



Ca' Foscari
University
of Venice

Master's Degree programme
in Economia e Finanza

Final Thesis

The big data environment

Opportunities, Threats, Blockchain integration and
Monetization

Supervisor

Ch. Prof. Lorian Pelizzon

Graduand

Andrea Vitulano

Matriculation Number 885160

Academic Year

2021 / 2022

Abstract

Big data has proven to be one of the most strategic assets for companies and institutions in recent years. Many refer to big data as the “fuel” of the digital economy and the statistics seem to confirm that. With an estimated market size of billions of dollars¹, big data positions among the most important markets worldwide. The reason of such importance is that nowadays, most of the internet activity is based on online advertising and product/service purchase. The nature of advertising is switching from mass communication to highly-tailored ads based on the interest and features of the target audience. Tailored advertising requires a process called “profiling” which is a data-intensive activity aimed at categorizing each user into pre-defined “profiles”. Companies and entities who gather and exploit data in an intensive manner are the one which are able to gain and sustain a key competitive advantage. Throughout intensive data exploitation, companies are able to better understand the interest and preferences of their customers. Moreover, thanks to profiling activities, whole new business models are being created based on personalized advertising.

In this thesis, we will analyze the big data environment under different aspects and with different purposes. The final goal is to understand the importance of big data in the digital economy, to analyze the issues and concerns which may arise, to explore possible fruitful integration with other technologies, and to learn about data monetization.

In the first chapter is introduced the big data environment, its definition, purposes, and key aspects. Moreover, the process of data gathering is analyzed together with an analysis of the main data giants. With respect to data giants, we will learn how big data allowed new business models to work properly and how companies convert user’s data in revenues.

The second chapter is more focused on threats and concerns regarding big data exploitation, we will analyze the “Cambridge Analytica” case study. Together with the issues we will discuss the main regulations which concerns big data and user privacy. We will adopt both a European and American perspective, in order to provide a full overview of the regulatory environment.

In the third chapter we will analyze the integration between big data and blockchain technology. To do that, we will devote the first sub-chapter to blockchain introduction and

¹ <https://www.statista.com/statistics/254266/global-big-data-market-forecast/>

definition. Consequently, we will discuss the main advantages that occur between the integration of big data and decentralized technology.

Finally, the last chapter will discuss about big data monetization and will analyze a case study of a British project called HUDI (Human Data Income). Such project aims at decentralizing the ownership of data by letting the data owners decide which information to share and which not to. We will discuss how big data can be used to generate a passive/active income and how the cryptocurrencies environment is linked to such activity.

Table of Contents

Abstract	2
Chapter 1: The big data environment	7
1.1 Big data definition	7
1.1.1. Types of big data	8
1.2 History and advantages of using big data	11
1.2.1 History of big data	12
1.2.2 Competitive advantage in using big data.....	13
1.3 How do companies collect data	16
1.3.1 HTTP Cookies and Web Beacons	17
1.3.2 Data selling market.....	20
1.3.3 Size of big data market	21
1.4 Data Giants	23
1.4.1 Meta (Facebook).....	23
1.4.2 Apple	25
1.4.3 Alphabet (Google)	27
Chapter 2: Improper use of data and new regulatory scenario	31
2.1 Data misuse	31
2.1.1 Cambridge Analytica case study: how not to use data	33
2.2 Big data regulation	38
2.2.1 Issues and features concerning big data regulation	38
2.2.2 The EU environment: General Data Protection Regulation (GDPR).....	41
2.2.3 The US environment: a fragmentated regulatory framework	46
2.2.4 Costs and implications of compliance with the regulatory environment	50
Chapter 3: Big data and blockchain technology: the ultimate combination.....	55
3.1 Blockchain 101	55

3.1.1 A simple definition	55
3.1.2 Basic elements of the blockchain (Nodes, Miners, Hash).....	56
3.1.3 How new blocks are added to the chain	59
3.1.5 A key distinction	61
3.1.6 Public vs Private blockchains	62
3.1.7 Consensus mechanism.....	64
3.1.8 Smart contracts, DApps and DAO	67
3.2 Integration of blockchain technology and big data	71
3.2.1 Storage and security issues	71
3.2.2 A fruitful integration	73
3.2.3 Real-life case studies of integration	78
3.2.4 GDPR regulation on blockchain-based data processing	81
Chapter 4: Data monetization, HUDI case study	84
4.1 The data monetization process	84
4.1.1 Business data monetization	84
4.1.2 Customer data monetization	86
4.2 The cryptocurrencies environment	89
4.2.1 Cryptocurrencies 101	89
4.2.2 ICO, IEO, IDO	93
4.3 HUDI – Human Data Income, data monetization	97
4.3.1 HUDI ecosystem	98
4.3.2 Roadmap and products	100
4.3.3 Security and GDPR compliance.....	103
4.3.4 HUDI token overview	104
4.3.5 DeFi functionalities	106
Conclusions	109
Bibliography	112

Sitography 114

Chapter 1: The big data environment

1.1 Big data definition

In this section we will introduce the definition of big data to better understand what does include and why it is gaining relevance in recent years.

We can think of the definition that the Journal of financial services provides for big data: “*Big data refers to large data sets, collected by firms and governments, that are so large and complex that traditional data processing methods are inadequate to deal with the calculations needed to make sense of the data.*”². This definition is pretty accurate; however, the truth is that we can find innumerable definitions for big data. Such topic is so complex and vast that finding the perfect definition would be an impossible work. While the definition of big data is not so clear cut, the features that define the topic are very objective and collectively accepted. Such features have been introduced by Doug Laney in a research work called “*3D data management: controlling data volume, velocity and variety*”³. In addition to the three original features of big data introduced by Laney, a fourth has been added, Veracity.

- **Volume:** refers to the quantity of data stored; such feature is crucially important to data collectors. In the past, the storage capacity of multiples entities such as businesses and governments were an issue and a barrier to big data. Nowadays such issue no longer exists, as new technologies have dramatically increased the storage capacity through new innovations such as clouds and powerful hard disk. The importance of volume in big data refer to all the instances in which data are non-normally distributed, skewed, non-linear or simply imbalanced. All such issues are commonly attributed to the size of the sample⁴ and therefore to volume⁵.
- **Velocity:** refers to the speed of data generation. Nowadays technologies allow data collectors to gather data at incredibly higher speed compared to few decades ago.
- **Variety:** refers to the type of data that is being collected. Examples of data types are text, audio, video, etc. While in the past only few types of data were easily collectible

² John E. Grable, CFP Angela C. Lyons, (2018), “*An introduction to big data*”, Journal of financial services professional.

³ Doug Laney, (2001), “*3D Data Management: Controlling Data Volume, Velocity and Variety*”, META group research.

⁴ The sample is defined as a portion of the population which is used to represent the whole population. Sample are often used because it would be impossible to analyze the whole population (surveying every citizen in a country).

⁵ Deepa N, Quoc-Viet Pham, Dinh C. Nguyen, Sweta Bhattacharya, B. Prabadevi, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, Fang Fang, Pubudu N. Pathirana, (2021), “*A Survey on Blockchain for Big Data*”.

(such as demographic, attitudinal, etc.) nowadays almost every type of data is easily collectible, processable and usable.

- Veracity: this feature was not originally included to the previous three and refers to the quality of data. Since data can come from heterogeneous type of sources, it is easy to collect noisy data⁶ which makes it difficult to separate good information from worthless one.

1.1.1. Types of big data

Before proceeding with further analysis, is necessary to give some definitions that will be useful to understand further topics.

Data gatherers: every entity which need to collect and process data to operate decisions for their businesses. Most of nowadays firms and governments are data gatherers, they exploit the data of users to better implement strategic decisions such as selling product and services.

Data analysis: is the process of examining data to find patterns, facts, relationships, insights, and trends inside the sample. The aim of data analysis is to support the decision-making process of data gatherer such as companies, governments, or other entities⁷. An example of data analysis is studying the relationships between ice cream sales and daily temperatures, in order to understand the consumer's behavior and predict future expenditure patterns.

Data analytics: is a broader term that encompasses data analysis, it includes the complete lifecycle of data which accounts for collecting, cleansing, organizing, storing, and analyzing data⁸. Data analytics include all the techniques of data analysis that allow data gatherers to process high volumes of information from heterogeneous sources. Through data analytics, one can retrieve key patterns that allow users to make factual-based decisions instead of simple guess based on past experience. There are four categories of data analytics that can be distinguished based on the result that they produce:

⁶ Noise in the data refers to the typical statistical issue of data gathering. It is usually caused by measurement tools used to gather data. Random noise in data is almost always inevitable. An example of noise in the data is the presence of "outliers" (observations extraordinarily far from the average of a distribution that if are not accounted for, can distort the data analysis process).

⁷ Thomas Erl, Wajid Khattak, Paul Buhler, (2015), "*Big data fundamentals*", Prentice Halls.

⁸ Thomas Erl, Wajid Khattak, Paul Buhler, (2015), "*Big data fundamentals*", Prentice Halls.

- Descriptive analytics;
- Diagnostic analytics;
- Predictive analytics;
- Prescriptive analytics;

Descriptive analytics is based on events that are happened in the past, this form of analysis contextualize data to generate information. The vast majority of analytics is of this type, that is because, some entities need to acknowledge past information to evaluate their performance. We can think of how firms define the variable remuneration of their managers, through achieving goals or pre-defined levels of productivity. Another example could be a car manufacturer that investigates where it is focused the core of its sales or again which of its products generate higher profits in terms of revenues minus costs. Such type of analytics is fundamental to entities because it allows to evaluate their past performance and gives an idea of where there is some place for improvement.

Diagnostic analytics is focused in finding the cause of a phenomenon that is happened in the past by inquiring the determinants of such fact. This type of analytics is essential to understand the key factors that determined an event in which the entity is interested. An example could be understanding why a productive division of a firm did not work as it was supposed to. Another example could be the investigation of two different performances in two years and the reasons that caused such difference. In most cases, diagnostic analytics is more useful than descriptive analytics because it does not just limit to analyze data to present an output, but it uses tools to investigate the reasons behind some phenomenon. By so doing, it provides useful information to solve some problematic situation or to keep enhancing some procedures that proved to be more efficient than others.

Predictive analytics is different from the aforementioned techniques because it focuses on different time lapses. While descriptive and diagnostic analytics focuses on past data, predictive analytics aim to forecast the outcome of an event that might occur in the future. Of course, no technique is able to exactly forecast the future, therefore, such models must rely on past data. In such a way, past data is gathered and processed to create models that try to forecast future events. However, since such models are based on past data, they must be periodically updated in order to account for most recent events. Due to its nature, predictive analytics requires more advanced skillset than diagnostic and descriptive ones.

To conclude, prescriptive analytics is based on the results of predictive analytics. More precisely, prescriptive analytics uses forecasts obtained with predictive analytics and add the necessary actions that need to be taken to implement pre-defined strategies. This type of technique can be used to gain advantage or mitigate risk. An example of quest that prescriptive analytics can try to solve is “Which is the best time to trade a particular stock?”. This type of analytics is the most useful and requires the most advanced skillset as well as specialized software and tools⁹.

The aforementioned techniques are needed to elaborate big data for different purposes. We must also consider that, since the type of analytics changes depending on the type of result that an entity wants to achieve, also the type of data available will imply different strategies. To this extent, we could distinguish between an infinite series of type of data (qualitative, quantitative, nominal, ordinal, etc.). However, since this section is about big data, we prefer to present the typical distinction between structured, unstructured, and semi-structured data. The most useful and rare type is structured data, which is already organized with dimensions defined by set parameters. Since this type of data is already organized, it is the simplest to work with. Usually, structured data is composed exclusively by numerical (or quantitative) variables, as an example we can think about an excel spreadsheet that represent the price of a stock in the previous six months. The data is already organized with temporal criteria, there are already columns with date, trading volume, closing price, etc. Structured data often conforms to a data model or scheme.

On the other hand, unstructured data represents roughly 80% of all available data in the market and is basically data which it has not yet been organized¹⁰. This data represents almost every transaction that people do with computers or phones. Since users tend not to organize their data but instead, they simply operate transactions (phone calls, audio message, website clicks, etc.) this data is more difficult to process. Examples of unstructured data could be audio, video, binary, text, and many others type of raw data.

Lastly, semi-structured data stands between structured and unstructured data, usually, this is the case of unstructured data with metadata¹¹ attached to it. An example of semi-structured data could be time, location, ID address, or email address (metadata) attached to others type of data¹².

⁹ Thomas Erl, Wajid Khattak, Paul Buhler, (2015), “*Big data fundamentals*”, Prentice Halls.

¹⁰ <https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/>

¹¹ Metadata provides information about a dataset’s characteristic and structure.

¹² <https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/>

1.2 History and advantages of using big data

It is common knowledge that nowadays, globalization and technology are dramatically changing the way that people communicate, travel, purchase and think. New technologies are rapidly increasing and changing people's need, as a consequence to that, firms and entities are facing more dynamic challenges in supplying customers with different product and services. Twenty years ago, it was highly unusual for someone to buy something from a different continent or country with no major complications. These days, customers have no issue in ordering products that are in other markets or simply that are not nearby their location. Product and service providers are now facing a whole new level of competitiveness. As an example, we can think of a firm that sells bikes in a specific region of a country. If some time ago, such firm faced the competition of bike sellers in that specific region and that specific country, now the competitors have exponentially increased. Today's customers can choose among multiple types of markets to buy any type of product or service. To stick with the previous example, a hypothetical customer can choose to go to the shopping center, to the local bike seller, to buy the bike online or also to rent the bike from a bike rental. Such conditions are caused by the increasing advancement in the technological field and from progressive globalization. Of course, the main driver of such revolution is internet, that triggered a domino effect on new other types of technologies based on internet and on other innovations. To cite only few of such new innovations, we can mention Artificial Intelligence (AI), Internet of Things (IoT), blockchain-based services, and many more. As a consequence, firms and other entities are facing the pressure to adapt and evolve in order to be able to sustain their business model through this technological revolution. To this end, big data plays a major role in sustaining the competitive advantage of firms and other entities in such periods.

1.2.1 History of big data

When we discuss about big data, it is normal to think about internet, mobile phones, computers, or other hi-tech instruments. In fact, the largest develop of big data market happened in recent years together with other technologies. Although, recent times provide more interesting point of views about big data, it is useful to briefly analyze the development of such sector through time to better understand the likely direction of future developments. One of the first concrete applications of big data dates back to 1880, when the U.S. Census Bureau estimated that it would take approximately eight years to process the data collected. Moreover, it forecasted that it would have taken ten years to process the data collected from the 1890 census. In such circumstance, a man named Herman Hollerith created the “Hollerith Tabulating Machine”, an invention based on “punch cards” which used inputs and outputs instead of raw information. Thanks to this innovative solution, the work that was expected to be done in eight years was finished in approximately three months¹³. Another fundamental example is the one of Social Security Act in 1937, when Franklin D. Roosevelt’s administration had to keep track of twenty-six million Americans and more than three million employers. Roosevelt’s administration charged IBM to develop punch card-based machines to help and sustain the gigantic bookkeeping work that was needed¹⁴. Later applications date back to World War II, when through “Colossus” (a data-process machine), the British were able to analyze and identify hidden patterns in huge amounts of apparently nonsense communications sent by the Nazis.

