National Research University Higher School of Economics

Institute for Statistical Studies and Economics of Knowledge

Master Thesis

IMPLEMENTATION OF BIG DATA ANALYTICS IN CONSULTING: KPMG CASE STUDY

Student: Oleksandr Herashchenko

Group: MYH191

Supervisor: Alexey Ponomarev

Moscow, 2021

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
BI	Business Intelligence
D&A	Data & Analytics
DB	Database
DM	Data Mart
DMBS	Database Management System
DW	Data Warehouse
ETL	Extract, Transform, Load
ICT	Information and Communication Technologies
IoT	Internet of Things
IT	Information Technology
LoS	Line of Service
ML	Machine Learning
PA	Public Administration
PM	Project Manager
PMO	Project Manager Officer
TB	Terabyte

List of Tables

List of Figures

Abstract

This research expands the knowledge of the academic literature on the world of Big Data Analytics through the deepening of how consulting firms implement these technologies in companies that rely on them for this purpose. Through interviews to consultants of the Italian network of KPMG, top experts in this field, some shared practices and technological patterns used in projects that encompass such technologies are illustrated. In addition, some insights about the critical parts of the technology and the critical factors which consultants should pay attention to be successful when implementing these technologies are also provided, as well as the possible direction of future developments in this field.

Therefore, this article starts with an introduction of the Big Data Analytics topic, underling the gap found in the current academic knowledge and explaining accordingly the goals of this research. Then, the main text is divided into three chapters. The first chapter provides the theoretical background based on the literature review: starting from the definitions regarding the dimensions of Big Data and the difference between the data itself and Big Data Analytics; continuing with the impact of Big Data Analytics on business in terms of opportunities and barriers for organizations, fields of application and benefits in exploiting such technologies; thus, ending with the non-technological aspects that are required to achieve these benefits. The second chapter explains the methodology and approach used in this research to answer the research questions, from the choice of a single case study methodology and the sources of information, to explain the two criteria used for the selection process of respondents and the design of the interview guide. The third chapter discusses the findings of this research: starting from the compliance of the Big Data Analytics projects in consulting with the knowledge gained by the literature review; continuing with the more technical aspects of these projects such as the infrastructure, platforms and tools involved as well as the critical parts of the technology; and ending with the critical aspects of such projects for consultants and the possible future directions of Big Data Analytics. In the concluding chapter is provided an overview on the research conducted in this article underling the key findings and the results achieved; finally, the limitations of this research are highlighted and consequently further research in this context are suggested.

Introduction

Nowadays, technological developments are improving exponentially and this leads to the emergence of several digital trends that could have a significative impact on businesses (Boston Consulting Group & University of Virginia, 2017). In this regard, one of the most important factors is that companies are now able to gather a huge amount of extremely detailed information about their stakeholders, from within and outside companies, in the form of structured and unstructured data. Thus, better storage solutions, decreasing storage costs, and the availability of algorithms that create meaning from data allow firms to extract more benefit from large volumes of data (Ghasemaghaei & Calic, 2020; Lee, 2017).

Actually, more data cross the internet every second than were stored in the entire internet just 20 years ago (McAfee & Brynjolfsson, 2012). For example, more than 98,000 tweets are written every sixty seconds, 695,000 status updates are posted on Facebook, 11 million instant messages are written, 685,445 Google searches are lunched, more than 169 million emails are sent, more than 1820 TB of data are created, and there are 217 new mobile web users (Raguseo, 2018). Basically, this means that those companies which are able to transform this data into value could gain a competitive advantage over their competitors (Gupta et al., 2020; Erevelles et al., 2015, Raguseo, 2018; Wamba et al., 2015; De Mauro et al., 2019). Indeed, through innovation, Big Data may improve firm performance. Specifically, it can help firms collect and process market information to better understand consumers' preferences, which can play a critical role in innovation performance (Ghasemaghaei & Calic, 2020).

Gupta et al. (2020) claim that information stands out as the most potent fuel from which an organization can derive success. Thus, in the "Era of Big Data", data itself is seen as a new class of economic asset, like currency or gold (World Economic Forum, 2012). In fact, there are several fields of application in business such as the personalization of offerings, fraud reduction, predictive maintenance, logistic optimization, marketing, business strategy and the exploitation of consumer analytics has become a key issue in the competition of several markets (McAfee & Brynjolfsson, 2012). Big Data Analytics even affects the overall economic and business policies (Amankwah-Amoah, 2016).

Despite that the academic literature in the past decade provided several studies around the world of Big Data regarding the definitions and challenges, as well as the opportunities and the barriers, the fields of application and the benefits of such technologies; little progress has been made to understand how these technologies are implemented in companies. Since these processes are quite complex, the majority of companies prefer to delegate these projects to consulting firms. In fact, in spite of the growing number of firms that are relying to consulting firms in order to launch Big Data initiatives, there is still limited understanding on how the consulting firms do the processes of design and implementation of the infrastructures, platforms and tools used for this scope.

Therefore, the aim of this research is to cover the gap in knowledge of the current literature about the processes of design and implementation of Big Data Analytics in consulting, in particular, answering the following research questions:

How do consulting firms design and implement Big Data Analytics in companies? What are the best practices and critical aspects for consultants in such projects?

In order to enter the black box of the implementation dynamics of Big Data Analytics technologies by consulting firms, the research proposes a case study focused on a single consulting company among the Big Four, which are the four largest professional services networks in the world by revenue. In particular, KPMG is the company in question, and the case study is limited to the Italian network.

The main source of information of the case study is represented by semi-structured interviews to Big Data consultants which are experts in this field and can provide a unique knowledge to better understand the implementation of these technologies in companies. In this regard, due to the complex functional structure of the company and the different career levels of consultants, there were used two criteria for the selection process, and as a result, three top expert consultants were chosen among all the Italian KPMG consultants. During the interviews a flexible approach was adopted, with the support of an interview guide to conduct the interviews but, at the same time, with the flexibility to deviate from it to gather as much relevant information as possible. Finally, the insights from the experts coupled with online document analysis led to a comprehensive overview of this theme.

Chapter 1. The Big Data Revolution

This chapter provides an overview of the academic literature that has already covered the Big Data Analytics topic and is structured as follows. The aim of the first sub-chapter is to explain what Big Data is and how it works. Thus, after a brief introduction on the relatively recent origins of the Big Data notion, the first two sections provide a detailed overview of several definitions found in the academic literature regarding the dimensions that differentiate Big Data from simply large datasets: from the 3Vs paradigm, in which each V represents a single dimension, additional dimensions are gradually proposed and explained. Then, the next section is focused on Big Data Analytics, underling the difference between the data itself (Big Data) and the tools and techniques (Analytics) that allow you to generate effective value from such data. Finally, the last section of the first sub-chapter gives an overview on the most popular Big Data tools used for different services.

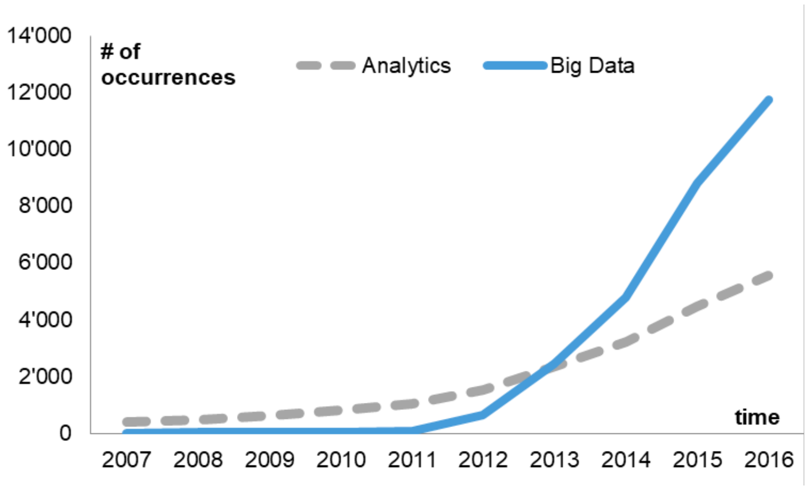
In the second sub-chapter the focus is on the impact of Big Data Analytics on business. After a brief introduction on the power with which Big Data Analytics are transforming the entire economy, the first section provides the several new opportunities that arise for the business as well as the barriers that organizations should overcome to be successful. The second section gives an overview on the fields of application in the business of such technologies which are changing the processes of decision-making within the companies and the competition over their competitors. Finally, the third section underlines the benefits for those companies which are able to translate Big Data into value and consequently gain a competitive advantage.

Finally, in the third sub-chapter some cultural challenges for organizations related to Big Data Analytics are provided. Indeed, since the technology itself it is not enough to extract effective value from Big Data, this sub-chapter investigates the other aspects involved in these processes such as the company culture and employee' skills that are also necessary to achieve the benefits that arise from these technologies.

1.1 Definitions of Big Data

As *Figure 1* below shows, Big Data is a relatively recent term that has met a dramatic rise in popularity of scholarly articles and on the web as of the beginning of the year 2011.

Figure 1: Occurrences of 'Big Data' and 'Analytics' in Academic Literature between 2007 and 2016



Source: De Mauro et al., 2019

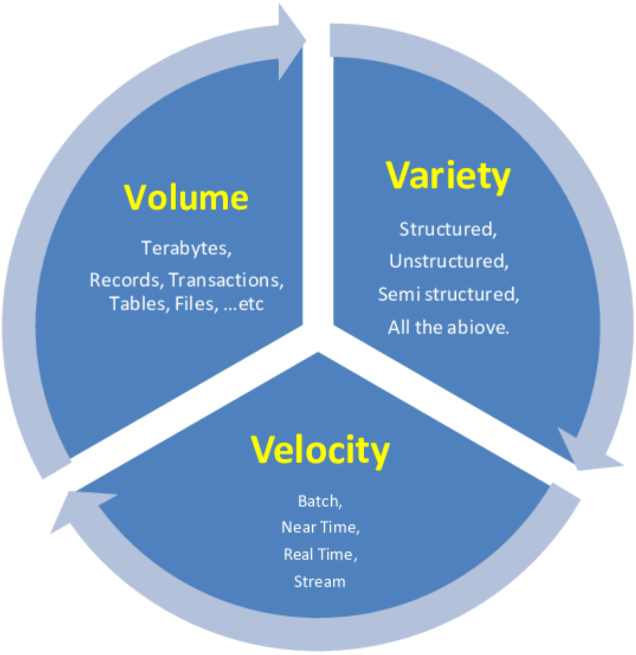
The notion of Big Data draws its origins from a multitude of more consolidated concepts, such as "Analytics", "Data Mining", "Business Intelligence" and it is progressively making its way as a new umbrella term that encompasses multiple aspects of Information Technology, Sociology, Business and Statistical Modeling (De Mauro et al., 2019). Indeed, as it can be observed, in the past decade the volume of data has immensely increased along with the development of data-generating, extracting, addressing, storage, and output technologies, driven by the consumers' use of social media tools, IoT devices, shopping apps/websites and online communities (Wang & Wang, 2020; Wamba et al., 2015). Lee (2017) states that Big Data represents a new technology paradigm for data that are generated at high velocity and high volume, and with high variety. Hence, Big Data can be simply seen as a huge amount of unstructured and fast-moving data.

After that, the academic literature provides some more precise definitions which are illustrated in the following sections.

1.1.1 The 3 Vs

First of all, a widely used definition of Big Data is in terms of 3Vs in which each V represents one dimension, as *Figure 2* below shows.

Figure 2: The 3 Vs



Source: Al-Barhamtoshy & Eassa, 2014

The first V refers to 'Volume', the large amount of data that either consume huge storage or entail of large number of record data. It has been noted that the world's data volume is anticipated to grow by 40% annually and expected to be 50 times by 2020 (Hajli et al., 2020). In this regard, Internet of things (IoT) contributed significantly on the explosive growth in volume. For instance, having access to large amounts of relevant data about consumers' behavior helps firms identify the products that could meet future market needs (Ghasemaghaei & Calic, 2020). Although this dimension is a primary distinguishing characteristic of Big Data, some firms possess massive data sets that lack the other characteristics of Big Data (Erevelles et al., 2015).

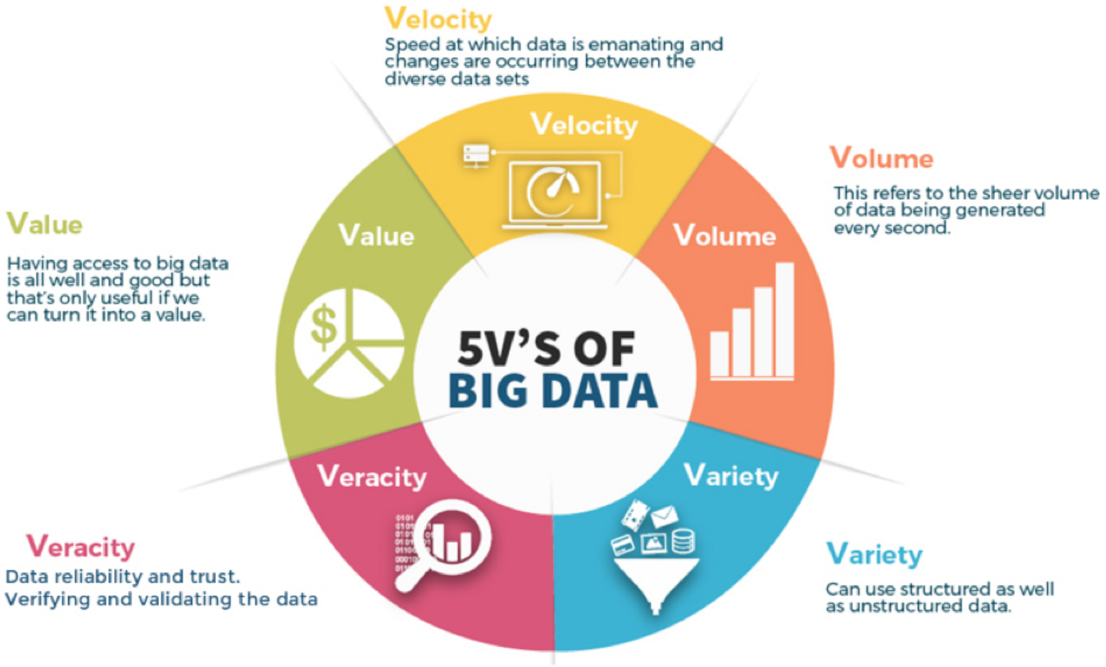
The second V refers to 'Velocity', which is the frequency or the speed of data generation and/or frequency of data delivery. In fact, for many applications, the speed of data creation is even more important than the volume. Indeed, in their study, Ghasemaghaei & Calic (2020) show that data velocity plays a more important role than other Big Data characteristics in enhancing firm innovation performance. Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors (McAfee & Brynjolfsson, 2012). For example, IoT devices produce large amounts of data in real-time, which allows managers to make rapid changes to production processes or services based on market loads or expectations (Sestino et al., 2020). So, to make appropriate decisions, firms must develop continuous processes for analyzing and interpreting data in real time to quickly generate new insights.

