Big Data in Accounting

Virginia Athanasia Sotiropoulou

A dissertation submitted for the partial satisfaction of the requirements for obtaining a Postgraduate Diploma of Specialization

School of Business Administration
Department of Economics

M.Sc.in Applied Economics and Data Analysis

August 2018
Three-member Supervision Committee

**Supervisor:** Dimitrios Tzelepis  Associate Professor

**Member of the Commission:** Konstantinos Kounetas  Assistant Professor

**Member of the Commission:** Emmanouil Tzagarakis  Assistant Professor

The dissertation titled:

«Big Data in Accounting»

Was prepared by Virginia Athanasia Sotiropoulou, AM 1018840, for the partial satisfaction of the requirements for obtaining a Postgraduate Diploma in Applied Economics and Data Analysis from the University of Patras and was approved by the members of the three-member Supervisory Committee.
I would like to dedicate my diploma to my family.
Thanksgiving

I would like to thank my supervisor, Dimitrios Tzelepis, Associate Professor, for his trust, his patience and his valuable help. I would also like to thank the members of my supervision committee, Konstantinos Kounetas and Emmanouil Tzagarakis as well as all of my professors in this M.Sc. program. Finally, I would like to thank my family for their continuous support.
Περίληψη

Στόχος της παρούσας διπλωματικής εργασίας είναι να διερευνήσει τη σχέση των Μεγάλων Δεδομένων με τη Λογιστική. Για να διερευνηθεί αυτό, αναφέρουμε τις αλλαγές που έχουν επέλθει στη λογιστική από την τεράστια αύξηση του όγκου των πληροφοριών λόγω της ολοκληρωματικής και επικοινωνιακής ανάπτυξης των τελευταίων δυο-τριών δεκαετιών. Αυτές οι επιδράσεις είχαν τέτοιο τεράστιο αντίκτυπο ώστε επηρεάσαν και τα λογιστικά πρότυπα. Η νέα δυναμική στη λογιστική λόγω της ενσωμάτωσης των Μεγάλων Δεδομένων, επέφερε μεγάλες αλλαγές στην ελεγκτική. Αναλυτικότερα, έχουν διαφοροποιηθεί οι πηγές των δεδομένων που χρησιμοποιούνται στη λογιστική και στην ελεγκτική. Οι αλλαγές στις πηγές αφορούν και την ποσότητα των πληροφοριών και την ποιότητα. Η επεξεργασία δεν παρέμεινε στις παλιές κλασικές μεθόδους της λογιστικής και της ελεγκτικής αλλά επεκτάθηκε και τεχνολογικά (hardware και software). Οι αλλαγές αυτές στη λογιστική και στην ελεγκτική βοήθησαν πολύ από ποσοτική άποψη καταλήγοντας στα εξής μοντέλα: Α) μοντέλο οικονομικής δυσπραγίας, Β) μοντέλο οικονομικής απάτης και Γ) μοντέλο πρόβλεψης της αγοράς μετοχών και ποσοτική μοντελοποίηση.

Στη συνέχεια της διπλωματικής αναφέρονται τέσσερις αξιόλογες μελέτες περίπτωση που εστιάζουν στο πώς οι «Big 4 Accounting firms» αξιοποιούν τα Μεγάλα Δεδομένα στην ελεγκτική και κατ’ επέκταση στη λογιστική. Η πρώτη αφορά την εταιρεία Deloitte Touche Tohmatsu Limited η οποία είναι η μεγαλύτερη από τις «Big 4 Accounting firms» και το φάσμα των λογιστικών δραστηριοτήτων της περιλαμβάνει φορολογικές, συμβουλευτικές, ελεγκτικές υπηρεσίες καθώς και υπηρεσίες για συγχωνεύσεις και εξαγορές, για ιδιωτικές εταιρείες ή και συμβουλευτικές υπηρεσίες για θέματα ρίσκου και χρηματοοικονομικά θέματα. Διαχειρίζεται τεράστιο όγκο δεδομένων και χρησιμοποιεί εκσυγχρονισμένα συστήματα επεξεργασίας. Η δεύτερη εταιρεία είναι η PricewaterhouseCoopers (PwC) της οποίας το αντικείμενο αφορά επίσης την ελεγκτική αλλά και τη συμβουλευτική για θέματα φορολογίας και ιδιωτικών. Επίσης, ασχολείται με θέματα διακυβέρνησης και διαφάνειας κινδύνου. Πρόκειται για μια πολύ μεγάλη εταιρεία με παγκόσμιες λογιστικές δραστηριότητες. Στη συνέχεια της τρίτης μελέτης περίπτωσης αφορά την Ernst &Young Limited (EY), η οποία ασχολείται με φορολογικές, ελεγκτικές και συμβουλευτικές υπηρεσίες και έχει εξίσου μεγάλο επιστημονικό ενδιαφέρον. Η τελευταία από τις «Big 4 Accounting firms», είναι η Klynveld Peat Marwick Goerdeler LLP (KPMG) και η οποία αντίστοιχα χρησιμοποιεί εκτός από τις κλασικές και διαφοροποιημένες πηγές δεδομένων σε τεράστιες ποσότητες και καινοτόμες μεθόδους επεξεργασίας δεδομένων. Όπως οι προηγούμενες τα εξελίσσει σε συμπεράσματα παρέχουν υπηρεσίες σε τρεις βασικούς τομείς της λογιστικής: ελεγκτική, συμβουλευτική και φορολογία.

Λέξεις κλειδιά: Μεγάλα Δεδομένα, Λογιστική, Ελεγκτική.
Summary

In this dissertation we try to examine the relationship between Big Data and Accounting. In order to reveal this relationship, we initially refer to the changes in accounting due to the increased volume of information that has occurred because of the progress of communication and IT systems during the last two or three decades. These radical changes have also affected the accounting standards. Additionally, the integration of Big Data in accounting curriculum has an important impact in auditing.

Specifically, the data sources have changed both in accounting and auditing. The quantity and the quality of the information are totally different because of Big Data. The processing did not remain the same, as the accounting and auditing processing methods have evolved technologically (hardware and software). The progress in accounting and auditing was also achieved by quantitative modeling taking into account the following models: A) model of financial distress, B) model of financial fraud and C) model of stock market prediction and quantitative modelling.

Then, the dissertation examines four very significant case studies that demonstrate the way that the «Big 4 Accounting firms» use Big Data in auditing and consequently in accounting. The first case study is about Deloitte Touche Tohmatsu Limited which is the biggest firm from the «Big 4 Accounting firms» and it provides tax, consulting, audit and assurance, mergers and acquisitions and private company services as well as risk and financial advisory services. This firm deals with a great volume of data and it uses modern processing systems. The second firm examined is PricewaterhouseCoopers (PwC) which also has a leading role in accounting by providing audit and assurance, consulting and tax services as well as services for private companies. It also deals with board governance and risk assurance issues. It is a giant company with international status. Ernst &Young Limited (EY) is the third case study. Its scope of accounting activities includes: advisory, assurance and tax services and as a result it demonstrates great scientific interest. Last but not least, it is the forth from the «Big 4 Accounting firms», Klynveld Peat Marwick Goerdeler LLP (KPMG). The firm harnesses huge amounts of data from various sources and innovative methods of data processing. KPMG (as the above mentioned companies) uses Big Data in accounting in order to gain useful insights and to improve its audit, advisory and tax activities.

**Keywords**: Big Data, Accounting, Auditing.
## Contents

Summary ......................................................................................................................... 7
Introduction ....................................................................................................................... 11
Part 1 .................................................................................................................................. 13
Theoretical Background ................................................................................................. 13
Section 1.1 ....................................................................................................................... 13
Chapter 1.1.1 Sources of Big Data in accounting ......................................................... 13
  1.1.1.a Sources of Big Data in accounting research ..................................................... 13
  1.1.1.b Sources of Big Data in Business ..................................................................... 15
Chapter 1.1.2 Processing of Big Data ........................................................................... 21
Chapter 1.1.3 Models of Big Data analytics in accounting ............................................ 24
  1.1.3. A Model of financial distress ............................................................................ 24
  1.1.3. B Model of financial fraud ................................................................................ 26
  1.1.3. C Stock market prediction and quantitative modelling ..................................... 29
Section 1.2 ....................................................................................................................... 33
Chapter 1.2.1 Opportunities and changes for accounting ............................................ 33
Chapter 1.2.2 Changes in accounting standards due to Big Data ................................. 40
Chapter 1.2.3 Big Data in auditing ................................................................................. 41
Part 2 .................................................................................................................................. 49
Case studies: How “Big 4 accounting firms” use Big Data in audit? ......................... 49
Chapter 2.1 The case of Deloitte Touche Tohmatsu Limited .................................... 50
Chapter 2.2 The case of PricewaterhouseCoopers (PwC) ........................................... 61
Chapter 2.3 The case of Ernst &Young Limited (EY) .................................................. 75
Chapter 2.4 The case of Klynveld Peat Marwick Goerdeler LLP (KPMG) ............... 84
Conclusions and suggestions for future research ......................................................... 95
References and webservces ............................................................................................. 96
Appendix ......................................................................................................................... 101
Chapter 1. Definition of Big Data ................................................................................. 101
Chapter 2. Big Data Technologies ................................................................................. 106
Chapter 3. Applications of Big Data in society, government and business ................. 114
Content of figures

FIGURE 1. Emerging Sources of Data 14
FIGURE 2. Accounting challenges and its most important information technology answers. 36
FIGURE 3. Effective Big Data Analytics in Continuous Auditing. 43
FIGURE 4. Enhanced audit integration model. 55
FIGURE 5. Enhanced insights-driven audit methodology. 56
FIGURE 6. Data analytics competency model. 57
FIGURE 7. Business and technology service spend on technology. 58
FIGURE 8. A closer look at the business’s spend. 59
FIGURE 9. Delivering the message through visual cues. 60
FIGURE 10. Delivering the message through visual cues. 60
FIGURE 11. Are infrequent users causing potential quality and efficiency issues? 62
FIGURE 12. Are there unexpected journals posted to key accounts? 63
FIGURE 13. Is someone outside of the finance department posting journals? 64
FIGURE 14. Are people wasting time posting entries twice, or reversing them? 64
FIGURE 15. Percentage of audit clients for which Halo for Journals is utilized. 65
FIGURE 16. Measure your success. 67
FIGURE 17. Retail Store Loss Prevention Audit Shrinkage % - High Level Analytic. 68
FIGURE 18. Sample of stores layout. 69
FIGURE 19. Sample of email recommendations. 69
FIGURE 20. The five core Es of breaking through analytics’ arrested evolution. 70
FIGURE 21. Analytics can help auditors review unstructured data in minutes instead of hours. 71
FIGURE 22. Embedded analytics enables continuous risk assessment. 72
FIGURE 23. Data types that are accomodated by EY analyzers. 76
FIGURE 24. The IA analytics maturity model. 79
FIGURE 25. Fraud triangle analytics. 83
FIGURE 26. Intelligent Automation. 85
FIGURE 27. An overview of maturity levels. 89
FIGURE 28. Audit Methodology-based Maturity Model. 90
FIGURE 29. Internal audit plan development at various maturity levels. 91
FIGURE 30. Big Data interest over time (Google trends, 2016). 95
FIGURE 31. Definitions of Big Data on an online survey of 154 global executives in April 2012. 96
FIGURE 32. The 4 Vs feature of Big Data. 100
FIGURE 33. Cloud computing logical diagram. 101
FIGURE 34. Data deluge: the increase of data size has surpassed the capabilities of computation. 102
FIGURE 35. Key components of cloud computing. 103
FIGURE 36. Cloud Storage. 104
FIGURE 37. The timeline of the Big Data processing and technologies. 106
FIGURE 38. Big Data in healthcare sector. 109
FIGURE 39. The correlation between Tweets about rice price and food price inflation. 114
Content of tables

| TABLE 1. Illustration of data expansion and measurement consequences. | 16 |
| TABLE 2. Types of Big Data and their impact on audit approach. | 47 |
Introduction

This dissertation consists of two parts. The first part examines the impact of Big Data in accounting and auditing theory and methodology. The second part of this dissertation examines four case studies that highlight the usefulness of Big Data in extracting valuable information in the auditing sector.

The first section of the first part contains the following chapters:

Sources of Big Data in accounting: In this chapter, we focus on the new sources of data that have emerged and they concern both business and the academic community. There are plenty Big Data sources that can offer insights to business and accounting research.

Processing of Big Data: Big Data need a specific processing in order to add value to the business. There are certain tools that can be used for this purpose and some of these are referred to this chapter. Specifically: Cloud environment provides Big Data storage and processing solutions to accountants. Similarly, tools like Hadoop can be extremely useful when it comes to Big Data.

Models of Big Data analytics in accounting: The models that are examined in this chapter are the following: model of financial distress, model of financial fraud and model of stock market prediction and quantitative modelling. Several data mining techniques are suggested for Big Data analysis in each model.

The second section of the first part includes the following chapters:
Opportunities and changes for accounting: This chapter presents various changes that will occur in the accounting profession because of Big Data. These changes are radical and they will transform the accounting procedure as well the role of the accountant in the future.

Changes in accounting standards due to Big Data: Standards and rules in accounting will also change. It is obvious that Big Data will alter the financial reporting and the disclosure rules and in this chapter some of these changes are demonstrated.

Big Data in auditing: Big Data will enhance auditing. Even though auditors have not fully adopted Big Data analytics yet, there are many future benefits. In this chapter, we present how auditors can harness Big Data in order to gain insights and we also underline the challenges that may occur due to the use of Big Data in the auditing profession.

In the second part, the case studies try to enlighten the role of Big Data in auditing and accounting. This is the reason why the case studies focus on four major accounting firms which are: Deloitte Touche Tohmatsu Limited, PricewaterhouseCoopers (PwC), Ernst &Young Limited (EY) and Klynveld Peat Marwick Goerdeler LLP (KPMG). In each case we examine the analytical tools and applications that each firm uses. Furthermore, we refer to internal audit analytics. Finally, for each accounting firm we indicate certain examples of Big Data analytics implementetion in auditing to their clients.
Part 1

Theoretical Background

Section 1.1

Chapter 1.1.1 Sources of Big Data in accounting

1.1.1.a Sources of Big Data in accounting research

Nowadays, there is no doubt that those who want to do accounting research by using Big Data, deal with many challenges. However, they have to use Big Data during their research detouring the existing difficulties. The data that are used by researchers can be real-time, at the exact time that the companies disclosure their financial statements or information about the prices and the quantities of the products (Moffitt and Vasarhely, 2013). That way, new accounting models can be created or the existing can be enhanced. These models would offer valid information.

U.S government and specifically SEC, uses data for the Audit Quality Model (Moffitt and Vasarhely, 2013). Similarly, accounting researchers can find Big Data in EDGAR and WRDS databases. Wharton Research Data Services (WRDS) is a platform that can be used as a Big Data analysis tool. It contains more than 250 terabytes of data from various sectors, including accounting. Of course, there are other trustworthy sources, for example the websites of the companies where they reveal
their data, Yahoo-Finance that has many data about enterprises, data from the press (magazines or newspapers) and social media (Moffitt and Vasarhely, 2013).

**Figure 1**: Emerging Sources of Data


In the future, accounting researchers may have more access in Big Data through open-source software and commercial hardware (Moffitt and Vasarhely, 2013). However, nowadays there is a difficulty for accounting researchers to find Big Data from companies, governments or non-profit organizations and as a result, accounting is not as data-intensive as other sciences (medicine, physics etc.) (Moffitt and
Apart from access to Big Data, another challenge is that most of these data need to be processed before they are ready for use. This pre-processing requires a lot of money. A great progress can be made in the accounting sector in case that these problems will be solved in the future.

