total	75	100,0	

According to table 8, the majority of the participants (78,7%) use Microsoft Excel as a software for data analysis, the 26,7% use SQL, the 25,3% use CRM, the 13,3% use SPSS, the 10,7% use R, another 10,7% use Python and another same percent use a custom-made web application.

Table 8.What software or tools you are using for data analysis? (Q11)

	Yes %	No %
1. Microsoft Excel	78,7	21,3
2. R	10,7	89,3
3. Python	10,7	89,3
4. SPSS	13,3	86,7
5. SQL	26,7	73,3
6.Apache Spark	6,7	93,3
7. Tableau	5,3	94,7
8. SAS	6,7	93,3
9.CRM	25,3	74,7
10. ArcGis	2,7	97,3
11.None	10,7	89,3
12. Teradata	2,7	97,3
13. Custom-made web application	10,7	89,3

According to table 9, the majority of the participants (45,3%) said that their company employs a data analyst a/or data scientist, a/or data engineer, a 37,3% don't employ a data expert and the rest 17,3% don't know.

Table 9.Does your company employ data analyst, data scientist, or data engineer?(Q12)

	N	%	Cumulative Percent
Yes	28	37,3	37,3
No	34	45,3	82,7
Don't know	13	17,3	100
Total	75	100	

According to table 10, the majority of the participants (44%) said that their team maybe utilizes modern predictive modeling, analytics, or machine learning, a 40% utilizes modern predictive modeling, analytics, or machine learning and the rest 16% don't utilize them.

Table 10. Does your team utilize modern predictive modeling, analytics, or machine learning? Do you think it could benefit from using it?(Q14)

	N	%	Cumulative Percent
Yes	30	40,0	40,0
No	12	16,0	56,0
Maybe	33	44,0	100,0
Total	75	100,0	

According to table 11, the majority of the participants (53,3%) said that the IT is the department of their organization that is involved in using data technologies and data analytics, the 40% referred the Business Development Department, the 34,7% referred the Customer Service Department, the 28% referred the HR department, the 24% referred the Research Department and the 21,3% referred the Operation Department and another same percent referred the Executive management Department.

Table 11. Which departments in your organization are involved in using data technologies and data analytics? (Q15)

	Yes %	No %
1. IT	53,3	46,7
2. HR	28,0	72,0

3. Logistics	16,0	84,0
4.Operation	21,3	78,7
5. Research	24,0	76,0
6. Marketing	21,3	78,7
7. Customer Service	34,7	65,3
8. Business Development	40,0	60,0
9. Executive Management	21,3	78,7
10. Don't know	14,7	85,3
11. Analytical teams	13,3	86,7

According to table 12, the majority of the participants (60%) said that one of the sources their organization collects data now is from the transactions, the 58,7% from emails, the 46,7% from PSI, the 41,3% from Phone usage, the 36% from Social media, the 34,7% from events and the 33,3% from Audio / Images/ Videos. The majority of the participants don’t know if their organization collects data from sensors (50,7%), free-form tests (40%) and external feeds (34,7%).

Table 12. From what sources does your organization collect, or expects to collect, data? (Q16)

	Collects now	Expects to collect in 5 years	No plans although it could be useful for our organization	No plans. We think that we do not need so advanced technology	Don't know
1. Transactions	60,0	5,3	12,0	5,3	17,3
2. Events	34,7	20,0	9,3	9,3	26,7
3. Emails	58,7	8,0	6,7	4,0	22,7
4. Social media	36,0	13,3	13,3	9,3	28,0
5. Sensors	20,0	10,7	10,7	8,0	50,7
6. PSI	46,7	8,0	17,3	2,7	25,3
7. Phone usage	41,3	12,0	14,7	6,7	25,3
8. External feeds	33,3	14,7	14,7	2,7	34,7
9. Free-form text	36,0	8,0	12,0	4,0	40,0
10. Audio / Images/ Videos	33,3	18,7	16,0	2,7	29,3

According to table 13, the 33,3% of the participants said that they don't know if their organization has the right analytical tools to handle big data, the 32% said dad their organization has the right analytical tools to handle big data now, the 16% expects to have the necessary tools in 5 years and the 13,3% said that they have no plans although they are aware that it could be useful for their organization. The rest 5,3% said that that they have no plans and believe that they do not need so advanced technology.