Finally, the third V refers to ‘Variety’, to highlight the fact that data are generated from a large variety of sources and formats, and contain multidimensional data fields including structured and unstructured data (McAfee & Brynjolfsson, 2012). Indeed, Big Data includes information from sources such as social networks, tweets, blogs and cellphones (Amankwah-Amoah, 2016; Liu et al., 2019). In fact, a major difference between contemporary Big Data and traditional data is the shift from structured transactional data to unstructured behavioral data (Ghasemaghaei & Calic, 2020). In particular, unstructured data include textual data (e.g., from blogs and text messages) and non-textual data (e.g., from videos, images, and audio recordings). In fact, handling both structured and unstructured data helps firms view innovation problems from different perspectives. Much unstructured data are captured through social media, where individuals share personal and behavioral information with friends and family (Erevelles et al., 2015). For the sake of achieving sustainable competitive advantage, a company is expected to combine all the data from different sources (Hajli et al., 2020). For example, the Big Data sources are extremely important for marketing purposes, since they include customers’ posts on social media that enable the companies to know their customers’ preferences in real time considering their profile features and locations; this information enables the companies to know, for example, when a customer has entered a store or a website (Raguseo, 2018). Also, firms that extract consumers’ comments about their products on social media websites and combine these with consumers’ purchasing histories can better identify consumers’ preferences, which may help them develop new products that match their needs (Ghasemaghaei & Calic, 2020).

1.1.2 Additional Dimensions

Despite the model of the three Vs is considered as the main definition in order to distinguish the Big Data from simply large-scale datasets, many academics mention also other dimensions which are listed below. *Figure 3* below shows that there are also other two dimensions (Value and Veracity), in addition to the previous three (Velocity, Volume and Variety), that make up the definition of 5Vs.

Figure 3: The 5 Vs


Velocity
Speed at which data is emanating and changes are occurring between the diverse data sets
Volume
This refers to the sheer volume of data being generated every second.
Value
Having access to big data is all well and good but that's only useful if we can turn it into a value.
Veracity
Data reliability and trust.
Verifying and validating the data
Variety
Can use structured as well as unstructured data.
Velocity
Value
Volume
5V'S OF
BIG DATA
Veracity
Variety

Source: Khalid & Rachid, 2019

The fourth V refers to 'Value', because the task is to eliminate unimportant and irrelevant data, so that the remaining data are useful. Indeed, it is noteworthy that the use of even most sophisticated analytical system is meaningless if inappropriate data is in place or poor-quality data is used (Wamba et al., 2015). Though data aggregation technology is improving and reducing the transmission size for the companies to get meaningful data, still many companies struggle to keep that data in a more structured way. Even the recent data tools are very limited and insufficient that raise many complications (Hajli et al., 2020).

Then, the fifth V refers to 'Veracity', to highlight the importance of quality data and the level of trust in various data sources. Indeed, not all Big Data about consumers is accurate. Uncertainty and unreliability arise due to incompleteness, inaccuracy, latency, inconsistency, subjectivity, and deception in data. Statistical tools and techniques have been developed to deal with uncertainty and unreliability of Big Data with specified confidence levels or intervals (Grover & Kar, 2017; Lee, 2017). In addition, the veracity of Big Data is a major issue at a time where the volume, velocity, and variety of data are constantly increasing (Erevelles et al., 2015).

Moreover, another one V refers to 'Variability', meaning the variation in data flow rates, because in addition to the increasing velocity and variety of data, data flows can fluctuate with unpredictable peaks and troughs (Lee, 2017).

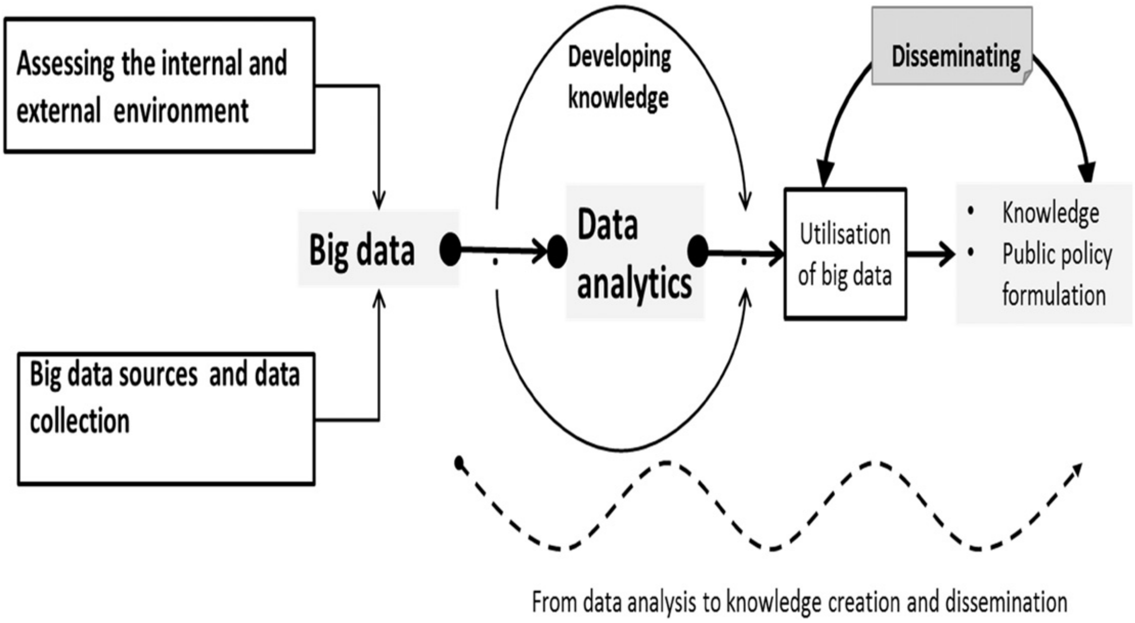
Also, another one dimension refers to 'Complexity', in particular the complexity related to the number of data sources. Since Big Data are collected from numerous data sources, complexity makes it difficult to collect, cleanse, store, and process heterogeneous data. So, it is necessary to reduce the complexity with open sources, standard platforms, and real-time processing of streaming data (Lee, 2017).

Finally, Lee (2017) proposes an interesting additional dimension: 'Decay' which refers to the declining value of data over time. He explains that in a time of high velocity, the timely processing and acting on analysis is all the more important. IoT devices generate high volumes of streaming data, and immediate processing is often required for time-critical situations such as patient monitoring and environmental safety monitoring. Wearable medical devices such as glucose monitors, pulse oximeters, and blood pressure monitors worn on or close to the body produce a stream of data on patients' physiological conditions. Thus, in the "Era of Big Data", the decay of data will be an exponential function of time.

1.1.3 Big Data Analytics

The Big Data concept refers to the data itself and the dimensions it carries, while Big Data Analytics refers to the data itself, as well as the technological infrastructure that enables the data to be obtained, the software that provides inferential inferences and the tools that provide the analysis (Hazirbaba & Yalcintas, 2019). In other words, Big Data Analytics is related to the field of business intelligence and analytics, in which companies try to make sense of huge piles of data. *Figure 4* below illustrates this process.

Figure 4: From Big Data to knowledge dissemination



Assessing the internal and external environment
Big data sources and data collection
Big data
Developing knowledge
Data analytics
Disseminating
Utilisation of big data
• Knowledge
• Public policy formulation
From data analysis to knowledge creation and dissemination

Source: Amankwah-Amoah, 2016

Various technological advances have led to the rise of tools that enable companies to make sense of data by filtering, correlating and reporting insights. Some of the traditional marketing analytics techniques include regression modelling, mapping/multidimensional scaling, diffusion modelling, stochastic processes, math programming modelling and optimization that turned out to be very useful for the companies (Hajli et al., 2020).

The literature provides also definitions of Big Data Analytics as new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high velocity capture, discovery and/or analysis to support decision making and action taking (Mikalef et al., 2019; Wang & Wang, 2020). Hence, Big Data Analytics refers to all the tools and techniques that analyze the vast and varied business data to generate meaningful insights for decision making. In fact, the complex nature of Big Data makes it hard to handle and decode (Gupta et al., 2020), and that's why the effective functionality of such data comes only from the analysis and consequential insights for decision making (Gupta et al., 2019). In this regard, Erevelles et al. (2015) state that the physical, human, and organizational capital resources moderate the following three processes: the process of collecting and storing records of consumer activities as Big Data, the process of extracting insights from Big Data, and the process of utilizing insights to enhance dynamic/adaptive capability. In a similar way, Gunasekaran et al. (2017) conceptualizes Big Data predictive analytics assimilation as a threefold process involving acceptance, routinization, and assimilation.

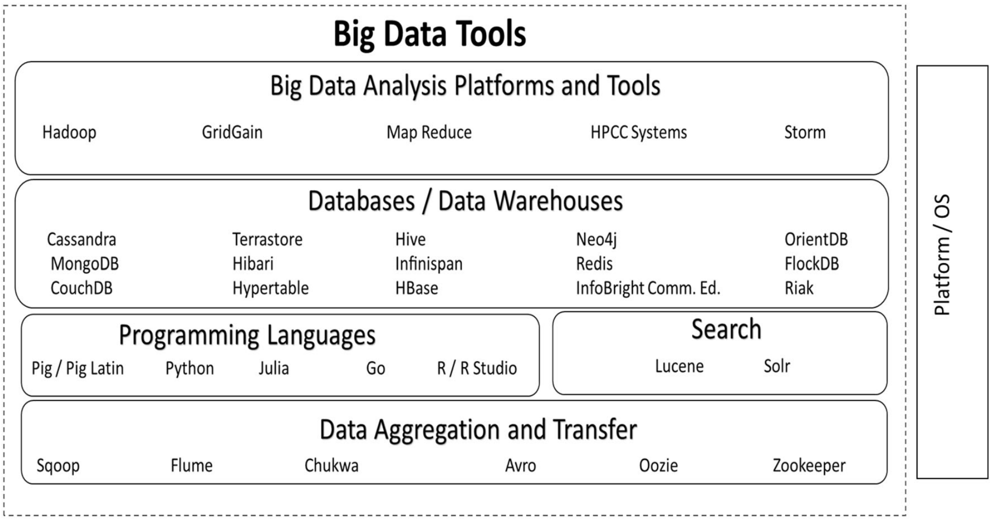
Big Data Analytics is also defined as the process of analyzing unstructured data set to provide undiscovered inferences, capture inter-data correlations and other useful information (Hazirbaba & Yalcintas, 2019). In his study, Raguseo (2018) demonstrate that the most diffused technologies are visual analytics technologies, scripting languages, and in-memory analytics software. Instead, the Big Data sources that companies use the most are online portal contents, POS data and smart meter data. This result is interesting because, unlike many others, highlights that companies are still more likely to use Big Data that are proprietary than to buy other data on customers from third parties, such as those that can be produced by the social media.

After all, Wamba et al. (2015) suggest thinking about Big Data not only in terms of Analytics, but more in terms developing high-level skills that allow the use of new generation of IT tools and architectures to collect data from various sources, store, organize, extract, analyze, generate valuable insights, and share them with key firm stakeholders for competitive advantage co-creation and realization.

1.1.4 Big Data Tools

Grover & Kar (2017) give the overview of the Big Data technologies for various services. *Figure 5* shows that the five popular Big Data analysis platforms and tools are Hadoop, GridGain, MapReduce, HPCC systems and Storm. Then, there are a lot of databases and data warehouses for the Big Data storage such as Cassandra, MongoDB, CouchDB, Terrastore, Hibari, Hypertable, Hive, Infinispan, HBase, Neo4j, OrientDB, FlockDB, Riak, Infobright Community Edition, Redis. Also, the programming languages for the Big Data are Pig, Python, Julia, Go and R. Among the Big Data searching tools there are Lucene and Solr. Finally, among Data aggregation and transfer tools help in transporting the data from one system to another and summarizing the data there are Sqoop, Flume, Chukwa, Avro, Oozie, Zookeeper.

Figure 5: An overview of various Big Data Tools



Source: Grover & Kar, 2017

In particular, Hadoop is an open-source framework that enables the distributed processing of data by leveraging clusters of dispersed machines and special computer programming models. The main components of Hadoop are its distributed file system, HDFS, which allows access to data disseminated over multiple machines without having to deal with the complexity inherent to their dispersed nature; and MapReduce, a programming model designed to efficiently implement distributed and parallel algorithms (Lee, 2017; De Mauro et al., 2019). Then, Cloud platforms let the system designer delegate technological complexity to external resources. By using this approach some crucial elements of a Big Data system such as storage and computational power for data processing become services that can be simply bought on the market (De Mauro et al., 2019).

1.2 Impact of Big Data Analytics on Business

Big Data growing bigger every second, every day, particularly driven by the consumers' use of social media tools, IoT devices, shopping apps/websites, and online communities (Wang & Wang, 2020). That’s why, according to McAfee & Brynjolfsson (2012), the Big Data revolution is a fundamental transformation of the economy and almost no sphere of business activity will remain untouched by this movement.

Thus, according to Raguseo (2018) the benefits associated with technological investments can be classified in four typologies: strategic, informational, transactional and transformational. The strategic benefits are those that can alter the way companies compete or the nature of their products. The informational benefits are those that provide information and communication that can be used to improve decision making in a company. The transactional benefits refer to investments that support operational management and which are able to cut the costs sustained by companies. Finally, transformational benefits refer to the results of changes that a firm has to make to the structure and to the capacity of implementing a technological investment.

After that, the academic literature has dealt in more detail with the opportunities and barriers for organizations, the fields of application of such technologies and the benefits for those companies that are able to exploit Big Data, which are illustrated in the following sections.

1.2.1 Opportunities and Barriers

The trend of increasing data volume has brought in a great number of potential opportunities (Raguseo, 2018; Wang & Wang, 2020; Gupta et al., 2019). Indeed, Lee (2017) explain that Big Data provides great potential for firms in creating new businesses, developing new products and services, and improving business operations. Since there are a large amount of available data and advanced technologies to process them, firms can quickly exploit new information to create and implement new ideas (Ghasemaghaei & Calic, 2020). Moreover, effective use of data analysis tools directly lead to new product success and has a direct effect on customer agility (Hajli et al., 2020). For instance, according to Erevelles et al. (2015), the magnitude of the data generated, the relentless rapidity at which data are constantly generated, and the diverse richness of the data are transforming marketing decision. In fact, one of the most important advantages that Big Data Analytics provide to companies is facilitate better informed decision-making (McAfee & Brynjolfsson, 2012; Gupta et al., 2019).