1.1.1.1 Sources of Big Data in Business

An enterprise can easily collect Big Data through various sources nowadays. ERP systems can provide a great amount of valuable information as they collect different kinds of data. Except from data that are derived from financial statements, data from other sources may be very useful for accountants (Moffitt and Vasarhely, 2013).

Data from enterprises can be used in many ways. The EDE (Enterprise Data Ecosystem) is growing bigger and bigger as the years go by. Consequently, the characteristics of this environment are changing every day. At figure 1 we can see the sources and the usage of the data into the business ecosystem. First, there are the traditional data that are produced mainly by the ERP system of the company as well as the legacy data (Moffitt and Vasarhely, 2013). Initially, this kind of data was obtained in a manual way. Today, scanners are used by accountants and this provides more information about the transactions. Accountants can report more detailed information, so company can improve inventory management or sales after the analysis of these data. The combination of manually and automatically acquired data can be very useful for the business. For instance, business can extract data from the World Wide
Web that would contain “URL, click-path, and content data” (Moffitt and Vasarhely, 2013). Specifically, the click-path data can be valuable for companies as they indicate the behavior and the preferences of the customer and in what way a client chooses products.

Table 1: Illustration of data expansion and measurement consequences

<table>
<thead>
<tr>
<th>Sources, Content, and Enhanced Content (Content)</th>
<th>Parameters (Meta Data)</th>
<th>Meta-Meta Data</th>
<th>New Sources of Business Data</th>
<th>New Sources of Financial Data</th>
<th>New Linkages Facilitated by IT and Analytic Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>News pieces</td>
<td>Source</td>
<td>RFID data</td>
<td>B2B transaction data</td>
<td>XBRL/FR data</td>
<td>Use news mentions, or emails, or click-path analysis to predict sales</td>
</tr>
<tr>
<td></td>
<td>Time</td>
<td>Detailed transaction data</td>
<td>Blog postings</td>
<td>Automatic classification and addressing of emails</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Publication Topic</td>
<td>B2B transaction data</td>
<td>Emails</td>
<td>EDGAR data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frequency over time</td>
<td>Change in nature over time</td>
<td>Social media postings</td>
<td>FDIC call reports</td>
<td></td>
</tr>
<tr>
<td>Emails</td>
<td>To whom</td>
<td>Emails linked to transactions with attachments</td>
<td>Paths of fraudulent behavior</td>
<td>Identify fraudulent user groups</td>
<td></td>
</tr>
<tr>
<td></td>
<td>From whom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Media</td>
<td>Groups of people that behave similarly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>To whom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>From whom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Click-path</td>
<td>Purchase paths by product</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Website visited</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pages visited</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Another type of data that can reveal many information about customer’s behavior and future trends are the mobile phones. They can provide a variety of data, for example location data (Moffitt and Vasarhely, 2013).
Important data are: video and audio data. Audio data can be produced by phone calls, media or surveillance cameras. After the recording they can be transformed into textual data. These data can reveal a lot about the feelings of the clients and they are produced all the time, as a result their volume is huge (Moffitt and Vasarhely, 2013). Similarly, video data can be included in the enterprise data ecosystem. Automatically obtained data can add value to a company. For instance, a bank can enrich the “traditional” valuation of the applications of loan candidates that are manually completed and processed by the system with data from sources that we mentioned above like phone calls, location, online questionnaires, website visits etc. (Moffitt and Vasarhely, 2013). That way, banks will reach to a safer conclusion about their clients.

At the table above we can see expanded sources of data and their possible uses into the business environment in the accounting and auditing sectors. It is obvious that the amount of sources is increased and as a result it depends on each company how it will use them. Furthermore, these sources provide various data that were not used before and they can be valuable for the business. The traditional data can increase their efficiency as they can be more detailed, they can be combined with non-traditional data and there can be further explained by meta-data (Moffitt and Vasarhely, 2013). Specifically, “sources of data” include the sources that the companies traditionally use and the sources that they can additionally use in no cost. Except from data entry, approval, and execution of transactions, accountants can also use e-mails or data from social media (Moffitt and Vasarhely, 2013). “Meta-
“Data” are data that provide additional information about data fields. Accountants can use this information if they standardize it. For instance, accountants can store or obtain data that are not internal by using automatically names of fields that they have stored. “Meta-meta-data” are based on meta-data and they provide further details about products (quality, etc.). “New Sources of Business Data” and “New Sources of Financial Data” reveal the opportunities and the potential sources that in combination with IT and analytic technologies can add value in case that they will be integrated in the EDE of the company (Moffitt and Vasarhely, 2013).

Internet of Things also produces Big Data useful for accountants. Chui et al. (2010) stated about The Internet of Things: “sensors and actuators embedded in physical objects (from roadways to pacemakers) are linked through wired and wireless networks, often using the same Internet Protocol (IP) that connects the Internet” (O’Leary, 2013). A variety of “things” such as sensors, databases, software or devices are connected to the internet and consequently there is also connection between them. When we refer to sensors we include pacemakers, GPS (and other systems that provide information about location) and RFID tags. RFID tags are radio-frequency identification tags that can provide information about place or time. Since there is connection between “things” they can exchange useful information. This information can be in a local warehouse or can be stored in a cloud environment, where can also be analyzed by the appropriate applications (O’Leary, 2013).
There is no doubt that there is a strong relationship between IoT and Big Data, taking into account that IoT produces Big Data but without being the main source of production. However, HP indicates that by 2030 IoT will be the major producer of Big Data since the amount of sensors will be 1 trillion (Min C. et al, 2014). According to Intel, IoT has three attributes that ensure that it generates Big Data: first, it is a source of a great amount of data, second the data that are produced are either semi-structured, or unstructured and finally these data must be first analyzed in order to offer some useful information (Min C. et al, 2014). In addition, there is no doubt that sensors can produce data all the time, so we can refer to the velocity of data that IoT generates which is an important feature of Big Data.

A very significant part of IoT is continuous monitoring that is a source of Big Data, since the data that are produced by this procedure are massive and they are generated constantly. Furthermore, various types of data are produced by sensors. An example is the monitoring of inventory goods where RFID tags are used (O’ Leary, 2013). Companies also use monitoring of social media for their interest by analyzing the data. This kind of information is very useful because it reveals a lot about the popularity of the company and the preferences of the consumers.

Continuous monitoring and Internet of Things give many solutions to the accountants as they improve the quality of the accounting information. Specifically, thanks to continuous monitoring companies can detect easier frauds or potential accounting mistakes that may
happen by accident. According to O’ Leary (2013), three are the main sources that produce accounting information about the transactions of the company: first, sensors that are part of the Internet of Things and an example is the RFID tags that inform about the location and the sales of the inventories as we mentioned above. The second source of information is transaction entries that can be entered automatically according to timing or specific events by the software of the ERP (enterprise resource planning) systems of the company. Finally, there is a more traditional way and that is to enter the human-based transactions manually. According to the theory of the IoT the methods that are used in auditing and monitoring of the digital transactions (the first two types of transactions that are recorded automatically) differ from those that are used in transactions that are entered by people. Apart from the traditional data that an accountant can get from the transactions of the company there is also the ability to audit accounting information using a wide range of data that have the features of Big Data. For instance, O’ Leary (2012) made the suggestion of monitoring financial information continuously by including data from blogs or message boards (O’ Leary 2013). This analysis would integrate not only Big Data from typical accounting sources but also a great variety of data from alternative sources. Certain examples would be: “social media discussions by employees in finance and accounting, social media discussions by employees that mention financial information or even social media discussions by anyone that mentions financial information of the particular company” (O’ Leary 2013).
Finally, company can extract value from the continuous monitoring of location that is a part of the Internet of Things and generates Big Data. An enterprise can be informed about location from RFID tags (on products, assets and badges), GPS and mobile phones (O’ Leary 2013). An example of continuous monitoring of location is the monitoring of badge location that provides information about the place that the employees are or the monitoring of the company’s vehicles (O’ Leary 2013). All these Big Data that are produced can be collected for various purposes. A possible suggestion is: the collocation of tagged assets and people’s badges or phones in order to achieve an effective control on assets and reduce potential losses (O’ Leary 2013). That way the asset management system of the business becomes more efficient.

Chapter 1.1.2. Processing of Big Data

According to De Mauro et al. (2016): “Big Data is the information asset characterized by such a high volume, velocity, and variety to require specific technology and analytical methods for its transformation into value” (Saggi and Jain, 2018). There are various technologies and techniques that operate as a tool in Big Data analysis in order to create value. Nowadays, it is true that companies can gather more data than ever and they can extract valuable information if they analyze them properly. Therefore, they need many tools for processing and storage of Big Data from different sources.

According to NIST’s definition for cloud computing: “cloud computing is a model for enabling convenient, on-demand network
access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” (Demirkan and Delen, 2013). Big Data and cloud are highly connected with each other. Data oriented cloud systems provide storage and computing in a distributed and virtualized environment (Demirkan and Delen, 2013). A cloud system is necessary for companies that deal with Big Data because it offers great storage opportunities. Additionally, it provides solutions to the processing of Big Data.

Cloud is becoming very popular in accounting. The “in-house” or “on-premise” accounting software is gradually replaced by online accounting software. Nowadays, thanks to cloud and web-based systems an accounting application can execute multiple tasks and instances at the same time in a common environment. This is very useful for companies taking into account that they can have high quality of services, upgrades that happen easier and more often and cheaper ownership. Something that is really important is that stakeholders have instant access in the accounting and financing information of the company and they can have real time reporting with only condition to have access to internet. There are various products in the market such as: NetSuite Financials, Intacct Financials and Accounting System, SAP ERP Financials, Microsoft Dynamics GP, Epicor Financial Management or SAGE (Belfo et al. 2013). These products offer many services for the accountants: “core accounting, project accounting, fund accounting,
inventory management, billing & invoicing, work order management, budgeting and forecasting, fixed asset accounting, financial reporting, payroll management or human resources” (Belfo et al. 2013).

There is no doubt that the massive amounts of different kinds of data cannot be stored and managed by the traditional relational databases. The NoSQL (not only SQL) databases provide a solution not only in the challenge of data storage but in data management, as well. The main advantage of NoSQL databases is that data storage and data management are different procedures and that way the database gains flexibility. Some of these databases are: Apache Hadoop, MapReduce, Amazon’s Dynamo system (which is extremely available and expandable), Voldemort (which is used by LinkedIn), Google’s BigTable (processing petabytes of data in thousands of commercial servers), HBase and HyperTable (they are identical to BigTable but they are open-source in contrast with BigTable), Cassandra, MongoDB (which is also an open-source development) (Min et.al, 2014).

Obviously, Hadoop has an extremely important relationship with Big Data and it provides batch processing. According to Chen and Zhang (2014): “Hadoop provides platforms and infrastructures for other specific Big Data applications”. Hadoop includes: Hadoop kernel, Map/Reduce and Hadoop distributed file system (HDFS) and other similar platforms like Apache Hive, Apache HBase, etc. (Chen and Zhang, 2014). Today Hadoop is used in various Big Data applications such as spam filtering, network searching, clickstream analysis, and social recommendation
(Min et.al, 2014). As we mentioned above Big Data storage systems use thousands of commercials servers. Consequently, it was necessary to develop more powerful parallel programming models in order to increase the effectiveness of NoSQL. MapReduce’s goal is to provide automatic parallel processing and distribution by using a lot of clusters on commercial PCs (Min et.al, 2014).

**Chapter 1.1.3. Models of Big Data analytics in accounting**

**1.1.3. A Model of financial distress**

Companies can discern and forecast financial distress by using data mining techniques. These techniques can also benefit auditors during the valuation of a company (Gepp et. al, 2018).

A very widespread technique is decision tree models. There are various studies that have used decision tree based models in order to forecast financial distress. Sun and Li (2008), indicate “theoretical feasibility and practical effectiveness” using a prediction model with thirty five financial ratios and 135 company pairs (Gepp et. al, 2018). Koyuncugil and Ozgulbas (2012), predict financial distress for small-to medium sized firms (Gepp et. al, 2018). They used 7000 companies and they created “risk profiles, risk indicators, early warning systems, and financial road maps” in order to warn for financial failure (Gepp et. al, 2018). Similar studies have been conducted for Chinese and U.S companies as well, by using classification and regression tree methods (Gepp et. al, 2018).
Neural networks can also be used in order to forecast financial distress (Gepp et. al, 2018). Chen and Du (2009) used 37 ratios in 68 firms and they demonstrated that their modelling was valid (Gepp et. al, 2018). Geng, Bose, & Chen, 2015 underline that neural networks outperform in comparison with decision tree models or other techniques such as support vector machines (Gepp et. al, 2018).

There are also researches, (Zhou et al. (2015), Zhou, Lu, and Fujita (2015)), that compare the effectiveness of data mining techniques for forecasting financial failure with models that are defined from accounting and finance professionals and they notice no substantial difference (Gepp et. al, 2018). According to Zhou et al., (2015) and Linand and McClean (2001), when professionals use both their own ideas and data mining techniques they achieve much better results (Gepp et. al, 2018).

Moreover, financial distress modelling can be very useful to corporate stakeholders, as well (Gepp et. al, 2018). Khandani, Kim, and Lo (2010) designed models to detect the credit risk of consumers with machine learning techniques. They examined these models in an individual or customer level and not in a company level (Gepp et. al, 2018). They gained knowledge from clients’ transactions and information for credit behavior and they enhanced classification rates in order to prevent credit card defaults and delinquencies (Gepp et. al, 2018).
As a result, it is obvious that auditors can use data mining techniques and along with their experience and their professional judgment they can reach to safer and more accurate conclusions. Data mining techniques enable auditors to forecast effectively the financial situation of a company and customers, as well. According to Gepp et al., 2018: “Incorporating Big Data models should help avoid the costly error of issuing an unmodified opinion prior to bankruptcy”. This phenomenon occurs often in non-Big 4 audit firms within the first five years of their operation (Read and Yezegel, 2016), cited to: Gepp et al., 2018. The reason why this happens maybe is that smaller audit companies are afraid of losing clients. Consequently, they issue “modified going concern opinion difficultly” (Gepp et. al, 2018). If they use Big Data, they will have more objective results to demonstrate to the customers so they would support their opinion in a strong way and they will be more independent.

Forecasting can be short-term (one year ahead), or long-term, as well. It is more possible for internal rather than external auditors to make long-term forecasts (Gepp et. al, 2018). Board of Directors and senior management can gain great advantage from longer-time forecasts as they can alter their strategy in order to avoid the predicted financial distress (Gepp et. al, 2018).

1.1.3. B Model of financial fraud
This model is very useful for auditors who want to detect financial fraud. In Section 200 of the Statement on Auditing Standards No.
122/123 is clearly stated: “external auditors must obtain reasonable assurance about whether the financial statements as a whole are free from material misstatement, whether due to fraud or error” (Gepp et. al, 2018). Data mining techniques will enable auditors to provide this “reasonable assurance”.

It is true that financial fraud is a major problem for both governments and enterprises around the world. Big Data can be very useful for auditors in order to detect, forecast and prevent fraud. A typical organization has a revenue loss of 5% per year due to financial fraud according to Association of Certified Fraud Examiners (2016) (Gepp et. al, 2018). There are various methods proposed for dealing with financial fraud. For example, Chang et al. (2008) proposed a feasible and effective method which controls millions of bank wire transactions with visual data analytics (Gepp et. al, 2018). Another suggestion is “meta-learning” by Abbasi, Albrecht, Vance, and Hansen (2012), that indicates a kind of machine learning model which increases the accuracy by “combining the outputs of multiple machine learning techniques in a self-adaptive way” (Gepp et. al, 2018).