Table 13.Does your organization have the right analytical tools to handle big data? (Q17)

	N	%	Cumulative Percent
Has now	24	32,0	32,0
Expects to have in 5 years	12	16,0	48,0
No plans although it could be useful for our organization	10	13,3	61,3
No plans. We think that we do not need so advanced technology	4	5,3	66,7
Don't know	25	33,3	100,0
Total	75	100,0	

According to table 14, the 36% of the participants said that they don't know if their organization has the right analytical tools to handle unstructured data expressed in natural language, the 25,3% said that their organization has the right analytical tools to handle unstructured data expressed in natural language now, the 14,7% expects to have the necessary tools in 5 years and another same percent said that they have no plans although they are aware that it could be useful for their organization. The rest 9,3% said that that they have no plans and believe that they do not need so advanced technology.

Table 14.Does your organization have the right tools to handle unstructured data expressed in (a) natural language(s)?(Q18)

	N	%	Cumulative Percent
Has now	19	25,3	25,3
Expects to have in 5 years	11	14,7	40,0
No plans although it could be useful for our organization	11	14,7	54,7
No plans. We think that we do not need so advanced technology	7	9,3	64,0
Don't know	27	36,0	100,0
Total	75	100,0	

According to table 15, the 32% of the participants said that the percentage that is further processed for value generation, from all the data collected by their organization, is less than 10% currently whereas the 45,4% (cumulatively) said that the expected percentage of data to be processed in 5 years from now is ranges from 10-60%.

Table 15. From all the data collected by your organization, what is approximately the percentage that is further processed for value generation? (Q19)

	<10%	10-40%	41-60%	61-90%	>90%
Currently	32,0	28,0	24,0	13,3	2,7
Expected (in 5 years)	20,0	22,7	22,7	17,3	17,3

According to table 16, the 60% of the participants said that big data creates value in their organization now by Improved customer targeting, the 49,3% by better product design, the 46,7% by Improved customer loyalty and retention, the 44% by Efficiency increase, the 41,3% by new business model and the 37,3% by Risk/Financial Management.

Table.16 In which way does big data create, or is expected to create, value in the organization? (Q20)

	Creates value now	Expects to create value in 5 years	Does not create value	Don't know
Improved customer targeting	60,0	17,3	9,3	13,3
Improved customer loyalty and retention	46,7	29,3	10,7	13,3
Efficiency increase	44,0	36,0	5,3	14,7
Better product design	49,3	16,0	16,0	18,7
Risk/Financial Management	37,3	29,3	6,7	26,7
New business model	41,3	24,0	12,0	22,7

According to table 17, the 61,3% of the participants said that the timeliness is a very important big data-related challenge, the 58,7% referred to Privacy concerns and regulatory risks, the 54,7% to Data quality, the 52% referred to the Access rights to data, another same percent referred to the Privacy concerns and regulatory risks, the 49,3% to Availability of data, the 45,3% the Cost of data, the 42,7% referred to the Lack of facilities, infrastructure, the 41,3% the Data ownership issues, the 40% the Shortage of talent/skills, the 36% referred to the Overwhelming volume, the 33,3% to the Managing unstructured data and another same percent the Corporate culture.

	Important	Very Important
1. Timeliness	26,7	61,3
2.Overwhelming volume	34,7	36,0

3.Managing unstructured data	22,7	33,3
4.Data quality	21,3	54,7
5.Availability of data	34,7	49,3
6.Access rights to data	26,7	52,0
7.Data ownership issues	34,7	41,3
8.Cost of data	37,3	45,3
9.Lack of facilities, infrastructure	32,0	42,7
10.Shortage of talent/skills	36,0	40,0
11.Privacy concerns and regulatory risks	30,7	52,0
12.Security	20,0	58,7
13.Corporate culture	29,3	33,3

According to table 18, the 49,3% (cumulatively) of the participants said that the data-driven add a lot to the competitive advantage of their company and the 29,3% said that they add moderately.