But the implementation of data-driven decision-making into companies is not always too simple, for example, organizational inertia may occur: the negative psychology and the fear of losing authority could erect, especially in larger corporations, barriers for cooperation with analytics departments (Mikalef et al., 2019). Then, companies often mistakenly make substantial investments to acquire data before investing in technology and without acquiring and retaining the right human capital (Gupta et al., 2020). About this, De Mauro et al. (2019) explain that data scientists alone are not sufficient in enabling organizations to have a real competitive advantage using Big Data. Multiple role families have been identified as related to an effective exploitation of Big Data, namely: business analysts, data scientists, Big Data developers, and Big Data engineers. Another factor contributing to the failure of IT systems in organizations is that organizations often become biased while choosing investment options and end up investing considerable resources in technological infrastructure and giving less importance to acquiring technical skills (Gupta et al., 2020). Then, another very important factor, highlighted by Mikalef et al. (2019) is that there are path dependencies that act as rigidities when considering the fusion of Big Bata Analytics into corporate strategy.

Moreover, companies have to take into account privacy and security issues before using Big Data technologies (Raguseo, 2018). People are increasingly concerned about how companies use their personal data, and it falls on companies to take actions that stimulate the kind of trust that fosters loyalty. For this reason, decision-makers need to involve users when developing digital ethics practices (Sestino et al., 2020). It is important that firms build an image as a trustworthy entity for their customers to consent to provide data and allow them to leverage this data appropriately and within what they believe is an ethically correct approach (Mikalef et al., 2019). Therefore, firms and customers need to strike a balance between the use of personal data for services and privacy concerns. It is noted that there is no one-size-fits-all measure for privacy, but the balance depends on service type, customers served, data type, and regulatory environments (Lee, 2017).

1.2.2 Fields of Application

In the past, although not so long ago, data analysis was based on historical records or customers surveys. Due to the path-dependent nature of cause-and-effect relationships, historical data have limited usefulness in illuminating the current and future causal structure of choices that determine firm success, particularly for choices concerned with generating and implementing entirely new ways of doing things (Ghasemaghaei & Calic, 2020). Now, the so-called Big Data revolution, potentially, will lead to entirely new ways of understanding consumer behavior and formulating marketing strategy (Erevelles et al., 2015; Hajli et al., 2020). By strategically collecting and interpreting Big Data in the context of existing businesses systems, companies can achieve a more realistic and useful vision in the decision-making process (Sestino et al., 2020). In other words, Big Data Analytics is perceived to be a facilitator for decision making (Gupta et al., 2019; Mikalef et al., 2019; Gupta et al., 2020). Indeed, Big Data enables managers to decide on the basis of evidence rather than intuition. That's why basically it has the potential to revolutionize management (McAfee & Brynjolfsson, 2012; Wamba et al., 2015).

Once shopping moved online, though, the understanding of customers increased dramatically. Online retailers could track not only what customers bought, but also what else they looked at; how they navigated through the site; how much they were influenced by promotions, reviews, and page layouts; and similarities across individuals and groups (McAfee & Brynjolfsson, 2012). For instance, companies such as Netflix, Facebook, and Google collect huge customer data, which is the most valuable asset of their business status. From the past search and purchase records, they are able to provide attractive advertisements to the customers. At the same time, those data are intelligently used by the marketing content providers to decide when, where, and who should be targeted for the promotion (Wang & Wang, 2020). In fact, Netflix collects enormous amounts of data and analyzes customer watching habits not only to generate personalized recommendations and offerings but also use this data as a basis to produce their own series (Boston Consulting Group & University of Virginia, 2017; Ghasemaghaei & Calic, 2020). Big Data have also a significative impact on Luxury's industry, in this regard, Liu et al. (2019) explain that Luxury brand managers may benefit from utilizing Big Data to obtain more accurate understanding of customer engagement on social media and consequently formulate more effective customer engagement strategies. In this context, social media is an important source of up-to-date brand information because customers consider it to be a more trustworthy source of information than traditional instruments of marketing communications such as press releases or advertising. That's why platforms such as Facebook and Twitter are able to record the interaction of users, as well as their habits and interests (De Mauro et al., 2019).

The public sector is another fertile terrain for Big Data. Governments can counteract their highly variable performance by leveraging the huge amount of transactional and census data collected from citizens. Transparency, in particular, can be a critical enabler of improved efficiency and productivity for public institutions. By opening and sharing data with citizens, governments can withstand the lack of competitive pressure in the sector and restrain public expenditure, avoid frauds and increase citizen's sense of ownership towards common good assets and funds (De Mauro et al., 2019).

Also, some years ago Amazon filed a patent for anticipatory shipping, in which the company uses Big Data, including order history, product search history, and shopping cart activities, to predict when a customer will make a purchase and begins shipping the product to the nearest hub before the customer submits the order online (Erevelles et al., 2015). Even modern scientists adopt Big Data technologies and methods in order to manipulate empirical data in a view of yielding research answers. One of the most notorious examples of such implementation can be found at CERN, the European Organization for Nuclear Research, based in Geneva, Switzerland. By storing and processing an extensive amount of data researchers were able to contribute to our understanding of the origin of the mass of subatomic particles, including the proof of the existence of the Higgs boson in 2012 (De Mauro et al., 2019). Another one field of application is for those firms operating in an extremely compound environment such as emergency services. For example, accessing the accurate information can have huge impact on 'when' and 'how' to evacuate the population of a given region during a flood or a bushfire. Indeed, the costs of making an erroneous decision can have significant implications at the management and political levels (Wamba et al., 2015). Thus, businesses such as Honda, Walmart, Samsung, and several others of varying scales in various sectors have significant activities that are amenable to Big Data technologies (Hajli et al., 2020). Overall, it seems that firms that do not develop the resources and capabilities to effectively use Big Data will be challenged to develop sustainable competitive advantage and to survive the Big Data revolution (Erevelles et al., 2015; De Mauro et al., 2019). The evidence is clear: data-driven decisions tend to be better decisions. Leaders will either embrace this fact or be replaced by others who do (McAfee & Brynjolfsson, 2012).

1.2.3 Benefits of Exploiting Big Data

Gupta et al. (2019), identify the various ways in which Big Data Analytics can be utilized to facilitate the adoption of circular economy paradigm. From their study emerge that Big Data functionalities can be utilized to generate insights for integrating processes and sharing resources. Moreover, they reduce uncertainties and can have a positive impact in decision areas such as daily production and maintenance variability; manpower performance; health, safety and environment; and critical raw material availability status.

Then, dynamic pricing enables an organization to implement a flexible pricing strategy based on changing consumer demand (Erevelles et al., 2015). Actually, McKinsey & Company reported that they have seen companies in industries as diverse as software, chemicals, construction materials, and telecommunications achieve impressive results by using big data to inform better pricing decisions (Baker et al., 2014). For example, eBay uses open-source Hadoop technology and data analytics to optimize prices and customer satisfaction. To achieve the highest price possible for items sellers place for auction, eBay examines all data related to items sold before (e.g., a relationship between video quality of auction items and bidding prices) and suggests ways to maximize results to sellers (Lee, 2017).

Companies are also using new technological solutions to understand their own operations and behavior at a much finer level of detail (Raguseo, 2018). In this regard, Big Data predictive Analytics can assist in reducing supply chain costs and achieving efficiency, responding faster to changing environment, providing more power in supplier relationships with suppliers and enhancing sales and operations planning capabilities (Gunasekaran et al., 2017). Capable Big Data predictive analysis techniques also equip organizations to efficiently manage the inflow of data so that it can accurately predict market requirements. This means that organizations can align their business processes and business strategies to cater to the current market needs at their highest potential and help gain a better understanding of market aspirations to deal with future demands (Gupta et al., 2020). Indeed, customer needs to keep on changing with the passage of time. It is of utmost importance for the company to sense and respond the customer needs accordingly. Data through its unique analytical and predictive capabilities can enable the companies to sense and respond to customer needs in an appropriate manner (Hajli et al., 2020).

Companies could also get benefits from the data generated from the trading partners in upstream, downstream, and horizontal collaborators to look into the opportunities. These virtually linked entities, while some could be partially owned by the target organization, can potentially share databases and information resources and to make mutually agreed decisions with the output of Big Data Analytics (Wang & Wang, 2020). For example, the data produced by Fitbit products, they enable the health of people to be improved by tracking their activity, exercise, food, weight and sleep, and these data at the same time can be sold to insurance companies in order to allow them to understand the profiles of different people and provide different insurance packages according to their profile. In this way, thanks to Big Data, new business opportunities can arise between two different companies that operate in different industrial sectors (Raguseo, 2018).

Furthermore, Big Data can be also useful in the field of business strategy. For instance, the differentiation strategy requires to offer different solutions that meet customers' needs and expectations. Thus, Big Data Analytics is an effective tool to contribute to the differentiation strategy (Hazirbaba & Yalcintas, 2019). In fact, the strategic benefits exploiting Big Data are those that can alter the way companies compete or the nature of their products (Raguseo, 2018). In order to capture value from Big Data the literature suggests technology roadmaps as a tool for strategic planning which entails linking resources and expertise to future courses of action. It can be seen as a unique framework for mapping processes and approaches towards achieving particular short- and long-term objectives. Consequently, roadmapping can equip decision makers with the tools and approaches to make better investment decisions (Amankwah-Amoah, 2016).

1.3 Data-Driven Culture

Wang & Wang (2020) recognize that Big Data Analytics could be a key role to achieve firm growth and trigger innovation in today's business environment. Companies can leverage consumers' real-time data flows to continually improve their products and marketing campaigns (Sestino et al., 2020). Moreover, Gunasekaran et al. (2017) highlighted that Big Data predictive Analytics assimilation is positively related to a firm's supply chain performance, as well as organizational performance.

Meanwhile, Big Data also faces challenges. The main challenge of obtaining strategic value from Big Data consists in the difficulty of creating an integrated Big Data infrastructure which supports the agile development of customer analytics in such a way that spending in worthless data or not spending enough is avoidable (Hajli et al., 2020). Often organizations do not place enough importance on human interpretative skills and rely primarily on machine output. This stands out as one of the significant factors for failure when deploying Big Data Analytics (Gupta et al., 2020). Indeed, Big Data's power does not erase the need for vision or human insight (McAfee & Brynjolfsson, 2012). In order to reap the full benefits from Big Data, managers need to align existing organizational culture and capabilities across the organization (Wamba et al., 2015). Indeed, strategic alignment is regarded as a major success in organizational design literature. The strategy directly affects the technology itself and other dimensions such as organizational structure, process, organizational culture and human resources (Hazirbaba & Yalcintas, 2019). So, Big Data Analytics should not be perceived as a solely technical challenge, but rather, an organizational one which requires fusion with the firm's business strategy (Mikalef et al., 2019).

In addition, (Jacques Bughin, 2011) shows that consumers capture a large part of the economic surplus that Big Data generates: lower prices, a better alignment of products with consumer needs, and lifestyle improvements that range from better health to more fluid social interactions.

Then, Mikalef et al. (2019) find that more technological and technical resources contribute towards performance gains in moderately uncertain environments, while organizational aspects and managerial skills are of greater importance in highly uncertain conditions. Ghasemaghaei & Calic (2020) suggest also that in order to implement new ideas successfully, firms should pay attention to the speed of processing and analyzing different types of data, rather than focusing primarily on collecting huge amounts of data.

Overall, in the regard of gain a competitive advantage, technology may not be a differentiating element of performance but rather a commodity, instead, strong structural and procedural practices, as well as a firm-wide data-driven culture are critical components (Mikalef et al., 2019). Indeed, technology is of no use if the manager is not able to extract insight and take strategic decisions by using his/her intellectual skills (Gupta et al., 2020). Thus, the ability to translate the data into insights, knowledge and value it can be considered as a new type of organizational capability (Wamba et al., 2015).

Chapter 2. Methodology and Approach

This chapter explains the methodology and approach used in this research and it is structured as follows. In the first sub-chapter are explained the reasons that led to the choice of a qualitative research method, consisting in a single case study methodology, to answer the research questions. Then, the sources of information used to develop the case study are explained, that are the semi-structured interviews supplemented by online documents analysis.

Consequently, the second sub-chapter is focused on the selection of the case study. In this regard, after a brief introduction on the Big Four networks that stand out over the other consulting firms, the reasons that conducted to the choice of a single company among the Big Four are first provided and then, given the enormous geographical dispersion, the considerations that led to the choice of a single nation, with its limit, are highlighted.

Then, the third sub-chapter explains the criteria used for the selection of experts to interview. In particular, due to the complex functional structure of KPMG and the different career levels of consultants, two criteria based on experience level and area of competence were formulated and used in the selection process. As a result of the selection, three top expert consultants were chosen among all the Italian KPMG consultants. Despite that the names of consultants are hidden, the relevance of their positions and their backgrounds are exposed to justify the choices.

Finally, the design of the interviews is provided in the fourth sub-chapter. A flexible approach was adopted during the interviews, with the support of an interview guide to lead the interviews. The interview guide is conceptually divided into three section, with two open questions for each section. In this regard, a detailed overview of the topics covered in each section is thorough.

2.1 Case Study Methodology

Considering the research questions and consequently the exploratory nature of this research, it is clear the need of a qualitative research method. Indeed, it is impossible to answer at these questions through quantitative analysis because it is not possible to find such information in any database. Among the qualitative methods, the case study provides descriptive details about how our workplaces function, and can increase understanding of a particular phenomenon (Brown, 2008). Indeed, the in-depth focus on the particular within a bounded system can help provide a holistic view of a situation (Brown, 2008). What makes a case study a case study is the unit of analysis; that is, a case study is an in-depth description and analysis of a bounded system (Merriam, 2010). Therefore, taking into account the previous consideration, the development of a single case study focused on a particular consulting firm was chosen in order to enter into the black box of the implementation dynamics of Big Data Analytics technologies.

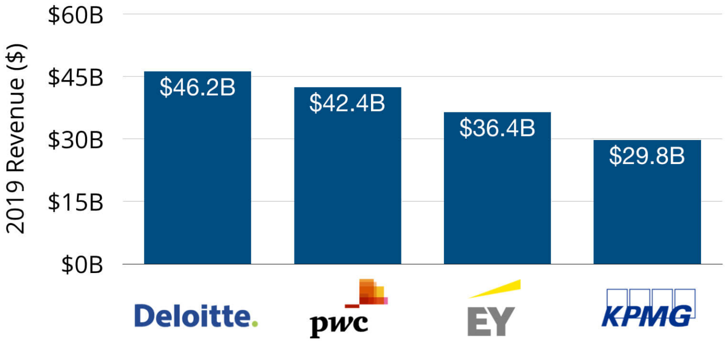
Then, according to Merriam (2010), interviews are the most common source of data in case study research. In fact, since the case study is a research strategy that focuses on understanding dynamics in a particular environment, it is recommended to use interviews as a source of information due to the explanatory nature of the method (Yin, 2003). Therefore, the principal source of information used in the case study analyzed in this research was based on a semi-structured interview approach, where the interviewer followed a guide but at the same time, he has the flexibility to deviate from it in order to catch more information as possible. Despite that using interviews is a time-consuming methodology, for example compared to questionnaires, through this approach it is possible to conduct an in-depth analysis on the causes and facts under study (Yin, 2003). Moreover, data can also be collected online. Indeed, web pages, papers available online, etc. can be considered documents (Merriam, 2010). Therefore, as a second source of information, several online searches were conducted to supplement the information extracted from each interview. In fact, according to Merriam (2010), simultaneous data collection and analysis allow the researcher to make adjustments to collect the best data.