There are also models based on supervised or unsupervised neural networks. The method of unsupervised neural networks that was suggested by Huang, Tsaih, Lin, et al. (2014) is the following: at first the statistically significant variables are gathered, then they are divided into smaller groups created by the significant variables and the last stage is the principal components analysis which demonstrates the key-features
of each group (Gepp et. al, 2018). This approach is effective according to the implementation of this method to 144 companies. There are various other approaches proposed for detecting fraudulent activity such as: neural networks, support vector machines, and genetic programming implemented by Ravisankar, Ravi, Rao, and Bose (2011) who used a genetic algorithm to detect fraud in financial statements (Gepp et. al, 2018). Kirkos, Spathis, and Manolopoulos (2007) proposed a Bayesian network (Gepp et. al, 2018).

There is a controversy about the best model. Lin et al. (2015) underline the differentiation between the judgment of professionals and data mining techniques and conclude that the rate of correct classification by neural networks and decision trees is more than 90% (Gepp et. al, 2018). On the other hand, Perols, (2011), suggested that logistic regression and support vector machines are more effective than neural networks and decision trees (Gepp et. al, 2018). According to Chen (2016) a combination of the above mentioned approaches in two stages gives better results (Gepp et. al, 2018). Zhou and Kapoor (2011) indicated that the best approach is to combine the experience of the professionals and Big Data techniques (Gepp et. al, 2018).

Detection of fraud can be implemented in the text of financial statements, as well (Gepp et. al, 2018). Purda and Skillicorn (2015) suggested a “language-based tool” that can detect financial fraud. At first, it uses a decision tree in order to process financial statements from companies that have committed financial fraud. That way it detects
words that demonstrate financial fraud and it makes a list of words that indicate fraud instead of non-fraud (Gepp et. al, 2018). This is the training period. Afterwards, the fraud is forecasted with a certain probability by vector order machines. The rate of correct classification of this approach is higher than 80% (Gepp et. al, 2018).

According to the approaches stated above, financial fraud is a binary variable, no matter what kind of fraud is or what the cost of each fraud is. These methods can be more precise if they include social characteristics of fraud according to Free and Murphy (2015) (Gepp et. al, 2018).

External and internal auditors could be benefited by Big Data techniques in detecting financial fraud. By combining Big Data techniques and traditional techniques such as logistic regression (F-score model), they can obtain more accurate results by using results from previous frauds (Gepp et. al, 2018). Even though auditors do not easily implement novel techniques, as they prefer their own judgment, it is certain that Big Data can create new opportunities for auditing.

**1.1.3. C Stock market prediction and quantitative modelling**

This genealogy is not connected with auditing directly, nevertheless it can add valuable knowledge and techniques in auditing profession. It indicates the data mining techniques that are used in order to predict the situation in stock market and inform managers and investors.

Lam (2004) forecasts the returns of the market by using neural networks, variables from the macroeconomic environment and financial
ratios (Gepp et. al, 2018). An approach using a genetic algorithm-base was proposed by Zhang, Hu, et al. (2015) in order to define rules of trading in stock market and they demonstrated that this model operates more efficiently than a decision tree and a Bayesian network (Gepp et. al, 2018).

Moreover, there are studies that indicate the connection between stock market and web searches. For instance, Curme et al., 2014 showed that market returns decrease when there is an increase in searches about business or politics in Google and Wikipedia (Gepp et. al, 2018). Another study of Li, Ma, Wang, and Zhang (2015), revealed a significant relationship between Google search volume index (which measures the interest of the investors), and the positions of the traders and upcoming changes of oil price (Gepp et. al, 2018).

Furthermore, Sun, Shen, and Cheng (2014) designed a model about the decisions of stock investors based on information about individual behavior of each stock investor (Gepp et. al, 2018). Eventually, they created trading networks that can predict the returns of each stock investor. Shapira, Berman, and Ben-Jacob (2014) created a model of stock market as: “a network of many investors” (Gepp et. al, 2018).

Stock market predictions can also be based in news in financial newspapers, magazines or online sources. There are various studies that indicate an association between stock market and what the press releases. For example, Tetlock (2007) used Wall Street Journal and revealed that when media negativity prevails, stock prices initially
decrease but afterwards they return to normality (Gepp et. al, 2018). Moreover, media negativity in an extreme high or low level, affects the predictions about trading volume. Alanyali, Moat, and Preis (2013) revealed that the volume of stocks per day is positively correlated with the amount of times that Financial Times refers to the stocks and this is observed both before and during the mention from the press (Gepp et. al, 2018). Furthermore, there are various studies that focus on the sentiment of the news such as: Smales, 2014a, 2014b, 2015) (Gepp et. al, 2018). Jensen, Ahire, and Malhotra (2013) indicated that “news sentiment and mainly bad news affect seriously intraday volatility persistence” (Gepp et. al, 2018). Online news are used in order to forecast the situation in financial markets, nevertheless the benefits are less than 5% (Gepp et. al, 2018). There is an effective method that combines news sentiment with online data, suggested by Ranco et al. (2016) (Gepp et. al, 2018). Another approach by Dhar, 2014; Kao et al., 2015; Zheludev, Smith, & Aste, 2014 integrates data from various sources like “social media, blogs, and forums” (Gepp et. al, 2018). Finally, there are many areas that quantitative modelling is implemented such as: “high-frequency financial data mining (Sun & Meinl, 2012), identifying drivers of firm value (Kuzey, Uyar, & Delen, 2014), sentiment analysis for predicting economic variables (Levenberg, Pulman, Moilanen, Simpson, & Roberts, 2014), volatility of returns (Sun, Chen, & Yu, 2015), option pricing (Thulasiram, Thulasiraman, Prasain, & Jha, 2016; Xiao, Ma, Li, & Mukhopadhyay, 2016), and market basket analysis (Videla- Cavieres & Rios, 2014) (Gepp et. al., 2018). Lately, Big
Data techniques are extensively used both in quantitative modelling and stock market prediction and their use is more extensive in sentiment analysis and online textual information (Gepp et. al, 2018).

Auditing can harness sentiment analysis and as a result, Big Data. Sources such as news and social media can give valuable information for products or customers. In other words, when there are negative sentiments in news concerning products or clients, auditors may be more skeptical and this will affect their decisions.

Sentiment analysis in the emails of a firm may be very helpful. There is a variety of implementation of sentiment analysis of the emails that can reveal possible fraud to the auditors. For example, in case that a unit of the company performs inconsequent sentiment in the emails that may indicates a fraud or some kind of anomaly in this unit (Gepp et. al, 2018). Moreover, sentiment analysis can be conducted in a business unit in comparison with the other units inside the company. Frauds conducted by managers may be revealed by the examination of the sentiments of emails of lower employee levels in comparison with higher levels (managers) in the company. If there is a dissatisfaction among the employees, while there is a complete harmony in management levels that is a signal of managers’ fraudulent behavior (Gepp et. al, 2018). Furthermore, Ravi and Ravi (2015) conducted a research based on sentiment analysis in Enron emails (Mohammad, 2012) and they observed how male and female emails differentiate in words about sentiments and mainly trust (Gepp et. al, 2018). Both
internal and external auditors could be benefited by sentiment analysis and extract valuable information.

Section 1.2

Chapter 1.2.1. Opportunities and changes for accounting

It is true that if accountants harness Big Data they can gain valuable information.

First of all, according to Moffitt and Vasarhely, 2013: “The current basic posting element is the journal entry but some entries in large corporations have hundreds or thousands of lines in one posting.” Consequently, in case that data are collected automatically, accountants will have more information available. Journal entries can be directly connected to data from invoices, objects that the company receives and their RFIDs (Moffitt and Vasarhely, 2013). Additionally, the names of the accounts and what they include may change since they remain the same for many years, specifically since the manual accounting was implemented (Moffitt and Vasarhely, 2013). Nowadays, account names can provide detailed information at low cost. For instance, information about the inventory such as location, physical age and type of product can add value to the company. Furthermore, company can gain enriched, low-cost estimates about sales, costs, product mix, human resource turnarounds (Moffitt and Vasarhely, 2013). Not to mention that data drilling can be used by accountants. By using drill-down, accountants can gain insight about the transactions or the customer preferences as they have detailed information. Drill-down means that
accountants start to analyze a general category and then they move to more detailed levels of information. They have access to various kinds of data by using Big Data techniques. The more granular the data, the more the amount of the available information. That way the company can gain a competitive advantage. Of course, there are plenty other changes in reporting that can be implemented. Process management can add value to the company since it reveals the path that each customer follows for its transactions. Process mining using Big Data can enhance process management (Moffitt and Vasarhely, 2013). Moreover, accounting professionals can use alternative data such as pictures, audio data or data from GPS except for invoices and receipts in order to confirm the transactions.

As for the disclosure rules, it is true that they are not really important since the whole database is available. As a result, the disclosure rules can be obtained from the initial data (Moffitt and Vasarhely, 2013). Consequently, accounting standards will be interested in raw data and not in specific accounting rules (Moffitt and Vasarhely, 2013). There are several technologies that can make this possible such as: relational databases, linkages with textual data, textual data analytics, drill-downs, census-like filtering of details and XBRL (Moffitt and Vasarhely, 2013).

Chartered Institute of Management Accountants (CIMA) and the UK’s University of Bath conducted an online survey in June 2010 (Belfo and Trigo, 2013). This survey addressed to 5,426 senior finance and
senior non-finance professionals. These finance professionals came from various countries across the world and they were responsible for a wide range of accounting operations such as: “transaction processing, financial reporting for internal use, financial reporting for external use, forecasting cash flow management, risk management and many others” (Belfo and Trigo, 2013). According to this survey, accounting professionals except for their typical responsibilities have to contribute in a more active way in a company. They must have the ability to provide substantial support and guidance to managers. In other words, lately (especially after the financial crisis of 2008), there is an urgent need for accountants who do not only report to the managers but they can gain an insight in the business situation, as well (Belfo and Trigo, 2013). There is no doubt that Big Data era can contribute to this. As we can see in the figure below, Big Data can provide many answers to the modern accounting challenges. Specifically, accountants have the opportunity to offer real-time as well as interactive reporting and “not one report for all”. A huge amount of a variety of data is available to accountants, and that way they have more information about the company and they obtain a more critical role into the company. They can also utilize Big Data in order to improve internal auditing and risk management as they can reach to safer conclusions by using Big Data. Additionally, accountants can combine current cost accounting and past cost accounting in order to provide useful information for the company.
Figure 2: Accounting challenges and its most important information technology answers

As a result, the use of Big Data, along with other technologies as well, creates what we call “accounting intelligence” (Belfo and Trigo, 2013). Accounting intelligence is about the technologies that are used to produce, process and present the accounting information. It is obvious that nowadays, accounting professionals who use Big Data obtain a critical role in the company that is beyond their traditional responsibility to simply report the transactions. According to Faye Chua, head of Future Research, ACCA, (2014): “Big data offers the finance professional the possibility of moving into a more strategic, proactive role in business (Bhimani and Willcocks, 2014).
Of course there is no doubt that the volume and the structure of data affect the information provided by the accountants. Large datasets enable accountants to implement hypothesis testing and experts’ judgment (Bhimani and Willcocks, 2014). Big data reveal current trends more efficient rather than focusing on past events and that way the information available to managers is up-to-date (Bhimani and Willcocks, 2014). Additionally, patterns from large databases, after they are analyzed, can reveal significant details for managers. As a result, Big Data and Big Data techniques make the transformation of data to useful information more effective. Taking into account that click path data are now available, and as a result we gain insight about the web visits of the customers, accountants can process this information and provide valuable knowledge about the pricing policy of a company or strategies about cash and working capital management or even the reduction of operational costs (Bhimani and Willcocks, 2014).

Until now managers have made their decisions about companies according to the transactions that accountants had confirmed and their own judgment about the current situation and the future of the company. However, today thanks to Big Data, companies are benefited as they can possess data from various sources and that makes their decisions safer and real-time (Bhimani and Willcocks, 2014). As we have mentioned above, the Internet of Thing generates Big Data all the time. After this, data that are processed and analyzed can enhance accounting in a great extent. Rifkin, 2014 stated: “The ‘Internet of Things’ may possibly generate Big Data that will undergo analysis and be
transformed into ‘predictive algorithms, and programmed into automated systems to ...dramatically increase productivity, and reduce the marginal cost of producing and delivering a full range of goods and services to near zero across the entire economy’ (Bhimani and Willcocks, 2014). Through RFID tags and sensors that are placed in products the data collection can be real-time. That way the products and the processes provide the information automatically. Under this condition the decisions about working capital management differentiate. A radical change for accounting information systems is that the finance function can receive and evaluate information that will be used for real-time processing and decision-making. In addition, accounting information systems, by collecting information continuously, they have the possibility to “understand the path dependency of product purchases and inform executives about cost incursions which may be distinctly separable or common, given the pricing strategy adopted” (Bhimani and Willcocks, 2014).

Of course, the data intensive application within a company requires specific changes in accounting information systems. Besides this change, Big Data alter the way that the managers operate. Managers depend highly on coordination instead of taking the decisions on their own and giving orders (Bhimani and Willcocks, 2014). Managers who work in a Big Data environment need a constant real-time information about the impact of their decisions. Thus, it is obvious that accounting professionals have a crucial role in the company, taking into account that they can evaluate managers’ actions. Accounting information
systems must conform to this situation by generating results that are “qualitative, quantitative, graphical, interactive, text-based and which show varying degrees of structure” (Bhimani and Willcocks, 2014). There is no doubt that this will affect the way that accountants work and what they include in the accounting information.

Big Data are available all the time. Real-time analysis offers the opportunity of parallel thinking and acting. Managers are able to make a decision and at the same time implement it. As a result, the previous situation of thinking and afterwards acting is left behind thanks to Big Data implementation and the provision of real-time, continuous information by accountants. Decision-making and acting happen the same time and sometimes by the same person. There is a strong relationship between Big Data and the way that decisions are made and executed (Bhimani and Willcocks, 2014). Accounting systems must adopt this new way of processing and consequently performance management systems must become “closed loop”, (Bhimani and Willcocks, 2014). While companies implement new responsibility and reward structures, accountants must understand how to evaluate the performance, the decisions and the motives of managers (Bhimani and Willcocks, 2014).

Nevertheless, accountants must be skeptical when analyzing Big Data. They should use their own experience and judgment. Because of that, changes in the structure or the operation of the company due to Big Data must be implemented after serious consideration. It is true that
large datasets contain a lot of valuable information, but smaller datasets must not be ignored as they are equally useful. The content of this huge amount of data is not always important. As a result, the quality of the information that Big Data contain must be judged by accounting professionals. Real-time data are not always more valuable than historical data, especially when are biased (Bhimani and Willcocks, 2014).

**Chapter 1.2.2 Changes in accounting standards due to Big Data**

If Big Data are included in the accounting curriculum, some radical changes in standards must happen. It is true that accounting standards are not yet aligned with the technology progress. They still put great emphasis on “presentation, aggregation, and sampling” (Gepp et al, 2018).

As for the financial reporting:

A very important element of disclosure rules is comparability (Moffitt and Vasarhely, 2013). Comparability is very useful when it comes to stakeholders’ valuation or decision for the distribution of the company’s resources.

Moreover, we must underline the significance of the main data and specifically, according to Moffitt and Vasarhely, 2013: “content, timing, and level of aggregation”. Disclosure rules will have to provide a much more detailed disclosure of the company, for example: company, company unit, subsections, and product.
Additionally, different disclosure rules must be followed in order to deal with different investors.

It is obvious that there must be some kind of differentiation in the reporting. Reports must not be common for all, but they must differentiate depending on who they are going to inform.