These are just few examples of “big data ancestors”, and as we approach to recent times, the number of examples and applications of big data increase exponentially. In 2005, Roger Magoulas coined the term “big data” referring to large quantities of data which are impossible to deal with if not with the appropriate technology.

¹³ <https://www.dataversity.net/brief-history-big-data/#>

¹⁴ <https://datafloq.com/read/big-data-history/239>

1.2.2 Competitive advantage in using big data

As we stated in the previous section, companies are facing a whole new level of competitiveness and if they do not join the digital revolution, they risk being defeated by more advanced competitors. The adoption of more “smart” solutions in the business model of a firm is becoming more and more a necessity in recent times. To this extent, big data provides the opportunity to join and adapt to such digital revolution in multiple ways. Moreover, the adoption and use of big data technologies also constitute the basis for creating a competitive advantage with respect to other firms that do not operate such solution. It can be therefore stated that big data enable organizations to increase their value and grow faster. In an interesting paper published by the McKinsey Global Institute, are defined five areas in which big data can create value with respect to firms and other entities¹⁵.

The first field which can be improved by big data is transparency. Improved transparency can increase a firm’s value in multiple ways, stakeholders will appreciate the largely available amount of data and will be more likely to support/invest in the company/entity. An excellent example is the public sector, if citizens were able to get, analyze, and use data in a simpler, more transparent way, this would lead to the reduction of search and process time that would translate in a reduction of costs. Another example could be the increased transparency of a public company towards corporate social responsibility. If such firm is able to get vast amount of unstructured data and to convert it into clear and readable information, that would largely improve the image of the company, thus increasing the value of it.

Another area that can be improved substantially through the use of big data is performance experimentation. As we know, companies are entities in which different types of stakeholders have interests. By using big data, a firm can analyze different aspects of its performance and provide an objective evaluation to its stakeholders. For instance, a company can analyze data from its inventory levels and investigate the presence of variability of inventory amounts in certain periods of the year. Or it can analyze in which periods its personnel is more likely to be ill in order to try and forecasts periods with less working force. These types of analysis are useful not only to evaluate the type of variability or performance, but for instance, they can be used to link the salary of a manager to the firm’s performance in a specific quarter. Or also, variability investigation can be useful to understand the root of a specific problem and try to solve it in the most efficient way.

¹⁵ McKinsey Global Institute, (2011), “*Big data: The next frontier for innovation, competition, and productivity*”

The key area that can be improved in the business environment is population segmentation. This area is gaining enormous importance in recent years due to the strong necessity of user-tailored ads and services. Big data is the primary ingredient for this type of segmentation analysis, once a company/public entity has the necessary data to analyze, it can perform various types of investigation to achieve its goal. In most of the cases, data is used to create different segment of population (population segmentation) which will be then targeted by specific customized actions. Most of this area is exploited by marketing and advertisement services providers. Once data on some defined population has been collected, categories or user profiles are created, and each element of a population is assigned to a specific profile. For instance, a company may be able to gather data of a population of a specific region, after such data has been analyzed, cleansed, and organized, each user is assigned to a profile. Examples of users' profiles can be athletic people, highly educated people, students, politically interested people, etc. Then, once every user has been assigned to a profile, he will be targeted with specific ads and marketing announces based on its "features". For instance, in the profile of "athletic users" will be more frequent advertisements of gym equipment, low-calories food, marathon programs, etc. In such a way, the likelihood that a user will click on the ad and buy a specific product/service will be much higher than in other cases. This type of data is usually gathered through cookies and is now representing one of the most interesting applications of big data for companies. This type of application of big data can increase the firm's value in an exponential way.

The McKinsey Global paper¹⁶ mentions another area which can be improved by the use of big data, that is, the replacement of human decisions with automated algorithms. Specific types of analysis can allow to minimize the risk and to improve decision-making by introducing data-driven automated algorithms. The application of this technology can be in every field, from a tax agency that can use automated risk engines to flag candidate for further examination, to retailers that can use automated systems to optimize decisions related to inventory levels. The application of automated algorithms overcome the problem of human error that is always possible when an employee must manually check some type of complex information. Of course, in some fields, automated algorithms cannot totally substitute the human action, however, they can prove to be very efficient and time-saving solutions to pair with human work. Such automated algorithms are feasible only with large amounts of data

¹⁶ McKinsey Global Institute, (2011), "*Big data: The next frontier for innovation, competition, and productivity*"

that can serve as a base in which one can search for patterns and trends. An example of application of such technology is the recent development of “robot-trading” which is based on automated analysis of some assets carried out by complex algorithms. Such analysis is then used to invest in stocks, bonds, or other financial assets or to provide “signals” to other users upon payment.

Lastly, the fifth area mentioned by the McKinsey paper¹⁷ is business model, product, and service innovation. Big data can be used to analyze the appreciation, problems and key features of a specific product or service that has been sold in the market. Manufacturers can exploit such information to enhance their product line or service portfolio by trying to correct eventual problems discovered during the analysis. Such mechanism can be exploited on an even higher level, because based on the product/service improvement process, can be built a whole new type of business models. Big data can be useful also to create after-sales services, based on the feedback received by clients. In such a way, a company can be able to change its strategic position towards the competition and to gain additional market shares.

The paper of McKinsey Global (2011) also argues about how firms will be able to gain and sustain their competitive advantage mainly thanks to the use of big data. They state that big data will be the main driver of the digital revolution. Firms are expected to invest and develop organizational models based on the collection and exploitation of big data in order to maximize efficiency and market penetration. The paper also mentions that these kinds of investments will require time and resources, but the firms that will not adapt to the digital revolution will be “left behind”.

Today, ten years after the publication of that paper, we cannot disagree with such statements as we are approaching a new type of competitive market mainly driven by the intensive exploitation of big data. Nowadays companies invest in technologies to gather huge amounts of data or pay specialized companies in such sector (data providers) to get the data. Moreover, the digital revolution changed dramatically the core business of many firms. Because as we will see, many big players who based their business models in a specific sector in which they are able to collect huge amounts of data are progressively switching the main sources of revenues from other fields to data sales.

¹⁷ McKinsey Global Institute, (2011), “*Big data: The next frontier for innovation, competition, and productivity*”

1.3 How do companies collect data

Until now we underlined the key role that big data plays in companies' environment, especially when we talk about competitive advantage. In this section we will analyze what are the main sources of data for companies, and how firms can gather data about users, customers, population, etc. This process represents a crucial point in the strategy of a company because the quality of the data, will define the quality of the output and analysis that will be generated afterwards.

First of all, we can define different types of data collected by companies, these categories of data differ from the ones defined in section 1.1 (Introduction to big data) because we are now analyzing different aspects of big data.

We can distinguish four main categories of consumer data¹⁸:

To begin with, personal data includes personally identifiable data such as, social security number, gender, age, etc. It can also be included non-personally identifiable information such as IP address, telephone number, browser cookies, etc. This type of data is crucially important for data gatherers, that is because it is the type used to carry out the so-called population segmentation technique discussed in the previous section.

Conversely, engagement data refers to all data that define how users interact with a specified website, social media, customer service, app, etc. This type of data is fundamental to understand how broad the audience for a specified entity is and how many users can be reached with a specified instrument. This category has gained substantial importance in recent years with the diffusion of social media that allows companies to create public profiles for their businesses. With such tools, companies can post and advertise their product and services on social media and can check in real time the number of accounts reached with their posts (impressions, tap, etc.).

Thereafter, behavioral data includes data which is slightly more complex to get and analyze. Behavioral data refers to a broad set of information such as purchase history, product usage information, and qualitative data. This information is then used to create profile of users in order to segment the market and to provide specifically tailored ads and services to different categories.

¹⁸ <https://www.businessnewsdaily.com/10625-businesses-collecting-data.html>

Finally, attitudinal data includes metrics on consumer satisfaction, post-sales feedback from clients, purchase criteria schemes, etc. Still, together with personal and behavioral data, this information is used to profile each user in order to create customer segmentation.

Now that we briefly analyzed the different types of data that can be collected by data gatherers, we will discuss the main ways in which data can be retrieved.

Firstly, companies can obtain data on its customers by simply asking for it¹⁹. This method is mainly employed through a series of tools such as questionnaires, login procedures, or feedback reviews. This type data collection is more efficient with users that do not worry too much about their data privacy or trust the party which is gathering information. A typical example is the fulfillment of a questionnaire which asks specific questions such as the gender, status, employment, age, etc. Another example is the typical login procedure that is required to use some functions of a website. For instance, in most of the cases, to order an item from a website a new registration will be required, and the data gatherer will acquire our personal data.

Another way in which companies can retrieve data, is through tracking customers during their web browsing. This procedure requires, in most of the cases, the use of HTTP Cookies, and Web Beacons²⁰.

To conclude, the third way in which companies can get data on their customers is by buying it from data providers. In the last decade the market of data providers increased enormously, that is mainly due to the increasing relevance of big data for companies and entities.

1.3.1 HTTP Cookies and Web Beacons

Cookies first appeared in 1994 when Lou Montulli, a 23-years-old engineer and Netscape²¹, invented a way for websites to remember users and facilitate their experience while browsing²². This feature was initially named “Magic Cookies” and allowed the user to resume interaction with a website at the point in which it was left in the previous session. This

¹⁹ <https://www.businessnewsdaily.com/10625-businesses-collecting-data.html>

²⁰ Janice C. Sipior, Burke T. Ward, Ruben A. Mendoza, (2011), “*Online Privacy Concerns Associated with Cookies, Flash Cookies, and Web Beacons*” Journal of Internet Commerce.

²¹ A popular web browser during the ‘90

²² <https://qz.com/2000350/the-inventor-of-the-digital-cookie-has-some-regrets/>

allowed websites to remember the shopping chart and other preferences of users. This is possible thanks to the exchange of small text characters from the website server to the user hard drive and back when the user revisits the site. This exchange allows the *stateless hypertext transfer protocol* (HTTP) to retain state information²³. That is why they are also called HTTP cookies or server cookies.

Without the use of cookies, a web browser can establish a connection with a web server, however, when the connection is closed (the user exits the website) the server retains nothing concerning the session. Therefore, the use of cookies enables the server to remember information regarding the session on the website by assigning a unique ID to the user. To this extent, cookies serve multiples purposes. Firstly, they can be used for session management, that means that every time that we open a website that requires a login, the cookies enable our browser to remember the username and password for a quick access. Moreover, it enables the website to remember our preferences and settings for the session (layout, language, etc.)²⁴.

The second, and more important, purpose of cookies is personalization. In fact, whenever we visit a website for buying, reading or even play a game, if marketing cookies are active, the server will remember our preferences and use them to build personalized ads. That is the reason why, for instance, after we look for a bracelet that we might want to buy, we will be targeted by ads on bracelets for a period.

The third, and last purpose of cookies is tracking. This function allows websites to remember previous transactions of a user in order to propose new purchases or suggest items to put in the shopping chart.

HTTP cookies can be of two types, session cookies and persistent cookies. Session cookies are used only when navigating a website and they are stored in random access memory (RAM), never in the user's hard disk. When the session ends, cookies are deleted in order to maintain privacy. On the other hand, permanent cookies remain on a computer indefinitely, and are much more difficult to delete (Ex: flash cookies). The main aim of permanent cookies is authentication (session management) and tracking²⁵.

²³ Janice C. Sipior, Burke T. Ward, Ruben A. Mendoza, (2011), "Online Privacy Concerns Associated with Cookies, Flash Cookies, and Web Beacons" Journal of Internet Commerce.

²⁴ <https://www.kaspersky.com/resource-center/definitions/cookies>

²⁵ <https://www.kaspersky.com/resource-center/definitions/cookies>

Depending on the browser that is used, non-essential cookies are usually rejectable. However, web browsing experience can be impacted by doing so or even, some websites will not allow access to the user unless he accept all cookies.

While cookies exchange text information between the user's browser and the web server; Web beacons are small graphic data also known as clear GIFs, that recognize user activity²⁶. Web beacons are usually the size of a pixel and therefore, are practically invisible to human eye. They are usually delivered through a web browser, or an email and they tag the user and transfer information to the sender. For instance, a typical application is to send a web beacon attached to an email, so that the sender will be able to know when the recipient will open the message and on which device the message will be read. Web beacons are commonly used together with cookies in order to provide a better browsing performance and retrieving data on the user.

Until now we introduced cookies and web beacons as instruments provided by websites aimed at enhancing the browsing experience. These instruments simplify the life of users by remembering crucial data such as username, password, or the items that were previously added to the shopping chart. However, we must also consider that not all cookies are much safer, and customer oriented as one might think. To elaborate, we must distinguish between first-party cookies and third-party cookies. First-party cookies are directly created by the website the user is browsing in. Therefore, as long as we are navigating on safe and reliable websites we should not worry too much. On the other hand, third-party cookies can be more dangerous. Third-party cookies are created by websites different from the one the user is browsing in. That means that if we are browsing on a website with 5 ads, we can generate 5 different cookies even if we do not click on the ads. Third-party cookies are exploited by advertisers for tracking down customers that navigate on websites that contain those ads. One types of third-party cookies are the so-called "zombie cookies" or flash cookies. Those are permanently installed on the user's computer even if he chooses not install cookies.

²⁶ <https://www.ntt.com/en/about-us/hp/webbeacon.html>

1.3.2 Data selling market

As we stated at the beginning on section 1.3, the third way in which a company can get data on customers is by buying it from data sellers. The data market increased exponentially with respect to ten years ago (see next section for figures), and the growth is expected to persist for future periods. The main fuel of such increase is, of course, the increasingly importance of internet-related services and social media. To cite few statistics from 2021, Google gets over 3.5 billion searches daily²⁷, WhatsApp users exchange up to 65 billion messages daily²⁸, in 2020 every person generated 1.7 Megabyte per second²⁹, data interactions went up by 5000% between 2010 and 2020³⁰.

With such high figures, it is natural that companies that want to survive the competition will gather all possible data on their customers. And for those companies that for the nature of their business have no or few access to user data, they can rely on the supply of data from third parties.

Apart from the popular data giants such as Facebook, Google, etc., the main sellers of data in the market nowadays are not so popular as one might think:

- Acxiom: it is one of the largest data companies in the world, it began operating in 1969 under the name of “Demographics”, its main purpose was to gather data for political uses. Later on, it focused its business on collecting data from people and process it to make it usable and sellable for marketing purposes. Nowadays Acxiom claims to possess data on nearly all US households, and this data is said to be used to make 12% of US’s marketing direct sales³¹.
- Nielsen: in the data market since 1923, it specialized in market research and ratings. It became popular for its pioneering role in the audience measurement techniques for TV. Moreover, it provides statistics on US consumer behavior, it operates in 100 countries. It is one of the largest data gatherers in the market worldwide with annual revenues of \$4.27 billion in 2021³².