Overall, there are many reasons that support the choice of doing interviews to consultants in order to answer the research questions: since this topic is almost new, the data needed is not available in any other source of information, and Big Data consultants could provide a unique knowledge to better understand the implementation of these technologies in companies. Finally, the insights from the experts coupled with online document analysis leads to an inclusive and thorough understanding of this topic.

2.2 KPMG Case Study

Around the world there are many consulting firms but there are four of them that stand out over the others, the so called Big Four. The Big Four is the nickname used to refer collectively to the four largest professional services networks in the world by revenue: Deloitte, Ernst & Young, KPMG and PricewaterhouseCoopers. *Figure 6* below shows the aggregate global revenue of the Big Four.

Figure 6: Big Four Revenue



Source: Hacking the Case Interview, 2021

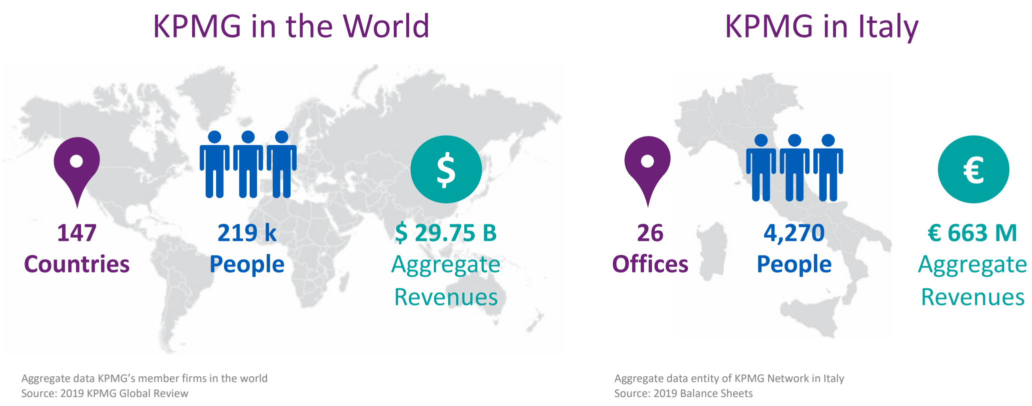
Actually, none of the "firms" within the Big Four is actually a single firm; indeed, they are professional services networks. Each is a network of firms, owned and managed independently, which have entered into agreements with the other member firms in the network to share a common name, brand, intellectual property, and quality standards. Each network has established a global entity to co-ordinate the activities of the network.

The four networks are often grouped together for several reasons: they are each comparable in size relative to the rest of the market, both in terms of revenue and workforce; they are each considered equal in their ability to provide a wide scope of professional services to their clients; and they are considered equally attractive networks to work in, because of the frequency with which these firms engage with Fortune 500 companies. Therefore, since there are several similitudes among the Big Four, this research is developed through a case study focused only on a single company among the Big Four, that is KPMG. However, due to the reasons listed previously, many results can also be extended to other consulting firms.

Present in almost 150 countries around the world, with nearly 219,000 professionals, KPMG is a multinational business company offering professional service to its clients in three specialized fields: Audit (40%), Advisory (38%), Tax (22%). The acronym "KPMG" stands for the initials of its founders "Klynveld Peat Marwick Goerdeler", after the last merger in the year 1987.

However, due to the multitude of countries in which KPMG is present, in this research the focus is only on a single country, that is Italy. Again, taking into account the structure, vision and values that are shared in each country, many findings can be extended to other countries as well.

Figure 7: KPMG in the World vs KPMG in Italy



Aggregate data KPMG's member firms in the world
Source: 2019 KPMG Global Review

Aggregate data entity of KPMG Network in Italy
Source: 2019 Balance Sheets

Source: the author, based on the official website of KPMG Italy[1]

As the *Figure 7* shows, with more than 4,000 professionals, 26 offices, 6,000 customers and a complete portfolio of services that meets the needs of the national and international market, KPMG is one of the most important professional services networks active in Italy. Indeed, thanks to a federal and integrated operating model, KPMG's Italian network can count on the ability to mobilize the thinking and skills available on a global scale in real time, while being able to operate in full strategic and managerial autonomy on the national market.

[1] https://home.kpmg/it/it/home.html

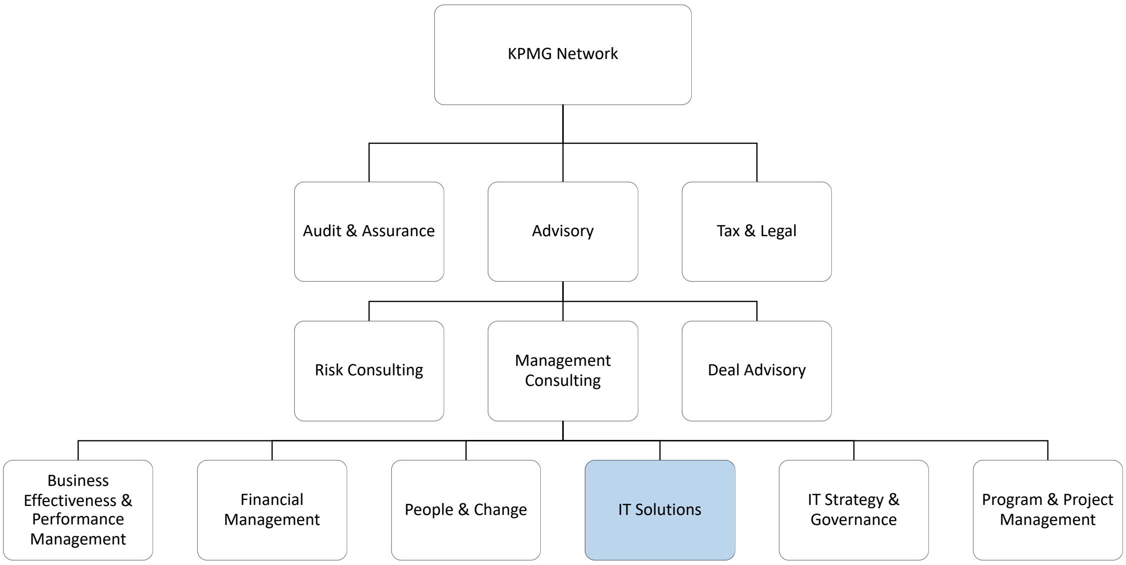
2.3 Selection of Consultants

In order to answer the research questions, there is a need of consultants that have a strong expertise in Big Data Analytics design and implementation. In this regard, it would be perfect if such consultants have at least ten-year's experience in this field, with a significative background on different companies and industries. Taking into account these considerations, the consultants are selected on the basis of the following two criteria:

- Area of Competence
- Experience Level

As regard the first criteria, since the KPMG Network provide a lot of different services, it is necessary to focus only on a specific area of competence related to Big Data Analytics. In this regard, *Figure 8* shows the complex structure of KPMG: the network is composed by independent entities affiliated with KPMG International Limited, the part more related to consulting is provided by "KPMG Advisory S.p.A." which have inside three departments: "Risk Consulting", "Deal Advisory" and "Management Consulting". For the scope of this research, the more relevant department is that of "Management Consulting" which is in turn divided into six Areas: "Business Effectiveness & Performance Management", "Financial Management", "People & Change", "IT Solutions", "IT Strategy & Governance", "Program & Project Management".

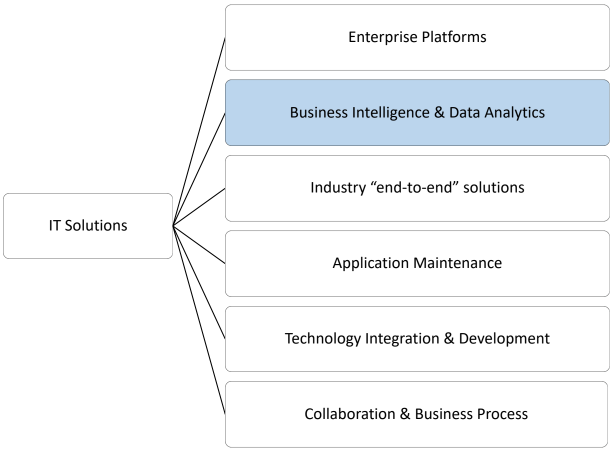
Figure 8: The KPMG's Structure



Source: the author, based on the official website of KPMG Italy

Therefore, within the "Management Consulting" department, the most relevant area is certainly that of "IT Solutions". Then, *Figure 9* shows that inside the IT Solution Area there are also several LoS: "Enterprise Platforms", "Business Intelligence & Data Analytics", "Industry end-to-end solutions", "Application Maintenance", "Technology Integration & Development", "Collaboration & Business Process".

Figure 9: Inside the IT Solution Area

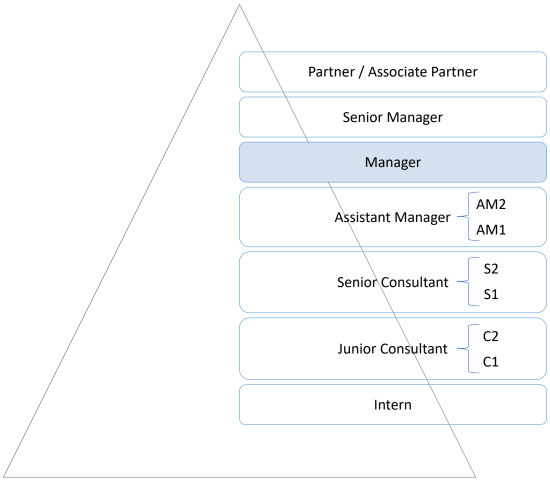

IT Solutions
Enterprise Platforms
Business Intelligence & Data Analytics
Industry "end-to-end" solutions
Application Maintenance
Technology Integration & Development
Collaboration & Business Process

Source: the author, based on the official website of KPMG Italy

Thus, the most appropriate profile to interview is those of consultants which work in the "Management Consulting" department, in the "IT Solutions" area and finally in the "Business Intelligence & Data Analytics" Line of Service.

As regard the second criteria, *Figure 10* shows that for each area of competence there are different levels of consultants, generally based on the years of experience.

Figure 10: KPMG Career Hierarchy



Source: the author, based on the official website of KPMG Italy

Thus, the most appropriate profile to interview is those of consultants that have the “Manager” level because this position ensure that they have a significative experience in their field. In fact, it is difficult to find people who have a deeper knowledge of such technologies.

Therefore, the consultant selection process was conducted considering these two criteria and finally the three top experts in this field were selected among all the KPMG’s Italian consultants. *Table 1* below explains the choice of these consultants by providing their positions and summarizing their backgrounds. To avoid any ethical issues, the names of the consultants are hidden.

Table 1: Profile of Respondents

N°	Position	Background
1	Manager BI and D&A Rome, Italy	The first consultant has more than 15 years of experience in Management and Digital Consulting for major global organizations. As a PM, he has been responsible for many projects in the fields of Business Intelligence, Business Analytics, Big Data and Advanced Analytics. During his career he has acquired more than 10 certifications per year and has experienced a significant growth in projects number. He has experience in several sectors such as Energy & Utilities, PA, Telco & Media, Manufacturing; both national and international contexts.
2	Manager BI and D&A Rome, Italy	The second consultant has gained experience in the design and development of 360° Business Intelligence applications in the Big Data & IoT sector, in the Enterprise Data Warehouse, both on-premises and cloud, Technical Team Leader in the Data Engineer field, design of solutions in the Data Governance field and Project Management activities. The skills acquired during the work experience are development, definition of enterprise architectures in traditional BI / Big Data & IoT, Data Governance, analysis of functional requirements, Project Management, PMO activities.
3	Manager BI and D&A Rome, Italy	The third consultant got his PhD from Imperial College London and worked as a researcher at the Sapienza University of Rome. He started his consulting career focusing on Data and Statistical analysis carried out by scripts and coding in Python and R. He was a senior consultant on data analysis with a strong focus on Big Data and Machine Learning and became an Analytics area manager. He has held the role of coordinator in the Analytics sector, spanning from Big Data to ML with strong focus on high level Math’s modelling.

Source: the author

2.4 Interviews Design

All the interviews were carried out in Italian on the Microsoft Teams communication platform, subsequently they were transcribed and then translated in English. Finally, the most relevant parts have been extracted and presented in the next chapter.

The interview guide was used to drive the interviews with consultants, but a flexible approach was adopted, leaving respondents the freedom to wander among the topics as they wish and without limiting the addition of further relevant information. The interview guide consists of six open questions, conceptually divided into three sections, with two questions for each section.

The first section aims to introduce the interviews and understand if the knowledge resulting from the literature review actually corresponds to the experience gained by these experts during the many projects of their careers. The main topics covered in this section are related to the fields of application of Big Data Analytics, the compliance with the definitions of Big Data and consequently the data sources, opportunities and barriers as well as the challenges of such technologies. *Table 2* below shows the questions of the first section.

Table 2: Questions in the First Section

N°	Question
1	What are the most relevant projects you have worked on concerning Big Data Analytics? For which companies and for what purposes?
2	How were these projects implemented? In terms of compliance with 3Vs, data sources, activities performed, difficulties encountered, customer requirements...

Source: the author

The second section is more focused on the technical aspects of these projects: from the layers of infrastructure commonly used for these technologies, to the differences among the many platforms and tools, the main programming languages and personal skills needed, the critical parts of the current technologies, the minimum requirements in terms of knowledge for those consultants interested in entering in this field. *Table 3* below shows the questions of the second section.

Table 3: Questions in the Second Section

N°	Question
3	How are these infrastructures made? How many layers are there? Where are the critical parts?
4	What platforms and tools are used for each layer of the infrastructure?

Source: the author

The third section is the most open because it consists of personal questions that leave respondents the freedom to add whatever they deem relevant to this topic. The main objective of this section, in fact, is to get some insights from the experts on the critical issues they encountered during the many projects of their career and on how they managed them. Finally, consultants were asked for personal opinions regarding the future directions of Big Data Analytics both in term of technology developments and customer needs. *Table 4* below shows the questions of the third section.

Table 4: Questions in the Third Section

N°	Question
5	What are the critical factors for consultants in the implementation of such technologies? Are there any limitations in these technologies?
6	What are the possible future developments of Big Data Analytics? In terms of technology standards and customer requirements...

Source: the author

Overall, all the interviews were rich in content and complementary to each other, in fact, they all lasted more than one hour. Moreover, other KPMG colleagues took part in these meeting and commonly asked some additional questions at the end of the interviews, thus completing the overview. Therefore, thanks to their contributions, the interviews were for some respects similar to the format of the focus groups.