**Chapter 1.2.3 Big Data in auditing**

Auditors do not use Big Data in a great extent. Actually, they are reluctant to adopt Big Data techniques and a reason for that may be that their clients have not incorporated Big Data in practice. However, it is obvious that further research must be done in this field and Big Data must become common practice for auditors. Big Data will enhance audit profession and the way that auditors detect frauds or errors in financial statements (Gepp et. al, 2018).

The notion that auditors must use the same techniques with their clients is quite widespread and has also stated by Alles (2015) who claims that this way auditors will protect their reliability (Gepp et. al, 2018). Nevertheless, this statement is not absolutely correct as there are many clients who already use Big Data. According to International Auditing and Assurance Standards Board auditors’ customers already use Big Data and as a result they’d rather to see auditors use Big Data techniques, as well. Additionally, auditors in the past have adopted techniques that were not used by their clients, such as random sampling (Gepp et. al, 2018). Of course, Big Data techniques should be combined with the judgment of experts.
It is true that Big Data give a great advantage to audit professionals as they can test the whole population and not just a sample as they do now. Consequently, they can gain a greater insight and of course this is very important for their clients, too. It is obvious that auditing standards must change in order to adopt Big Data and not to focus that much in “presentation, aggregation, and sampling” (Gepp et. al, 2018).

It can be stated that Big Data is a new source of evidence for auditors (Moffitt and Vasarhelyi, 2013). Nevertheless, Big Data must be incorporated carefully and after serious consideration in order to confirm that the standards of “sufficiency, reliability, and relevance” are met (Gepp et. al, 2018).

Auditors can now use video, audio and textual data and that way they will enhance the information that they provide. According to (Gepp et. al, 2018): “Process mining, which analyses the event logs of business systems (Jans, Alles, & Vasarhelyi, 2014), has been shown to improve audit results when tested on real world data sets (Werner & Gehrke, 2015)”. Big Data will improve the audit profession but they will also enable auditors to gain a more crucial role into the company than just verifying the transactions.

An important challenge posed in auditing is the real-time information that is provided by accountants (Gepp et. al, 2018). Big Data techniques enable accountants to inform immediately their clients and as a result this information must be audited real-time. Quarterly and annual financial statements are now available in higher frequency.
Auditors must follow this progress and harness Big Data techniques in order to audit all this accounting information much faster than in the past. Continuous auditing along with Big Data techniques can solve this problem and process vast amounts of data that contain missing values and other kind of data that is difficult to be analyzed by traditional audit techniques.

**Figure 3: Effective Big Data Analytics in Continuous Auditing**

![Figure 3](image)


The figure above demonstrates the 4 characteristics of Big Data (the 4 V’s), the gaps that currently exist in data analytics because of the 4 V’s and the challenges that are created by these Big Data gaps. As for the Big Data qualities: “massive Volume or size of the database, high
Velocity of data added on a continuous basis, large Variety of types of data, and uncertain Veracity” (Zhang et al, 2015).

The first gap is data consistency that is caused by the huge variety of data sources which inevitably creates data conflicts (Zhang et al, 2015). Big Data systems are a mosaic of various systems so there are either partial or full replication of data, “overlap” of information and storage of “derivative” data (Zhang et al, 2015). This creates problems to data consistency. Zhang et al, 2015 refers to three problems in consistency which are namely: issues with data formats (difference in data structure), with data synchronization into the organization and with data contradiction (that occurs when there are data from different sources) (Zhang et al, 2015).

Second, the issue of data integrity is about data that are either altered or not in a complete form (Zhang et al, 2015). Data can be incomplete because “an employee enters partial records of unfinished transactions or records with missing fields into an enterprise database” (Zhang et al, 2015). Modified data may occur by accident because of the procedures in Big Data environment or on purpose by employees. However, due to complexity of Big Data environment it is really difficult to reveal data modification or incompleteness and their cause. That creates several issues in continuous monitoring procedure such as the rejection of reliable data (Zhang et al, 2015).

Third, there is the issue of data identification. Data identification is about: “records that link two or more separately recorded pieces of
information about the same individual or entity” (Zhang et al, 2015). The issue of data identification rises because of the unstructured form of Big Data. For instance, continuous audit process has no problem to identify revenues from a certain sale but sometimes it cannot associate successfully sales terms and conditions (which are textual and unstructured) with this information (Zhang et al, 2015).

The forth issue is data aggregation. Massive amounts of data require corresponding computing power. Additionally, the occurrence of large variances may create excessive amount of “red flags” in audit or may decrease the ability to audit properly (Zhang et al, 2015). As a result, data aggregation helps continuous auditing to use Big Data in an effective way. However, audit of aggregated data is challenging as “normality comes at the price of missed detections at the detail level. So, the choice of aggregation levels has to be made on a case-by-case basis taking into account the inherent characteristics and risk level of the underlying transactional data.” (Zhang et al, 2015).

Finally, data confidentiality concerns continuous auditing when it comes to Big Data. Data confidentiality is about sensitive, non-public data (Zhang et al, 2015). The issue with Big Data environment is that if sensitive data leak they will be known very quickly and they will be associated with a lot of data that are relevant. A common practice for this issue is encryption (Zhang et al, 2015). Nevertheless, encryption creates two problems: data searching and audit of the encrypted data (Zhang et al, 2015).
The presentation of financial statements that contain real-time information, is going to be improved in order to be more understandable for all and to demonstrate changes from various periods. This will mean modern visualization and dashboards that require new methods of auditing (Gepp et. al, 2018). Auditors must be able to verify “existence, completeness, classification, and understandability, and accuracy and valuation” of this new form of financial statements (Gepp et. al, 2018). According to Gepp et. al, 2018: “real-time financial reporting to the public would necessitate a fundamental change for auditors, from providing assurances about numbers to assurances about real-time systems (that subsequently produce numbers). Nevertheless, real-time accounting information to the public and the managers is not a commonplace in many countries, yet. At first, real-time reporting and real-time auditing can be implemented to senior managers, and then to the public. Specifically, auditors can obtain very important information from the dashboards that are available to the managers (Gepp et. al, 2018). They can reveal potential frauds or breaches in the company. Besides that, auditors can obtain substantial information about the company, for example about the inventories (Gepp et. al, 2018). In order that to happen, auditors must turn to the assurance of systems that is mentioned above.

Another major challenge of audit is peer-to-peer market places (Gepp et. al, 2018). Due to Big Data, it is possible for suppliers and consumers to communicate via collaborative platforms (Gepp et. al, 2018). An example of that kind of firm is Airbnb. As a result, auditors
must use new methods in order to confirm the validity of the activities that include a lot of data, from various sources that change real-time.

**Table 2:** Types of Big Data and their impact on audit approach

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Current Practice</th>
<th>Potential Future Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-financial Data (NFD) or</td>
<td>Used only marginally on audits, or used with significant auditor</td>
<td>Tools developed to run models or predictive analytics to aid auditors in identifying business risks and areas of focus during planning; aid in fraud detection, and help evaluate and assess going concern.</td>
</tr>
<tr>
<td>Non-financial Measures (NFM)</td>
<td>judgment required to interpret.</td>
<td></td>
</tr>
<tr>
<td>Financial Data (FD)</td>
<td>Auditors collect and test a sample of transactions and use judgment on those areas that are difficult to test (such as management estimates).</td>
<td>Tools can test 100% of transactions. Will identify anomalies/unexpected patterns in client-provided transaction data. This will guide additional testwork, possibly uncovering fraudulent transactions. Judgment used in assessing next steps after anomalies are uncovered.</td>
</tr>
</tbody>
</table>


The table above demonstrates non-financial and financial data along with their current and future use in auditing. Non-financial data are internal data that are produced by “customers, human resources and marketing, etc.” and can give many insights to auditors (Early, 2015). The table indicates that the appropriate analytical tools that can process and analyze Big Data will improve the audit procedures and provide more insights. Auditors will be able to detect fraudulent trends and risks easier as they will have more evidence in hand.

The advantages from the integration of Big Data in audit processes can be summarized as follows: audit professionals can examine the entire population of transactions, the quality of audit will be improved as auditors will have a clear image of client’s procedures, the detection of fraudulent activity will be easier and finally auditors will be able to
use data from various sources, both internal and external (external data can be from social media, from industry peers, from the macroeconomic environment etc.), and that way they can be more effective in audit planning and risk assessment and they can provide predictive analytics (Early, 2015).
Part 2

Case studies: How “Big 4 accounting firms” use Big Data in audit?

The Big Four accounting firms are: Deloitte Touche Tohmatsu Limited, PricewaterhouseCoopers (PwC), Ernest &Young Limited (EY) and Klynveld Peat Marwick Goerdeler LLP (KPMG). They provide most of the world’s auditing services. In the USA, they audit 80% of the public companies. They also provide other services such as consulting, tax and financial advisory, assurance and management services, actuarial, corporate finance and legal services.

Big Data era has emerged so it is crucial for these companies to harness data analytics in order to add extra value to the services that they provide. In the following cases studies it will be examined how each of these firms implement data analytics in the auditing procedure now and what are the future directions.
Chapter 2.1. The case of Deloitte Touche Tohmatsu Limited

“We used to joke that auditors were drowning in documents,” says Panos Kakoullis, Deloitte Global Audit & Assurance Business Leader. “Now, dynamic technology has had an enormous impact on the profession. Thanks to new tools and resources, auditors are able to work smarter and more effectively – with each other and with clients.”

In the digital era there are many tools available for auditors such as natural language processing or remote-operated drones. Deloitte uses machine-learning tools in order to process a vast amount of documents in very little time.

Private companies are also benefited by changes in audit as audit innovation is “scalable” and as a result can be adjusted in various firm sizes (Nanney and Raphael, 2016).

According to Joe Ucuzoglu Chairman and Chief Executive Officer Deloitte & Touche LLP 2016: “At Deloitte we’re investing several hundred million dollars in data analytics and artificial intelligence with some cutting-edge applications” (Deloitte, 2016). Deloitte is currently implementing advanced audit analytics in external auditing in order to keep up with their clients (Deloitte, 2016).

Deloitte’s Argus can provide real-time analytics (Forbes, 2017). It uses natural language processing and artificial intelligence and it searches in entire populations of e-documents for “learned key-concepts”. It can “identify key differences across like documents instantaneously” (Nanney and Raphael, 2016). Argus is the first
cognitive audit application that “learns” from people’s interaction (Deloitte, 2016). It can obtain very important accounting data from a variety of electronic sources.

**Deloitte’s Icount** is a “proprietary tablet-based application that enables auditors to conduct counts, capture results and share the information in real time with all members of the audit team” (Forbes, 2017). Inspections of assets such as inventory counts and property, plant and equipment have drastically changed (Forbes, 2017). Asset verification is not as time-consuming as it was, thanks to Icount. Auditors can also use their smartphones besides their tablets, in order to scan and consolidate inventory count results and then they consolidate and analyze them automatically in an online portal (Nanney and Raphael, 2016). It is very important that “while conducting the count, the auditor could use a voice-to-text capability to create documentation, take pictures of the inventories and produce automatically audit working papers” (Nanney and Raphael, 2016).

**Deloitte’s Illumia:** Illumia is a platform of data analytics. It contains Spotlight that transforms the data in an analyzable format. A very useful application is the testing of journal entries. It also provides visualization of the relationships among data and that facilitates regression analysis (Deloitte, 2017).

**Deloitte’s Magnia:** This is an international audit delivery platform. It distributes an understandable and streamlined audit all around the
world. Of course this increases the quality of the audit services (Deloitte, 2017). Magnia includes: Icount, Iconfirm, Argus etc. (Deloitte, 2017).

**Deloitte’s Cognia:** This is a “single collaborative global repository of innovative auditing tools and leading practices” (Deloitte, 2017). It provides various solutions to auditors.

Furthermore, auditors can have access in a huge database that contains numerous financial statements and that way they can inform their clients about industry trends in real time as they conduct competitive analysis (Nanney and Raphael, 2016).

**Deloitte’s Iconfirm:** It is an online platform where the confirmation is automated and streamlined. It contains various activities such as: “preparing, sending, receiving, and monitoring activities” (Deloitte, 2017).

**Cognitive technology:** Deloitte’s auditors harness cognitive technology for examining contract terms and other documents. Additionally: “The reviews can also incorporate segmentation of documents — for example, separating contracts that include escalation clauses from those that do not” (Davenport and Raphael, 2017). The process contains five steps. Each of these steps alone can add value to the audit procedure. These steps include standardization, digitization, automation, analytics and finally cognitive technology (Davenport and Raphael, 2017).

**An example** from the implementation of advanced analytics in external auditing is the following: H&R Block is a firm that provides tax
services. Deloitte in collaboration with H&R’s audit adopted advanced analytics techniques in the revenue information (Deloitte, 2016). Jeff Brown, Chief Accounting and Risk Officer H&R Block notices that “it was an ideal area to apply analytics because of the breadth of metrics and data that H&R Block had available that could be correlated to expected revenue outcomes” (Deloitte, 2016).

Of course, this procedure took a long time as analytics require a lot of data that must be transformed in a proper way. After this procedure had taken place, auditors could notice “data points that fell outside of hypotheses” (Deloitte, 2016). That way, according to Brown, the quality of auditing increases and auditors have access to the entire population, not just a sample, so they understand better the situation prevailing in the firm. Analytics are very helpful for auditors as they provide information about “outliers and unexpected trends in data, and visualize areas of risk that sampling methods couldn’t do” (Deloitte, 2016).

Deloitte states that there are two main advantages from the implementation of data analytics in auditing.

The first is about the “core” audit procedure, in other words the fact that advanced analytics enable auditors’ access to entire datasets, not just sample and the identification of outliers. Nevertheless, the judgment of auditors is very important and must not be neglected during the audit procedure by Deloitte. Auditors will enhance their performance by using analytics as they will provide high-quality services but under no circumstances they will be replaced by analytics. Furthermore, they will
have more time available in order to provide their services as with data analytics they save time (Deloitte, 2016)\textsuperscript{2}.

The second advantage is that analytics can process a vast amount of data from various sources and as a result auditors can provide extremely useful information to managers.

Lara Abrash, Deloitte & Touche LLP’s National Managing Partner for Audit Innovation and Client Service Delivery, 2016 claims that: “Our audit professionals will be able to discuss the business with client executives in a much more meaningful way. They’ll be able to point out where we are seeing trends in customer behavior, operations, and other key business factors. We’re building industry specific advanced audit analytics for our audits that teams can apply to client data. Over time, we can consistently provide differentiated insights and benchmarking observations to benefit our entire audit delivery process” (Deloitte, 2016)\textsuperscript{2}.

Of course auditors must have certain \textit{skills} in order to handle these new technologies (statistical packages, visual analytics systems, workflow automation, and artificial intelligence tools) and as a result, they must have knowledge about science, technology, engineering, and mathematics (Deloitte, 2016)\textsuperscript{2}.

Deloitte collaborates with organizations and colleges in order to train future audits in advanced audit analytics. Deloitte hires professionals that are experts in analytics in order to improve auditing
and at the same time it trains auditors in order to obtain the appropriate knowledge in these innovative technologies.

**Figure 4:** Enhanced audit integration model

![Enhanced audit integration model](image)


Auditing can gain benefits from professionals in advisory and consulting businesses that implement advanced analytics (Deloitte, 2016).