²⁷ Internet live stats, (2021), <https://techjury.net/blog/big-data-statistics/#gref>

²⁸ Connectiva Systems, (2021), <https://techjury.net/blog/big-data-statistics/#gref>

²⁹ IBM, (2021), <https://techjury.net/blog/big-data-statistics/#gref>

³⁰ Forbes, (2021), <https://techjury.net/blog/big-data-statistics/#gref>

³¹ <https://bernardmarr.com/where-can-you-buy-big-data-here-are-the-biggest-consumer-data-brokers/>

³² <https://bernardmarr.com/where-can-you-buy-big-data-here-are-the-biggest-consumer-data-brokers/>

- DataSift: it is smaller compared to other data sellers, but it plays a major role when we consider social media data. DataSift specialized in human data intelligence and gather data from more than 20 sources such as Twitter, Facebook, Instagram, etc. With its services it helps companies to acquire and exploit social media data to become more visible and attract new customers³³.
- Corelogic: it focuses its business model on gathering data to sell to financial institutions. The main focus concerns the mortgage industry and the real estate sector. Financial institutions such as banks, purchase data from Corelogic to have more information on their customers³⁴.

1.3.3 Size of big data market

To properly understand the relevance of big data market and how is it changing the dynamics of companies, it is necessary to make an estimate of the size of such market. Moreover, it is interesting to analyze how companies such as Facebook, Apple, and Google use customer's data.

According to Statista³⁵, The global big data market is expected to grow up to \$103 Billions in revenues by 2027, while the figure for 2022 is expected to be \$70 Billions. These are astonishing numbers if we compare them with the nearly \$7.6 Billions of revenues of 2011. All the money that companies spend to buy and use all these data are aimed are creating specifically tailored ads, profiling customers, etc.

The interesting part is that most of internet users is not aware of how they contribute to this market, how their data is used, and how much do they “pay” with their data. However, the younger generation is showing signs of increasing awareness regarding their data security. In fact, in a survey by CISCO, 61% of individuals which are active about their privacy are under the age of 45³⁶. Moreover, there are some signs of awareness in the social media sector, in which 79% of users have adjusted their privacy-related settings on their social media

³³ DataSift LinkedIn profile and <https://bernardmarr.com/where-can-you-buy-big-data-here-are-the-biggest-consumer-data-brokers/>

³⁴ <https://bernardmarr.com/where-can-you-buy-big-data-here-are-the-biggest-consumer-data-brokers/>

³⁵ <https://www.statista.com/statistics/254266/global-big-data-market-forecast/>

³⁶ Cisco cybersecurity series 2019 - consumer privacy survey

accounts³⁷. Nonetheless, still a lot of users is not sufficiently aware of the implicit cost they are paying when they transmit their private data. In fact, Blis, a digital advertising company carried out a survey on 2000 US adults in November 2018 regarding data privacy and awareness. They found out that 83% of users were not aware of companies tracking their locations and the majority of consumers would charge at least \$10 to access their personal data³⁸. These statistics suggest that people worldwide are not yet completely aware of the importance of their own data. Furthermore, they are not aware of how much money data giants make thanks to it.

³⁷ DuckDuckGo Privacy research

³⁸ <https://martech.org/83-percent-of-consumers-now-aware-of-marketers-tracking-their-locations-study/>

1.4 Data Giants

In this section we will discuss some brief cases of companies that possess huge amounts of data and built their success with a data-driven business model. The main aim of this section is to explain how much big data can be valuable to companies, and why today's entities are willing to allow discounts to users who grant access to their personal data. The interesting point is that, more companies than one might think have built their hegemony thanks to big data. Moreover, the only way in which a company can maintain its strategic advantage is by continuing to exploit its data resources.

1.4.1 Meta (Facebook)

The first and most relevant example that should come to mind when we talk about data giants is Meta (also known as Facebook). Meta was born in October 2003 under the name of "FaceMash", it changed its name to Facebook in 2004, and changed it again to Meta in 2021. As almost everybody knows, Facebook is a social media platform that consent people to chat with each other, share posts, photos, ideas, and whatever content that can take the form of text, image, or video. Moreover, among its most relevant acquisitions, Meta acquired Instagram in 2012 and WhatsApp in 2014. It is now estimated that the overall number of users of Meta is around 3 billion, which is almost half the size of the world's population. The interesting part of Meta is about its business model, because as everybody know, the services that Meta offers are very valuable (connecting with people, video editing, document exchange, etc.). However, Meta does not ask any type of payments to its users, nor for Facebook, Instagram, or WhatsApp. So, how does Meta earn money and sustain its business model? The answer is, mainly through advertisement which is highly effective thanks to the enormous amount of data available to Meta. To prove that, it is sufficient to check Meta's Income statement and earning reports. Firstly, we can download Meta's income statement for Q3 of 2021 and check the amount of total revenues.

FACEBOOK, INC.			
CONDENSED CONSOLIDATED STATEMENTS OF INCOME			
<i>(In millions, except for per share amounts)</i>			
<i>(Unaudited)</i>			
	Mar 31, 2021	Jun 30, 2021	Sep 30, 2021
Revenue	\$ 26.171	\$ 29.077	\$ 29.010

39

As can be noticed, at the end of Q3 total revenues for Meta are about \$29 Billion. To check the share of revenues that comes from advertisements we must look at the closing report for Q3 2021.

In millions, except percentages and per share amounts	Three Months Ended September 30,		Year-over-Year % Change
	2021	2020(1)	
Revenue:			
Advertising	\$ 28,276	\$ 21,221	33%
Other	734	249	195%
	29,010	21,470	35%

40

The report suggests that at the end of Q3 2021, advertising expenses constitute around 98% of Meta's revenues, with the remaining 2% composed of the Oculus VR device⁴¹ sales and payment fees from developers.

Moreover, if we want to analyze the revenue composition through annual report (and not quarterly) we can look at Meta's 2018 Form 10-K filed to SEC. At page 64 section "Revenue Recognition" we can observe the percentage of revenues that comes from advertising at the end of 2016, 2017, 2018.

³⁹ Image 1: <https://investor.fb.com/financials/default.aspx>

⁴⁰ Image 2: https://s21.q4cdn.com/399680738/files/doc_news/Facebook-Reports-Third-Quarter-2021-Results-2021.pdf

⁴¹ In 2014 Meta acquired the "OCULUS VR" start-up for developing its project in the Metaverse.

	Year Ended December 31,		
	2018	2017 ⁽¹⁾	2016 ⁽¹⁾
Advertising	\$ 55,013	\$ 39,942	\$ 26,885
Payments and other fees	825	711	753
Total revenue	\$ 55,838	\$ 40,653	\$ 27,638

(1) As noted above, prior period amounts have not been adjusted under the modified retrospective method.

42

As can be easily noticed, advertisement represented the quasi-totality of Meta’s revenues through years. More precisely, advertisement contributed for 97% in 2016, 98% in 2017, and 99% in 2018 on the total revenues. This brief analysis suggests that Meta does not directly sells user’s data to third parties. However, Meta uses its huge amount of data to profile and create precise categories of users. Thanks to that, it is possible to place an advertisement on its platform to specifically tailored customers, depending on the necessity.

For instance, a company that sells books can pay Meta to acquire an advertisement spot which will be shown to people that is interested in books. That specific category of users could be interested in culture, reading, academic articles, etc. Such profiling is created by Meta thanks to the enormous amount of data, which is available on its users such as interests, likes, age, gender, job, etc.

The take-home message here is that the biggest resource of Meta is its data. Moreover, users should carefully consider that whenever they accept “terms and conditions” to freely join a platform such as Facebook or Instagram, they are not really joining for free, they are paying with their data.

1.4.2 Apple

Apple was founded by Steve Jobs and Steve Wozniak in 1976 under the name of Apple Computers⁴³. The main goal of the company was to produce a computer small enough, to be brought at home, and which was sufficiently user friendly. After years of ups and downs the company launched the first iPhone in 2007, which turned out one of the most successful products in the mobile industry worldwide. Apple released several versions of the iPhone, iOS (the operating system of apple devices), iPad, and many other products. In 2015 Apple

⁴² Image 3: https://s21.q4cdn.com/399680738/files/doc_financials/annual_reports/2018-Annual-Report.pdf

⁴³ <https://guides.loc.gov/this-month-in-business-history/april/apple-computers-founded>

launched the first Apple Watch, which later revealed to be very important for its data collection process.

Although Apple was enjoying a strong success after the first launch of the iPhone, its data exploitation process was too outdated compared to its competitors. Therefore, during recent years, Apple carried out an impressive work to level up with the competition, and the results were incredible. Although Apple is very secretive about its data management process, some information is publicly available for analysis. Apple exploits data in a different way than Meta (Facebook), that is because its business model is based on the sale of real products that generate a huge stream of cash flows. Nonetheless, the success of Apple's products is mainly due to the huge exploitation of big data. The first field in which Apple exploits big data analytics is application design. Apple analyzes the way in which its customers use the apps and by doing so, it is able to design and upgrade them to make them more user-friendly⁴⁴. Moreover, after the introduction of the Apple watch in 2015, the data collection process of Apple stepped to another level. Through the Apple watch, and in a collaboration with IBM, Apple was able to project health-care-related apps with new functionalities. For instance, if a customer wears its Apple watch during the night, he will be able to monitor and schedule for achieving an optimal sleep level. This will also allow Apple to gather data on millions of users and use it for enhancing some functions. Such functions could be the time at which is optimal to send notification or the suggested time to set an alarm in the morning. Another new feature introduced by the Apple watch is the automated emergency call in case of "hard fall". Thanks to the health data that Apple can gather and analyze, the iWatch can detect whether the user have taken a dangerous fall. In case of no movement, it will automatically contact the emergency numbers and send a GPS location to receive help⁴⁵. Finally, another functionality of Apple that massively exploits and produces big data is Siri. Siri is the voice-controlled personal assistant of Apple, through vocal commands is able to answer questions or execute functions such as "call mom". Through Siri, the voice data captured by the machine is uploaded to its cloud analytics platforms, which compare them alongside millions of other user-entered commands. This will allow Siri to become better at recognizing speech patterns, and more accurately match users to the data they are seeking⁴⁶. It must be noted that Apple keeps the vocal data for 2 years disassociated from the real identity of the user.

⁴⁴ <https://www.analyticssteps.com/blogs/how-apple-uses-ai-and-big-data>

⁴⁵ <https://support.apple.com/it-it/HT208944>

⁴⁶ Bernard Marr, (2016), "*Big Data in practice*", Wiley.

From this brief analysis we can already understand how important big data has become for Apple. First of all, we can notice how Apple exploits data in a very different way than Meta (Facebook). Apple does not sell its users data to third parties, not directly, nor indirectly. Instead, it uses the data to project and design more effective apps and functionalities for its products. We can fairly say that, thanks to big data Apple has been able to diversify from the competition, by providing specifically tailored functions to customers, based on their experience.

1.4.3 Alphabet (Google)

To conclude this overview on the main “Data Giants” that built and maintained their success thanks to big data, we should talk about Alphabet (referred to as Google for simplicity). The popular company was founded in 1998 by Larry Page and Sergey Brin who met at Stanford University in 1995. Google focused on internet-related services such as cloud computing, search engines, software, etc. The company went public in 2004 as Larry and Sergey own 13% of the company. However, their stocks have increased voting powers in order to allow them to keep the majority of the votes. In 2015 the company was reorganized, and Alphabet Inc. was created in order to legally own google as its subsidiary⁴⁷.

If we refer to Google search engine, anyone who carry out a web search, is actually manipulating big data. Google’s index is expected to contain 100 petabytes (100 million gigabytes) worth of data which is easily available to all users with an internet connection. Google’s goal was to overcome the problems of web browsing in such a vast amount of data. Already in Stanford, Larry and Sergey created an algorithm called “PageRank” which established that the higher the number of pages that link to a specific page, the higher will be that page “authority”. Such algorithm has been used later on for Google search engine. The higher the number of links of a page with others, the greater will be the authority, and the more visible will be the page. Such principle is basically transforming unstructured data (the content of web pages) into structured one. Google exploits the huge amount of big data which possess to ease the web searching process for its users. Instead of randomly wandering the internet in search of a specific information, Google allows users to organize the main data available based on popularity and reliability of the pages. Moreover, Google built its web

⁴⁷ <https://about.google/our-story/>

indexes by using the so-called “spiders”, a software robot which gather all type of information from web pages and transfer it to its archive⁴⁸. The pros of having all the information in just one archive is that the research process will be far quicker. Along with its intensive data-driven search engine, Google also tracks the searches of its users and collect data on preferences, interests, age, gender, etc. This huge amount of information is then used to profile the customers and then to sell advertisement spots to companies who wish to get visibility. As Meta, Google offers the vast majority of its services freely, therefore, to sustain its business model, it needs a way to generate revenues. To check where Google’s revenues come from, we can take a look at the 10-K form deposited at SEC for Alphabet Inc. At page 32 we find the executive overview of the income statement for years 2019 – 2020.

Executive Overview

The following table summarizes our consolidated financial results for the years ended December 31, 2019 and 2020 (in millions, except for per share information and percentages).

	Year Ended December 31,	
	2019	2020
Revenues	\$161,857	\$182,527
Increase in revenues year over year	18 %	13 %
Increase in constant currency revenues year over year	20 %	14 %
Operating income ⁽¹⁾	\$ 34,231	\$ 41,224
Operating margin ⁽¹⁾	21 %	23 %
Other income (expense), net	\$ 5,394	\$ 6,858
Net Income ⁽¹⁾	\$ 34,343	\$ 40,269
Diluted EPS ⁽¹⁾	\$ 49.16	\$ 58.61

⁽¹⁾ Results for 2019 include the effect of the \$1.7 billion EC fine. See Note 10 of the Notes to Consolidated Financial Statements included in Part II, Item 8 of this Annual Report on Form 10-K for further information.

- Total revenues were \$182.5 billion, an increase of 13% year over year, primarily driven by an increase in Google Services segment revenues of \$16.8 billion or 11% and an increase in Google Cloud segment revenues of \$4.1 billion or 46%. Revenues from the United States, EMEA, APAC, and Other Americas were \$85.0 billion, \$55.4 billion, \$32.6 billion, and \$9.4 billion, respectively.
- Total cost of revenues was \$84.7 billion, an increase of 18% year over year. TAC was \$32.8 billion, an increase of 9% year over year, primarily driven by an increase in revenues subject to TAC. Other cost of revenues were \$51.9 billion, an increase of 24% year over year, primarily driven by an increase in data centers and other operations costs and content acquisition costs.

49

As we can notice the total revenues for 2019 are around \$161 billions while for 2020 the amount is \$182 billions. To better check where does these figures have been generated and more precisely, thanks to which service, we shall look at the revenue breakdown scheme which is available at the same link.

⁴⁸ Bernard Marr, (2016), “*Big Data in practice*”, Wiley.

⁴⁹ Image 4:

https://www.sec.gov/Archives/edgar/data/1652044/000165204421000010/goog20201231.htm#id55be7992b374e1a9a2bc48887ddb3f_4

Financial Results

Revenues

The following table presents our revenues by type (in millions).