Chapter 3. Findings

This chapter discusses the results of this research and is structured as follows. The aim of the first sub-chapter is to assess the correspondence between the experience gained by Big Data Analytics consultants in their various projects and the knowledge offered by the current academic literature regarding this topic. Thus, after a concise introduction on Big Data Analytics projects, in order to evaluate the compliance with the fields of application of Big Data Analytics and the benefits that derive from these technologies, the first section focuses on the customers and the purposes of such projects. Then, the next section compares the dimensions of Big Data detected by consultants' experience with the definitions found in the literature review. Finally, the third section sets out the key steps of these projects and the activities that are generally performed by consultants.

The second sub-chapter focuses on the more technical aspects of these projects. Hence, the first section provides a detailed illustration of the different layers that together composes the infrastructure, the main platform of the Big Data field and the several tools that are typically used in such projects. Then, the second section points out the critical parts of the technology based on the difficulties encountered by experts during their several projects.

The third sub-chapter groups and exposes the various suggestions and insights that emerged during the interviews with the experts. Therefore, the first section underlines the critical factors for consultants in implementing Big Data Analytics projects as well as the minimum requirements for those consultants who want to enter in this context. Finally, the second section analyzes the different opinions of the experts about the future directions of Big Data Analytics, both in term of technology developments and customer needs.

3.1 Big Data Analytics Projects

First of all, in these projects the consulting team typically interfaces with the client's ICT team, which in turn interfaces with the company's business. Then, as regard the number of resources involved in such projects, it depends in large part on the specific case, and therefore it is impossible to get some general conclusion in this direction.

"…It depends a lot on the projects but in large projects you can create consulting teams with even more than 100 people, perhaps divided between different consulting companies…"

Consultant 1, Manager – BI and D&A

Another very common pattern that emerged from the interviews is that there is never a unique solution in these projects. Indeed, since the speed of development is incredible and constantly evolving, in each project it is possible to implement different solutions that could lead to the same results:

"… there is no single solution, there are several solutions…"

Consultant 2, Manager – BI and D&A

After that, the next section focuses more deeply on the customers and the purposes of such projects, in order to assess the correspondence with the literature review about the fields of application of Big Data Analytics and the benefits that derive from these technologies.

3.1.1 Customers and Purposes

From the interviews to Italian consultants of KPMG, accordingly with the literature review, it emerged that Big Data Analytics projects are present in several industries and sectors, with different purposes. Among the industries, the public administration bodies stand out as well as players in the energy sector, but also actors in the pharmaceutical and manufacturing sectors. Instead, regarding the purposes of these projects, from the experience of consultants emerged that in most of the cases the main aim of these projects is to create a central platform in which all the business data converge, thus allowing analysis and reporting for the various needs of companies.

For example, one of the most interesting projects related to Big Data Analytics in which one of the experts worked on is for the Italian National Statistical Institute (ISTAT):

"…The customer is an Italian public administration body, the National Statistical Institute (ISTAT). The goal of the project was to create a unique platform with a single and valid corpus data for all offices, in order to structure the experimental work in a more industrialized way through the Cloudera platform…"

Consultant 1, Manager – BI and D&A

In addition, another project for the Italian PA is also cited by another consultant:

"…In the project for the National Institute for Insurance against Accidents at Work (INAIL), a body of the Italian public administration, the aim was to create an Enterprise Data Hub that would act as a collector of all the various areas in a single centralized point. The goal was to centralize the entire Data Analytics world on a single Big Data technology platform such as Cloudera…"

Consultant 2, Manager – BI and D&A

Moreover, from the interviews emerged that these projects are also present, for example, in the Energy sector:

"…Another very interesting project is for Enel, the company that sells electricity in Italy. The customer's need was to have a single Business Intelligence platform, to carry out analysis and reporting, usable by every department of the company…"

Consultant 1, Manager – BI and D&A

As also mentioned by another consultant:

"…Another project in which I was involved was to design the IoT architecture for the predict maintenance and monitoring of the pylons of Terna, the company that manages

the Italian energy market and therefore monitors the electric current and the daily needs of the Italian system…"

Consultant 2, Manager – BI and D&A

Therefore, more in general, from the experience of consultants it is emerged that Big Data Analytics projects are present in several industries and for different customers:

"…I have been involved in various projects of various types such as, for example, in the pharma world to manage all patient data, for INPS and therefore a platform for the public sector, A2A is an energy company for which trading algorithms have been developed, Prysmian for the outlayer detection…"

Consultant 3, Manager – BI and D&A

In particular, one of these projects for example encompass the planning and forecasting of practices and activities:

"… the INPS project consisted in planning and forecasting practices or planning future activities: from estimating the number of files in a certain period to sorting them in the various offices throughout the country on the basis of available resources…"

Consultant 3, Manager – BI and D&A

3.1.2 Big Data Dimensions

Regarding the conformity of Big Data dimensions in real projects with the definitions found in the academic literature, such as the Big Data paradigms of 3Vs or 5Vs, there are different opinions between the consultants. In fact, in some projects the compliance is completely verified, as for example stated by one of the consultants:

"...In the INPS project, the compliance with the 3/4Vs is definitely verified: the data source was made up of internal databases containing extensive information on the activities carried out by employees rather than information relating to the management of individual practices at the online counter. So, a fairly large amount and wealth of data..."

Consultant 3, Manager – BI and D&A

Or by another consultant that confirmed the compliance, for example, in the INAIL project:

"... compared to the 3Vs paradigm of Big Data, all 3 the dimensions in the INAIL project were quite respected, with the main constraint on Velocity due to the network speed..."

Consultant 2, Manager – BI and D&A

But on the other hand, there are other projects where compliance doesn't seem to be properly verified:

"...In all the projects I have faced, the three dimensions have never been completely satisfied, in fact, almost always the only dimension respected was that of Volume..."

Consultant 1, Manager – BI and D&A

In addition, the consultant underlined the importance of considering the different sources for the Velocity dimension because not always all the data coming from the multiple sources have the same updating speed and this means that it is necessary to wait the alignment to be able to propose consistent results:

"...The Variance tends not to exist because it is often structured data and as regards the Velocity it depends, because the various data sources that flow into the Data Lake have different updating speeds and therefore, even if the data itself is updated practically in real-time, the usefulness of the data for reporting purposes proceeds at a slower speed..."

Consultant 1, Manager – BI and D&A

Therefore, from the experience of consultants emerges that in Big Data Analytics projects the dimensions that differentiate Big Data from simply large datasets in most of the cases are respected, but not always. Indeed, there may be cases where one or more dimensions are not adequately met.

3.1.3 Key Steps and Activities

In Big Data Analytics projects there are some key steps and activities that are typically carried out by consultants. Starting from the process of collecting the necessary data, they move on to the data processing phase which can take place in different ways such as data cleaning or conversion, checking for inconsistencies or empty values, analyzing any duplicates, to finish with the phases of data aggregation and presentation. Moreover, in some projects the necessity of IoT devices lead to another step that is the field work.

For example, one of the consultants explained that the Italian National Statistical Institute (ISTAT) keeps track of the trend of the country's retail prices and therefore the national inflation, acquires the retail sales data of consumer goods, defines that some must stay within the so-called basket and with respect to these prices it tracks the country's macroeconomic performance. Therefore, the activities involved in this project were collecting data from selected supermarkets and processing them. The processing consisted of about seven steps, each step had a deconstructed data at the input and a calculated (normalized and clean) data at the output. A typical processing was done when consistency problems were encountered, and in this regard, the consultant proposed an example taking tuna cans:

"…at the entrance there was a list of rows with an identification code for each type of can, its price and a series of characteristics such as weight. If in the next update the product in question, for various reasons, was removed from the market and replaced by a similar one, for example a certain can of tuna was replaced with one of a different weight, the identifier remained the same and therefore consistency problems were created and had to be managed…"

Consultant 1, Manager – BI and D&A

Then, a further example of processing was carried out on empty input data:

"…rows could arrive with values but for example with the price field empty and therefore unusable…"

Consultant 1, Manager – BI and D&A

In addition, an analysis of the waste was carried out to avoid duplication. Subsequently, the lines were aggregated by identification and at the end of the process a structured list of the weekly average price was created, made available to the next platform which tracked the price of the so-called basket. Furthermore, the consultant underlines a particular customer request:

"...in this project a customer request was to have the possibility to analyze the data punctually for each step, thus creating an incoming and an outgoing view for each step..."

Consultant 1, Manager – BI and D&A

Big Data Analytics are also very linked to IoT devices. Thus, from the Terna project emerged an additional component compared to the other projects, that is the field work:

"... go and install the IoT sensors, understand how to make the sensor communicate with the platform (e.g., through the time), installation of the Hubs that acted as gateway between the various sensors and therefore knowing how to manage and optimize the position in which to place these Hubs..."

Consultant 2, Manager – BI and D&A

In addition, the consultant underlined that in these scenarios the real-time is essential.

3.2 Technical aspects of the Projects

From the interviews to experts, several shared practices regarding the more technical aspects of Big Data Analytics projects are emerged. In effect, these are the most important aspects for a consultant. Therefore, during the interviews, the experts mentioned many software products, tools, and platforms. In this regard, a software product is a tool, or a collection of tools, packaged together by a software company. A tool is a standalone piece of software designed to perform a single specific function or a small set of functions. It lives in a world all its own and serves one purpose. A platform refers to substantial piece of software that other companies can access for content distribution purposes or use as a foundation on which to build their own products. Platforms are designed to allow third parties to use the platform infrastructure to deliver value to users via data and process integrations.

For instance, one of the consultants gave a concise overview about the technical aspects of these projects:

"… The platform we mainly use is Cloudera, also as a SQL web interface. The most used programming language is Python as an interface for Hadoop and Spark nodes. For some customers instead the classic R is used. Two other low-code tools that interface well with Big Data platforms are Knime and Alteryx. Then Impala, Kafka, Pub/Sub for queue management. Also, on the Google side there are two Big Data systems called BigTable and BigQuery, both manageable with Python…"

Consultant 3, Manager – BI and D&A

After that, the next section delves into the layers of the infrastructure, platforms and tools used in these projects.

3.2.1 Infrastructure, Platforms and Tools

Starting from the infrastructure typically used in Big Data Analytics projects, from the interviews emerged some common patterns regarding the different layers. So, to explain the various layers of the infrastructure, it is necessary to first provide the difference between Data Warehouse (DW), Database (DB) and Data Mart (DM). The Data Warehouse is a software tool, or set of tools, with the purpose of storing all business data, whether they come from management systems or from external sources. For this reason, data structures alternative to those of operational databases have been conceived; while these are based on concepts and relational rules (Entity-Relationship), Data Warehouses are generally based on the dimensional model or Star Schema, optimized to respond quickly to various types of queries. The activity of Data Warehousing, that is the construction and management of a Data Warehouse, includes various phases:

- identification of starting data
- data conversion, extraction and cleaning
- use of a DBMS to manage the Data Warehouse
- use of Business Intelligence tools to access it

The Database, instead, indicates a set of structured data that is homogeneous in terms of content and format, stored in an electronic computer and interrogated via the terminal using the access keys provided. The databases are created, managed, and queried by a software system called Database Management System (DMBS) which represents a data management system that guarantees a level of data security, allowing users to share it safely and reliably. This system is interposed between the user and the data in the database; thanks to this layer of software, the user does not have direct access to physically stored data, but only to a logical representation of them, thus allowing a high level of independence between physical data and applications. Current database applications allow access to data by multiple users at the same time: this is made possible by the fact that there are DBMS developed in such a way that, using a single copy of the data, they allow the creation of more logical representations of these, reducing their redundancy and inconsistency.

Finally, a Data Mart is a specialized data collector on a particular business segment or area. It contains an image of a portion of the data and allows you to formulate strategies based on past trends. It is placed downstream of a DW and is fed from it, of which it constitutes, in practice, an extract. In more technical terms, a data mart is a logical or physical subset of a larger DW. The fundamental difference consists in the fact that the creation of the Data Warehouse takes place first in a generalized manner and then undergoes modifications to be able to adapt to specific needs, while the data mart is generally created to meet an already determined need. The need to create a separate system for the DM compared to the DW can be summarized in the following reasons:

- need to use a different scheme
- improve performance by separating the dedicated computer
- ensure greater security by having to authorize access to a smaller set of data

Hence, the infrastructure is commonly composed of the following three layers and each layer has one of the previous technologies within it, as also shown in *Figure 11*:

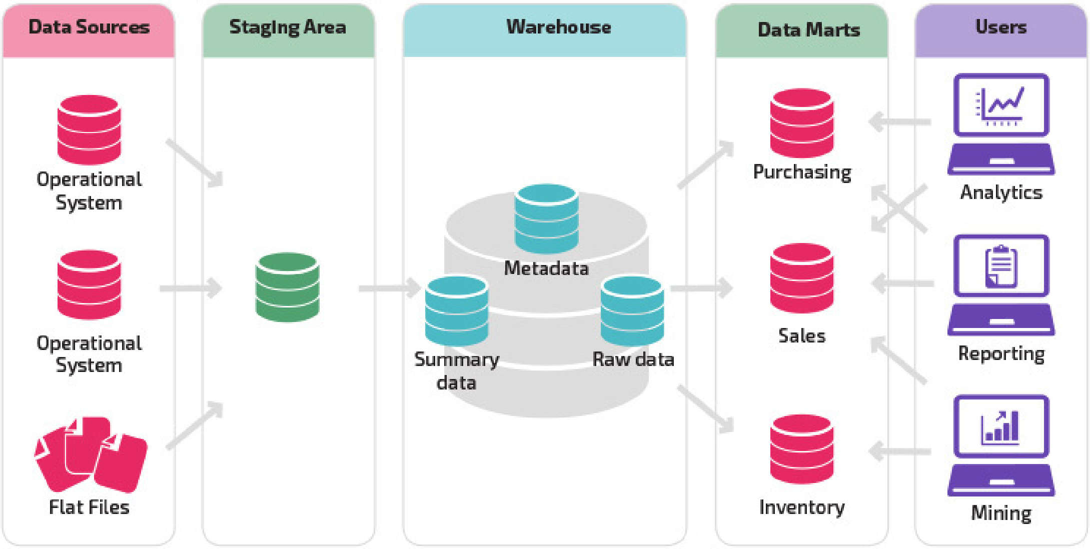
- Staging Area or Data Lake (DB)
- Data Warehouse (DW)
- Data Marts (DM)

In effect, one of the experts explained that, for example in the Enel project, the infrastructure was basically composed by these three layers:

"…The data infrastructure followed the classic Data Warehouse paradigm, consisting of three layers: Staging area, in which the data were reported as they were (DB); the processing phase in which the Data Model was created and the data are normalized according to the needs (DW); and Data Mart (DM), where the navigation axes were identified for the purposes of the analyzes requested by the customer…"

Consultant 1, Manager – BI and D&A

Figure 11: Layers of the Infrastructure



Source: Lavecchia, 2019

In addition, the consultant continued to explain what specific tool was used in each layer:

"…the data was structured but came from extremely different sources such as CRM, SAP for billing, telephone contact center. A Data Lake was therefore created, always on Cloudera technology, in which all the data were entered in a flat manner, then the data passed to the AWS where there was a normalized environment and a Data Model connected to the customer, finally the last layer was made up from the Data Mart for analysis or the part of BI in which there are Analytics tools such as Qlik Sense or Microsoft Power BI for graphical representation of the data and navigation…"

Consultant 1, Manager – BI and D&A

Then, one of the most widely used platform infrastructure for Big Data solutions is the Hadoop open-source framework, which is used for storing and processing Big Data in a distributed manner on large clusters of commodity hardware. Basically, it is a platform or framework which solves Big Data problems. It can be considered as a suite which encompasses several services for ingesting, storing, and analyzing huge data sets along with tools for configuration management.