Table 18.To what extent does data-driven innovation add to the competitive advantage (CA) of your company? (Q22)

	N	%	Cumulative Percent
Small CA	7	9,3	9,3
2	9	12,0	21,3
3	22	29,3	50,7
4	19	25,3	76,0
Large CA	18	24,0	100,0
Total	75	100,0	

According to table 19, more than half of the participants said that they consider very important having consistent, integrated security and governance for their data and the 18,7% consider it moderately important.

Table 19.How important is having consistent, integrated security and governance for your data? (Q23)

	N	%	Cumulative Percent
Very important	38	50,7	50,7
Important	9	12	62,7
Moderately Important	14	18,7	81,3

Of Little Importance	6	8	89,3
Not Important	8	10,7	100
Total	75	100	

According to table 20, the 81,3% of the participants said that big data provide valuable insights on risk decision making.

Table 20.Do you believe that big data provide valuable insights on risk decision making? (Q24)

	N	%	Cumulative Percent
Yes	61	81,3	81,3
Maybe	14	18,7	100,0
Total	75	100,0	

4.3. Research Questions Analysis

- *How businesses adopt big data analytics and use them in decision making?*

The majority of the participants (56%) agreed with the opinion that the collection and analysis of data relating to their organization operations is highly important for their companies, a 14,7% consider them important, a 17,3% consider them neither important nor unimportant, whereas a 12% percent consider them of little importance or not important at all.

- *Which big data analytics techniques and implementation tools are used?*

Most of the participants said that they use Microsoft Excel as a software for data analysis, SQL, CRM, SPSS, R, Python and custom-made web applications.

- *How big data analytics are used in the economy of Cyprus and what are the risks of using big data analytics?*

The majority of the participants (60%) said that one of the sources their organization collects data now is from the transactions, the 58,7% from emails, the 46,7% from PSI, the 41,3% from Phone usage, the 36% from Social media, the 34,7% from events and the 33,3% from Audio / Images/ Videos. The majority of the participants don't know if their organization collects data from sensors (50,7%), free-form tests (40%) and external feeds (34,7%).

- *How big data analytics help businesses in risk detection?*

The 81,3% of the participants said that big data provide valuable insights on risk decision making

4.4. Research Hypothesis

1st.Does the size of the organization the employees work for, affects their opinion about the importance of the collection and analysis of data relating to their organization operations?

In order to confirm or not the above hypothesis, we conducted crosstabulation among the variables with chi-square test.

Table 21.Size of the organization (number of employees) * and How important is for your company the collection and analysis of data relating to your organization operations?

		6.How important is for your company the collection and analysis of data relating to your organization operations?					
		Not at all	2	3	4	Fully	Total
Very Small (1-9 employees)	Count	0	4	1	0	11	16
	% of Total	0,0%	5,3%	1,3%	0,0%	14,7%	21,3%
Small (10-49)	Count	1	2	5	2	4	14
	% of Total	1,3%	2,7%	6,7%	2,7%	5,3%	18,7%
Medium (50-149)	Count	1	1	2	2	4	10
	% of Total	1,3%	1,3%	2,7%	2,7%	5,3%	13,3%
Large(150 or more)	Count	0	0	5	7	23	35
	% of Total	0,0%	0,0%	6,7%	9,3%	30,7%	46,7%
Total	Count	2	7	13	11	42	75
	% of Total	2,7%	9,3%	17,3%	14,7%	56,0%	100,0%

From table 22, we see that among our variables we have a statistical importance as p<0,05. More analytically, we see from table 21, that most of the participants that worked in very small companies, as well as medium and large companies, consider very important the collection and analysis of data relating to their organization operations whereas most participants that worked in small companies consider it moderately important.

Table 22.Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22,980[a]	12	,028
Likelihood Ratio	27,438	12	,007
Linear-by-Linear Association	3,616	1	,057
N of Valid Cases	75		

a. 14 cells (70,0%) have expected count less than 5. The minimum expected count is ,27.

2nd.Does the Age of the participants, affects their general understanding they have of big data?