“For example, Deloitte has a predictive analytics tool that analyzes publicly available financial information to predict and assess scores for fraud, going concern, and restatement potential. To continuously recalibrate, refine, and improve the models, Deloitte leverages data scientists from its consulting business to challenge the correlations and
expand the inputs to develop a leading edge predictive model” (Deloitte, 2016).2

**Internal auditing and analytics:** ““Analyzing data is important!” is an understatement given the deluge of Big Data and where traditional notions of testing are anachronistic to providing assurance” (Deloitte, 2016).1

**Figure 5:** Enhanced insights-driven audit methodology

Integration makes analytics more effective. Consequently, a strong collaboration among core internal auditors, data analysts and subject matter specialists must exist. That makes internal auditing more effective (Deloitte, 2016).1

The advantages of the implementation of analytics in internal auditing are that auditors perform improved audits in lower cost and
higher quality (Deloitte, 2016). Deloitte suggests three steps so as to succeed in implementing data analytics: assessment that is the procedure of “analyzing current analytics capabilities both within IA and across the business and rapidly develop proof of concepts to identify challenges and opportunities” (Deloitte, 2016). Then, auditors must design a strategy concerning the application of analytics and finally, the implementation of the program and the monitoring. Very significant factors are: “Analytics strategy, process, technology, data and people” (Deloitte, 2016).

**Figure 6**: Data analytics competency model


An example of analytics in internal auditing: The CIO of a leading company wanted to audit the company’s expenditures for technology. The purchases were from certain suppliers of the IT firm that were “key
vendors”. “Since the IT organization relied on strategic partnerships with a few key vendors to supply its hardware, it seemed that it would be difficult for employees to go outside the appropriate procurement channel” (Deloitte, 2016). Internal auditors used analytics and a sample of individual purchase orders and data from both internal and external sources (Deloitte, 2016).

**Figure 7**: Business and technology service spend on technology

![Graph showing business and technology service spend on technology](https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-risk-internal-audit-analytics-pov.pdf)

The figure above indicates that purchases outside the IT company had occurred. Moreover, in some cases the overpayment was 300
percent (Deloitte, 2016). The orange line was expected to be flat so this is an indicator of an anomaly.

The figures below provide extra details about the units inside the firm that spent more and why they did so.

**Figure 8:** A closer look at the business’s spend

In the treemap, each box demonstrates a purchase and inside the box there are the dollars spent.

This procedure could possibly reveal many insights such as: the reason why business buyers bypassed the defined suppliers, how they escaped the control mechanisms and what is the cost of this detour (Deloitte, 2016).
Figure 9: Delivering the message through visual cues

Drill-downs into departmental level spend help to give context to this issue.

Using a variety of visual design cues the team effectively communicated big data issues.


Figure 10: Delivering the message through visual cues

The treemap helps to highlight that troubling spending patterns are systemic throughout the organization.

Chapter 2.2. The case of PricewaterhouseCoopers (PwC)

According to Maria Castañón Moats, US Assurance Leader: “The ability to process and interrogate vast amounts of data is helping automate the audit in ways that were unimaginable a decade ago. This in turn is changing how, where, and when our work gets done” (PwC, 2017).

**Halo:** Halo provides data auditing. It is a platform that gives auditors the ability to obtain, visualize and audit huge amounts of data. That way auditors can observe easily the areas that include higher risk into a company and gain insights into the data. It enables automated tests and confirmation of a variety of data.

There is Halo for Journals that “analyzes journal entry data and identifies journal entries to test in order to address the risk of fraud” (PwC, 2017). It uses algorithms in order to detect risky transactions or anomalies in the financial procedures of the company. The procedure is the following: First the data are extracted and renewed every month or quarter. Then, the data are transformed into the appropriate “Halo” form and several tests are conducted in order to reveal certain trends (PwC, 2012-2016). It can handle data from any company no matter the amount of the journals (PwC, 2012-2016). Journals could be controlled real-time which is very important in order to confirm that there is no fraud or error. If the procedure is done manually requires many time due to the fact that journals are too many and from different sources (PwC, 2012-2016). “With Halo, you can interrogate journals to reveal relationships and patterns in account codes, individual users, months, times of day,
types of journals and amounts” (PwC, 2012-2016). It can answer various questions about companies’ journals that reveal certain trends. Below there is an example of a company that uses Halo for Journals and we can see the relationships that were discovered for the fiscal year 2014.

**Figure 11:** Are infrequent users causing potential quality and efficiency issues?


The diagram above indicates the top five financial statement line items (FSLI) in value and volume. Then it shows the top and the bottom five users that post journal entries as far as volume and value are concerned. It also provides various information about journals as their total number, the number of users, the users that have posted less than ten journals, the average line per manual journal and the journals that
their value is less than ten pounds. That way if an infrequent user causes troubles it can be revealed.

**Figure 12**: Are there unexpected journals posted to key accounts?

The diagram above indicates that in payrolls there are a lot of unexpected journals posted. It compares it with many accounts and it finds that it has the largest volume of such journals.

As we can see below, there is the ability to select a person outside the finance department (here is Tyler) and observe the value and the number of the journals that he has posted during the year. This is very important as that way managers are able to control the journal entry procedure and detect any odd or even fraudulent activity. They can obtain details very useful for the company.

Figure 13: Is someone outside of the finance department posting journals?


Figure 14: Are people wasting time posting entries twice, or reversing them?

We can see the debt and the credit (with pink and yellow color respectively) and their duplicates and reversals by each employee. The value and the number of the journals can be noticed, too. Consequently, useful conclusions concerning the employees that waste time in this procedure can be reached.

The above diagrams answer to some questions about the journals and reveal important patterns about the company that without data analysis could not be revealed.

**Figure 15:** Percentage of audit clients for which Halo for Journals is utilized

![Percentage of audit clients for which Halo for Journals is utilized](image)


In 2017, Halo for Investments was introduced. It processes and then visualizes data about investments and “price testing results” (PwC, 2017). Halo for Employee Expenses analyses and visualizes data from employees using algorithms in order to manage costs and detect any potential fraud (PwC, 2015-2018). Expense reports are not so detailed so they may contain some errors. As a result, Halo for employee expenses enables managers to make better decisions about cost management (PwC, 2015-2018). Halo for ERP enables automated testing and risk
management by using algorithms (PwC, 2016). Halo Performance Analyser analyses the business performance and notices any patterns and anomalies that shape business’ performance (PwC, 2016). Halo for Mid-tier is a tool for smaller companies and provides audit tests, risk assessment and reports (PwC, 2016).

PwC Extract was also implemented in 2017 for the first time. It is a tool that auditors use in order to extract proprietary data (PwC, 2017). With this tool the extraction, the protection and the request of customers’ data occurs in a standardized way (PwC, 2017).

Another tool is Audit360 for customers that are in asset management industry. It mines and analyzes data and it enables automated tests of “certain accounts and aspects of financial statement presentation through direct data feeds from the clients’ service providers or through reports uploaded by engagement teams” (PwC, 2017). In 2017, 3000 funds of 75 customers implemented Audit360.

Below we can see the way that data analytics evolve and what happens in auditing and in business, in general. As data analytics are integrated into the company, the procedures that auditors must follow change and actually they are improved.

We can see the five levels of progress and what each level includes. Companies have to try a lot in order to reach the final level which is the transformational one. In the final stage analytics operate as a real game-changer for the audit function in the business. Actually, the execution and the values of auditing are totally transformed.
Figure 16: Measure your success


"Retail Store Loss Prevention Audit": PwC conducted an audit in order to discover the trends in future losses in retail stores (PwC, 2014). It used Big Data and data analytics and predicted several patterns of loss in retail stores. The situation of the industry is the following: There is high inequality of shrinkage and theft levels among retail stores in a national level. Retailers have not the ability to predict the percentage of the loss and as a result they cannot mitigate it. The model was based in certain characteristics: store, location, workforce, economic, customer, multimedia messages (PwC, 2014). The data were gathered from both traditional and Big Data sources. Then, the key data elements in
attributes in stores were profiled. Afterwards, the hypothesis was generated in order to “identify data elements for key attributes prior to loss increase and decrease” (PwC, 2014). Finally, analytics forecasted loss. Eventually, the model observed the changes in key attributes and forecasted loss attributes. Consequently, some suggestions for the mitigation of loss were made.

**Figure 17:** Retail Store Loss Prevention Audit Shrinkage % - High Level Analytic

There were various data categories for modeling such as: customer, demographic and economic, store location, store layout, workforce and email suggestion attributes (PwC, 2014).
**Figure 18:** Sample of stores layout

![Graph showing theft amount and layout](image1)


**Figure 19:** Sample of email recommendations

![Graph showing theft amount and sentiments](image2)

Using all this information PwC created a predictive model and forecasted the shrinkage for more than 1000 retail stores. For this reason it used six variables with their significance in order to conclude.

**Internal auditing and analytics:** Internal auditing must change and integrate analytics in order to face important challenges such as Big Data (PwC, 2018). Analytics must progress so as to adopt this change.

**Figure 20:** The five core Es of breaking through analytics’ arrested evolution


Indeed, internal audit function is progressively changing and it incorporates data analytics and data mining techniques. However, this attempt must be systematical in order to be successful because right now according to PwC: “they get stuck at a point of arrested evolution”
According to PwC the way in order this to happen is the “five Es” as they are demonstrated above:

The five Es in detail:

**Enable:** It is important for internal auditors to have the ability to use data analytics in their daily audit routine.

**Figure 21:** Analytics can help auditors review unstructured data in minutes instead of hours.


Here is an example: Auditors sometimes have to deal with unstructured data, such as equipment-lease accounting and assurance (PwC, 2018). According to the Financial Accounting Standards Board and the International Accounting Standards Board: “lessees will be required to recognize lease-related assets and liabilities on their balance sheets” (PwC, 2018). Auditors extract those data from the contracts that most of the times are in a pdf form, in other words auditors must review unstructured data. If they examine them manually they must waste a
lot of valuable time that could be saved in case that text analytics were implemented (PwC, 2018).

**Embed:** Data analytics must be integrated into every phase of auditing and not just play a small role in the whole procedure implemented (PwC, 2018).

**Figure 22:** Embedded analytics enables continuous risk assessment

Empower: Data analytics could be efficient in case that the roles and the obligations in internal auditing are well-defined (PwC, 2018).

According to PwC there are four significant responsibilities for auditors as far as data analytics is concerned: a dedicated analytics leader who oversees analytics, an analytics manager who suggests solutions for the daily issues that come up, developers who can create applications and business analysts who combine data analytics with audit implementation (PwC, 2018).

Enhance: It is vital for internal auditors to obtain the necessary education and training. Both internal audit analytics team and business auditor should obtain skills in analytics, technology and data, business and audit implemented (PwC, 2018).
Execute: According to PwC: “It’s about taking lessons in agile development from the software industry and applying them to internal audit by conducting short phases of work, or sprints, with frequent review and iteration” (PwC, 2018).  

For example, if auditors want to detect fraud in accounts payable they will use as indicator of fraud a vendor that includes only post office box address and invoices that overpass one million dollars in a certain period of time (PwC, 2018).  

Text analytics are used in order to find the post office boxes’ address. Afterwards, they make a conversation about the results of the analysis and they indicate possible issues such as “It’s keying off the wrong data field, using check amounts when it should be using invoice amounts?” (PwC, 2018). They also discuss the exception list and the business team explains the reason why invoicing was really high during this period (PwC, 2018). The audit and the analytics team collaborate to create a dashboard that will present the results in a user-friendly way (PwC, 2018). The business team searches for exceptions and before the procedure is over the next sprint begins by the audit and analytics team.  

An example of the use of Big Data in internal auditing: The case of Eli Lilly:  

In 2018, PwC conducted a study that was about the Internal Audit Profession. Eli Lilly, according to PwC, achieved to integrate Big Data and Big Data analytics in internal auditing. Specifically, the company
attempts a cross-functional collaboration and as a result a common data warehouse will be created (PwC, 2018). 

Sectors as: Compliance and Ethics, Finance and Internal Audit are independent. However, they cooperate when it comes on risk assessment management issues (PwC, 2018). For example, Internal Audit combines its ability for data extraction with the ability of Compliance and Ethics sector to analyze the data in order to obtain some useful information. Consequently, it was suggested that the internal audit function should collaborate with other functions and a shared warehouse must be created (PwC, 2018). Data analytics will be improved through this initiative.

According to Kathy St. Louis, CAE: “There are three phases in auditing: risk assessment, planning and execution” and internal auditing can enhance all these three phases thanks to data analytics (PwC, 2018). St. Louis states that all the internal auditors should know data visualization techniques thoroughly. In order to improve company’s performance Louis intents to recruit external professionals that will add value to this procedure (PwC, 2018).

According to PwC the benefits from implementing Big Data analytics in audit are: “Increased risk coverage, efficient audit cycle time, insight in real-time or “right-time”, manage risk / return, build a learning organization” (PwC, 2014). These benefits will be created by the progress of visualization abilities, analytics techniques that can save valuable time and the use of unstructured data (PwC, 2014).
Chapter 2.3. The case of Ernst & Young Limited (EY)

EY states: “It's a massive leap to go from traditional audit approaches to one that fully integrates Big Data and analytics in a seamless manner” (EY, 2015). The current progress in Big Data and analytics will transform audit processes. Auditors can test the entire population instead of a single sample and they can raise the quality bar by using intelligent analytics (EY, 2015). Big data combined with analytics give auditors the opportunity to detect fraud and risks and to identify financial reporting (EY, 2015). EY extracts vast amounts of data from its clients and afterwards EY auditors analyze them. However, in the future it is possible that the analysis could be inside each company’s environment (EY, 2015).

EY contains a network of 550 data analysts in order to process and analyze Big Data. These data analysts collaborate with EY auditors and the firms that EY audit (EY, 2017). According to EY, auditors of the future must necessarily have “ability to analyze large data sets and analytical mindset” among other characteristics (EY, 2017).

**EY Helix:** EY Helix is a suite of analytical tools (EY, 2017). It enables auditors to analyze a vast amount of data and to gain better insights. Thanks to Helix auditors have access to the total amount of transactions rather than a part of them. They can also detect several patterns, trends and anomalies through the tools that Helix provides (EY, 2017). Helix provides the ability to mine and analyze data about clients activities (EY, 2018). For instance, annual sales invoicing activity, the impact of credit memos and how the invoices are paid off are detected by using analytics.
That way a better picture of revenue and trade receivables is formed and that benefits the audit procedure (EY, 2018). Helix contains a library that includes all the data analytics from all over the world (EY, 2018).

**Figure 23:** Data types that are accommodated by EY analyzers


In the graphic above the various analytical tools that are accommodated in the Helix suite are presented. According to EY there are future analyzers to come such as: “a warranty tool for use in automotive audits, a trade spending analyzer for use in audits of consumer products companies and a claims analyzer for use in audits of insurance entities” (EY, 2017). **General Ledger Analyzer:** “examines and evaluates GL journals of any data size and volume” (EY, 2018). **Trade payables and receivables analyzer:** analyzes payables and receivables activities in order to conclude about the “purchase-to-pay cycle” and “revenue-to-cash cycle” (EY, 2018). **Group Scope Analyzer:** facilitates auditors in
displaying financial data and defines the “group audit scoping strategy” (EY, 2018). **Inventory Analyzer**: implements analytics to inventory data.

**EY Canvas**: It is an international platform for audit that is used by all US auditors in order to detect and deal with the risk and handle the daily audit responsibilities (EY, 2017). It provides real-time monitoring of the auditing procedures. Additionally, Canvas enables the communication between EY and the firms that EY audits from all over the world and consequently auditors receive faster the information that they need. Auditors are able to connect to Canvas from their mobile devices, as well (EY, 2017). Mobile apps of EY Canvas are: EY Canvas Pulse that gives auditors the opportunity to update the status real time and demonstrate it immediately to the clients (EY, 2018). There is also EY Canvas Inventory and EY Canvas Engage that demonstrates the responsibilities of each auditor along with their deadlines (EY, 2018). According to EY, Canvas offers many benefits such as: “better risks identification and response, centralized planning and monitoring of the global audit and improved project management and ability to monitor key milestones in the audit” (EY, 2018). There is also EY Canvas client portal that enables smooth communication between EY and client firms and facilitates the auditing procedures (EY, 2018).