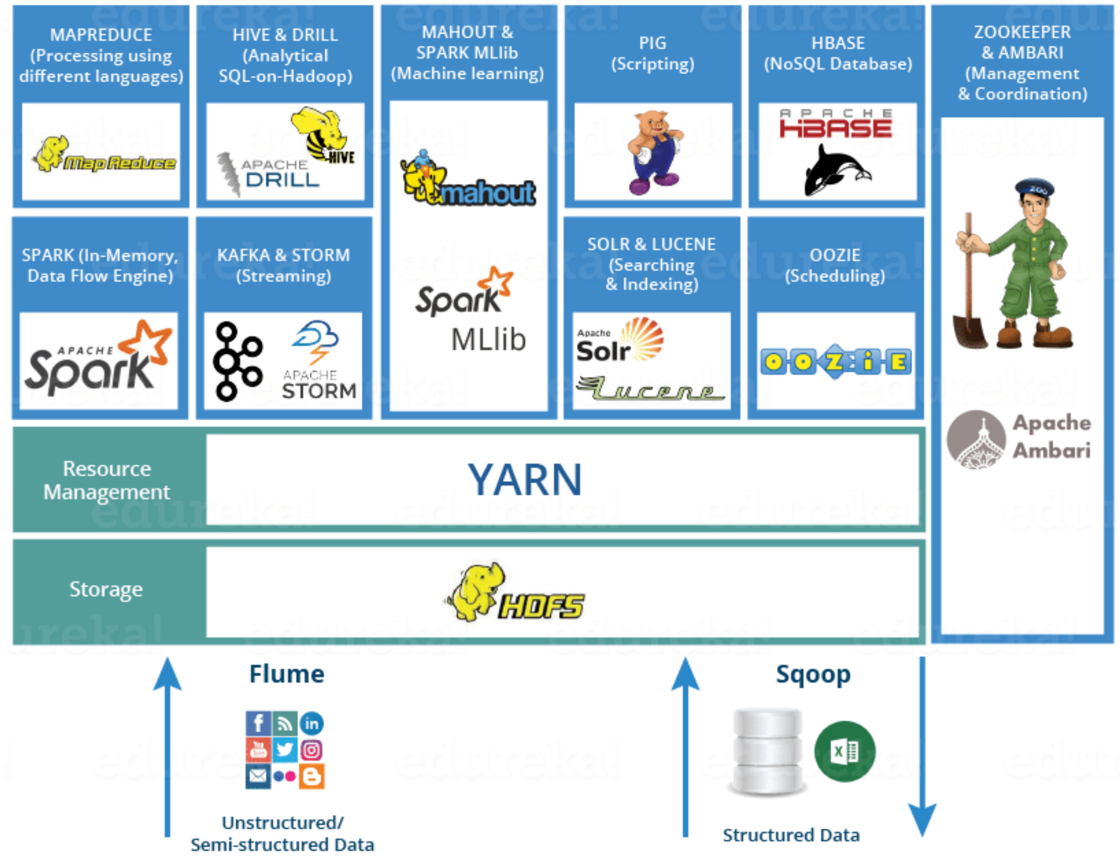
"…Hadoop is an ecosystem for Big Data used for example by Google and AWS, a project open to everyone and with a series of free tools that communicate with each other…"

Consultant 1, Manager – BI and D&A

Below are listed the Hadoop components, that together composes a Hadoop ecosystem, as also shown in *Figure 12*:

- HDFS: Hadoop Distributed File System
- YARN: Yet Another Resource Negotiator
- MapReduce: Data processing using programming
- Spark: In-memory Data Processing
- PIG, HIVE: Data Processing Services using Query (SQL-like)
- HBase: NoSQL Database
- Mahout, Spark MLlib: Machine Learning
- Apache Drill: SQL on Hadoop
- Zookeeper: Managing Cluster
- Oozie: Job Scheduling
- Flume, Sqoop: Data Ingesting Services
- Solr & Lucene: Searching & Indexing
- Ambari: Provision, Monitor and Maintain cluster

Figure 12: Hadoop Ecosystem


MAPREDUCE (Processing using different languages)
Map Reduce
HIVE & DRILL (Analytical SQL-on-Hadoop)
HIVE
APACHE DRILL
MAHOUT & SPARK MLlib (Machine learning)
mahout
Spark MLlib
PIG (Scripting)
HBASE (NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI (Management & Coordination)
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM (Streaming)
APACHE STORM
SOLR & LUCENE (Searching & Indexing)
Apache Solr
Lucene
OOZIE (Scheduling)
OOZIE
Apache Ambari
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data
edureka!

Source: Sinha, 2016

Furthermore, from all the interviews it emerged that currently the main platform for Big Data industry is Cloudera. There are several reasons that lead Cloudera to dominate the Big Data landscape, one of these is the following:

"…At the beginning of the INAIL project an analysis was made to understand which platform was the best, as a result of this analysis the Cloudera platform was chosen mainly because it was the best in all the administration part, in fact it had made a tool of its own called Cloudera Manager, done much better and more intuitive than competitors…"

Consultant 2, Manager – BI and D&A

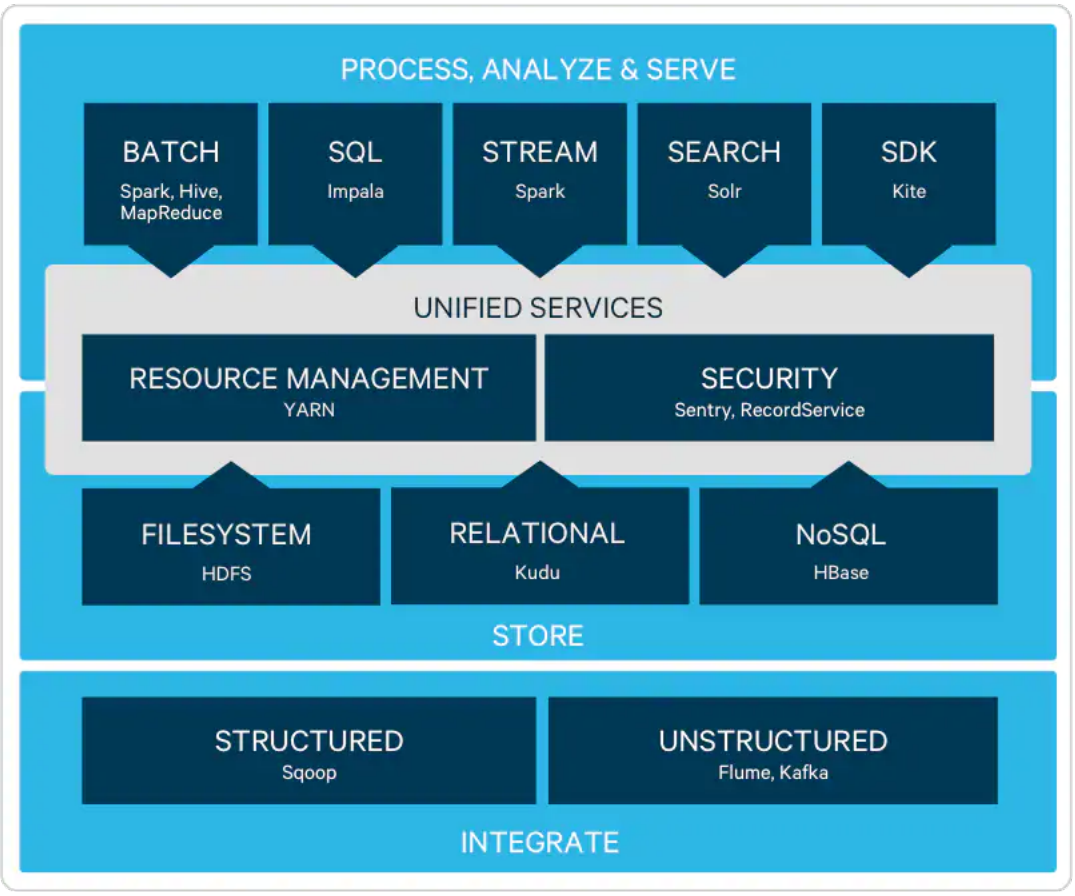
Cloudera has substantially incorporated Hadoop and added its own tools that it deemed necessary for the use of this ecosystem, such as, for example, a graphical interface for configuring and maintaining the platform. Within this platform, there are then at least ten tools, each specific for a specific activity, as shown in *Figure 13*.

"…This ecosystem allows you to manage what Big Data is in the most flexible and effective way possible. Among the main tools we mention Kafka to maintain the indexing of where we have come to read in the data source, Spark which is a programming

environment, HDFS a distributed file system for storage that allows you to write in parallel on multiple machines, HBase with an SQL interface for viewing…"

Consultant 1, Manager – BI and D&A

Figure 13: Cloudera Platform



Source: Cloudera, 2021

3.2.2 Critical parts of Technology

Interviews with experts revealed that there are several limitations in infrastructure and platforms that consultants need to be able to overcome.

"…When it comes to Big Data, customer expectations and demands are very high but there are also limitations in the infrastructure…"

Consultant 1, Manager – BI and D&A

For example, in the INAIL project, a constraint dictated by the customer's needs was that the end user had to use certain data visualization tools that they did not want to change. As a result, a series of difficulties arose related to compatibility between the different levels of the infrastructure, and it was necessary to adopt workarounds to overcome these incompatibilities. Then, also the update tool in Big Data technologies has some issues:

"…the data update tool in Big Data technologies does not work very well, there are solutions that allow you to perform the operation but are not as efficient in terms of performance as on Oracle databases…"

Consultant 2, Manager – BI and D&A

Another goal of the INAIL project was the dispose of the Data Warehouse and in this regard, there have been various discussions because the DW technology had little fit with Big Data technologies:

"…there is this trend of bringing everything to Big Data technologies, even if it is not always the best way … for example, I see the Data Warehouse alongside Big Data, not to replace it but to support it…"

Consultant 2, Manager – BI and D&A

In other words, the consultant explain that Big Data technologies are basically used to process large amounts of data in real-time, when there is no strict need, technical efficiency is also lost.

Among the difficulties encountered in the INPS project, instead, there was a reduced historical depth of data due to a migration of databases with different data encodings. Then, the customer knew little about the Machine Learning and Analytics part and therefore it created a difficulty in communicating the results and this often translated into mock-up activities to exemplify the results to the customer and therefore make a clear idea of the direction in which is going. But making mock-ups is an activity that requires time that must be subtracted from the time allocated for the development part.

"…In this specific case, also given the size and power of the customer, it was not a big problem in terms of time and costs but when it comes to a customer dimensionally smaller it can become a problem because you are taking away time from developments…"

Consultant 3, Manager – BI and D&A

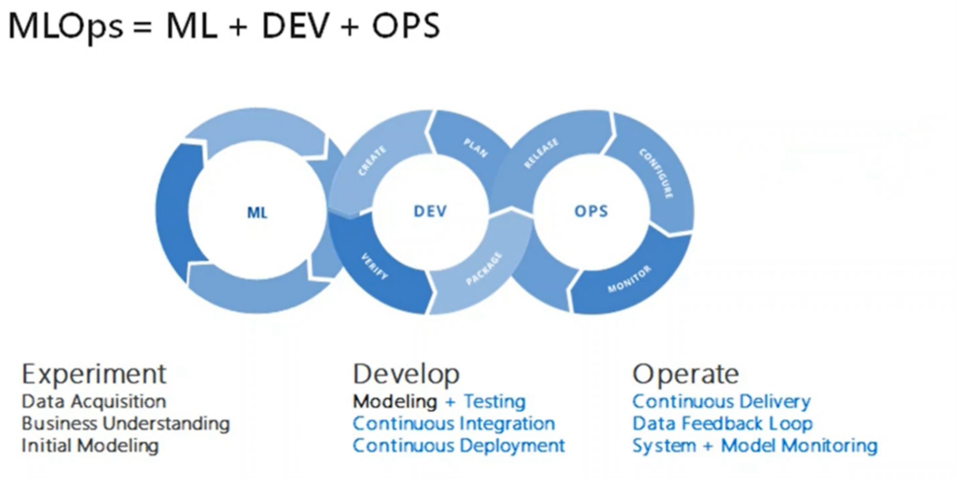
Basically, this means that in projects for smaller companies making mock-up activities could become a problem in terms of cost and time. In addition, the consultant revealed that sometimes methodological difficulties are also created because:

"…you work within a larger project and the MLOps model is very different from other Data Warehouse management models and therefore clearly there are slightly different rhythms than those of all the other teams … for example, one of the components is the management of ETL processes which takes place in a much faster time frame than the development of a mathematical model…"

Consultant 3, Manager – BI and D&A

Figure 14 below explains the MLOps Model, that is known as DevOps for Machine Learning and empowers data scientists and app developers to help bring ML models to production. It is able to audit, certify and re-use assets in the ML lifecycle.

Figure 14: MLOps Model



Source: Zahra, 2019

After that, an important aspect that emerged from the interviews concerns the need for programming to enhance the effectiveness of Big Data tools:

"…In my experience, a Big Data product is more effective if you program code by hand because in addition to having carte blanche to develop the necessary functionalities, you also have the possibility to manage the calculation in parallel and in a Big Data context this is an essential necessity (contention for resources in the multi-tenant) …"

Consultant 1, Manager – BI and D&A

In fact, the expert explain that a lot of consultants often try to avoid these solutions by using graphic products that produce an over-cost and do not always prove to be effective. Therefore, the need for programming emerges because any graphic type of product will introduce some rigidity or some modules that severely limit its use.

Furthermore, among the databases there are the more classic ones, the Oracle ones and the Hadoop ones which are the most suitable for this area. Thus, if a Hadoop database is used there are no problems from the point of view of performance, but otherwise:

"... if the customer has never made the transition to a Big Data paradigm but continues to use a traditional database, then in doing Analytics and Machine Learning, a bottleneck is created in data extraction because you want to have precise information without having the power of adequate parallel computing which is fundamental in this area..."