Table 23.Age and 8.Do you have a general understanding of big data? Crosstabulation

		8.Do you have a general understanding of big data?					
		Not at all	2	3	4	Fully	Total
18-29	Count	3	1	8	3	2	17
	% of Total	4,0%	1,3%	10,7%	4,0%	2,7%	22,7%
30-39	Count	3	1	12	11	7	34
	% of Total	4,0%	1,3%	16,0%	14,7%	9,3%	45,3%
40-49	Count	1	3	0	4	2	10
	% of Total	1,3%	4,0%	0,0%	5,3%	2,7%	13,3%
50-59	Count	4	1	1	1	1	8
	% of Total	5,3%	1,3%	1,3%	1,3%	1,3%	10,7%
60+	Count	4	0	0	1	1	6
	% of Total	5,3%	0,0%	0,0%	1,3%	1,3%	8,0%
Total	Count	15	6	21	20	13	75
	% of Total	20,0%	8,0%	28,0%	26,7%	17,3%	100,0%

From table 24, we see that among our variables we have a statistical importance as p<0,05. More analytically, we see from table 23, that most of the participants that were from 18-29 years old, said that they have moderate general understanding of big data, most of the participants that were from 30-49 years old, have full or good general understanding of big data general understanding of big data, whereas most participants that were from 50 years old and up, said that they don’t a general understanding of big data at all.

Table 24.Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	31,824a	16	0,011
Likelihood Ratio	31,635	16	0,011
Linear-by-Linear Association	3,994	1	0,046
N of Valid Cases	75		

a 21 cells (84,0%) have expected count less than 5.
The minimum expected count is ,48.

CHAPTER 5

Discussion - Conclusion

According to what mentioned above from the results' analysis, the majority of the participants (56%) agreed with the opinion that the collection and analysis of data relating to their organization operations is highly important for their companies, a 14,7% consider them important, a 17,3% consider them neither important nor unimportant, whereas a 12% percent consider them of little importance or not important at all.

The majority of the participants consider almost all kinds of data very important. More specifically, the 68% of the participants consider most important data Customer data, then Financial Data (66,7%), Sales Data (52%), Personnel Data (44%) and Reference Data (34,7%). The 29,3% of the participants consider Social media data moderately important as also the majority of the participants (26,7%) fully agree with the opinion that they have a general understanding of big data, a 17,3% agree as well, a 28% consider them neither important nor unimportant, whereas a 28% percent consider them of little importance or not important at all.

Moreover, the majority of the participants (25,3%) fully agree with the opinion that their organization has experience with big data, a 16% agree as well, a 22,7% said that their organization is neither experienced nor unexperienced with big data, a 21,3% said that their organization has little experience and the rest 14,7% has no experience at alla s also the majority of the participants (45,3%) currently analyze big data in their company, a 30,7% don't currently analyze big data and the rest 24% don't know whether their company analyze big data at the moment or not.

The majority of the participants (78,7%) use Microsoft Excel as a software for data analysis, the 26,7% use SQL, the 25,3% use CRM, the 13,3% use SPSS, the 10,7% use R, another 10,7% use Python and another same percent use a custom-made web application, the majority of the participants (45,3%) said that their company employs a data analyst a/or data scientist, a/or data engineer, a 37,3% don't employ a data expert and the rest 17,3% don’t know and the majority of the participants (44%) said that their team maybe utilizes modern predictive modeling, analytics, or machine learning, a 40% utilizes modern predictive modeling, analytics, or machine learning and the rest 16% don’t utilize them.

The majority of the participants (53,3%) said that the IT is the department of their organization that is involved in using data technologies and data analytics, the 40% referred the Business Development Department, the 34,7% referred the Customer Service Department, the 28% referred the HR department, the 24% referred the Research Department and the 21,3% referred the Operation Department and another same percent referred the Executive management Department as also the majority of the participants (60%) said that one of the sources their organization collects data now is from the transactions, the 58,7% from emails, the 46,7% from PSI, the 41,3% from Phone usage, the 36% from Social media, the 34,7% from events and the 33,3% from Audio / Images/ Videos. The majority of the participants don’t know if their organization collects data from sensors (50,7%), free-form tests (40%) and external feeds (34,7%).