**EY Atlas**: Both auditors and auditing companies can obtain the most recent accounting and auditing information from Atlas (EY, 2018). Atlas is a platform based on cloud (EY, 2018). It examines accounting
and auditing information by using: “external standards, EY interpretations and thought leadership” (EY, 2018)².

**An example of data analytics:** Analytics tools give the opportunity to explain the correlation between “revenue journal entries and cash collections” (EY, 2017). When it comes to retail firms that indicates that auditors can analyse all the retail locations in order to acquire information about the revenues. This information is more reliable because it includes the whole population (EY, 2017).

**Internal auditing and analytics:** According to EY: “By looking through unprecedented amounts of data from internal and external sources — for instance, individual items at the transaction level — IA can identify and focus on attributes that previously were out of reach, and discern relationships and correlations that were never before visible” (EY, 2014)¹.

However, there are some challenges following the integration of Big Data and analytics in the internal audit curriculum. Firstly, audit professionals must know what kind of information they are looking for as they deal with vast amount of information (EY, 2014)¹. Another major challenge is the cooperation between auditors and data analysts in order to produce the appropriate outcome. There is no doubt that the contribution of IT team is valuable. Finally, it is crucial that the auditors can gain some insights from data analysis and reach to some useful conclusions for the business (EY, 2014)¹. These three challenges are referred by EY as: “define, produce and consume” (EY, 2014)¹.
In the maturity model above we can see the evolution of a firm as far as the internal audit analytics is concerned. Each phase indicates the extent that analytics are used by the company for internal audit. The transition from one phase to another becomes smoothly (EY, 2014).1

**An example of integration of internal audit analytics:** A large multinational technology company decided to create a “leading-class IA function” (EY, 2014). EY’s responsibility was to advise the company as well as to create and implement a data analytics plan. EY demonstrates this implementation by using journal entry analytics as an example (EY, 2014).1

Initially, analytics were used in a pre-defined way and their use was dependent on each occasion. However, it was obvious that the efficiency would be increased if there was “central analytics database with repeatable scripts” (EY, 2014).1. After the transition to the
“repeatable phase” the firm should go to the “defined phase”. In order to do so it must define “significant accounts and relevant assertions”, the appropriate strategy for tests and the necessary information for detecting anomalies (EY, 2014). The next phase, “managed” was achieved by investments in technology tools like server space, allocating resources and creating a standard methodology for reporting and documenting (EY, 2014). Entering into this phase means that the company has integrated analytics into the auditing curriculum and this provides company better risk coverage, increased efficiency and effectiveness, as well (EY, 2014). The final phase, “optimized”, required progress in continuous monitoring of key risks. This would be achieved with the in-time detection of journal entries that include high-risk (EY, 2014). Furthermore, any “accounting surprises” should be detected in an early stage and it was very important to be achieved: “Continuous targeted auditing of all transactions flowing through the general ledger” (EY, 2014). Moreover, governance, that ensured that the team had the right skills and resources for analytics, contributed in order to obtain appropriate tests objectives and outcomes (EY, 2014).

The benefits from this transformation of the auditing procedure were: reduced turnaround analytics time and better adaptation to the business environment, among others (EY, 2014).

**Case study: Moving toward sustainable analytics.** Here is another example of integrating analytics in internal audit function. An
international industrial company had the following goal: “to own a world-class audit organization” (EY, 2014). The firm followed this strategy:

Initially, the firm conducted pilot projects that demonstrated the need for analytics in audit. Afterwards, a three-year “roadmap” was created. In the first year, the goal was to construct an audit analytics function that included a well-defined strategy along with the appropriate technology and tools (EY, 2014). That way, the firm achieved to get to the “repeatable phase”. The next year’s goal was to enlarge risk coverage and incorporate analytics in audit delivery by implementing a certain methodology and process (EY, 2014). Eventually, the company entered the “defined” phase. Finally, at the third year, the risk coverage was optimized by implementing analytics during the risk assessment procedure, industrializing the analytics enablers and tools and promoting training in the function. Additionally, “the function instituted a formal analytics governance process” (EY, 2014).

The firm succeeded to implement advanced analytics in the audit function: in the annual risk assessment, advanced analytics were implemented in almost half of its audits and 4 to 5 projects per year (EY, 2014).

The major advantages according to EY are: audit intelligence, in other words the ability of the auditors to demonstrate the outcomes of their work by using interactive dashboards (EY, 2014). Furthermore, it is very important that the time needed for providing analytics and
monitors is reduced in half to the third and business can gain greater insight than previously (EY, 2014)\textsubscript{1}.

**Case study on forensic data analytics:** A company that conducts FDA, gathers and analyzes various kinds of data in order to deal with legal, compliance and fraud threats (EY, 2018)\textsubscript{1}. Firms can detect and forecast potential anomalies in business operations.

An international company that provides financial services decided to use forensic data analytics in its trading operations in order to detect suspicious activity or leakage of private information (EY, 2014)\textsubscript{2}. The chief compliance officer and the chief information security officer collaborated and they choose certain groups to examine. Specifically, they choose: “Selected portfolio managers in higher-risk security products”, “selected regional and local brokerage branch managers” and “selected employees from a recent acquisition” (EY, 2014)\textsubscript{2}.

The nickname of these analytics was “know your trader” (EY, 2014)\textsubscript{2}. Company harnessed resources so as to facilitate internal inquiries along with “human resource, legal, regulatory and internal affairs matters” (EY, 2014)\textsubscript{2}. Eventually, the firm found out that certain employees created problems in data privacy. The company then gave them the opportunity to take “immediate remedial action” (EY, 2014)\textsubscript{2}.

For example, FDA analytics used e-mails and instant messages to identify rogue activity of the employees as it is pictured in the graph. Emails were compared with a library with key-terms: A) opportunity terms, B) rationalization terms, C) and incentive/pressure terms in line
with the fraud triangle (EY, 2014). In the figure, the high-risk employee Russell Draper displays a triangle that has a lot of these terms in the April 2008, November 2008 and August 2009 which indicates fraudulent activity (EY, 2014).

According to risk director, UK: “We are in a high-risk industry given the prevalence of cash payments. Fraud, bribery and corruption risks are significant concerns. Raising awareness among senior management regarding how forensic data analytics can help mitigate these risks is one of my top priorities” (EY, 2014).

**Figure 25:** Fraud triangle analytics
As for the fraud triangle it is a model built by Dr. Donald Cressey and it gives the three major reasons why somebody would have a rogue activity: rationalization, opportunities and need for money (EY, 2014).

Chapter 2.4. The case of Klynveld Peat Marwick Goerdeler LLP (KPMG)

According to KPMG the audit quality is about making sense from all these tones of data and information in order to provide valid audit results (KPMG, 2018). Of course, auditors need the right tools so as to deal with these huge amounts of data.

KPMG Clara: According to Murray Raisbeck, D&A Leader, KPMG in the UK: “KPMG Clara client collaboration is also the way we will deliver outputs from the data and analytics capabilities, which provide greater transparency to the audit and access to deeper insights, which can ultimately enhance trust and confidence in financial reporting” (KPMG, 2018).

KPMG Clara client collaboration is an online platform that enables the smooth cooperation between KPMG and its customers. It provides “real-time, continuous visibility” into important audit activities (KPMG, 2018).

KPMG employees can obtain, through Clara, the most recent information about audit, about the situation of the prepared by the customer procedure (PBC) and about the reports that the component auditors create (KPMG, 2018).
Figure 26: Intelligent Automation


**Data analytics in detecting fraud:** Firstly, when an auditor tries to detect potential outliers or anomalies in the data it is very important to analyze the appropriate data. According to KPMG this is the quality factor that refers to the quality of the data that are analyzed when detecting fraud (KPMG, 2016). “The sources of data for analysis should include the processes in which an employee could possibly influence a transaction, such as employee expense reports, accounts payable and any transaction that includes the handling of cash” (KPMG, 2016). The data must be precise and not old and the auditor should understand and know the sources. The analytics program must be created in order to respond to the needs of the procedures, for example the transactions that are under investigation (KPMG, 2016). Moreover, it is very
important to realize first what is normal and then trying to find the outliers. Sometimes, when an analytics program fails it is because of the ignorance of normality which makes the procedure harder (KPMG, 2016). Another major concern in the procedure is “false positives” (KPMG, 2016). According to Gerben Schreurs, Global Head Forensic Technology Partner, KPMG in Switzerland, 2016: “A successful anti-fraud analytics process has to walk a fine line between generating too many and too few red flags. Refining the algorithm to achieve this balance is a process of trial and error” (KPMG, 2016). In case that too many red flags are detected, corporate leaders may be skeptical for the process. Furthermore, employees and stakeholders may be suspicious for the program and the employers. However, this case is much more preferable than have a few red flags, as it is safer (KPMG, 2016). For detecting fraud, it is very important to ensure that the operational controls are sustainable (KPMG, 2016). The algorithm that is initially created must be updated constantly. The operational control for long-term period must be “established” and “optimized” in order to have the expected outcome (KPMG, 2016). The ethical factor is also very significant in the whole process. The methods adopted for detection of fraudulent activity must be accepted by everyone in the company such the employees, the employers and the stakeholders (KPMG, 2016). For instance, except for analyzing transactions there is the analysis of the employees’ behavior that may create some issues (KPMG, 2016). A possible solution to that could be the anonymization of the employees’ data until a serious concern has arose (KPMG, 2016).
An example: “Deep pattern analysis algorithms for security anomaly detection”: An international retail company collaborated with KPMG because “its database systems were compromised a cyber-attack that leak non-financial data of its marketplace users” (KPMG, 2016). Employee credentials were harnessed in order to enter to these data. KPMG US was responsible for improving company’s Security Command Center (SCC) monitoring procedures (KPMG, 2016). This had many positive outcomes as real-time detection of network anomalies and deep understanding of network activities. “The client is better able to recognize anomalous network activity, inappropriate applications or applications using unusual ports by tracking network traffic in real time” (KPMG, 2016).

How does a firm implement data analytics in audit? Thanks to data analytics entire datasets can be processed and analyzed in a great speed. For instance, a firm’s transactions can be virtually analyzed and examined in a greater detail (KPMG, 2017). If we want to examine the implementation of data analytics in a firm’s audit procedure, we can focus to the primary “layer” of company’s organization which is the general ledger (KPMG, 2017). The general ledger is the main bookkeeping system of the firm. There are also sub-ledgers such as sales, purchases, inventory, etc. For example, when we analyze the general ledger, we compare the content of the journal entries with what the accounting rules and regulations indicate. That way, a potential material misstatement in the financial statements can be detected (KPMG, 2017). After the quantitative analysis, the qualitative analysis of the books is
possible as well as the analysis of the sub-ledgers. Data-analytics enable auditor to drill-down the data. For example: “with a company’s revenue or sales, the auditor can analyze not just the postings in their financial management system, but the underlying documentation itself, such as actual invoices and bank feeds” (KPMG, 2017). Auditors can find out transactions that contain high risk of misstatement or occasions that the automated processes were skipped and someone intervened in the procedures manually. Predictive analysis is also a very significant part that must not be ignored. For instance, predictive analytics can indicate if forecasts about future revenues are going to occur (KPMG, 2017). “This is important because if a company is overestimating future prospects, it could lead to an impairment or write-down further down the line - something that could hit its share price, cut the value of investors’ holdings and damage market confidence” (KPMG, 2017). KPMG UK in cooperation with McLaren created an “impairment modeling tool” by using “advanced modeling techniques” and “simulation” (KPMG, 2017).

**Internal audit analytics:**

As the company moves from traditional auditing to ad hoc integrated analytics, it implements analytics but in a sporadical way so it does not have the desired outcomes. In this phase, a company does not fully exploits data analytics across the entire audit function so the analytics results do not affect auditing activities significantly (KPMG, 2015). The final maturity level which is the continuous assurance indicates that the
A company has integrated analytics and related technologies and has achieved a great level of automation in the audit procedure.

**Figure 27:** An overview of maturity levels

<table>
<thead>
<tr>
<th>Maturity Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level I</td>
<td>Traditional Auditing</td>
</tr>
<tr>
<td>Level II</td>
<td>Ad Hoc Integrated Analytics</td>
</tr>
<tr>
<td>Level III</td>
<td>Continuous Risk Assessment &amp; Continuous Auditing</td>
</tr>
<tr>
<td>Level IV</td>
<td>Integrated Continuous Auditing &amp; Continuous Monitoring</td>
</tr>
<tr>
<td>Level V</td>
<td>Continuous Assurance of Enterprise Risk Management</td>
</tr>
</tbody>
</table>


“By starting with the phases of a common internal audit methodology and identifying the characteristics at different levels of maturity, an organization can identify logical integration points for repeatable and sustainable data analytics, continuous auditing, and other related initiatives” (KPMG, 2015). It is very important for the firm to recognize its own level of maturity.

The table below indicates in what extend data analytics are used in each internal audit methodology and in each maturity level. That way company can realize in what maturity level its internal audit operates (KPMG, 2015). Strategic analysis is necessary for understanding the industry situation and the organizational structure of the business. Enterprise risk assessment is also vital in order to notice possible risks and dangers for the company (KPMG, 2015).
Let’s take for instance the methodology of internal audit plan development. When an auditor assess the enterprise risks must create the internal audit plan taking into account the risks that he has spotted (KPMG, 2015). According to KPMG, 2015: “Internal audit plan development involves defining the operational, financial and strategic risks that need to be addressed through the execution of the internal audit plan, including the approximate resources necessary to accomplish the scope, and provides a basis for an organization to monitor progress and performance”.

When the internal audit plan development is in the second maturity level, auditors are able to use certain “high level quantitative measures” and find out special patterns in financial statements (KPMG, 2015). The results from the quantitative method are combined with those from the qualitative method. In the third maturity level: “the quantitative and qualitative measures are aligned with priority business risks and internal audit evaluates these quantitative and qualitative measures regularly throughout the year on a quarterly or monthly basis” (KPMG, 2015). In this level the company achieves almost real time reaction to possible changes of the industry situation. In addition, in this level...
internal audit can reveal risks that do not come from the ordinary sources (KPMG, 2015). In the next level, auditors analyse data from both internal and external sources (KPMG, 2015). Furthermore, “Data analytics are system generated from within the business units to enable audits to be added, accelerated, dropped, or deferred (i.e., dynamic audit planning)” (KPMG, 2015). Finally, in the fifth maturity level data analytics are used in a great extent and the company examines risk and performance indicators in a constant basis under the prism of its goals. Both internal and external risk factors are examined in order to identify obstacles to business success. “the analysis of the changes in risk drives the prioritization of audit areas on a continuous basis at predetermined intervals (e.g., daily, weekly, monthly, etc.) (KPMG, 2015).

Case study: Big Data on forensic audit. A public organization collaborated with KPMG in order to find out if the facilities manager had fraudulent activity. Specifically, the organization wanted to investigate whether the manager had skipped certain procedures so as to offer maintenance work to professionals of his own choice (KPMG, 2018). The main issue with this practice is that the firm was paying for imaginary work.

The company extracted billing data of certain vendors from previous years. The company analysed them and also created indicators of fraud such as: “total daily billings divided by hourly rates” (KPMG, 2018).
When KPMG took a look at the organization’s investigation, it found out two main concerns: first, there was the false assumption that all invoiced costs are relevant to labor (KPMG, 2018). The second issue was the assumption that all the labor was executed at the date of the invoice. The days that the labor had occurred were described in detail in “the invoice line item descriptions or in supporting timesheets” and in some cases plenty days of work corresponded to one and only invoice (KPMG, 2018). That created several issues and did not point to the right direction and as a matter of a fact the organization miscalculated the hours. As a result, the client’s analysis provided a little in the procedure of forensic audit basically because the auditors did not have knowledge about the way that the contractors worked (KPMG, 2018).