Consultant 3, Manager – BI and D&A

Also, despite that in this context it is theoretically even more necessary to have proportional development, test, and production landscapes, because each step should be tested with something very similar to the production one, a big difficulty highlighted by the consultants is that the development and test environment is much smaller than the production one:

"...Typically, there is a partial test environment in which the tests are performed and then a multiplication factor is applied to predict the impact it will have on the production environment..."

Consultant 1, Manager – BI and D&A

In these scenarios, another common constraint is linked to the issue of licenses that limit the use and consumption of the various platforms and tools:

"...there are commonly upper bounds that must not be exceeded, an overload can occur and therefore we must also pay attention to this aspect and find remedies..."

Consultant 2, Manager – BI and D&A

Moreover, the consultants explained that a possible bottleneck inherent to Cloud is represented by the network speed:

"...there were significant difficulties in transporting data from the Data Source, which in this case was a Mainframe, to Azure, because it was very slow and there was not much possibility of parallelizing the operations..."

Consultant 2, Manager – BI and D&A

For this reason, it seems that the best practice is to do everything on the Cloud or everything on-premises:

"...my personal advice is to do everything on the Cloud or everything on-premises because there is a bottleneck in the network speed..."

Consultant 2, Manager – BI and D&A

Also, another critical factor is related to the costs of scalability on Cloud solutions. For example, on Google's BigTable it is possible to have information that has a very high temporal depth and, also incredible spatial depth as for the identification of geographic places.

"... The speed of the solution obviously depends on the nodes below and on the type of infrastructure that has been chosen, but you pay a lot for use and a lot for waiting time..."

Consultant 3, Manager – BI and D&A

In other words, it is theoretically possible to have everything immediately but the costs scale proportionally and therefore it is created this saddle point with respect to how much you want to spend and how much you want to get. In more quantitative terms:

"... for example, a virtual machine you get on Azure can cost about € 2/3 k per year, then if you have to do Machine Learning you put some RAM, a fairly powerful CPU and finally some GPU graphics cards. The machine goes from € 3k per year to € 8k per year, so the cost rises exponentially..."

Consultant 3, Manager – BI and D&A

Finally, the consultant stands out a personal concept related the limitations of the technology:

"... The limit of the technologies in my opinion is relative to the Volume ... the quantity of data, the cost of data transmission and the management of the data mean that we try to limit what information is on the channel..."

Consultant 3, Manager – BI and D&A

He explained that they try to make the machines that transmit more and more intelligent but, the one who has the vision of everything is the central Data Lake: when the system is limited, on the one hand it is intelligent but on the other the central brain that should make decisions has information limited because the single device sends so-called polarized information.

3.3 Insights from the Experts

The experience of consultants revealed several tips and insights during the interviews. For example, one of the consultants explained one of the reasons that led to the explosion of Big Data Analytics projects in recent years:

"…The customer does not want to spend months of analysis to understand from what source they have to get the data they need for reporting purposes; they ask to take everything that comes from the systems and then they will understand what they need…"

Consultant 1, Manager – BI and D&A

Then, another interesting insight concerns cloud solutions. Although the technology is already quite advanced to date, for customers there are still doubts about the implementation of these solutions, especially in the bodies of the Italian PA:

"…in the INAIL project the scenario was on-premises, in fact, in Italy the public administration is still a bit skeptical about Cloud solutions…"

Consultant 2, Manager – BI and D&A

Furthermore, as regard the minimum requirements for those consultants who want to start a career in Big Data consulting, some experts consider enough a good knowledge of main programming languages used in this field, such as Python, Java, and SQL:

"… among the minimum requirements to approach the world of Big Data, programming skills such as Python, Java and SQL are certainly necessary … these basics are enough to learn practically any type of Big Data tool…"

Consultant 2, Manager – BI and D&A

On the other hand, if with only the Big Data part it is possible to talk about IT, the part more related to Analytics is closer to mathematics than to information technology. Therefore, to become a Big Data Analytics consultant, transversal skills are necessary because IT alone is not enough.

After that, the next sections first provide the critical factors for consultants in such projects and then analyzes the possible future directions of Big Data Analytics.

3.3.1 Critical factors for Consultants

The critical factors for a Big Data consultant are many, starting with the high competence level required, because working on a certain amount of data, each mistake could cost a lot:

"...working with certain amounts of data, a small mistake can have a huge effect both in terms of time and resources ... a lot of attention and high competence is in fact required..."

Consultant 1, Manager – BI and D&A

Indeed, due to the delicateness of these projects, a high attention to detail is required:

"...These are extreme projects, very delicate, the data must be handled with care..."

Consultant 1, Manager – BI and D&A

Thus, another consultant exposed that there are two types of difficulties for a Big Data consultant:

"... In my opinion, the difficulties for a consultant in this type of projects are of two types: the first is of a technical-professional nature while the second is more related to soft-skills or how the client perceives certain things..."

Consultant 3, Manager – BI and D&A

Indeed, the consultancy can be divided into two different aspects: technical aspect and strategic aspect. In technical consultancy the customer tends to already know the technology and needs the consultant who knows that specific technology to implement it. From a strategic point of view, however, the customer does not know the technology and therefore the consultant must direct and educate the customer towards a certain technology.

Then, among the critical factors there is also the difficulty of treating many different technologies together. In fact, Big Data is not a single product:

"...The Big Data world is not a single block but more like a puzzle, for each piece you have to understand if it fits together and therefore there are many possible combinations of scenarios..."

Consultant 2, Manager – BI and D&A

Consequently, a good consultant must at least know for each tool what it does, the pros and cons of each of the competitors in order to be able to choose optimally and build the puzzle in the best possible way.

"... this allows you to have ample flexibility and therefore when there is difficulty on one path to take another..."

Consultant 2, Manager – BI and D&A

For instance:

"...Impala has a limit on RAM and therefore it can happen that it sometimes goes out of memory and returns error on the query result. Hive on the other hand does not have this type of problem, maybe it takes a lot longer, but it will never go out of memory and will always complete the result, for this reason it is recommended for batch type uploads..."

Consultant 2, Manager – BI and D&A

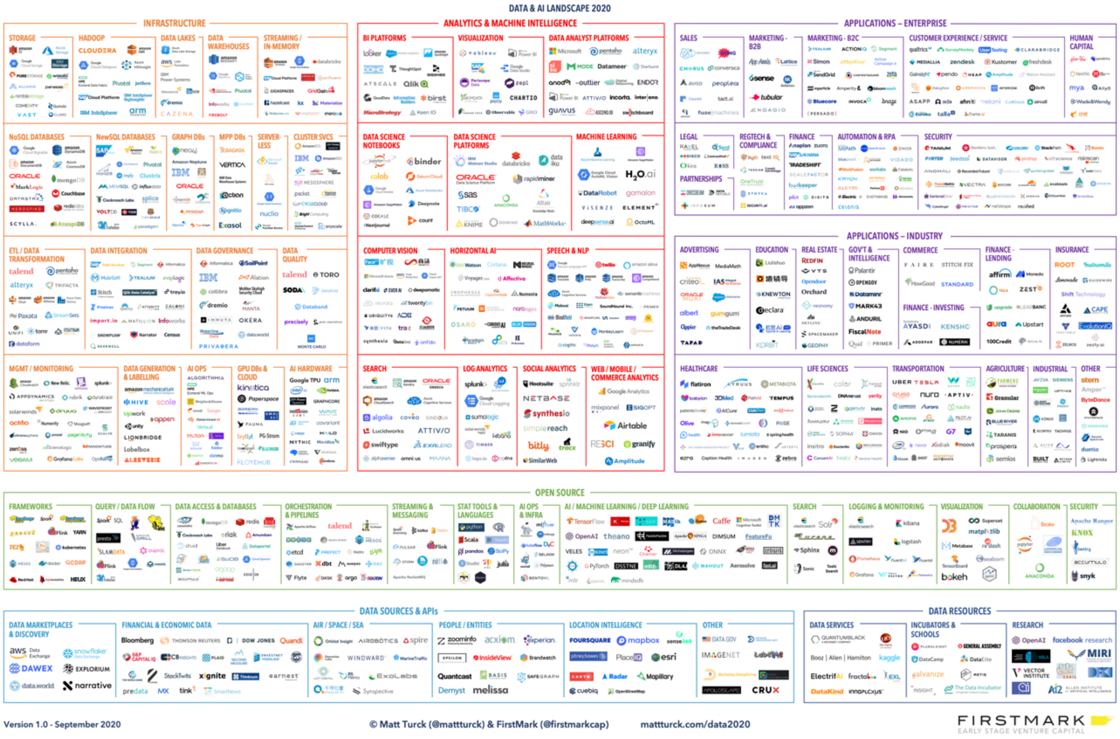
To do all this it is essential to keep up with the technologies that evolve every day. One of the consultants defined these continuous evolution as "continuing shock". So, as for how to update on new technologies, there are several sources on the internet and during the interviews one of the experts suggested one of them:

"...I suggest you look for Best of Brand by Matt Turk which for each year groups the best technological developments and divides them by area and sector..."

Consultant 2, Manager – BI and D&A

Figure 15 below shows the source suggested by the consultant.

Figure 15: Big Data & AI Landscape 2020


DATA & AI LANDSCAPE 2020
INFRASTRUCTURE
ANALYTICS & MACHINE INTELLIGENCE
APPLICATIONS – ENTERPRISE
APPLICATIONS – INDUSTRY
OPEN SOURCE
DATA SOURCES & APIs
DATA RESOURCES
Version 1.0 - September 2020
© Matt Turck (@mattturck) & FirstMark (@firstmarkcap)
mattturck.com/data2020
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Source: Matt Turck, 2020

3.3.2 Future of Big Data Analytics

From the experience of the consultants, it emerged that for some aspects the future of Big Data Analytics is still very uncertain:

"...The future is still very uncertain because currently there is no market standard, typically in these types of projects Data Lakes are made..."

Consultant 1, Manager – BI and D&A

In addition to the lack of a market standard, the consultants have highlighted that in recent years Big Data projects have become as a fashion, although these solutions not always are the better:

"...It is difficult to imagine what future developments in this area could be because until a few years ago it was a fashion, everyone wanted to do Big Data projects or use the appropriate infrastructures such as Cloudera that had become a must, or NoSQL databases, but these infrastructures do not solve all critical issues ... the future, though I don't know when it will happen, will be understanding that this type of infrastructure only responds well to certain types of projects such as those of Data Lakes..."

Consultant 1, Manager – BI and D&A

Therefore, the hope of consultants is that the software producers will be good at making new releases that will optimize the current technology:

"...it is difficult to go back and change the practices that are currently used with Big data, software producers will have to be good at making releases that actually optimize the use of these infrastructures, otherwise it will always be very difficult projects to achieve..."

Consultant 1, Manager – BI and D&A

After that, as for the on-premises solutions compared to those on Cloud, from the experience of the consultants there seems to be no doubt about the dominance in the future of Cloud:

"...In my opinion, the future is tending towards everything on the Cloud and the on-premises will disappear ... Big Data in five years will become the norm ... A possible evolution will be related to the automation of tasks through AI, less code and more user-friendly..."

Consultant 2, Manager – BI and D&A

In effect, another consultant explained that the reasons that lead the future to be all on Cloud are related to costs:

"… for example, when it is necessary to carry out forecast analyzes, there is a need to have great computing power but for a very short time (e.g., 30 knots but for 48 hours). The Cloud is scalable and therefore allows you to implement such a temporary infrastructure while obviously a solution of this type is not possible on-premises…"

Consultant 3, Manager – BI and D&A

And of course, same conclusion can be extended for memory too. Moreover, another important theme underlined during the interviews about the future of Big Data Analytics is related to the growing need of Data Governance:

"… in the future an aspect that will take root will be in the field of Data Governance, as there are more and more technologies, even heterogeneous ones, all together. I foresee a strong development in the areas of Data Lineage, Data Catalog and Semantic Layer…"

Consultant 2, Manager – BI and D&A

Finally, according to experts, in the medium term, there are at least two aspects to consider. The first is the increase in Machine Learning and consequently the data will have to become increasingly structured within NoSQL systems; this means that much more advanced data acquisition logic will be created:

"… the device will no longer filter the data, or it will filter them in a different way so that they arrive at the Data Lake with a set of more information…"

Consultant 3, Manager – BI and D&A

It is likely that many systems in the future will be of the Big Data type but that they will use widespread information in the sense that it will no longer be the company's employees who will interface with the system and enter the information of their customers but rather that it will be the customers themselves that will interface directly to the system through, for example, online forms and will enter their information that will be validated through AI control systems. And second, the consultant is sure that there will be a more automatic management of all Big Data processes with the birth of all these so-called low-code tools:

"… the scripting part will be eliminated, and the code will continue to remain but will be replaced by a graphical system…"

Consultant 3, Manager – BI and D&A

In this way, people in any area of the company will be increasingly able to carry out Big Data analyzes independently without the support of the IT area, and this means that a widespread data culture will be created.

Discussion and Conclusion

The first section of this chapter provides an overview of the research conducted in this article: starting from the gap in knowledge found in the academic literature and consequently the goals of this research, continuing through the discussion of the key steps that led to answering the research questions, and ending with the key findings and the relevance of the results achieved. Finally, in the last section the limitations of this research are highlighted and the opportunities for future developments of the study in this context are suggested.

Results of the Research

This research started with a literature review on the Big Data Analytics field. In particular, it was first explained what Big Data is, through several definitions regarding the dimensions that differentiate Big Data from simply large datasets: from the 3Vs paradigm, to the 5Vs paradigm, ending with further dimensions. Then, it was explained how Big Data Analytics work, underling the difference between the data itself (Big Data) and the tools and techniques (Analytics) that allow you to generate effective value from such data. Subsequently, the impact of Big Data Analytics on business was proposed: starting from the new opportunities that arise for the business and the barriers that organizations should overcome to be successful, continuing with the fields of application of such technologies and ending with the benefits for those companies which are able to exploit Big Data Analytics. Finally, some challenges for organizations were highlighted regarding the other non-technological aspects that are required to achieve these benefits.

Thus, although the academic literature in the past decade provided several studies around the world of Big Data Analytics, from the literature review emerged that little progress has been made to understand how these technologies are implemented in companies. Since most of them rely on consulting firms in order to launch Big Data Analytics initiatives, there was a gap in the academic knowledge on how the consulting firms do the processes of design and implementation of such technologies as well as the critical parts of the technology and critical aspects for consultants in these projects.

Therefore, the aim of this research was to enter the black box of the implementation dynamics of Big Data Analytics technologies by consulting firms, developing a case study focused on the KPMG consulting company, limited to the Italian network. The main source of information of the case study was represented by semi-structured interviews to Italian Big Data Analytics consultants. For the selection process there were used two criteria based on the experience level and area of competence of the consultants, and as a result, three top expert consultants were chosen among all the Italian KPMG consultants. During the interviews a flexible approach was adopted, with the support of an interview guide to conduct the interviews but, at the same time, with the flexibility to deviate from it to gather as much relevant information as possible. Finally, to achieve a comprehensive overview, the insights from the experts was coupled with online document analysis.