Moreover, the 33,3% of the participants said that they don’t know if their organization has the right analytical tools to handle big data, the 32% said dad their organization has the right analytical tools to handle big data now, the 16% expects to have the necessary tools in 5 years and the 13,3% said that they have no plans although they are aware that it could be useful for their organization. The rest 5,3% said that that they have no plans and believe that they do not need so advanced technology and the 36% of the participants said that they don’t know if their organization has the right analytical tools to handle unstructured data expressed in natural language, the 25,3% said that their organization has the right analytical tools to handle unstructured data expressed in natural language now, the 14,7% expects to have the necessary tools in 5 years and another same percent said that they have no plans although they are aware that it could be useful for their organization. The rest 9,3% said that that they have no plans and believe that they do not need so advanced technology.

Furthermore, the 32% of the participants said that the percentage that is further processed for value generation, from all the data collected by their organization, is less than 10% currently whereas the 45,4% (cumulatively) said that the expected percentage of data to be processed in 5 years from now is ranges from 10-60%, the 60% of the participants said that big data creates value in their organization now by Improved customer targeting, the 49,3% by better product design, the 46,7% by Improved customer loyalty and retention, the 44% by Efficiency increase, the 41,3% by new business model and the 37,3% by Risk/Financial Management.

In advance, the 61,3% of the participants said that the timeliness is a very important big data-related challenge, the 58,7% referred to Privacy concerns and regulatory risks, the 54,7% to Data quality, the 52% referred to the Access rights to data, another same percent referred to the Privacy concerns and regulatory risks, the 49,3% to Availability of data, the 45,3% the Cost of data, the 42,7% referred to the Lack of facilities, infrastructure, the 41,3% the Data ownership issues, the 40% the Shortage of talent/skills, the 36% referred to the Overwhelming volume, the 33,3% to the Managing unstructured data and another same percent the Corporate culture and the 49,3% (cumulatively) of the participants said that the data-driven add a lot to the competitive advantage of their company and the 29,3% said that they add moderately and more than half of the participants said that they consider very important having consistent, integrated security and governance for their data and the 18,7% consider it moderately important and the 81,3% of the participants said that big data provide valuable insights on risk decision making.

As to w*hich big data analytics techniques and implementation tools are used,* most of the participants said that they use Microsoft Excel as a software for data analysis, SQL, CRM, SPSS, R, Python and custom-made web applications. As to *how big data analytics are used in the economy of Cyprus and what are the risks of using big data analytics,* the majority of the participants (60%) said that one of the sources their organization collects data now is from the transactions, the 58,7% from emails, the 46,7% from PSI, the 41,3% from Phone usage, the 36% from Social media, the 34,7% from events and the 33,3% from Audio / Images/ Videos. The majority of the participants don't know if their organization collects data from sensors (50,7%), free-form tests (40%) and external feeds (34,7%).

Finally, as to how big data analytics help businesses in risk detection, the 81,3% of the participants said that big data provide valuable insights on risk decision making.

As to the basic **Research Hypothesis,** if *does the size of the organization the employees work for, affects their opinion about the importance of the collection and analysis of data relating to their organization operations,* in order to confirm or not the above hypothesis, we conducted crosstabulation among the variables with chi-square test.

Moreover, it was found that most of the participants that worked in very small companies, as well as medium and large companies, consider very important the collection and analysis of data relating to their organization operations whereas most participants that worked in small companies consider it moderately important as also most of the participants that were from 18-29 years old, said that they have moderate general understanding of big data, most of the participants that were from 30-49 years old, have full or good general understanding of big data general understanding of big data, whereas most participants that were from 50 years old and up, said that they don't a general understanding of big data at all.

Therefore, and based on the above results, the continuous evolution of technology and especially in the management of large data volumes have made Big Data a technological challenge of our time for many companies and organizations. The purpose of the dissertation was to explore the Big Data technology, the technologies that help in the analysis and storage of large volumes of data as well as the roles and opportunities that exist for Greek companies through the use of Big Data in the Cypriot economy, of research show the importance and power of Big Data in today's Cypriot business.

Although Big Data is not yet so well established in our country, according to the answers of the majority of respondents who state that Big Data is rarely used in Greek business, there is a hopeful and encouraging attitude of Greek companies towards them. With the largest percentage of respondents, who do not use Big Data, having answered that they will definitely invest in this technology in the near future and with most believing that the use of Big Data will be very important in the functionality of the business in the next 5 years, it shows us that in the future this technology will be an important factor for companies that want to grow, thrive and innovate.