According to Samir Seyed, Senior Manager, KPMG Forensic Services Canada, 2018: “KPMG used Optical Character Recognition (“OCR”) and unstructured text analysis to index every contractor invoice and distinguish material from labor costs.” (KPMG, 2018). Furthermore, a labor database was developed and it contained invoices’ data so as to detect if contractors billed the actual days of work. Moreover, KPMG used the “log of all magnetic card swipes” (KPMG, 2018). KPMG examined whether there were days that had more than 24 labor hours, whether there was labor that occurred in the exact same date and if there were two invoices with the same description and finally if contractors charged the organization without inserting their magnetic card.
KPMG used “advanced tools to analyze years of invoice data, time sheets and access logs” and that way, besides analysing the data, it captured the “context” of the investigation (KPMG, 2018).3

Eventually, the organization gained the insight that needed and fired the facilities manager.
Conclusions and suggestions for future research

In this dissertation, we examine the impact of Big Data in accounting and auditing. There is no doubt that new procedures, tools and methods must be applied in order to integrate Big Data in the accounting curriculum successfully. Additionally, theory confirms that Big Data offer a variety of advantages to accountants and auditors. Of course there are certain changes that will occur to the profession. However, the implementation of Big Data is a decision that will enhance auditing and accounting and as a result it worths the try. Firms can save a lot of money and time by applying more automated procedures. Furthermore, they can drastically reduce errors in accounting and auditing procedures. Accountants and auditors will provide more useful and meaningful information to managers. Moreover, auditors can examine the entire population and not just a sample and that way the audit quality is improved and potential frauds are detected easily. The «Big 4 Accounting firms», that are the object of the case studies in this dissertation, provide certain examples of Big Data analytics implementation in the audit procedure. For this reason they have developed special tools and they can effectively use all these massive amounts of data that are produced from various sources. Real examples from businesses that cooperate with «Big 4 Accounting firms» prove the huge progress that has been made by these firms. However, the accounting sector must evolve a lot in order to achieve the optimum level and adopt Big Data analytics in the daily schedule.

The issue examined in this dissertation is very significant and in the future it will define accounting. Consequently, further research must be done so as to detect the problems and the gaps that are created and they possibly include several costs (for instance, financial or business costs). Furthermore, the models that are suggested by the literature could be examined and processed thoroughly. For instance, models that indicate the use of data mining techniques for provisioning can be extremely useful for businesses. These techniques enable firms to make forecasts based not only to financial statements data but also to Big Data analytics.
References and webservces


Deloitte, 2016,. Internal audit analytics: The journey to 2020. Insights-driven auditing. Available at:


Deloitte, 2017,. Aspire with assurance. Illuminating the audit of the future. Available at:

https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Audit/gx-global-audit-brochure.pdf

Deloitte, 2017,. 2017 Audit Quality Report. Advancing quality, for today and tomorrow. Available at:


PwC, 2018. 2018 State of the Internal Audit Profession Study. Moving at the speed of innovation. The foundational tools and talents of technology-enabled Internal Audit. Available at:


Appendix

Chapter 1. Definition of Big Data

Strawn (2012), described Big Data as “fourth paradigm of science”, while (Hagstrom, 2012) defined it as “new paradigm of knowledge assets”. According to McKinsey & Company, an international consulting agency: “Big Data is the next frontier for innovation, competition, and productivity” (Chen et al., 2014). Gantz and Reinsel, (2011) defined Big Data as “a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and analysis”.

Figure 30: Big Data interest over time (Google trends, 2016)

Although Big Data are everywhere today there is not only one definition. The origins of the term are probably in the midnineties, specifically: “big data . . . probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid-1990s, in which John Mashey figured prominently”, (Gandomi & Haider, 2015). The term became popular in 2011, according to the figure. There is no doubt that the major feature of Big Data is the size, or in other words, the volume. According to McKinsey Global Institute Big Data are “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze” (Moffit, 2013). Nevertheless, there are also other important characteristics of Big Data. Actually, there are the 3 V’s of Big Data: Volume, Velocity and Variety.

**Figure 31**: Definitions of Big Data on an online survey of 154 global executives in April 2012

For instance, Gartner, Inc. definition for Big Data is: “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”. There is a similar definition for Big Data by TechAmerica Foundation: “Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.” (Gandomi & Haider, 2015).

If we want to describe the three V’s in detail:

**Volume** is about the size of the data. The amount of data is huge and their scale is high. That means that Big Data are in terabytes or petabytes. Every hour Walmart gathers at least 2.5 petabytes of data that are relevant with its transactions with clients. The storage capacity of a terabyte equals with 1500 CDs or 220 DVDs and can store 16 million Facebook photographs. One petabyte equals 1024 terabytes and is one quadrillion bytes. According to various estimates Facebook stored 260 billion photos by using more than 20 petabytes. There are also the exabytes that are 1000 petabytes or one billion gigabytes. There are some examples of companies that deal with Big Data today: Tesco produces more than 1.5 billion new data per month (Wamba, 2015). Similarly, Dell created a database that includes 1.5 million records about sales and advertisements (Wamba, 2015).
Variety refers to different structures of data that are produced. There are structured data (that are traditionally used by companies), semi-structured data, unstructured data and multi-structured data. Structured data that are only 5% of all existing data (Gandomi & Haider, 2015), refers to the tabular data in spreadsheets or relational databases. Semi-structured data do not follow certain rules. An example of semi-structured data is XML (Extensible Markup Language) a textual language for exchanging data on the web. XML documents are machine-readable because they have data tags defined by the user. Unstructured data are difficult and sometimes impossible to be analyzed by machines. Such data are videos, images, text and audio. Multi-structured data are a combination of various kinds of data with different structures.

Velocity is about the frequency that data are generated. It is also about the real time analysis. Nowadays, companies generate high-frequency data and there is the need for high-speed analysis for the interest of the company. Real-time analysis can offer a significant competitive advantage to a company. Wal-Mart, processes more than one million transactions per hour (Gandomi & Haider, 2015). Another example is Amazon that has no delay in deliveries due to wise and real-time management of new products, suppliers, customers and promotions (Wamba, 2015). Smartphones and mobile apps provide numerable information about customers such as location and previous consumer behavior that give the ability to the company to create real time customer value by analyzing all these data in real time. The
immediate analysis of these data is a challenge that only Big Data analytics can confront in order to create “real-time intelligence”.

Except of the above mentioned wide-known features of Big Data, there are also some other characteristics such as:

**Veracity:** This is the fourth V of Big Data according to IBM. Although Big Data offer valuable information, a great amount of data is not accurate. As a result, there is the need for certain tools, for instance Big Data analytics and data mining in order to extract value from all these uncertain data. For example, customer data from social media need a certain processing before becoming useful for a company.

**Variability (and complexity):** These two attributes of Big Data are introduced by SAS. Complexity is about the fact that Big Data are produced by a great variety of sources and thus, it is a challenge to combine all these different sources and generate value for the company. Variability is about the fact that the data flow rate is not the same all the time, in other words, there are “ups” and “downs” in data velocity.

**Value:** This dimension of Big Data was stated by Oracle, which consider it as a critical feature. Value refers to the extraction of useful information from massive amounts of data for business interest. Big Data can make noise and outliers very informative and meaningful for the company. This important feature of Big Data was also underlined by IDC: “Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture,
discovery, and/or analysis.” Match.com noticed 50% increase in revenue in the past two years, thanks to Big Data analytics (Wamba, 2015).

All these attributes relate to each other and they affect each other to a great extent. However, the main features of Big Data (volume, variety and velocity), in other words the three V’s model, define if a company deals with Big Data or not. These characteristics are not identical for all businesses, they depend on sector, location and size of each company.

**Figure 32**: The 4 Vs feature of Big Data

![Figure 32: The 4 Vs feature of Big Data](source)


**Chapter 2. Big Data Technologies**

**Big Data and Cloud**

Cloud environment is very popular among companies because managers don’t have to worry about the maintenance and the administration of the hardware and the software. Companies can have access to their data anytime, no matter their location, even though the
place where the data are differs from the company’s location. Furthermore, companies have the ability to pay only for the cloud services that they use, after they have used them.

**Figure 33**: Cloud computing logical diagram

Another definition is: C.L.O.U.D: (1) Common infrastructure, (2) Location independence, (3) Online accessibility, (4) Utility pricing, and (5) on-Demand resources (Moffit, 2013). According to Weinman (2012): Cloud affects many sectors into a company such as the design of new products, research, and engineering. It also affects the relationship of the company with its customers (Moffit, 2013).

Thanks to cloud, Big Data usage becomes smooth because cloud systems use enormous storage and computing resources that enable Big
Data analysis. Big Data analytics in cloud can be achieved with various virtual analytical applications that provide data processing in different times, patterns and frequencies (Demirkan and Delen, 2013).

**Figure 34:** Data deluge: the increase of data size has surpassed the capabilities of computation.

As a result, it is obvious that the progress of cloud systems depends on Big Data and their increasing use by companies. Nowadays, Big Data are everywhere so it is urgent for cloud to be improved all the time in order to respond to the demand of companies for analyzing and storing these massive amounts of data. Various companies offer several solutions to businesses that store Big Data, for example: Amazon’s Web Services, AT&T’s Synaptic Hosting, AppNexus, GoGrid, Rackspace Cloud Hosting, the HP/Yahoo/Intel Cloud Computing Testbed, the IBM/Google
and MicroStrategy BI Cloud (Demirkan and Delen, 2013). Even though cloud computing and Big Data depend on each other, they are different notions. Big data, after their preprocessing and analysis, affect the decisions that are made by the business and as a result they concern the CEO (Chief Executive Officer). Cloud is about the technology that business use and concerns CIO (Chief Information Officer). However, the development of Big Data enhances cloud computing and vice versa. According to the President of EMC “the application of Big Data must be based on cloud computing”.

**Figure 35:** Key components of cloud computing

![Cloud computing applications and services](image)


Governments around the world also use cloud as a means of data storage and Big Data analytics in order to improve the operation of public sector. For instance, U.S government has introduced the “Cloud
First” proposal. The primary goal of this is to shut 1000 government data centers until 2015 and put 79 services to the cloud (Kambatla, 2014). A more recent example is the “Building a 21st Century Digital Government” initiative which was proposed by the US President in 23 May 2012 (Kambatla, 2014). Furthermore, there are several similar projects that were adopted such as: AWS GovCloud which is a commercial service created to move great amount of data to the cloud, US.gov, the official site that provides all the necessary information about US government and last but not least, the decision that was made by the Department of Treasury to move all the websites to Amazon’s Elastic Cloud (EC2) (Kambatla, 2014). Of course, all the above mentioned changes have improved the operation of government services since cloud makes the use easier and government can save time and money, as well.

**Figure 36: Cloud Storage**

Of course the use of cloud apart from advantages has also some disadvantages. For instance, there is a concern related to the fact that data is possible to leak because servers are not in the company and customers cannot choose the place that they are stored. Another major problem is that companies have to obey to the laws and the system of the country that the data are stored (the data of a company can be in different country than the company). For instance, the only choice that Amazon S3 gives to its clients is whether the location for data storage be in US or EU (Demirkan and Delen, 2013). An additional danger of cloud environment is data loss or data availability that can create a huge damage to the company. Nevertheless, it is sure that as the use of Big Data becomes more and more extensive, the problems of the cloud computing will be diminished.

**Big data and platforms**

The Apache Foundation Software was founded in 1999. A few years later, in 2003 batch processing of Big Data started. In batch processing the data are first stored and they are analyzed after a particular time. In 2004, Google published its papers about Google File System and Map Reduce. Nowadays, MapReduce is the most popular technology for batch processing (Saggi and Jain, 2018). Then in 2006 Hadoop was created. Yahoo begun the creation of Apache Pig, and Facebook created the Apache Hive which is an open-source platform that enables the processing of Big Data. These tools were officially introduced and started to operate in 2008 and 2009, correspondingly (Saggi and Jain, 2018).
Later, Twitter introduced an open source for real time processing, the Apache Storm. Other examples of Big Data technologies are Flume and Kafka by Cloudera and LinkedIn. These open source technologies have created an environment around Apache Hadoop that involves many useful tools for Big Data. Other elements of this environment are: Apache Spark, Mahout, Sqoop (it transmits data between Hadoop and various other systems), Oozie (it schedules the jobs on Hadoop and does the monitoring, too), Zookeeper (remoted process for configuration and coordination), and Apache Giraph (it creates graphs from data processing) in 2014 (Saggi and Jain, 2018).

**Figure 37**: The timeline of the Big Data processing and technologies

In fact, in June 2012 it was stated that Hadoop is used by Yahoo in 42,000 servers and 4 data centers to provide various services such as spam filtering, searching and so on. Similarly, the exact same period Facebook claimed that the ability of its Hadoop cluster is 100 PB that increased in 0.5 PB each day in November of the same year. Furthermore, there are many companies such as Cloudera, IBM, MapR, EMC, and Oracle that offer support for Hadoop (Min et.al, 2014). In addition, Bahga et al. introduced a combination of Hadoop and a cloud computing infrastructure called CloudView in order to analyze data produced from sensors. Cloud computing infrastructures, such as Amazon AWS and Microsoft Azure, along with Hadoop (or MapReduce, Microsoft DryadLINQ) can also be used for data processing in bio-medicine and bio-science, as well (Min et.al, 2014).

As we can understand from its name MapReduce offers to the user two choices: Map and Reduce. Both operations are programmed by the user. The procedure goes as follows: Map function produces intermediate key-value pairs from input key-value pairs. Then, MapReduce will create a combination from all the intermediate values that are relative to the same key and transfer them to the Reduce function. This function transforms the value set into a smaller one. This procedure complies with divide and conquer method (Chen and Zhang, 2014). In other words, it divides the initial complex problem into sub-problems in order to solve them easier and reach to a conclusion for the initial problem. The divide and conquer method is implemented by two steps: Map step and Reduce step. Hadoop cluster has two types of nodes,
master nodes and worker nodes. The input goes to the master node where it is divided into sub-problems. Then, in the Map step they are transferred to the worker nodes. In the Reduce step the output is created after the collection of the solutions from the master node. In order to start the processing in the Reduce, the Map processing must stop, because of that MapReduce is not appropriate for real-time analysis. Another parallel programming model is Dryad that has a different structure than MapReduce (Min et.al, 2014). According to Chen and Zhang, 2014: “It bases on dataflow graph processing”. Apache Mahout is an analytical tool that is very handful for companies that deal with Big Data because it provides machine learning algorithms for clustering, classification and so on. Mahout is on Hadoop platform and it uses MapReduce in order to operate.

Of course, there are many other platforms and software that provide batch processing as well as real time processing for Big Data analysis. All these developments are very important because they enable Big Data analysis and therefore they create value for the businesses that they become more agile.

Chapter 3. Applications of Big Data in society, government and business

Society

Healthcare Sector

Big data can drastically change the healthcare sector, taking into account that their analysis can provide valuable information to medical
professionals. The analysis of Big Data can transform many aspects of health management such as Research and Development (R&D), treatment, testing, and diagnosis (Saggi and Jain, 2018). Nowadays, hospitals and clinics are increased day by day in order to treat efficiently the growing number of patients and as a result the medication that is provided is increased, too. Therefore, there are issues in storing, processing and analyzing of data that come from the patients (Saggi and Jain, 2018).