	Year Ended December 31,	
	2019	2020
Google Search & other	\$ 98,115	\$ 104,062
YouTube ads	15,149	19,772
Google Network Members' properties	21,547	23,090
Google advertising	134,811	146,924
Google other	17,014	21,711
Google Services total	151,825	168,635
Google Cloud	8,918	13,059
Other Bets	659	657
Hedging gains (losses)	455	176
Total revenues	\$ 161,857	\$ 182,527

Google Services

Google advertising revenues

Our advertising revenue growth, as well as the change in paid clicks and cost-per-click on Google Search & other properties and the change in impressions and cost-per-impression on Google Network Members' properties and the correlation between these items, have been affected and may continue to be affected by various factors, including:

- advertiser competition for keywords;
- changes in advertising quality, formats, delivery or policy;

50

Under “Financial Results” we can observe the revenues breakdown for years 2019 – 2020. As anticipated before, advertisement produces the biggest part of Alphabet revenues for both 2019 and 2020. More precisely, “Google advertising” revenues amount at \$134 billions in 2019 which accounts for 83% of total revenues. In 2020 revenues from advertising amount at \$146 billions which represents 80% of total revenues. By looking at such figures we can finally understand the relevance of big data for Google. Without the tailored profiling of users which is possible thanks to huge amount of data available to Google, advertisement would be far less effective. Thus, Google would not be able to profit and earn revenues necessary for sustaining its business model. To better comprehend the range of advertisement services that Google provides to companies we can check out a statement in the aforementioned SEC 10-K form. At page 59 under the heading “Advertising revenues” we find:

“We generate advertising revenues primarily by delivering advertising on Google Search & other properties, including Google.com, the Google Search app, Google Play, Gmail and Google Maps; YouTube, and Google Network Members’ properties.

Our customers generally purchase advertising inventory through Google Ads, Google Ad Manager and Google Marketing Platform, among others⁵¹.”

To conclude this topic, few considerations are worth mentioning. Firstly, this section was not about describing some big companies in the market. The take-home message here is how

⁵⁰ Image 5:

https://www.sec.gov/Archives/edgar/data/1652044/000165204421000010/goog20201231.htm#id55be7992b374e1a9a2bc48887ddb3f_4

⁵¹ https://www.sec.gov/Archives/edgar/data/1652044/000165204421000010/goog-20201231.htm#id55be7992b374e1a9a2bc48887ddb3f_106

such companies have exploited their resources to achieve success. The common thread of such companies is that they all have gained enormous benefits from big data analytics. For companies such Meta and Alphabet, big data is the tool that allowed their business model to work since advertising is their main source of revenues. For Apple, revenues are not directly generated by big data but that it is the main source of their competitive advantage. Without big data analytics Apple would probably lose an important market share of users which pay a higher price for having a “premium” product. Moreover, if we search for the “Big Five” tech companies in the world we will find articles talking about GAFAM (Google, Apple, Facebook, Amazon, Microsoft)⁵². Of course, is no coincidence that among the world’s five biggest tech companies we find data giants. That is, all five companies⁵³ are extremely successful and they all exploit big data analytics to gain market share or sustain their business model.

⁵² <https://growthrocks.com/blog/big-five-tech-companies-acquisitions/>

⁵³ Amazon and Microsoft are other examples of data giants.

Chapter 2: Improper use of data and new regulatory scenario

In the previous chapter we introduced the concept of big data, analyzed the history and advantages of using big data and how companies can gather and use data. In this chapter we will focus on issues regarding the improper use of data. In the first part we will analyze a case study which regards data breaches and unfair behavior of companies collecting user's data. We will do so to understand how crucial it can be for a business to get and exploit big data, and how some companies have broken the law trying to access user's data improperly. This first part will naturally lead to the second section of this chapter which will concern data privacy regulation. In nowadays dynamic scenario, data privacy and safety are becoming a first-tier topic for both businesses and people. Therefore, we will go through the main aspects of the topic and analyze which have been the main implications of the regulation.

2.1 Data misuse

In an always more progressively data-driven world as ours, big data represents one of the most important resources. As briefly introduced in the previous chapter, data can be used for both good and bad purposes. The distinction between bad and good can depend on whether or not the information is used to facilitate one's life and improve its digital actions. For instance, when first-party cookies are used to remember which were the items in a user's shopping chart, that is a good purpose. On the other hand, when data is gathered without the user's knowledge for commercial/cultural purposes, that is a bad one. To this end, we will make some examples of improper uses of data.

A first possible misuse of data is improper profiling, it is a technique that we already discussed in the previous chapter. It consists in the collection of people's data for unfair purposes or without the knowledge of the user⁵⁴. When a user is browsing the web or visiting an app, he may be unconsciously profiled without its permission. The real issue concerns the

⁵⁴ <https://irishtechnews.ie/5-ways-big-data-gets-misused/>

type of data that may be collected from him. For instance, it could be created a dataset of ill people or a list of names of previously convicted peoples. Such information could then be collected by the so-called data brokers and sold to companies for commercial and strategical purposes.

A 2015 Google Ads study⁵⁵ revealed a second improper use of data. It seemed like Google were presenting discriminating online ads based on the user's gender. More precisely, the advertisement showed different job ads based on the gender of the user. Men were showed ads of higher salary jobs and for senior positions, more often than women⁵⁶. Such discrimination is possible only thanks to the exploitation of huge amount of data to profile user's activity.

Another possible issue of collecting user's data could be inaccuracies and collection errors⁵⁷. This type of issue can appear less dangerous, especially when it concerns data which is not important. As an example, web cookies can collect information from your mobile browsing about a car you searched for. However, that car may have been searched by a friend of yours using your mobile. When this type of issue concerns more relevant data such as health data, or private data, the situation can become more dangerous. There are reported cases of data-driven errors that caused serious problems such as, welfare cuts in the U.S.

Another very important issue regarding big data is cyber-attack and data breaches⁵⁸. In nowadays world, companies control and store their data in local or in-house data centers while others exploit third-parties services such as clouds. The real issue is, companies do not only possess their own private data (concerning the sole company, such as strategic plans or accounting files) but they also have data on their users. When a company is attacked by a group of hackers and some data is leaked, also user's data become relevant as they can try and sell it to the black market for unethical purposes.

Finally, another bad example of exploiting user's data is social and political manipulation. Even though it can appear less dangerous than the other examples, this one can represent a serious threat to society. When organizations possess huge amounts of data on a large share of population, they can influence the choice that individuals make. Big data providers have the tools and skills to fetch, analyze, prepare, and use the right data to build algorithmic models to influence the decisions that you may have to make. These decisions can regard

⁵⁵ <https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html>

⁵⁶ <https://irishtechnews.ie/5-ways-big-data-gets-misused/>

⁵⁷ <https://irishtechnews.ie/5-ways-big-data-gets-misused/>

⁵⁸ <https://irishtechnews.ie/5-ways-big-data-gets-misused/>

buying a product, picking a restaurant to dine, or chose which politician to vote at the elections. This type of data misuse will concern the next section as we are going to analyze a case in which big data have been improperly used to influence the outcome of elections.

2.1.1 Cambridge Analytica case study: how not to use data

The Cambridge Analytica case made a name for itself in recent years for the public scandal that caused. This case sets an extreme example of data misuse and what do data providers can do with huge amounts of data.

To begin with, Cambridge Analytica was a company controlled by SCL group (Strategic Communication Laboratories) which was focused on behavioral research, and strategic communication. Cambridge Analytica was founded in 2013 by Alexander Nix⁵⁹ as a London consulting firm that provided services to companies that wanted to “*change audience behavior*”. More precisely, the CEO (Alexander Nix) stated that the company wanted to “*to address the vacuum in the US Republican political market*”⁶⁰. Therefore, the real focus area of Cambridge Analytica was to sustain the Republican elections by exploiting data-intensive models to change people’s opinion on some hot topics. Nix also stated “*The Democrats had ostensibly been leading the tech revolution, and data analytics and digital engagement were areas where Republicans had failed to catch up. We saw this as an opportunity.*”⁶¹. Cambridge Analytica was principally owned by the right-wing donor Robert Mercer, a 75-years-old U.S. billionaire, which became one of the biggest republican party contributors for spending \$25 Millions in 2016 for baking Donald Trump campaign⁶². Mercer was also one of the main funders of Breitbart News, a conservative right-wing information site lead by Steve Bannon. Bannon was Trump’s counsellor and strategic advisor during the Trump’s campaign for president in 2016⁶³. As stated before, Cambridge Analytica was focusing on public image management, and it specialized in social media data. More precisely, it specialized in social media data collection., such as Facebook posts, likes, comments, etc. It used all this data to build some algorithmic models to try and profile users exploiting

⁵⁹ <https://www.theguardian.com/news/2018/mar/18/what-is-cambridge-analytica-firm-at-centre-of-facebook-data-breach>

⁶⁰ <https://www.theguardian.com/news/2018/mar/18/what-is-cambridge-analytica-firm-at-centre-of-facebook-data-breach>

⁶¹ <https://www.theguardian.com/news/2018/mar/18/what-is-cambridge-analytica-firm-at-centre-of-facebook-data-breach>

⁶² <https://www.forbes.com/profile/robert-mercero/>

⁶³ <https://www.ilpost.it/2018/03/19/facebook-cambridge-analytica/>

psychometric techniques. Cambridge Analytica did not obtain all its data by itself; It also bought data from data providers to have a better chance at building efficient models. Some valuable insights on the type of work carried out by Cambridge Analytica can be obtained from an interview⁶⁴ of Alexander Nix at the Concordia summit. The Concordia summit is “*a registered 501(c)(3) nonprofit, nonpartisan organization dedicated to actively fostering, elevating, and sustaining cross-sector partnerships for social impact.*”⁶⁵. The name of the speech was “The power of big data and psychographic in the electoral process”. Nix started by talking about how Cambridge Analytica contributed to Ted Cruz’s⁶⁶ presidential primary campaign. He underlined how before Cambridge Analytica’s work, Ted Cruz was one of the least favorite candidates in the run, and how he became Trump’s only serious contender in 2016. To sustain Cruz’s campaign, Cambridge Analytica embraced three methodologies: behavioral science, data analytics, and addressable ad technology.

For what extent behavioral science: Nix stated that their type of communication was based on the behavior of the people and not on demographic factors such as gender or age. Cambridge Analytica focused on psychographic, which aims at understanding the personality of a subject. They understood that it is personality that drives behavior, and behavior that influence how you vote. Therefore, their model was based on the detection of people’s personality through psychographic algorithms. By doing so, they would be able to influence people’s decision not only based on their preferences but based on their personality. Cambridge Analytica built a model called OCEAN⁶⁷, which used people’s data to assign some personality traits to individuals. OCEAN stands for Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. Thanks to the OCEAN model, Cambridge was able to profile voter’s personality and project specifically tailored ads based on their behavior. Alexander Nix made an example by proposing two different communication strategy on the second amendment regarding the guns right in America. The first strategy showed the possibility of a burglary and the deterrent of possessing a gun. This type of communication was aimed at leveraging highly neurotic and conscious people (fearful people which likes security). The second strategy focused on the tradition of hunting passed down from father to son. This strategy leveraged highly closed and agreeable people, which is very close to tradition.

⁶⁴ <https://www.youtube.com/watch?v=n8Dd5aVXLCc>

⁶⁵ <https://www.concordia.net/about/>

⁶⁶ Ted Cruz is a member of the American republican party who ran for president in 2016 against Trump.

⁶⁷ <https://retinacromatica.it/capire-l-uso-dei-big-data-cambridge-analytica/>

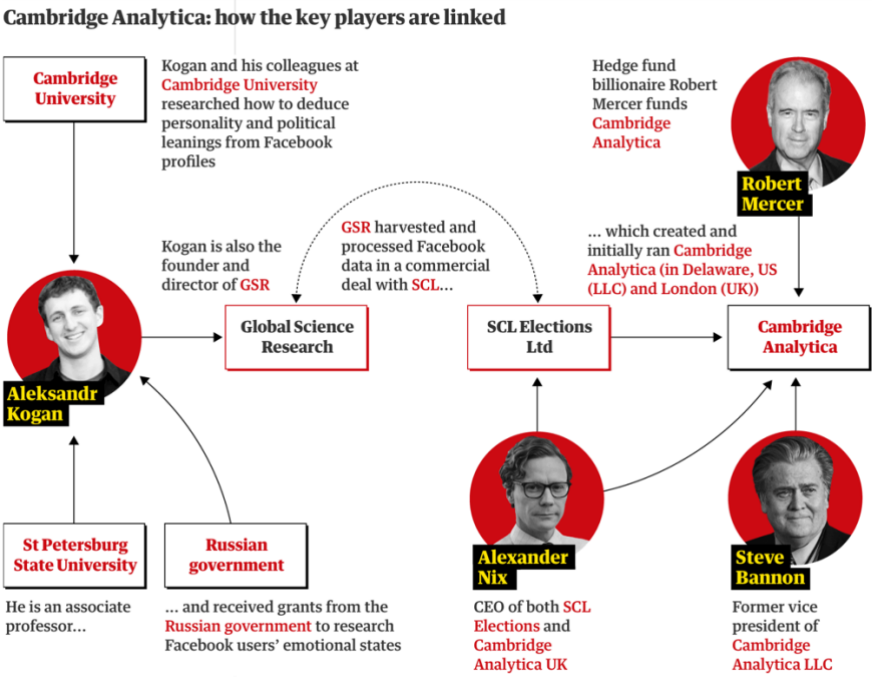
For what extent data analytics: Nix stated that their model will substantially change the way that companies communicate. He compared the “old-fashioned” communication based on creativity, to the new type of communication that they offer based on personality. Their goal was not to find something brilliant to be able to push people to act or buy, they wanted to propose a different style of communication to every different person, based on their personality. To do so they had to exploit the so-called data points on individuals, which are basically units of information. The data possessed by Cambridge Analytica was both demographic (factual based), psychographic (attitudinal), and behavioral (personality based). With such instruments, they were able to segment population, discover more persuadable voters, and target them with the right type of communication to steer them and vote for Cruz. This type of targeting differentiates from any type of “standard” procedure because it leverages people’s sentiment and behavior.

For what extent addressable ad technology: Nix underlined how the mass communication based on blanket advertising, TV ads, e-mail is coming to an end. He suggested how the future of communication will be based on highly targeted strategies. With such technology the personalized ads will not only concern web browsing through the use of cookies, also tv advertising, email marketing, and other types of communication will be tailored.

The results of this strategy on Ted Cruz’s campaign for primary elections are measured on public appreciation. Cruz started from less than 5% and steadily rose up to above 35% becoming Trump’s second most threatening contender in the race. The developer of Cambridge Analytica’s algorithm, Michal Kosinski, stated that the algorithm needed 70 Facebook likes to know more on a person than his friends, 150 likes to know more than his parents, and 300 likes to know more than his partner. These were the instruments used by Cambridge Analytica to profile and influence the electorate by exploiting their own behavior and personality.