The Big Data, although it has always existed in our lives, was discovered and began to be explored in recent years due to the evolution of technology. Their emergence has brought about radical changes in both the way human societies operate and the way businesses operate. Data that was considered useless or unmanageable is now considered extremely useful for extracting valuable information.

The constant evolution of technologies combined with the science of Big Data has a huge impact on today's business, helping companies, large and small, to have a powerful weapon for improved decision making as well as better real-time information for all parts of the business. It is certain that Big Data has a wide range of possibilities with great prospects for the future, helping businesses to constantly evolve and grow, in order to achieve the maximum possible business benefit.

By summarizing and based on what has been studied Big Data is and will remain important in the development of any business, as long as they are utilized in the right way. In addition, in the future there will be a great need for further research in the field of Artificial Intelligence but also its technologies, Machine Learning and Deep Learning, which are extremely important in Big Data science. Artificial Intelligence will play a key role in the development and improvement of techniques and applications of Big Data and will be an important success factor for many companies.

Therefore, the results of the primary specific analysis, they showed that most of the employees on a 56 percent of the total questioned believe that the collection and analysis of data is fully important for an organization to smoothly operate while only 2 have the opinion that big data are not important at all, the customer data are the most important data collected with a percentage of 68% and 5,3% believe that customer data are not important at all.

Secondly and most important data, there are financial data with only one vote less of the customer data and just 2 of the surveyed thinks that financial data are not that important. On the third place are the sales data with 52% of positive votes and a percentage of 6,7% believe that Sales Data are not that important for their organization. The Personnel Data get an 44% of the importance of data collected while 5,3% have the opinion that personnel data does not impact the efficiency of their organization.

Finally, 26 people think that if you collect Reference Data this could benefit the operations of their businesses and 3 people believe that the ranking of this data is of a little importance and the least important data according to the survey are the social media data that only 15 people trust that social media data are very important and 13 estimate that this kind of data are not important at all.

As to the relation of the above mentioned and the results of the primary research with the use of questionnaires, it could be said that in order to be able to identify and assess the risks that currently threaten the big data of each business entity, one must first take into account the modern environment of risk as well as its development trends.

It should be noted that the majority of violations have not been reported by companies. Big data breaches are becoming more widespread and attack trends are showing no signs of slowing down. They mainly target high value data such as social security numbers, health information, credit and debit card numbers, emails, passwords and other user access information. The following are the most important trends and technological developments that characterize the external environment of today's business, as well as the most critical threats to personal data arising from it:

✓ **Mobile telephony**: Undoubtedly, both the confidentiality and the integrity of information through mobile phones are fraught with many risks, technical and non-technical. Multimedia content stored on mobile phones, such as photos, videos and other related data files, can now easily be found in the possession of an attacker, while intercepting calls and messages is particularly common. Given that attacks on mobile devices are constantly increasing, more and more such breaches are expected in businesses as well, leading to particular concerns about the security of corporate data.

✓ **Critical infrastructures**: Critical infrastructures of vital importance are defined as goods, systems or subsystems, which become necessary for the maintenance of the vital functions of society, health, physical protection, security, economic and social prosperity. Communication and information systems are among the critical ones of the country, the security of which is now a major issue of national interest.

✓ **Internet of Things**: It is the communication network of a variety of electronic devices as well as any object that integrates electronic means, software, sensors and network connectivity, to allow the connection and exchange of data. Cybercriminals' networks often exploit the flexibility of the security of such devices, in order to spread malware, which infects the user, then demands money from him to remove the infection. Most of these attacks usually target shared devices, such as servers, routers, CCTV systems, network storage devices, and industrial control systems.

REFERENCES

Ahsaan, S. U., & Mourya, A. K. (2019). BIG DATA ANALYTICS: CHALLENGES AND TECHNOLOGIES. *ANNALS of Faculty Engineering Hunedoara – International Journal of Engineering*, 1-5.