**Figure 38:** Big Data in healthcare sector

It is obvious that a massive amount of data is collected from the diagnosis and the treatment of the patients. These data can be
structured and unstructured data, self-monitoring health data, real-time sensor devices, pictures, videos, several reports, and documents. All these kind of data can be extremely useful in various healthcare systems like: health-care management and innovation drug discovery (Saggi and Jain, 2018).

Apart from collecting data that can be used for the treatment and the diagnosis of the condition of the patient, Big Data can be very useful in storing individual information about each patient. For instance, this information can include: case history, physician notes, Lab reports, X-ray reports, diet rule, list of doctors, and nurses in a specific hospital, national health register data, medicine and surgical instruments expiry date identification based on RFID data (Saggi and Jain, 2018).

Furthermore, there are platforms that provide Big Data analysis for the healthcare sector: CHESS (Batarseh & Latif, 2016), EHR, LIMS, MQIC, CMS (Ward, Marsolo, & Froehle, 2014). There is also the Center for Health Analytics that is necessary in order to provide services to medical professionals all over the world. In addition, Big Data analysis is very useful in preventing mistakes, fraud and waste in the sector of health insurance (Saggi and Jain, 2018).

According to a thorough research that was carried out by McKinsey & Company the U.S medical sector can gain more than 300 billion dollars from the effective use of Big Data (Min et.al, 2014). Additionally, the cost of this particular sector can be reduced by 8%.
The field of biomedicine can gain a lot from Big Data. Lately, Big Data are extremely used in this sector after the HGP (Human Genome Project) and the development of sequencing technology (Min et. al, 2014). Gene sequencing can generate a lot of data that after their analysis can lead to immediate and accurate diagnosis as well as special treatment of a disease. According to Min et.al, 2014: “One sequencing of human gene may generate 100 600GB raw data”. Additionally: “In the China National Genebank in Shenzhen, there are 1.3 million samples including 1.15 million human samples and 150,000 animal, plant, and microorganism samples.”

It is obvious that the R&D that medical centers conduct provides a lot of information. For instance, the Medical Center of Pittsburgh University (UPMC) has 2 terabyte of data (Min et. al, 2014). American companies as Explorys and Practice Fusion maintain huge amounts of data about patients’ records and they provide various services in order to manage them. IT companies as Google, Microsoft, and IBM do a thorough research in order to find out applications that can extract value from Big Data in the bio-medical sector. IBM in the 2013 Strategy Conference predicted: “with the sharp increase of medical images and electronic medical records, medical professionals may utilize Big Data to extract useful clinical information from masses of data to obtain a medical history and forecast treatment effects, thus improving patient care and reduce cost.” (Min et.al, 2014). In 2015, every medical center stores 665TB of data.
The flu of 2009 and the contribution of Google

Big data analysis was proven valuable when the flu pandemic occurred in 2009. Google by analyzing all these massive amount of data informed citizens more efficiently in comparison with prevention centers (Min et.al, 2014). The procedure that prevention centers of this new type of flu followed was more complicated and slow taking into account that people did not visit medical professionals the moment that they had the virus. Additionally, after doctors examined patients they sent the diagnosis to the prevention centers and there they needed some time in order to analyze the data. So, it was almost impossible for these centers to prevent the disease. Google after analyzing the data gained insight in this phenomenon. First of all, it was noticed that during the influenza people searched different things than usual. There were various frequencies and locations related with the quest about this type of flu. Specifically, people in Google spotted 45 search groups about this disease (Min et.al, 2014). Then, by using mathematical models they forecasted the location and the frequency of the flu.

Health application

In 2007 Microsoft, released the HealthVault application. This application processes medical Big Data of individuals and their medical devices (Min et.al, 2014). Information can be obtained by devices such as mobile phones and third-party agencies that have the medical records of the patients.
Smart cities

The application of Big Data in smart cities is not just about the architecture of the city. It includes many aspects such as the transportation system, education, energy, health and many others. Big Data can be stored and analyzed and that way the daily life of the citizens will be improved.

For example, there is a collaboration between Miami-Dade County in Florida and IBM. This collaboration has multiple purposes: improve the use of water resources, provide a safe environment to the citizens and manage traffic efficiently. The County has saved a lot of money thanks to this project and an example is the case of Department of Park Management of Dade County that saved 1 million dollars from watering as it had the ability to repair immediately the damages (for example: leaking of water pipes) (Min et.al, 2014).

Another example in U.S.A is the Big Data analysis in the Santa Cruz Police Department. It uses Big Data in order to forecast trends about crime behavior. Furthermore, it has the ability to predict the number of crimes in important areas.

Astronomy and physics

Sloan Digital Sky Survey (SDSS) which is the most significant survey in astronomy has stored 25 terabyte of data from 1998 to 2008 (Min et.al, 2014). Taking into account the improvement of the telescope, the number of data that are generated every night is increased. Specifically, in 2004 the number of data was more than 20 terabyte.
Similarly, in the field of physics the number of data generated every day is huge. The Atlas experiment of Large Hadron Collider (LHC) of European Organization for Nuclear Research produces 2 petabyte of data and has a warehouse of 10 terabyte per year.

**Social behavior**

In April 2013, Wolfram Alpha a computing company analyzed data from Facebook for more than 1 million Americans. That way it found out patterns about the social behavior of Americans and particularly when they get engaged, when they get married, when their relationship status changes etc. (Min et.al, 2014). Another company, Global Pulse, analyzed Twitter messages from U.K, Japan and Indonesia from 2010 to 2011 (Min et.al, 2014).

**Figure 39:** The correlation between Tweets about rice price and food price inflation

They observed: the increase and decrease of a reference to a specific topic in order to notice unusual events. They collected public discussions in Twitter so as to observe the trends and understand how the behavior of people changes by analyzing the various sub-topics of conversations. An important finding was the increase in Tweets about the price of rice that was observed simultaneously with the inflation in Indonesia.

**New South Wales state emergency service**

The New South Wales state emergency service in Australia uses Big Data in order to provide effective services. Specifically, the NSW SES deals with many emergent situations such as floods, cyclones, storms, tsunamis, and other natural and human disasters (Wamba et. al., 2015). In order to improve its services it must process data from different sources such as videos, pictures, weather data (from the Bureau of Meteorology), etc. In other words, it has to deal with both structured and unstructured data. Of course, the volume of these data is huge and they require real-time analysis. In addition, in order to prevent various disasters the service must send data to its volunteers as soon as possible. For this reason, it uses many technologies and platforms. The procedure is the following: Bureau of Meteorology sends data that reveal important information on a dashboard in the service’s headquarters. Then, the service shares this information with its volunteers, social services and the police through many kinds of sources like social media or smartphones. The NSW SES uses an improved IT infrastructure that
enables Big Data analysis. That way not only can manage in a more effective way the natural crisis but it can also predict future disasters and as a result respond in an immediate way.

**Government**

McKinsey & Company during their research in developed countries in E.U discovered that the European governments can save more than 100 billion euros if they take advantage of Big Data (Min et.al, 2014).

Of course there are specific applications in governance that can improve the current situation in many ways. Some of these applications are presented below:

**Surveillance:**

There are various systems that are used by governments and they include Big Data. In U.S.A there is the xKeyscore system that collects e-mails and web traffic data. In addition, there is the National Security Agency (NSA) that gathers data from phone calls (not the substance of the phone calls) all over the U.S, from U.S companies and then keeps them for five years. The companies do not maintain the data for this long due to the massive amount of them. This model that is named “collect first” provides useful data to national security investigations (Moffitt and Vasarhely, 2013). It is obvious that the amount of data is large, taking into account that daily phone calls are 500 million. Another huge database is the PRISM that includes data from emails and social media of foreigners. Of course, these systems service government’s purposes and as a result only a few details are known to the public.
However, there is no doubt that all these systems can be very handy for businesses, too. For example, they can offer a lot in accounting reporting, assurance, marketing analysis and sales (Moffitt and Vasarhely, 2013). The reasons are obvious: when a business knows the identity of those who call or the frequency and the number of phone calls can gain valuable knowledge. For instance, they can gain insight about competition, their own performance and the preferences of their customers as well as the general economic situation. Consequently, big databases can be very useful even when not used for the initial purpose.

**The SEC Audit Quality Model**

In 2013 the SEC (U.S Securities and Exchange Commission), announced that it would apply something new in order to find out and decrease fraud (Moffitt and Vasarhely, 2013). This would be the Audit Quality Model (AQM), or according to several journalists: “Robocop”. It will follow this procedure: first, it will select the XBRL data that are available on the EDGAR database. According to Financial Accounting Standards Board: “XBRL, or eXtensible Business Reporting Language, is an XML standard for tagging business and financial reports to increase the transparency and accessibility of business information by using a uniform format. Furthermore: “XBRL US, the US-based organization of XBRL International, supports the implementation of XBRL in the United States through the development of the digital dictionary (taxonomy) relevant for use by US public and private sectors”. As far as for the EDGAR database SEC states that: “EDGAR, the
Electronic Data Gathering, Analysis, and Retrieval system, performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the U.S. Securities and Exchange Commission (SEC)”. So, the AQM will cross-check the reports from all the competitors of an industry and then will point out those who are very different from their peers. That way any peculiar behavior of the business will be revealed especially on the management of its discretionary accruals (Moffitt and Vasarhely, 2013). Even though at the time that the implementation of this model was announced it did not have the requirements for Big Data (taking into account that the database was in gigabytes) it is sure that during the next years the database will be expanded in terabytes.

Moreover, the data will not be necessary XBRL. SEC will have the ability to gather various kinds of data for the companies such as: phone calls, data from the web or the social media of the company, what a company buys (Moffitt and Vasarhely, 2013).

Additionally, it will be very useful for SEC to know how the clients as well as the employees of the business act. Actually, it is true that in many cases the U.S Justice Department needed these kind of data. Consequently, SEC will have to deal with various and plenty of data and because of that it will be crucial to use Big Data tools. Thus, the amount of information will be increased and that will be beneficial for the government and the companies, too.
**Big Data R&D initiative**

In 2012 the Obama administration announced the Big Data R&D initiative that has the purpose to solve several governance problems. For this reason, 84 different programs were created by 6 departments (Chen P. and Zhan C-Y., 2014).

**Business**

Big data applications in enterprises can improve their productivity, make them more agile and as a result give them a competitive advantage. Big data analysis can be applied in many aspects of the business. For example, in marketing they can understand the preferences of the consumers and the current consuming trends. Furthermore, if enterprises analyze Big Data effectively they can enhance their supply chain in many ways. They can improve the management of the inventories and the logistics and consequently they can improve the relationship between supply and demand.

**Smart agriculture**

Agriculture firms can gain a lot from Big Data analysis as they can improve their performance in many ways. They have already used Big Data extensively in order to gain knowledge. Internet of Things, as we have mentioned above, produces a large amount of information that agricultural companies can use for their interest. Consequently, they use Big Data applications for processing and analyzing Big Data to obtain real-time information by analyzing different structures of data, data from various sources or data from previous periods.
Specifically, IoT generates agricultural information from sensors that exist in fields. These information are about: the moisture level of soil, trunk diameter of plants, micro-climate condition, and humidity level, as well as to forecast weather (Saggi and Jain, 2018). After all these data are collected they are processed and analyzed. That way they enable agriculture firms to control or apply several techniques in the production procedure. Some examples are: on time watering, control of the humidity in order to avoid the development of fungus and climate control during the harvesting of crops (Saggi and Jain, 2018).

According to a case study presented by Khestri (2014), agriculture can have many benefits, chances and challenges by using Big Data analysis. He also proposed Big Data analytics to farmers in order to get information about changes in the weather, the condition of soil, topography and innovative methods of planting. In addition, Xie et al. (2015), suggested that a hierarchy of agricultural information system can be obtained by a technology that process Big Data. Specifically, it will be about collecting, storing, analyzing and visualizing Big Data of the agricultural sector. The technology that will provide Big Data services is Map Reduce. That way, farmers can save money from fertilizers and pesticides and thus, increase their efficiency and productivity. Finally, Bendre et al. (2016), introduced an application for predicting rainfall by using Big Data forecasting methods (Saggi and Jain, 2018).
Retailing, e-commerce and finance

The above mentioned research that was conducted by McKinsey & Company also predicted that retailers can increase their earnings over than 60% if they use Big Data in a fully productive way (Min et.al, 2014).

Taobao and Alibaba are using Big Data all the time. Specifically, the data that Taobao collects every day are thousands and they are about sex, origin, preferences of the consumers as well as about the time, the quantity and the frequency of each transaction. Taobao uses Data Cube which is a platform where businesses can get many information about the economic situation and the consumers’ behavior. As a result they can decide how to manage their inventories, their prices etc.

Alibaba gives a credit loan to a business after it analyses the data of its transactions. For this purpose, it uses Big Data. It is remarkable that even though it has lent more than 30 billion Yuan, the percent of failure is 0.3% (Min et.al, 2014).

China Merchants Bank (CMB) uses data analysis in order to rate its clients (Min et.al, 2014). This bank analyses the transaction record of the customers in order to reach to some conclusions for their credibility. That way it has achieved to decrease the ratio of those who did not pay the bank at 15% for Gold Card owners and 7% for Sunflower Card owners.
Cyber-physical systems

The use of computer security networks is very extensive in companies and in governments, as well. They use these kind of networks in order to create a protection for their sensitive information. They use Big Data tools so as to gather, analyze and store their data. Cyber defenders apply technology not only to protect data but to find out if there is any malware or cyber attackers. The role of Big Data analytics is very significant to ensure that the data of the business are secure and private and at the same time to permit to certain organizations to gain access to data in order to create an opinion for the situation of the company. According to Saggi and Jain, (2018): “The emergence of cyber-physical systems can be used for production, transportation, logistics, and other sectors to bring new challenges for simulation and planning, for monitoring, control, and interaction with machinery or data usage applications”.

Airlines

In airline industry it is very important for each company to know the exact time of landing. If the plane lands earlier, passengers will have to wait so they will complain and if it lands later the staff in the ground will have to wait for no reason. In case that the exact time is known the company saves money as it “uses” effectively its employees and it offers the best services to its clients. An American airline company found out that nearly 10% of its flights landed 10 minutes earlier or later and 30%
landed 5 minutes earlier or later. Consequently, it decided to use Big Data in order to improve its operation.

Until then, airports had to rely on the information provided by pilots shortly before they reached the ground (these estimates called ETAs). In 2001, PASSUR Aerospace offered its own information which called RightETA (McAfee A., Brynjolfsson E., 2012). It collected many data about the weather, flights, and many other data about each plane. In 2012, PASSUR Aerospace had more than 155 radar stations that collected data about the airplanes. A huge volume of data is generated every 4.6 second for each plane. Additionally, it stores all these data in order to study the conditions of landing and creating certain patterns. Based on previous data, it can reach a conclusion about the time of the landing. The airline company has saved millions of dollars by improving the procedure of decision making thanks to Big Data.

Sales promotion

A few years ago, Sears Holdings decided to combine data from Sears, Craftsman, and Lands’ End brands so as to improve its sales strategy (McAfee A., Brynjolfsson E., 2012). Because of that it was necessary to use Big Data to gain insights for the products that each customer would prefer. Of course, initially this procedure took many time because of the volume of the data. Furthermore, the data required were stored in different places and it was difficult to be obtained quickly. That was the reason why company decided to use Hadoop in order to store and analyze the data coming from different sources. All these
petabytes of data were processed at no time and with very low cost. Except of the cost and the time, the quality of the promotions was improved a lot. Sears Holdings could understand better what kind of products would prefer certain groups of clients. The company could make the right decisions about the sales of its products by using Big Data and the related technologies.