However, the issue here is not the model developed by Cambridge Analytica but instead how they collected the data. In 2014, Aleksandr Kogan was a PhD at Cambridge University, and he developed a Facebook app called “ThisIsYourDigitalLife”. Such app was designed to preview the personality traits of a people based on their online activity. Thanks to this app, and thanks to Facebook’s terms of use in 2014, Kogan was able to collect data not only on the 300 thousand users that installed the app, but also on all of their friends. By doing so, Kogan obtained behavioral data on 50 millions of users. The real scandal was triggered when

Cambridge Analytica illegally bought such data from Kogan to develop its algorithm. That was an issue because Facebook terms of use allow app developers to collect user’s data but not to sell them to third parties. Therefore, the data collected from Kogan was illegally sold to Cambridge Analytica for about \$1 million.⁶⁸ The 16th of March 2016 Robert Mueller, an American attorney and ex chief of the FBI asked Cambridge Analytica to present some documents on their activity. The suspect was that they somehow helped the Russian’s interest to defeat Hilary Clinton and to promote Donald Trump. That is because Cambridge Analytica not only helped Cruz campaign but also Trump’s by damaging the image of Hilary Clinton. They used some bots⁶⁹ to publish and divulge fake news regarding Clinton that would ruin her public image. To better understand the relationships and connections between Cambridge Analytica and other supporters it can be useful to exploit the next scheme created by “The Guardian”:



70

After the scandal of American’s primary elections came up, a series of shady actions of Cambridge Analytica started to be questioned. More precisely an undercover investigation by Channel 4 revealed some new videos which directly incriminate Cambridge Analytica and Alexander Nix. Among the allegations, Cambridge Analytica was accused to have influenced

⁶⁸ <https://www.ilpost.it/2018/03/19/facebook-cambridge-analytica/>
⁶⁹ “Bots” means robot used online for creating fake profiles and divulge fake news.
⁷⁰ Image 6: <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election>

not only American's elections but also others, among which Kenya's. Nix publicly denied any involvement with fake news and public manipulation towards Kenya's election and in an undercover footage released by channel 4 he contradicted himself⁷¹. The final blow on Alexander Nix was struck by Channel 4 when they released an undercover video in which Nix talked to a hypothetical client about ways to blackmail their political opponents. Nix suggested sending girls around the house of the target and illegally funding his political party to blackmail him afterwards. Of course, after the publishment of such undercover video, Alexander Nix was removed as the company's CEO and in 2018 Cambridge Analytica filed chapter 7 for bankruptcy.

To conclude this section, some comments are worth mentioning. Cambridge Analytica's case offers an interesting insight on today's big data market. The aim of presenting this case was not to sustain any conspiracy theory according to which all big data firms are bad. Conversely, the aim of this section was to give an idea of how much power can be derived from huge amounts of data. Without some firms that collect, manage, and use big data in a correct way, most of the progress achieved in advanced technology would not be possible. Although, we have seen how powerful big data can be and how much damage it can cause when exploited in an incorrect way. In order to clarify what measures have been taken from the EU and US regulators regarding data security and privacy, we will analyze the main features of the topic.

⁷¹ <https://www.youtube.com/watch?v=mpbeOCKZFfQ>

2.2 Big data regulation

Big data regulation is becoming an increasingly important topic nowadays, that is due to the progressive importance of big data in today's world. When a society starts to develop into a more advanced data-driven ecosystem, some issues concerning privacy may arise. It is fundamental to ensure that all personal data is processed in respect of people's right and in accordance with laws. Cambridge Analytica offers a perfect example of how data can jeopardize society, and where the regulators should intervene to protect people's privacy. With no doubts, social media platforms are at the center of the debate due to the huge amount of data that they possess. Moreover, since for most data giants, big data represents the main source of revenues (see sect. 1.4), the role of regulation assumes an even more important play. That is, an effective privacy regulation not only should ensure an adequate level of security to users, but it should also allow data-driven companies to continue sustaining their business models by exploiting data. Of course, a balancing is needed between data safety for users, and availability of data for companies which exploits big amount of data to survive. Big data regulation does not only interfere with data giants, but it also impacts the progress of research and development around the world. With an excessively conservative regulatory environment, the research process could slow down and that would have terrible consequences on the economy. However, the research and development field, should be able to analyze big data in accordance with the privacy regulations that we are going to analyze.

2.2.1 Issues and features concerning big data regulation

As stated before, privacy issues are a natural consequence of big data exploitation. To this extent, we will analyze the main issues concerning big data usage and how they can be avoided.

A fundamental distinction is between data privacy and data security, data privacy is more concerned with how sensible data is treated by data gatherers. It ensures that the process of collecting, cleansing, processing, and using sensitive data is compliant with the current regulation and has the authorization of the data owner⁷². Therefore, data privacy is a feature

⁷² <https://www.tokenex.com/blog/data-privacy-vs-security>

that concerns the process of data gathering from the beginning. Data privacy begins when the user is properly informed about the type of data that will be collected, how will it be used, for how long will it be stored, and whom will have access. That means that data privacy is very tied with the concept of transparency. On the other hand, data security concerns a slightly different concept. Data security is more focused on preventing unauthorized access to sensitive data from third parties. It focuses on the main issues that can affect personal data, such as data breaches, leaks, or hacker attacks. Therefore, to ensure data security entities should embrace technical tools such as firewalls, user authentications, and internal security practices⁷³. Some other techniques that companies can implement to ensure data security are encryption and tokenization. Such instruments allow personal data to become unreadable for anyone but the recipient of the data which will have the key to read and access it. Such topics will be briefly resumed in Chapter 3.

To recap, data privacy and security can appear similar but are in fact different concepts, whereas data security can be achieved without data privacy, the opposite is not true. That is because data security aims at protecting user's data while data privacy aims at protecting the user's identity⁷⁴.

Now that the difference between these two key concepts is clear, we can move forward to analyze the main features of data privacy. Therefore, we will discuss the main fundamentals of the process that ensures the proper treatment of personal data, that is identity protection: Data confidentiality: is the prevention of unauthorized access to sensitive data⁷⁵. This feature affects differently each type of personal data. Information that contains more sensible data such as health data or payment data is more sensible than preferences about food for instance. To better understand the concept of sensitive data, we can analyze an article of the General Data Protection Regulation (GDPR) that we will analyze in the next session: “[37] *Personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms. Those personal data should include personal data revealing racial or ethnic origin... [53] Member States should be allowed to maintain or introduce further conditions, including limitations, with regard to the processing of*

⁷³ <https://www.tokenex.com/blog/data-privacy-vs-security>

⁷⁴ <https://www.tokenex.com/blog/data-privacy-vs-security>

⁷⁵ <https://www.sdxcentral.com/security/definitions/what-are-the-data-privacy-fundamentals/>

*genetic data, biometric data or data concerning health.*⁷⁶” From this definition we can deduce that, according to the GDPR, sensitive data refers to race, ethnical and political background, genetic, biometric, health data, and sexual orientation.

Limiting data collection: in principle, data gatherers should collect only data that they intend to use for good purposes⁷⁷. Moreover, they should not collect any type of additional data in addition to the one that they need to process. Such principle is also cited in the GDPR Art 5 (1) (b), which states that personal data should be “*Collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes...* ”.

Transparency: whenever an entity plans in collecting some personal data, it should be transparent in the process of data gathering. That should include the authorization of the interested party in sharing information with the data gatherer. Also, it should be included the section “privacy policy” in which must be enlisted the purposes of the data collection, and all the party that will have access to it. Moreover, also a “disclosure” section should be made easily available in which are enlisted the privacy policy and eventual cookies and web beacons⁷⁸. For what extent transparency, all the information should not only be available, but it should also be easily accessible and understandable to all users.

Compliance: in addition to the guidelines provided, all entities that will incur in the process of data collection must abide to the current data privacy regulation. To this end, the competent regulation will depend on the geographic area of the user. For instance, in the European Union the current regulation is the General Data Protection Regulation (GDPR). In the United States, there are some laws at the federal level, but each member state can implement stricter privacy rules such as the California Consumer Privacy Act (CCPA) or the Health Insurance Portability and Accountability Act (HIPAA). If an entity in the United States wants to gather data on a user located in the EU, it should comply with the GDPR because it counts the location of the user, not the data gatherer.

⁷⁶ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

⁷⁷ <https://www.sdxcentral.com/security/definitions/what-are-the-data-privacy-fundamentals/>

⁷⁸ <https://www.sdxcentral.com/security/definitions/what-are-the-data-privacy-fundamentals/>

2.2.2 The EU environment: General Data Protection Regulation (GDPR)

The General Data Protection Regulation 2016/679 (EU) is the current law enforced in the European Union to ensure EU citizen's data privacy. It was drafted in 2016 and entered into force in 2018, by replacing the 1995 European Data Protection Directive which was too obsolete for nowadays' digital environment⁷⁹. The GDPR is believed to be the toughest privacy law in the world, also thanks to the huge fines (tens of millions of dollars) which are imposed to those who do not respect its standards.

For the sake of clearness, before enlisting the GDPR's main principles and features it is necessary to introduce some terms used in the regulation text:

- Personal data: “*Personal data is any information that relates to an individual who can be directly or indirectly identified*⁸⁰”. Therefore, personal data is every type of data that be conducted to the identity of the person, including email, phone number, pseudonym, ethnicity, gender, age, etc.
- Data processing: The GDPR text reports “*...Processing means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction*⁸¹”. Therefore, we can understand that by “processing” the EU regulators wanted to include any type of action carried out on people's data.
- Data subject: the person whose data is processed⁸².
- Data controller: The person that decides why and how personal data will be processed⁸³.
- Data processor: a third party that processes personal data on behalf of a data controller, GDPR has special rules for these individuals and organizations⁸⁴.

⁷⁹ <https://gdpr.eu/what-is-gdpr/>

⁸⁰ <https://gdpr.eu/what-is-gdpr/>

⁸¹ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

⁸² <https://gdpr.eu/what-is-gdpr/>

⁸³ <https://gdpr.eu/what-is-gdpr/>

⁸⁴ <https://gdpr.eu/what-is-gdpr/>

For what extent the principles of the regulation, we can analyze Article 5 of Chapter 2 GDPR, which enlists the fundamentals of processing personal data. The article states “*Personal data shall be:*”

- [a] “*Processed lawfully, fairly and in a transparent manner in relation to the data subject (‘lawfulness, fairness and transparency’);⁸⁵”*. This first article is mostly tied with the concepts of transparency and confidentiality analyzed in section 2.2.1. This implies that the data gatherer should take action to provide full understanding to the data subject and to comply with the law. For the process to be lawful, the data gatherer shall have received a consent given by the data subject, it should protect the vital interest of the individual. Moreover, the processing shall be aimed at achieving the legitimate interest of the organization, unless the interest of the individual would be prejudice. The following elements must be included, the identity of the data controller, the purpose for which the data will be collected, and any further information that could be useful to provide full understanding. To recap, the data processing shall not be deceiving and misleading⁸⁶.
- [b] “*collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes (‘purpose limitation’);⁸⁷”*. This principle suggest that data should be collected only for specified and legitimate purposes. However, when we consider public interests, historical, statistical purposes and other instances mentioned in article 89(1), an exception can be made.
- [c] “*adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed (‘data minimization’);⁸⁸”*. We deduce how the data collected by the data gatherer shall be adequate, relevant, and not excessive. It should be noticed how the regulators used the word “necessary”, that should avoid the

⁸⁵ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

⁸⁶ M.u.S.A, “*Data privacy and protection fundamentals*”, Hellenic open University.

⁸⁷ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

⁸⁸ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

collection of data for yet unspecified purposes⁸⁹. This article refers to the concept of “limiting data collection” analyzed in section 2.2.1.

- *[d] “accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay (‘accuracy’);⁹⁰”*. It is pointed out that the data gatherer should implement every reasonable step to ensure that data on the data subject is precise, updated and adjourned. Whenever the subject provides new data, the organization shall immediately provide at adjourning its archive. Moreover, the data gatherer must periodically clean the data to ensure that it is not incorrect and misleading⁹¹.
- *[e] “kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject (‘storage limitation’);⁹²”*. The first element of this article refers to the period of time for which the data will remain to the data gatherer. It is specified that the data should be kept for no longer that is necessary for the purpose of the collection. Organizations must regularly review the time length of such period to ensure compliance with the law⁹³. It is also specified that data can be held for longer than necessary whenever it will serve public utility purposes or other instances defined in article 89 (1).
- *[f] “processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures (‘integrity and confidentiality’).⁹⁴”*. This principle refers

⁸⁹ M.u.S.A, “Data privacy and protection fundamentals”, Hellenic open University.

⁹⁰ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

⁹¹ M.u.S.A, “Data privacy and protection fundamentals”, Hellenic open University.

⁹² (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

⁹³ M.u.S.A, “Data privacy and protection fundamentals”, Hellenic open University.

⁹⁴ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

more to the concept of data security. It is pointed out that the data gatherer shall implement the appropriate measures of security to ensure that the data of the subject is safe and protected. Moreover, the data gatherer must take measures that should minimize the risk of accidental loss or damage to the data in the interest of the subject's fundamental rights.

- [g] *“The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 (‘accountability’).”*. To meet the requirements of accountability the data gatherer should put in place some measures such as taking a data protection and design by default, maintaining the documentation of your processing activity, implementing the adequate security measures, etc.⁹⁵.

To conclude this overview on the GDPR EU regulation we will analyze the rights provided to the individuals with respect to data security (section 2,3,4):

- Article 13, “Right to be informed”: *“Information to be provided where personal data are collected from the data subject... [a] the identity and the contact details of the controller... [b] the contact details of the data protection officer... [c] the purposes of the processing for which the personal data are intended... [2a] the period for which the personal data will be stored... [2d] the right to lodge a complaint with a supervisory authority”*⁹⁶. To summarize, the EU regulator ensured that the data subject is always informed with respect to the purpose, length, and main aspects of the data processing.
- Article 15, “Right of access”: *“The data subject shall have the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data...”*⁹⁷. This article ensures that the data subject is always informed on whether or not its data is being processed and gives it the right to access it. Moreover, such access shall be given freely whenever the request is not excessive or unfunded.
- Article 16, “Right to rectification”: *“The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data*

⁹⁵ M.u.S.A, “Data privacy and protection fundamentals”, Hellenic open University.

⁹⁶ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

⁹⁷ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

concerning him or her. Taking into account the purposes of the processing, the data subject shall have the right to have incomplete personal data completed, including by means of providing a supplementary statement.⁹⁸”. This right is strictly connected with the aforementioned principle [d] “accuracy” which states that the data shall be complete and accurate.

- Article 17, “Right to erasure”: “*The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay...⁹⁹*”. This right ensures that the data subject can demand the erasure of its personal data whenever they have been unlawfully collected, he withdraws consent for processing, or personal data is no longer necessary.
- Article 18, “Right to restriction”: “*The data subject shall have the right to obtain from the controller restriction of processing where one of the following applies: ... [a] the accuracy of the personal data is contested by the data subject; ... [b] the processing is unlawful; ... [c] the controller no longer needs the personal data for the purposes of the processing.¹⁰⁰*”. Therefore, for similar instances to those presented in article 17 (right to erasure), the data subject can rightfully demand the restriction of the data processing.
- Article 20, “Right to data portability”: “*The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided¹⁰¹*”. This article states that in the instances presented in Articles 6(1)[a] and 9(2)[a], the data subject can require to be delivered his data. Moreover, the data subject will have the right to give his data to another controller.

⁹⁸ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

⁹⁹ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

¹⁰⁰ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

¹⁰¹ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

- Article 21, “Right to object”: “*The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her which is based on point (e) or (f) of Article 6(1)*¹⁰²”. Said point [e] and [f] of article 6(1) state: “[e] Processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller; ...[f] processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party...”.

2.2.3 The US environment: a fragmented regulatory framework

In the last section we discussed about big data’s increasing importance in the EU, and some instances in which companies gathered data in unlawful ways. The analysis of EU’s big data regulation was the natural continuous, after analyzing some instances of data misuse. We will now move our focus area in the United States, where the need for big data regulation seemed even more pressing in the past years. That is because, the US can be considered as the cradle of big data implementation and development, mostly thanks to the advanced technological level of the country. It is sufficient to think that the big five tech companies discussed in section 1.4.3 (Google, Apple, Facebook, Amazon, Microsoft) are all US-based companies. The number of US data-intensive entities is the higher among all the world, thus, the need for a solid consumer privacy regulation is strong.