Analytics, t. o. (2020, October 12). *SelectHub*. Retrieved from What are the Types of Big Data?: https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/

Abhay Kumar Bhadani, Dhanya Jothimani (2016). Big data: Challenges, opportunities and realities, Effective Big Data Management and Opportunities for Implementation, pp.1-24.

Ardavan Ashabi, Shamsul Bin Sahibuddin, Medhi Salkhordeh Haghighi (2020). Big Data: Current Challenges and Future Scope. Received from https://ieeexplore.ieee.org/document/9108826

Avita Katal, Mohammad Wazid, R. H. Goudar (2013). Big Data: Issues, Challenges, Tools and Good Practices. Received from https://ieeexplore.ieee.org/document/6612229/references#references

Almeida Fernando (2018). Big Data: Concept, Potentialities and Vulnerabilities, Vol. 2, No. 1.

Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih (2017). Big Data technologies: A survey, Journal of King Saud University – Computer and Information Sciences, 30(2018), pp.431-448.

Anna Brzozowska, Leszek Ziora, Robert Sałek, Anna Wiśniewska-Sałek (2016). The Possibilities of Big Data Solutions Application in Logistics. International Multidisciplinary Scientific Conference University of Miskolc.

Alexandre da Silva Veith, Marcos Dias de Assunção (2018). Apache Spark. Inria Avalon, LIP Laboratory, ENS Lyon, University of Lyon. Received from https://www.researchgate.net/publication/323447097_Apache_Spark

Apache CASSANDRA. (2016). Retrieved from Manage massive amounts of data, fast, without losing sleep: https://cassandra.apache.org/

Apache Hadoop. (2021, January 10). Retrieved from The Apache Software Foundation: https://hadoop.apache.org/

Apache HBase. (2021, 01 29). Retrieved from http://hbase.apache.org/

APACHE HIVE TM. (2014). Retrieved from https://hive.apache.org/

Apache Storm. (2019). Retrieved from Apache Software Foundation: https://storm.apache.org/

Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2015, March). An overview of big data opportunities, applications and tools. In 2015 Intelligent Systems and Computer Vision (ISCV) (pp. 1-6). IEEE.

Big Data Analytics: Challenges and Technologies, UI AHSAAN, Shafqat, MOURYA, Ashish Kumar, Nov2019, Vol. 17 Issue 4, p75-79. 5p.

Bhadani, A. K., & Jothimani, D. (2016). Big Data: Challenges, opportunities and Realities. *Effective Big Data Management and Opportunities for Implementation*, 1-30.

Chong, D., & Shi, H. (2015). Big data analytics: a literature review. *Journal of Management Analytics*, 175-201.

Bernard Marr (2016). BIG DATA IN PRACTICE. United Kingdom: John Wiley and Sons Ltd.

Balamurugan Balusamy, R Nandhini Abirami, Seifedine Kadry, Amir H. Gandomi (2021). Big Data: Concepts, Technology, and Architecture. John Wiley and Sons Inc.

Dr Mark van Rijmenam (2013, August 7). Why The 3V's Are Not Sufficient To Describe Big Data. Retrieved from https://datafloq.com/read/3vs-sufficient-describe-big-data/166

Daniel Trabucchi, Tommaso Buganza (2018). Data-driven innovation: switching the perspective of Big Data. European Journal of Innovation Management, Vol. 22 (No.1), pp.23-40.

Frank Ohlhorst (2012). Big Data Analytics: Turning Big Data into Big Money. Canada: John Wiley & Sons, Inc.

Garry Kranz, Dave Raffo (2018, May). Storage (computer storage). Retrieved from https://searchstorage.techtarget.com/definition/storage

Gema Bello-Orgaz, Jason J. Jung, David Camacho (2015). Social big data: Recent achievements and new challenges, Information Fusion, 28(2016), pp.45-59.

Goel, P., Datta, A., Mannan, M., O'Connor, M. K., & McFerrin, A. (2017). Application of Big Data analytics in process safety and risk management. *IEEE International Conference on Big Data (BIGDATA)*, 1-10.

Grover, P., & Kar, A. K. (2017). Big Data Analytics: A Review on Theoretical Contributions. *Global Journal of Flexible Systems Management*, 1-27.