Hence, the findings of this research follow the interview guide structure and consequently they can be divided into three main blocks. From the first part it emerged that in Big Data Analytics projects the consulting team typically interfaces with the client's ICT team, which in turn interfaces with the company's business. As regard the fields of application and purposes of this projects, the case study partially confirmed the knowledge gained from the literature review, especially for the public administration bodies, and at the same time complemented it, for example proposing projects in the energy sector. More in general, it is possible to conclude that accordingly with the literature review, such projects are present in several industries and sectors, with the main aim of creating a central platform in which all the business data can converge to allow analysis and reporting for the various needs of companies. Then, with respect to the definitions regarding the dimensions that differentiate Big Data from simply large datasets, such as the 3Vs or 5Vs paradigms, the case study revealed that in most real projects they are respected, but on the other hand there may be cases where due to certain constraints one or more dimensions could not be adequately met. Also, the activities typically performed by consultants in these projects encompass the analysis and collection of data needed, the conversion, the cleaning and the checking for errors or duplicates, the data aggregation, and the presentation. Besides, a further interesting step highlighted by the case study is the field work.

Then, from the second part it emerged that the infrastructure is commonly composed by three main layers: Staging area or Data Lake, in which data were commonly reported as they are from the Databases; the Data Warehouse in which a data model is commonly created and the data are normalized according to the customer needs; and finally, the Data Mart, where the data are commonly split across the different business segments for various customer purposes. Then, accordingly with the literature review, the case study confirmed that the most widely used platform infrastructure for Big Data solutions is the Hadoop open-source framework, that is a suite which encompasses several services for ingesting, storing, and analyzing huge data sets along with tools for configuration management. In addition, the case study revealed that currently the main platform for Big Data industry is Cloudera. Cloudera has substantially incorporated Hadoop, with its many tools for specific tasks (such as HDFS, YARN, MapReduce, Spark, HBase) and added its own tools that it deemed necessary for the use of the ecosystem. As regard the critical parts of the technology, the case study revealed that there are several limitations in infrastructure and in platforms that consultants need to be able to overcome. Sometimes it regards customer special requirements that could lead to constraint between the layers of the infrastructure, other times it is simply because certain tools have little fit with Big Data technologies. From the point of view of limitations, there is also the aspect of the customer understanding, these are very complex issues and consequently this often translates into mock-up activities to exemplify the results to the customer, thus taking away precious time from developments. Then, in these projects there could also be methodological difficulties as well as some rigidity in the functions of the tools and this is why the need for programming emerges. Among the difficulties, the case study revealed also that the development and test environment is commonly much smaller than the production one and furthermore, the issue of licenses that limit the use of the various tools. Above all, a best practice is to do everything on the Cloud or everything on-premises because there could be a bottleneck in the speed of the network.

Finally, from the last part it emerged that among the minimum requirements for those consultants who want to start a career in Big Data Analytics consulting, a good knowledge of main programming languages used in this field is necessary but not sufficient, indeed, transversal skills are required because IT alone is not enough. Thus, the case study revealed that the critical factors for a Big Data consultant are many, from the high competence level to the high attention to detail required because in this context each mistake could cost a lot. Then, since Big Data Analytics is not a single product, there is the difficulty of treating many different technologies together: knowing the pros and cons of each tool allow you to build the puzzle in the best possible way; but to do this, it is essential to keep up with the technologies that evolve every day. The future of Big Data Analytics is uncertain because there are still too many limitations in the technology and therefore the software producers should be good at making new releases. The case study also revealed that in the future it will be all on the Cloud, basically due to the costs of this solutions. Furthermore, the last aspects to consider for the future are the growing need of Data Governance, increase of Machine Learning and the expansion of the so-called low-code tools.

Limitations and Further Research

Since the purpose of this research was to enter the black box of the implementation dynamics of Big Data Analytics technologies by consulting firms, a case study focused on the Italian network of the consulting firm KPMG, one of the Big Four, was developed. The case study, being an in-depth description and analysis of a bounded system, as a bounded system has some limitations. Hence, in this research there are three main constraints which are discussed below and for each constraint further research are accordingly recommended.

First of all, the most important constraint of this research is that the case study is limited only to the Italian network of KPMG, due to the multitude of countries in which the company is present. Therefore, possible future research could focus on other nations of the KPMG network, so confirming the key findings of this research also for other nations, accordingly with the shared values and vision between the several nations of the KPMG worldwide network; or deny them and prove in this way that the results of this research could not be extended to the other nations of the network.

Then, another constraint of this research is related to the development of a single case study. Since it was chosen a single consulting company among the Big Four, another possible future research could develop a multiple case study: for example, one for each of these networks. In this way, it will be possible to compare the shared practices and patterns used in the different companies. Alternatively, if the development of four case studies proves to be too much effort, then it is suggested to focus only on one other company among the Big Four, thus allowing the comparison with the results of this research, that is limited only to KPMG.

Finally, the last constraint of this research concerns the choice of the consulting firm. In particular, for the selection of the company there were taken into account only the four networks that together make up the Big Four because they are the four most prestigious professional services networks in the world and because there are a lot of similarities between them. Therefore, possible future research could focus on other consulting firms that are not among the Big Four.

Bibliography

Al-Barhamtoshy, H., & Eassa, F. (2014). A Data Analytic Framework for Unstructured Text A Data Analytic Framework for Unstructured Text. *Life Science Journal*. https://doi.org/10.13140/2.1.4330.0485

Amankwah-Amoah. (2016). Emerging economies, emerging challenges: Mobilising and capturing value from big data. *Technological Forecasting and Social Change*, *110*, 167–174. Scopus. https://doi.org/10.1016/j.techfore.2015.10.022

Baker, W., Kiewell, D., & Winkler, G. (2014). Using big data to make better pricing decisions. *McKinsey Quarterly*, 4. https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/using-big-data-to-make-better-pricing-decisions

Boston Consulting Group & University of Virginia. (2017). *Digital Transformation*. Coursera. https://www.coursera.org/learn/bcg-uva-darden-digital-transformation/home/welcome

Brown, P. A. (2008). A Review of the Literature on Case Study Research. *Canadian Journal for New Scholars in Education/ Revue Canadienne Des Jeunes Chercheures et Chercheurs En Éducation*, *1*(1), Article 1. https://journalhosting.ucalgary.ca/index.php/cjnse/article/view/30395

Cloudera, © 2021. (2021). *Framework open source di Apache Hadoop*. Cloudera. https://it.cloudera.com/content/www/it-IT/products/open-source/apache-hadoop.html

De Mauro, A., Greco, M., & Grimaldi, M. (2019). Understanding Big Data Through a Systematic Literature Review: The ITMI Model. *International Journal of Information Technology & Decision Making*, *18*(04), 1433–1461. https://doi.org/10.1142/S0219622019300040

Erevelles, S., Fukawa, N., & Swayne, L. (2015). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904. Scopus. https://doi.org/10.1016/j.jbusres.2015.07.001

Ghasemaghaei, M., & Calic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research*, *108*, 147–162. https://doi.org/10.1016/j.jbusres.2019.09.062

Grover, P., & Kar, A. K. (2017). Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature. *Global Journal of Flexible Systems Management*, *18*(3), 203–229. https://doi.org/10.1007/s40171-017-0159-3

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., & Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research*, *70*, 308–317. Scopus. https://doi.org/10.1016/j.jbusres.2016.08.004

Gupta, S., Chen, H., Hazen, B. T., Kaur, S., & Santibañez Gonzalez, E. D. R. (2019). Circular economy and big data analytics: A stakeholder perspective. *Technological Forecasting and Social Change*, *144*, 466–474. https://doi.org/10.1016/j.techfore.2018.06.030

Gupta, S., Drave, V. A., Dwivedi, Y. K., Baabdullah, A. M., & Ismagilova, E. (2020). Achieving superior organizational performance via big data predictive analytics:

A dynamic capability view. *Industrial Marketing Management*, *90*, 581–592. https://doi.org/10.1016/j.indmarman.2019.11.009

Hajli, N., Tajvidi, M., Gbadamosi, A., & Nadeem, W. (2020). Understanding market agility for new product success with big data analytics. *Industrial Marketing Management*, *86*, 135–143. https://doi.org/10.1016/j.indmarman.2019.09.010

Hazirbaba, N., & Yalcintas, M. (2019). *Designing Strategy Dimension of the Organization Based on Big Data Analytics Capability*. 299–309. https://doi.org/10.15405/epsbs.2019.10.02.27

Jacques Bughin. (2011). The Web's €100 billion surplus. *McKinsey Quarterly*. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-webs--and-8364100-billion-surplus

Khalid, B., & Rachid, C. (2019). Big Data in Economic Analysis: Advantages and Challenges. *International Journal of Social Science and Economic Research*, *04*, 5196.

Lavecchia, V. (2019, June 10). Differenza tra Data Warehouse (DWH), Database (DB) e Data Mart (DM). *Informatica e Ingegneria Online*. https://vitolavecchia.altervista.org/differenza-tra-data-warehouse-dwh-database-db-e-data-mart-dm/

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business Horizons*, *60*(3), 293–303. https://doi.org/10.1016/j.bushor.2017.01.004

Liu, X., Shin, H., & Burns, A. C. (2019). Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing. *Journal of Business Research*. https://doi.org/10.1016/j.jbusres.2019.04.042

Matt Turck. (2020). The 2020 Data & AI Landscape. *FirstMark*. https://mattturck.com/tag/artificial-intelligence/

McAfee, A., & Brynjolfsson, E. (2012, October 1). Big Data: The Management Revolution. *Harvard Business Review*, *October 2012*. https://hbr.org/2012/10/big-data-the-management-revolution

Merriam, S. B. (2010). Qualitative Case Studies. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)* (pp. 456–462). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01532-3

Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics and firm performance: Findings from a mixed-method approach. *Journal of Business Research*, *98*, 261–276. https://doi.org/10.1016/j.jbusres.2019.01.044

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, *38*(1), 187–195. Scopus. https://doi.org/10.1016/j.ijinfomgt.2017.07.008

Sestino, A., Prete, M. I., Piper, L., & Guido, G. (2020). Internet of Things and Big Data as enablers for business digitalization strategies. *Technovation*, *98*, 102173. https://doi.org/10.1016/j.technovation.2020.102173

Sinha, S. (2016, October 28). Hadoop Ecosystem | Hadoop Tools for Crunching Big Data. *Edureka*. https://www.edureka.co/blog/hadoop-ecosystem

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

Wang, W. Y. C., & Wang, Y. (2020). Analytics in the era of big data: The digital transformations and value creation in industrial marketing. *Industrial Marketing Management*, *86*, 12–15. https://doi.org/10.1016/j.indmarman.2020.01.005

World Economic Forum. (2012). *Big Data, Big Impact: New Possibilities for International Development*. World Economic Forum. https://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development/

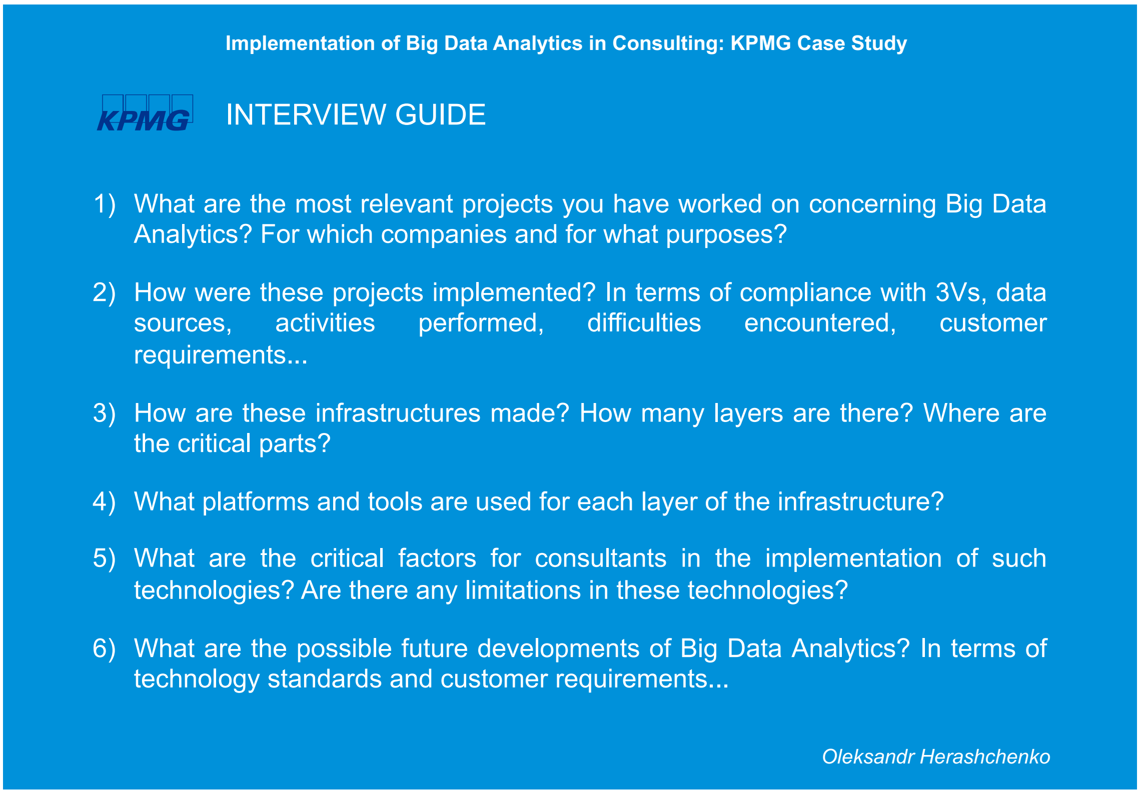
Yin, R. K. (2003). *Case Study Research: Design and Methods*. SAGE.

Zahra, A. B. (2019). MLOps Explained. *C# Corner*. https://www.c-sharpcorner.com/blogs/mlops

Appendix

Figure 16 below shows the slide of the interview guide projected by the author during the interviews on Microsoft Teams.

Figure 16: Interview Guide



Implementation of Big Data Analytics in Consulting: KPMG Case Study

KPMG INTERVIEW GUIDE

1) What are the most relevant projects you have worked on concerning Big Data Analytics? For which companies and for what purposes?

2) How were these projects implemented? In terms of compliance with 3Vs, data sources, activities performed, difficulties encountered, customer requirements...

3) How are these infrastructures made? How many layers are there? Where are the critical parts?

4) What platforms and tools are used for each layer of the infrastructure?

5) What are the critical factors for consultants in the implementation of such technologies? Are there any limitations in these technologies?

6) What are the possible future developments of Big Data Analytics? In terms of technology standards and customer requirements...

Oleksandr Herashchenko

Source: the author

Declaration in lieu of oath

by Oleksandr Herashchenko

This is to confirm my Master Thesis was independently composed / authored by myself, using solely the referred sources and support. I additionally assert that this thesis has not been part of another examination process.

Moscow, 11 May 2021

Oleksandr Herashchenko