As everybody knows, the United States are a federal republic, this implies that there can be laws both at a federal and state level. For what extent big data regulation, there is not a single, comprehensive, federal law which regulates the US as a whole. Instead, each state can implement differently stricter laws for each area of big data regulation. This causes the US to have a fragmented big data regulatory environment, which causes the companies that operates in more than one state, to have higher costs for law compliance.

To have an idea of the situation for big data regulatory environment in the US, we will analyze the main and most relevant laws:

¹⁰² (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

- California Consumer Privacy Act (CCPA): this act is aimed at providing customers with more control over their personal information and is effective only for California’s companies and citizens. Among the main points, we find:
 - “Right to know”, the right to be informed whenever a data gatherer is going to collect and process the subject’s data, and for which purposes it will be processed¹⁰³.
 - “Right to delete”, the right to have one’s personal data deleted from a data gatherer and the right to ask for deletion also to the service provider of the data gatherer. However, there are some instances in which the data gatherer can refuse¹⁰⁴.
 - “Right to opt out”, the right that a data subject has to ask the data gatherer not to sell its personal data unless he will tell to do so in the future. The data gatherer has to wait at least 12 months to ask the data subject to opt back in¹⁰⁵.
 - “Right to non-discrimination”, businesses cannot deny goods or services, charge you a different price, or provide a different level or quality of goods or services just because you exercised your rights under the CCPA¹⁰⁶.
- Virginia Consumer Data Protection Act (VCDPA): VCDPA is a broad privacy law concerning users of the Virginia state. The law is based on the consensus and rejection principles; data gatherers should ask permission before accessing personal data and at any time, the data subject can withdraw consent and ask for the deletion of data. VCDPA also requires the data gatherers to implement protection assessments to ensure the safety of consumers data concerning mainly personalized ads and targeted sales. The VCDPA differs from the CCPA mainly for its length, as it is only eight pages long. However, it is considered to be safe and tough in ensuring its citizens an adequate level of data privacy¹⁰⁷.
- Colorado Privacy Act (CPA): the CPA (2021) is a comprehensive Colorado’s privacy law inspired mainly by the EU’s GDPR, the California’s CCPA and the Virginia’s VCDPA. The CPA is very similar in content to CCPA and VCDPA, they differ for some details regarding the range of companies subject to legislation. “*The CPA applies to companies that conduct business in Colorado or sell product or services*

¹⁰³ <https://oag.ca.gov/privacy/ccpa>

¹⁰⁴ <https://oag.ca.gov/privacy/ccpa>

¹⁰⁵ <https://oag.ca.gov/privacy/ccpa>

¹⁰⁶ <https://oag.ca.gov/privacy/ccpa>

¹⁰⁷ <https://pro.bloomberglaw.com/brief/what-is-the-vcdpa/>

*intentionally targeted to residents of Colorado, and meet either of the following thresholds: (i) controls or processes personal data of 100,000 or more consumers during a calendar year; or (ii) derive revenue or receive discounts from the sale of personal data and control or process data of at least 25,000 consumers.”*¹⁰⁸.

- Health Insurance Portability and Accountability Act (HIPAA): HIPAA is a 1996 federal law which ensures the privacy of communications regarding people’s health data. The HIPAA regulation individuates the so-called “covered entities” which are subject to a set of “privacy rules”. Among said covered entities fall healthcare providers, health plans, healthcare clearinghouses, business associates. The goal of the privacy rules is to prevent the flow of health information for unnecessary purposes. Moreover, it is necessary to find a balance between the protection of the data subject and the necessary disclosure of data to ensure a high-quality healthcare¹⁰⁹. One last consideration is worth mentioning with respect to HIPAA; this regulation does not concern all types of health data, but only official data transmitted to “covered entities”. All types of health data such as the Apple watch measurement, cardiac rhythm, and other device-measured data, is not under the application of HIPAA.
- Electronics Communications Privacy Act (ECPA): the ECPA (1986) is a federal-level law that regulates the way in which the government and other entities can wiretap, follow, and monitor the communications of US citizens. The ECPA updated the Federal Wiretap Act of 1968 and included not only telephonic data, but also computer, mobile, and other technologies’ data. Moreover, this regulation also sets the rules for how the employers can control and monitor their employees’ communications at work. The ECPA was strongly criticized for being too outdated and not protecting against new high-technology techniques of surveillance¹¹⁰.
- Children’s Online Privacy Protection Rule (COPPA): the COPPA (1998) is a federal law aimed at protecting the online data collection of children under the age of thirteen. Such law requires website operators and other online entities to require parental or guardian consent for the collection of data concerning the child. Furthermore, after the parental consent, the data gatherer should communicate which kind of data will

¹⁰⁸ <https://www.natlawreview.com/article/and-now-there-are-three-colorado-privacy-act>

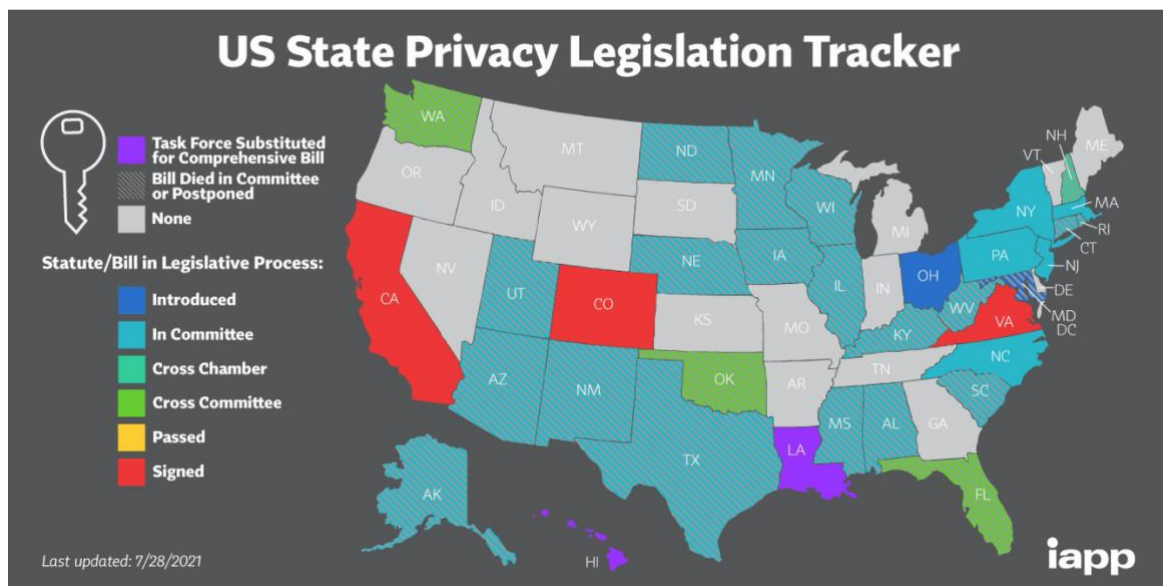
¹⁰⁹ <https://www.cdc.gov/phlp/publications/topic/hipaa.html>

¹¹⁰ <https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/1285>

be collected, for what purposes, and eventually to which third-parties subject will be shared to¹¹¹.

- Federal Trade Commission Act (FTCA): the FTCA is a federal-level regulation that empowers the Federal Trade Commission to check and control for entities to respect the privacy laws. The FTC have the power to investigate apps and websites that violates its own privacy policy; Furthermore, it can also investigate about violations of marketing language related to privacy.

As can be easily noticed, the United States do not have a unique and comprehensive regulatory framework. There are federal-level acts, that do regulate practices for all US citizens, but most of such laws are incomplete and outdated. There are only three US states that have their own comprehensive privacy laws, California with the CCPA, Virginia with the VCDPA, and Colorado with the CPA. As in an article of the New York Times about privacy legislation in the US, this is the situation country-wide:



112

“These laws have similar provisions that tend to give you some type of notice and choice in controlling your data. Essentially, a company operating under these regulations must tell you if it’s selling your data; you also get a choice in whether you’re okay with that or not,

¹¹¹ <https://usercentrics.com/knowledge-hub/childrens-online-protection-act-coppa/>

¹¹² Image 7: <https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/> / IAPP primary source, the world’s largest global information privacy community <https://iapp.org/about/>

*and you have the right to access, delete, correct, or move your data.*¹¹³”. Some have criticized these laws for being too permissive; yet these three countries have at least a base of comprehensive privacy regulation compared to other states that do not have proposed a new law. The experts suggest that among the three, California’s CCPA is the toughest and most safe. However, many believe that worldwide, the EU’s GDPR is the most effective and comprehensive data privacy regulation.

2.2.4 Costs and implications of compliance with the regulatory environment

At the beginning of this chapter, we stated that the natural consequence of the massive increase in use of big data was an increase in big data regulation. Of course, that is due to the fact that governments want to protect their citizen’s privacy and assure a minimum level of safety. We analyzed two different approaches to big data regulation, one based on a unique piece of legislation (EU), another based on multiple laws (US). The main aim of this section will be to analyze the effects, costs, and consequences of the implementation of such data regulations. We will begin by analyzing the effects of regulation in the EU environment.

As we stated in section 2.2.2, the EU’s GDPR requires any data gatherer that attempts to collect information on a subject to be endorsed by a legal basis such as an explicit consent, which must be “*freely given, specific, informed and unambiguous*”¹¹⁴. The GDPR demand both data controllers and processors to implement a process of privacy-by-design. That means being able to implement data-protection principles into their products and to consider both costs and risks of data processing for their customers. Since EU’s GDPR adopt a risk-based approach, the costs and efforts that a data gatherer has to make to be compliant with the regulation, vary depending on the riskiness of the activity of processing. The riskier will be the data processing, the harder and costly will be for the controller to be compliant. On the other hand, for low-risk activities, the efforts for being compliant will be far less great. The concept of riskiness of data processing of course refers to a risk for the data subject not

¹¹³ <https://www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/>

¹¹⁴ Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “*Cost of compliance and enforcement of data protection regulation*”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594.

data gatherer. Activities which are considered riskier are for instance processing based on new technologies, and extensive automated decision-making with legal effects¹¹⁵. When we consider the costs of compliance of an EU-based company we should consider the size of the firm. In a 2019 joint research study by the International Association of Privacy Professionals (IAPP) and Ernst & Young the costs of compliance to GDPR privacy rules amounted to \$1 million in 2018 (the year GDPR went into effect) and \$622'000 in 2019¹¹⁶. However, the study was not carried out on companies that were subject to GDPR alone but considered also other countries. In fact, another study conducted by Ponemon Institute¹¹⁷ in 2019 focused on GDPR compliance only and revealed an average budget of \$13.2 million for 2018 and \$13.6 million in 2019. Although such figures are already high, it must be considered that the amount of costs varied dramatically based on the size of the firm. In fact, the same research presented results for the big companies included in the FTSE 100 stock index of around \$84 million for banks, \$26 million for technology firms, and \$6 million for industrial companies. Overall, the greatest slice of the costs was allocated to hiring privacy-specialized personnel (more than 25%), and another important source of costs were technologies updates (12-17%). To conclude this analysis, the researcher observed that the overall privacy expenses increased as the number of employees of a form increased. To this extent, we present a table depicting the average privacy costs in relation to the form size (number of employees):

¹¹⁵ Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “*Cost of compliance and enforcement of data protection regulation*”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594.

¹¹⁶ Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “*Cost of compliance and enforcement of data protection regulation*”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594.

¹¹⁷ Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “*Cost of compliance and enforcement of data protection regulation*”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594.

Category	<5k Employees	5k–24.9k Employees	25k–74.9k Employees	75k+ Employees
Privacy Team Salaries	\$170,700	\$581,800	\$744,200	\$847,100
Privacy Team Technologies	\$23,500	\$47,100	\$39,700	\$115,600
Outside Privacy Team Technologies	\$38,700	\$30,500	\$57,500	\$814,200
Other Privacy Budget	\$24,700	\$84,500	\$82,000	\$106,200
TOTAL PRIVACY SPEND	\$257,700	\$743,800	\$923,400	\$1,883,200

118

As can be easily noticed, the average cost for privacy compliance increases as the size of the company (measured per employee) increases.

If we switch our focus area on the privacy compliance costs of the US companies, the situation is different. That is because, as we have seen in the previous section, the US do not have a unique and comprehensive privacy regulation for the whole federation. Instead, it has few sectorial laws and only three states have a complete regulation. For the sake of simplicity, we will briefly analyze the costs of implementation only for two US privacy laws.

The first law is HIPAA (Health Insurance Portability and Accountability Act), and the estimated costs of compliance in this sector (health) are believed to be much higher than for GDPR’s average. The Department of Health and Human Services estimated that industry-wide implementation would cost \$3.2 billion in HIPAA’s first year and \$17.6 billion for the first ten years¹¹⁹. However, different studies faced different sector niches and therefore provided different results. To this end, we provide a table depicting the costs of HIPAA compliance for the entire industry:

¹¹⁸ Image 8: Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “*Cost of compliance and enforcement of data protection regulation*”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594 (page 13).

¹¹⁹ Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “*Cost of compliance and enforcement of data protection regulation*”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594.

Research Entity	Affected Respondents	Estimated Cost of Compliance
Healthcare Consulting Companies (2003)	Health care providers (covered entities)	\$25-43 billion (first 5 years)
Department of Health and Human Services (2002)	Health care providers (covered entities)	\$3.2 billion (first year) \$17.6 billion (first 10 years)
Gartner Group (2003) ¹²⁸	Entire health care industry	\$3.8 - \$38 billion (2003-2008)

120

If we move to a different US privacy law, we can appreciate the difference in costs based on the sector. By considering the COPPA (Children’s Online Privacy Protection Rule), we can notice some relevant differences. In fact, the children’s privacy compliance costs appear to be far less costly than HIPAA’s. In fact, in 2000, the House of Representative’s Committee on Commerce estimated the cost of compliance with COPPA to range from \$115,000 to \$290,000 per year¹²¹. The House Committee drafted a table presenting the typical compliance costs for COPPA:

Activities	Cost
Legal (audits, construction of private practices and policy)	\$10,000 - 15,000 (one time)
Engineering costs to make the site compliant	\$35,000 (one time)
Professional chat moderators (price differs depending on training, hours of operation, and organization)	\$25,000 - \$10,000 per month
Personnel overseeing offline consent, responding to parents’ questions, reviewing phone consents, and reviewing permission forms	\$35,000 - \$60,000 per one person per year in charge of these activities
Personnel overseeing compliance, database security, responding to verification and access requests	\$35,000 - \$60,000 per one person per year in charge of these activities

122

¹²⁰ Image 9: Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “Cost of compliance and enforcement of data protection regulation”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594 (page 21).

¹²¹ Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “Cost of compliance and enforcement of data protection regulation”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594

¹²² Image 10: Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “Cost of compliance and enforcement of data protection regulation”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594 (page 25).