Grover, P., & Kar, A. K. (2017, June 13). Big Data Analytics: A Review on Theoretical Contributions and Tools used in Literature. *Global Journal of Flexible Systems Management*. Global Institute of Flexible Systems Management.

Hossein Hassani, Xu Huang, Emmanuel Sirimal Silva (2019). Fusing Big Data, Blockchain and Cryptocurrency. Springer Nature Switzerland AG.

IBM. (2021). Retrieved from Apache HBase: https://www.ibm.com/analytics/hadoop/hbase

Khvoynitskaya, S. (2020, January 30). *The future of big data: 5 predictions from experts for 2020-2025.* Retrieved from itransition: https://www.itransition.com/blog/the-future-of-big-data

Kopanakis, J. (n.d.). *mentionlytics*. Retrieved from 5 Real-World Examples of How Brands are Using Big Data Analytics: https://www.mentionlytics.com/blog/5-real-world-examples-of-how-brands-are-using-big-data-analytics/

Koyuncugil, A. S., & Ozgulbas, N. (2012). Financial Early Warning System Model and Data Mining Application for Risk Detection. *Expert Systems with Applications*, 6238-6253.

L, A. B. (2020). *BIG DATA MADE SIMPLE*. Retrieved from Top 8 programming languages every data scientist should master in 2019: https://bigdata-madesimple.com/top-8-programming-languages-every-data-scientist-should-master-in-2019/

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), 2032-2033.

Lekha R. Nair, Sujala D. Shetty, Siddhanth D. Shetty (2016). INTERACTIVE VISUAL ANALYTICS ON BIG DATA: TABLEAU VS D3.JS. Journal of e- Learning and Knowledge Society, Vol. 12, No. 4, pp.139-150.

Marr Bernard (2015, February 24). A Brief History of Big Data Everyone Should Read. Retrieved from https://www.linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr

Min Chen, Shiwen Mao, Yunhao Liu (2014). Big Data: A Survey. Mobile Networks and Applications, Vol. 19 (No.2), pp.171-209

Mousouleas Ioannis (2019). Applying big data in pharmaceutical industry: development, strategy and administration. University of Patras, Patras.

McCombes, S. (2019, February 25). *How to write a research methodology*. Retrieved from Scribbr: https://www.scribbr.com/dissertation/methodology/

Mehta, A. (2017, October 13). *Analytics Insight*. Retrieved from FOUR TYPES OF BUSINESS ANALYTICS TO KNOW: https://www.analyticsinsight.net/four-types-of-business-analytics-to-know/

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2017). *Big data analytics capabilities: a systematic literature.* Trondheim: Norwegian University of Science and Technology, Trondheim, Norway.

Morgan Stanley. (2020, May). *Morgan Stanley Barney LLC*. Retrieved from Fraud Detection and Prevention: https://www.morganstanley.com/what-we-do/wealth-management/online-security/fraud-detection-prevention/

Neville, S. (2020, January 26). *Financial Times*. Retrieved from Why big pharma sees a remedy in data and AI: https://www.ft.com/content/4743d76c-af9b-11e9-8030-530adfa879c2

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., & Smith, R. (2018). News and narratives in financial systems: exploiting big data for systemic risk assessment.

ProjectPro. (2021, January 25). Retrieved from Types of Analytics: descriptive, predictive, prescriptive analytics: https://www.dezyre.com/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209

Sena, V., Bhaumik, S., Sengupta, A., & Demirbag, M. (2019). Big data and performance: what can management research tell us? British Journal of Management, 30(2), 219-228.

Software Testing Help. (2021, January 18). Retrieved from Top 15 Big Data Tools (Big Data Analytics Tools) In 2021: https://www.softwaretestinghelp.com/big-data-tools/

The R Foundation. (2021). Retrieved from What is R?: https://www.r-project.org/about.html

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big Data Analytics: Applications, Prospects and Challenges. *Engineering and Communications Technologies*, 3-20.

Watson, H. J. (2014). Tutorial: Big Data Analytics: Concepts. Technologies and Applications. *Communications of the Association for Information Systems*, 1247-1268.

WIKIPEDIA. (2021, January 24). Retrieved from https://en.wikipedia.org/wiki/SQL

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, 350-361