To conclude this overview on the effects and costs of the implementation of the privacy rules concerning big data worldwide we make the next consideration. The EU's GDPR provide a solid legal bases which has the advantage of being unique and more easily applicable. A firm operating in the EU does not have to be compliant with every single sectorial regulation, it is sufficient to be compliant with GDPR overall. On the other hand, in the US the highly fragmentated regulatory environment forces US companies to spend more to be compliant. That is because, as one firm operates in a specific sector, it has to be compliant with a specific regulation, if it operates in more than one sector, it will have to comply with a huge number of regulations. However, despite the huge amount of costs that companies worldwide have to sustain, the benefits of being complaints with the regulatory environment are far greater considered the enormous amounts of fines which are imposed to non-compliant companies. Furthermore, a privacy-compliant company is believed to be more safe, reliable, and transparent leading to an increase in perceived value from the investors.

Chapter 3: Big data and blockchain technology: the ultimate combination

3.1 Blockchain 101

The aim of this section is to give a clear-cut definition of the blockchain technology. After reading this chapter, also a non-familiar subject will understand the topics presented. In the first sub-chapter, we will analyze the practical functioning of the blockchain and in the second one, we will present the integration between blockchain technology and big data. Finally, we will state some conclusions on the profitability of such integration.

3.1.1 A simple definition

We will begin by talking about the creation of the first blockchain, and the popular interest that has formed around this technology through time. The first appearance of the blockchain technology dates back to 2008, when a person (or group of people) named Satoshi Nakamoto¹²³ published a paper named “Bitcoin: A Peer-to-Peer Electronic Cash System”. Such paper argues about a new system of payments based on the absence of third-party intermediaries in transactions “*A purely peer-to-peer version of electronic cash would allow online payments to be sent directly from one party to another without going through a financial institution.*”¹²⁴ Such system would be based on the blockchain technology.

To begin with, we can define blockchain in the simplest way as a distributed database that a group of individuals controls and that store and share information¹²⁵. More often the word “database” is replaced with the term “ledger” which represents the book of accounts in which account transactions are recorded. Blockchain is built on the mechanism of the “public ledger” and all transactions occurred in the blockchain are registered in blocks. All information in each block is linked to the previous block in a tamper-resistant manner. The

¹²³ The real identity of the founder(s) of the blockchain has never become publicly known. The “initiator” of such technology named itself with an alias, namely, Satoshi Nakamoto.

¹²⁴ Satoshi Nakamoto, (2008), *Bitcoin, a peer-to-peer electronic cash system*, www.bitcoin.org

¹²⁵ Tiana Laurence, (2019), *Blockchain for dummies*, John Wiley & Sons.

data in each block is stored inside the so-called “nodes”. Although the name can sound “complex”, nodes are electronic devices such as computers, laptops, or even bigger servers¹²⁶. Each block can contain multiple transactions and each transaction has its own “reference number” called hash. The hash is a unique string of characters that refers to one and only transaction. Through the hash, one can verify the previous transaction as well as the information in the transaction itself. With such system, each node¹²⁷ has access to all previous blocks down to the first block of the chain called “genesis block”. The time stamp gives each block an immutable temporal position in the chain¹²⁸. A fundamental element of blockchain technology is cryptography, which is the science of secure communication. Cryptography consists in encrypting a message which will be readable only by the intended recipient through decryption. Such technique is based on complex mathematical principles. More recently, cryptography has evolved to include applications like proving the ownership of information to a broader set of actors, such as public key cryptography, which is a large part of how cryptography is used within blockchain¹²⁹.

Now that we defined in very general terms what the blockchain is, we will look at each of its core aspects separately. By so doing, we will be able to analyze further aspects/applications of the blockchain technology that will require previous basic knowledge.

3.1.2 Basic elements of the blockchain (Nodes, Miners, Hash)

As briefly mentioned before, nodes are essential to the existence and functioning of the blockchain. We can think of nodes as the ensemble of computers and other electronic devices that sustain the structure of the blockchain. Nodes are basically the hardware on which the blockchain runs. The main purpose of nodes is adding new blocks to the chain or validating transactions of other nodes that attempt to add a new block. We will see shortly that the concept of nodes is strictly linked to the one of miner. As mentioned previously, nodes are the element on which the blockchain is built, therefore, all the information present in the blockchain is stored inside nodes. We can distinguish between different types of nodes based on the role that they play inside the network. Starting from the most to the least important,

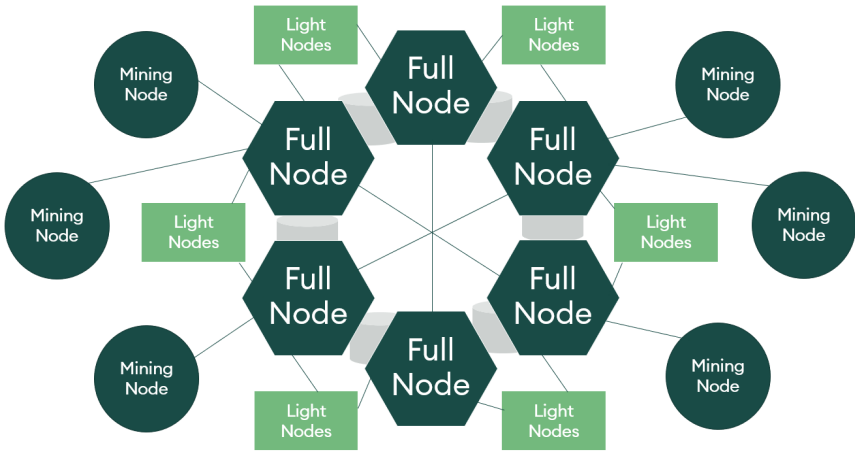
¹²⁶ <https://medium.com/coinmonks/blockchain-what-is-a-node-or-masternode-and-what-does-it-do-4d9a4200938f>

¹²⁷ More about different types of nodes in next sub-sections.

¹²⁸ Legal TechNews, (May 14, 2018), *A brief history of blockchain*, ALM publications.

¹²⁹ Chris Burniske, Jack Tatar, (2021), “*Cryptoassets: The Innovative Investor Guide to Bitcoin and Beyond*”.

we find miner nodes¹³⁰; This type of nodes are the ones which can add new block to the chain. Miner nodes do not need to hold the full history of the blockchain in order to validate a new block, they just need to acknowledge the data in the previous block. The second type of node is the full node, which is the node that contains all information in the network but that cannot add new blocks to the chain. Full nodes are essential to the survival of the blockchain because they hold the full history of the chain. The blockchain can theoretically run on just one full node, however, this would be highly risky because if such node happened to break down, the full chain would be jeopardized. Therefore, the higher the number of full nodes in the network, the safest is considered the blockchain. Lastly, light nodes are nodes that contain only part of the information in the network and usually they hold a block header. A block header is a detailed summary of a specific block that includes information relating to a particular previous block which is connected to. Light nodes fulfill the same purpose as full nodes; however, they do not hold a full copy of the blockchain and must rely on full nodes to check past transactions¹³¹. Since these concepts can sound a little harsh to a non-familiar user, here we present a graphical representation of how the nodes environment looks like:



132

To conclude this overview on nodes, we also distinguish between the status of a node, online-offline. On the one hand, online nodes receive, validate, and secure data as it is broadcasted to the network. They are always updated on the state of the chain. On the other hand, offline nodes are not updated periodically because of their status. Therefore, as they come back online, they must download all the blocks that have been added to the chain during their

¹³⁰ The concept of miner will be clarified in the next sub-section.
¹³¹ Seba Bank, (2020), The bridge, "<https://www.seba.swiss/research/Classification-and-importance-of-nodes-in-a-blockchain-network>"
¹³² Image 11: "Full, light and mining nodes illustrated on a blockchain", Seba Bank, (2020), The bridge, "<https://www.seba.swiss/research/Classification-and-importance-of-nodes-in-a-blockchain-network>"

“absence”. This process of downloading data and updating the status on the blockchain after an offline period is called *synchronizing with the blockchain*¹³³.

In order to fully understand the functioning of the blockchain, we must introduce the concept of “miner”. A miner is a computer or a group of computers that can add transactions to the blockchain (add new blocks) and verify blocks created by other miners¹³⁴. It becomes clear now that miners are what we previously defined as miner nodes. Basically, miners work to maintain the blockchain safe and functioning and get paid with “transaction fees” that will be discussed later on in this chapter. Whenever a miner attempts to add a new block to the chain, it broadcast the block to all the nodes present in the network. Depending on whether the transaction is valid (the hash of the transaction is legitimate and unique), other miners will validate/refuse the transaction and the block will be added/rejected to the chain. In case the block is found to be valid, the other full nodes/miners in the network will save and store the new transaction. In such a way, another block will be added “on top of the existing chain” and the other nodes will update the transaction history of the chain by adding the new transaction that has been validated. Such system of decentralized consensus needed for updating and changing the blockchain is called “consensus mechanism”. This mechanism implies that each action is approved by the (simple) majority of the validating network (miners) that agrees to the updated state of the ledger. Such principle is basically a set of rules and principles that allows the chain to remain coherent and unique. Once a new block is broadcasted to the network, every miner can attempt to validate the new transaction and getting rewarded with the “transaction fee”.

To validate a new transaction each miner attempt to solve highly complicated mathematical problem. Such quest involves finding a 64-digit hexadecimal number (hash) that is less than or equal to the target hash. To clarify the notion of hexadecimal number (or hash) we can present an example: a hexadecimal number is a 64-digit sequence of number and letters. The name hexadecimal is derived from the words “hex” which means “six” in Greek and “deca” which means “ten” in Greek. This suggests that every digit could assume 16 possible realizations (number or letter) and only one “nonce” (number only used once) is the correct hash (or the closest one).

Example of hash:

0000000000000000057fcc708cf0130d95e27c5819203e9f967ac56e4df598ee

¹³³ <https://medium.com/coinmonks/blockchain-what-is-a-node-or-masternode-and-what-does-it-do-4d9a4200938f>

¹³⁴ <https://www.igi-global.com/dictionary/has-bitcoin-achieved-the-characteristics-of-money/59928>

Miners attempt to find the correct solution by exploiting powerful computers that “guess” as many nonces as possible in a short window of time. Such complex numerical procedures require extremely strong computational power. The number of possible solutions to the mathematical problem increases as the number of miners in the network increase, this feature is known also as “mining problem”¹³⁵.

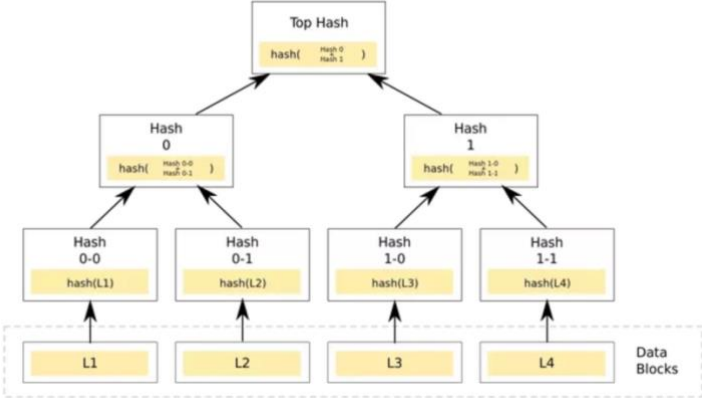
3.1.3 How new blocks are added to the chain

Now that the basic concepts of blockchain have been briefly introduced, we can analyze how (technically) new blocks are added to the chain and what ensures the safety and truthfulness of added blocks. Two fundamental elements of a block are the list of transactions and the block header. As briefly mentioned before, the block header contains information regarding the block itself and the previous block which is linked to the new one. More precisely, the block header is composed by the hash of the current block, the hash of the previous block, the date in which the current block has been hashed, and the target difficulty of the block¹³⁶. We can consider as an example, a hypothetical transaction to add a new block. Assume a block A which has already been successfully added to the chain and a block B which is trying to be added to block A (and therefore to the chain) by a miner. Past transactions have been added to block A and a 256-bit number that uniquely identifies A have been hashed (the hash of block A). In order to add block B, miners have to collect another set of transactions and add the hash of block A to block B. By so doing they hash the new set of transactions plus the hash of A to get the transaction for B done. The first miner able to get the new hash for B is considered the “winner” and is awarded the “transaction fee”. Now if anyone attempt to change the information contained in block A (already been added) that would change the hash of A and consequently the hash of B. If finding the “winning” hash to add a new block is a tremendously complicated work, finding the backwards solution to change hash A and B before any new block is added to the chain (every ten minutes in the Bitcoin blockchain for instance) is almost impossible. Another fundamental element of the blockchain technology

¹³⁵ <https://www.investopedia.com/tech/how-does-bitcoin-mining-work/>

¹³⁶ <https://medium.com/@blairlmarshall/how-do-miners-validate-transactions-c01b05f36231>

is the Merkle tree, which is basically what ensures the functioning of every decentralized network, overcoming the issue of transactions between untrusted parties. A Merkle tree is a “data structure that is used in computer science applications. In bitcoin and other cryptocurrencies, Merkle trees serve to encode blockchain data more efficiently and securely.¹³⁷”. The Merkle tree is also known as binary tree where the nodes stores hashes instead of storing chunks of data.



138

We can consider an example in which we have 4 data blocks, we apply the hash function to each chunk of data, and we obtain 4 “leaves nodes” also called “children nodes” because they are not tied to any previous node if not the original data (Hash 00; 01; 10; 11). If we concatenate the results of hash 00 and hash 01 and we apply the hash function, we will obtain a parent node (hash 0). This process can be repeated iteratively until we will obtain a final node (or Merkle root). The Merkle or final hash is unique and would change if any of the data in the parent or leaves nodes is changed. This ensures a fast and reliable checking process for verifying the authenticity of a block. The root hash assumes the role of address of a determined set of data.

To complete the discussion about “How new blocks are added” we must introduce the concept of difficulty of a block. As briefly stated before, the more the miners in the network, the more difficult will be finding the winning nonce that solves the mathematical problem discussed above. We already know that miners exploit extremely powerful computers to solve the problems. Therefore, the only way in which they can increase the possibility of finding the correct nonce is increasing the computational power of their computers. More

¹³⁷ <https://www.investopedia.com/terms/m/merkle-tree.asp>
¹³⁸ Image 12: <https://medium.com/@hmishfer17/blockchain-data-integrity-e70e17cac086>

precisely, the computational power of such computers is usually referred to as “hashing power/rate” or “hash per seconds”. Of course, the higher the hashing rate of a device, the higher will be the number of attempts to solve the problem that a computer can make in a pre-determined frame of time. Therefore, higher hashing rates will imply higher probabilities of being the first to solve the problem and being rewarded the famous “transaction fee”. That said, the difficulty of the block is decided based on the time that miners employ to successfully mine a new block. The difficulty is usually expressed as estimated mining power needed TH/s (Total hash rate per second). The harder for miners to get new blocks, the safer is considered the network. For instance, in the Bitcoin blockchain, the difficulty of blocks is adjusted every 2016 blocks (approximately every two weeks). The difficulty is set in such a way so that the time needed for mining a new block is approximately equal to ten minutes¹³⁹.

3.1.5 A key distinction

To clarify the topic “transaction fee” that we mentioned multiple times in the previous definitions we must state a simple but fundamental distinction. Since there are multiple types of blockchains, with different purposes, functioning and users (all based on the (almost) same principles) each one of them will function slightly differently and will give to miners different types of transaction fees. We will now present the typical transaction fee awarded to miners in a public blockchain like Bitcoin. Be careful not to confuse the Bitcoin blockchain with the famous bitcoin cryptocurrency. The Bitcoin blockchain is (as the name itself suggests) a blockchain on which the bitcoin (cryptocurrency) is based. To avoid any kind of confusion, the Bitcoin blockchain is usually referred to with the capital “B” (Bitcoin) and the bitcoin cryptocurrency is usually referred to with the lowercase “b” (bitcoin).

To continue the explanation on the “transaction fee” that miners receive when they successfully solve the so-called mathematical problems, we can take the example of the Bitcoin blockchain. In such case miners are awarded with cryptocurrency bitcoin, and the amount of such reward changes over time. When the Bitcoin blockchain first appeared online (2009), the transaction fee for successfully mining a block was 50 bitcoins, such reward today

¹³⁹ <https://www.blockchain.com/charts/difficulty>

would roughly be equal to 2.557.89,5€¹⁴⁰. However, roughly every four years the reward is halved, and that means that the reward today is equal to 6,25 bitcoins (approximately 319.328,13€¹⁴¹).

That said, the aim of this chapter is to introduce blockchain technology and to do so, we will try to generalize the concepts in order to provide an impartial definition. However, we will make some exceptions in some instances (talking about a specific blockchain) just to clarify the concepts.

3.1.6 Public vs Private blockchains

After the first appearance of the blockchain technology (2008), many others decentralized database inspired by Bitcoin have become popular. One important distinction that will be useful later on in this chapter is between public and private blockchains. Public blockchains, as the name suggests, are accessible by the general public, and anyone with an adequate hardware and software can join the network and acknowledge all past information present in the chain. Differently from public ones, private blockchains are accessible only by users that have the authorization to join the network. To simplify things, public blockchains are comparable to the internet and private blockchains are comparable to the intranet¹⁴². Both technologies benefit from the principle of the blockchain, but with different functioning and purposes. Starting from definitions, public blockchains are also called “permissionless” due to the lack of control of who can join the network. On the other hand, private blockchains are also called “permissioned” since members should have permission to join the network and sometimes, are also called DLT (distributed ledger technology). Examples of public blockchains are Bitcoin and Ethereum, while private blockchain are often used by corporations or consortiums that have trusted members and trade confidential information. Both types of blockchains use cryptography technology to allow each participant on any given network to manage the ledger in a secure way without the need for a central authority to enforce the rules¹⁴³.

¹⁴⁰ Value on the 19th of November 2021.

¹⁴¹ Value on the 19th of November 2021.

¹⁴² Chris Burniske, Jack Tatar, (2021), “*Cryptoassets: The Innovative Investor Guide to Bitcoin and Beyond*”.

¹⁴³ Tiana Laurence, (2019), *Blockchain for dummies*, John Wiley & Sons.

Another key distinction is between closed and open blockchains. Whereas the difference between public and private blockchains relates to the type of people that can access and write data in the ledger (everyone vs authorized), closed vs open distinguish who can read such data. Therefore, we can distinguish between four combinations of features of the blockchain, public and open, public and closed, private and open, private and closed¹⁴⁴.

- In a public and open blockchain, anyone can access and add data to the network and can also read all previous transactions in the chain. This type of blockchain is the one used by Bitcoin, Ethereum and many others blockchains. Such types of blockchains are perfect for currency transactions (cryptocurrencies) and videogaming purposes.
- In a public and closed blockchain, anyone can join the network and write on it (as in public and open blockchains). However, users cannot access and read all past transactions occurred in the network. An ideal application of such technology would be for voting purposes, users are granted anonymity and can only vote, without being able to acknowledge what other users have voted.
- In a private and open blockchain, only permissioned users can access the network and contribute to the ledger. However, any user can also read and access all information in the chain without asking permission. Such technology would be perfect for supply chain management purposes, where only authorized members can join the network (for instance, employees) but anyone can read all transaction/data in order to be fully updated on the state of the chain.
- In a private and closed blockchain, access is granted only upon permission by the “host” and users can only contribute to the chain, without being able to access past information. Such technology would be useful in military applications, as well as national defense or law enforcement. That is because, only authorized users could join the network (police agent or a cadet) and could not access all previous data in the chain (due to privacy reasons for instance).

As we have seen, different blockchains fulfill different purposes. One fundamental aspect that plays an important role in such distinction is identity. The peculiarity of public blockchains is anonymity, as already stated, anyone with an adequate hardware and software can access the network and interact with other participants. Such anonymity is vital to users,

¹⁴⁴ <https://medium.com/coinmonks/public-vs-private-blockchain-in-a-nutshell-c9fe284fa39f>

think of a person which holds one million dollars in cryptocurrencies, they would want to maintain private their name and personal data. In public-open blockchains users are usually recognized through addresses, which are not linked by any means to personal data. Anyone can access all data available in the network (including anyone's wallet of cryptocurrency for instance), but nobody knows the real identity of other users. However, this anonymity comes at a price, because if anyone can do anything without being publicly recognized, what prevents users to behave incorrectly? For that reasons, public blockchains developed principles that allow the safety and truthfulness of operations, exploiting economic incentives. Different types of public blockchains employs different consensus mechanisms to guarantee that rules are being followed and that nobody is able to behave badly. One example of consensus mechanism is called Proof of Work or (PoW) and is basically what we described in the sub-section "1.1.3 Miners". Mining nodes have the incentive to check the validity of every new block through complex and expensive mining techniques (finding the nonce) and in return, they get economically compensated (cryptocurrency in most of the cases).

While public blockchains have to adopt consensus principles to implement a safe and sound environment for users, private blockchains does not. Since the very nature of private blockchains is authorization to entrance, users' identities are known by the "host" and therefore users are "naturally" motivated to behave properly. Think of a private blockchain in which users are the employees of a firm, no reasonable person would try to "cheat" because he/she would be recognizable by the authority (the firm's division that manages the chain) and could suffer the consequences of such act.

3.1.7 Consensus mechanism

A consensus mechanism is "*a fault-tolerant mechanism that is used in computer and blockchain systems to achieve the necessary agreement on a single data value or a single state of the network among distributed processes or multi-agent systems, such as with cryptocurrencies.*¹⁴⁵". To further simplify the definition, a consensus mechanism can be seen as a set of rules that should be implemented in a blockchain, in order to maintain correctness

¹⁴⁵ <https://www.investopedia.com/terms/c/consensus-mechanism-cryptocurrency.asp>

and truthfulness of the public ledger. As we mentioned before, consensus mechanisms are peculiar of public blockchain due to their anonymous nature. However, also private blockchains exploit different types of consensus mechanisms.

The first (and most popular) type of consensus mechanism is called Proof of Work (PoW). As mentioned before, the PoW mechanism is the one which was first adopted by the Bitcoin blockchain and gained large popularity in later years. To explain how the Proof-of-Work mechanism works we must have clear in mind the concepts of sections 1.1.3 and 1.1.4. PoW basically is a way of proving that you have invested (computational) resources into a task¹⁴⁶. In the case of PoW-based blockchains, the computational resources are referred to the high level of electricity needed to sustain the mining node equipment such as powerful computers and servers. That is because, buying and maintaining such equipment is far more expensive than one can think. For instance, we can take the case of two siblings, Ishaan and Aanya, respectively with fourteen and nine-years-old. They were cited by CNBC¹⁴⁷ for their successful mining activity, in the article it is stated that they currently own approximately 97 Nvidia RTX 3090 graphic cards¹⁴⁸ each one of those costs approximately 3000\$, which would imply a total cost of 291.000\$ just for the hardware. Moreover, the article states that they pay electric bills of around 3.000\$ per month, plus the costs for renting a data center in Dallas. This is just one of the many examples that can prove how expensive mining could be. Going back to the Proof-of-Work mechanism, the high costs that miners bear, are the proof that they have invested computational and economic resources to operate. This mechanism implies a fair level of security over the network. That is because if any anonymous user ever tries to change a past transaction in the chain, he would challenge the computational power of the whole network which has incentives in not to change past transactions. As long as the simple majority of the validating network (51% of miners) maintain a fair behavior, the public ledger is considered immutable. The Proof-of-Work system is based on the simple concept that it should be difficult to validate new blocks, while it should be easy to verify the validity of it. PoW mechanism has been largely criticized due to the large amounts of energy that it consumes. For that reason, other types of consensus mechanisms have gained popularity, the most promising one is Proof-of-Stake (PoS).

The aim of a PoS consensus mechanism is the same as the PoW; however, the way to achieve the goal is different. Whenever a new block is to be added to the chain, each “validator”

¹⁴⁶ Horst Treiblmaier, Roman Beck, (2019), *Business transformation through blockchain*, Volume II, Palgrave Macmillan.

¹⁴⁷ <https://www.cnbc.com/2021/08/31/kid-siblings-earn-thousands-per-month-mining-crypto-like-bitcoin-eth.html>

¹⁴⁸ A powerful type of computer processor available in the market.

should stake a portion of his/her cryptocurrency in the blockchain. Of course, the crypto to be staked must be the native crypto of the blockchain of interest. The validators are the equivalent of the miners in the PoW ecosystem, the difference is, they do not have to invest in expensive equipment to have a chance of validating a block. To try to be selected to validate a new transaction, each validator stakes a portion of his/her crypto, a deterministic algorithm chooses the new validator based on the number of coins staked and other variables. In such a way, there is no need to consume huge quantities of energy and the system is less centralized than in a PoW ecosystem. Differently from a PoW consensus mechanism, the validators do not get a block reward for validating a new transaction (6,25 bitcoins in the Bitcoin blockchain), instead, they gain from transaction/network fees¹⁴⁹. For what extent security PoS mechanism is as safe as PoW, that is, for a hacker to crack the system, double spend a coin, or implement other malicious actions he will need to possess 51% of the cryptocurrency. Moreover, whenever such hacker will try and hack the system, he will lose the whole amount of cryptocurrency that he staked (51%). To conclude, the PoS consensus mechanism offers some advantages over the traditional PoW:

The main advantage of the PoS consensus mechanism is the incredibly smaller amount of electricity implemented to maintain the blockchain safe and functioning. Without the powerful servers exploited in the PoW ecosystem, the PoS mechanism can be considered eco-friendly. In fact, it has been estimated that the bitcoin mining activity (PoW-based) consumed 121,36 terawatt-hours per day in 2021¹⁵⁰, that is more than the electricity consumption of Switzerland, Argentina, Philippines or the Netherlands.

In the PoS ecosystem, validators are discouraged to validate fraudulent transactions, that is because, they would lose the whole amount of coins staked as a collateral.

Scalability is much higher in a PoS system, in a PoW environment the transactions are slower than in a PoS system and that is due to the huge computational power required. Therefore, transaction rate in a PoS consensus mechanism is much higher and faster.

¹⁴⁹ <https://www.investopedia.com/terms/p/proof-stake-pos.asp>

¹⁵⁰ <https://www.profolus.com/topics/pos-advantages-and-disadvantages-of-proof-of-stake/>

3.1.8 Smart contracts, DApps and DAO

Now that we have enlisted and explained some of the main features and peculiarities of the blockchain technology, we can discuss Smart Contracts and DApps. These elements will be a key knowledge requirement when we will discuss decentralized exchanges in chapter 4. We will start by analyzing Smart Contracts.

Smart contracts were firstly theorized by Nick Szabo in 1997 with an article named “*Formalizing and securing relationships on public networks*”¹⁵¹. The concept of smart contract already existed but it would have been necessary the rise of the blockchain technology to better appreciate the utility of such instruments¹⁵². The definition of smart contract according to Szabo is: “*An electronic transaction protocol that executes the terms of a contract. The general objectives are to satisfy common contractual conditions (such as payment terms, liens, confidentiality, and even enforcement), minimize exceptions both malicious and accidental, and minimize the need for trusted intermediaries. Related economic goals include lowering fraud loss, arbitrations and enforcement costs, and other transaction costs*”¹⁵³. Already from this definition we can understand the purpose and aim of a smart contract. They are designed to allow the safe execution of contracts between people without the need of a central intermediary. Under a practical point of view, smart contracts are programs that run on the blockchain under pre-defined circumstances. Since one of the aims of the smart contract is the execution of transactions without the need of a central authority, it comes naturally that the blockchain technology is the perfect tool to match with. To recap, smart contracts present the following features¹⁵⁴:

- Automatically executable: smart contracts do not need the human intervention to be executed. They abide to the code that has been written and operate in a fault-tolerant way¹⁵⁵.

¹⁵¹ <https://firstmonday.org/ojs/index.php/fm/article/view/548>

¹⁵² Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁵³ <https://firstmonday.org/ojs/index.php/fm/article/view/548/469>

¹⁵⁴ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁵⁵ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

- Enforceable: all transactions and contracts are automatically enforced by the code which is considered as law. Smart contracts automatically adapt to different situations and under pre-defined circumstances, enforce the contract¹⁵⁶.
- Secure: smart contracts are safe for definition; they are built on the blockchain technology which allows them to be tamper-resistant and immutable. However, it must be considered that in order to be completely safe and immutable the code must be properly written and bug-free¹⁵⁷.
- Deterministic: smart contracts are built in such a way that if a same input is entered, the same output will be returned. That is because the code provides a deterministic function that operates based on data¹⁵⁸.
- Unstoppable: this feature means that smart contracts are executed no matter what adversaries or problem comes up. Moreover, when the smart contracts execute, they complete their performance deterministically in a finite amount of time¹⁵⁹.

It is easily noticeable why smart contracts have gained massive popularity in recent years. They allow the execution of contracts relating to any possible field in a safe, transparent, fast, and reliable way. Smart contracts applications are multiple, among the most promising ones we find trade finance, records, property ownership, mortgages, insurance, medical research¹⁶⁰, etc.

As we already stated, smart contracts need a blockchain to operate in a safe manner. For the sake of completeness, it is worth mentioning that Ethereum is a popular blockchain that is known for hosting a huge number of smart contracts and decentralized apps. More precisely, smart contracts are executed thanks to the EVM (Ethereum Virtual Machine) which allows the execution of programs and smart contracts designed for the blockchain users.

Decentralized Application (DApps) are applications that run on a blockchain. More precisely are a part of the software that interact with the blockchain and regulates the state of all network members¹⁶¹. To better understand the difference between smart contracts and DApps we provide the followings: DApps are blockchain-based websites, whereas smart contracts

¹⁵⁶ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁵⁷ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁵⁸ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁵⁹ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁶⁰ <https://www.devteam.space/blog/10-uses-for-smart-contracts/>

¹⁶¹ <https://medium.datadriveninvestor.com/what-is-the-difference-between-smart-contracts-and-dapps-d252d88d32d3>

are API connectors¹⁶² which connect the DApps with the blockchain. DApps are a broad system that comprehend smart contracts, DApps are a combination of smart contract and front-end code as a complete computer program¹⁶³. Based on the functioning protocol of the DApps, we find three different types:

- Type 1: DApps that run on their own blockchain, in such category we count smart-contracts-based DApps running on Ethereum¹⁶⁴.
- Type 2: DApps that run on a blockchain which is not their own. They run on already existing blockchains and bear custom protocols of type 1 blockchain¹⁶⁵.
- Type 3: DApps that use protocols of Type 2 DApps. Type 3 DApps are basically DApps which run on a Type 2 DApps which are based on a Type 1 Blockchain¹⁶⁶.

To conclude the overview on software that runs on the blockchain we must introduce the DAO, Decentralized Autonomous Organization. To provide a definition: “A *decentralized autonomous organization (DAO)* is a software running on a blockchain that offers users a *built-in model for the collective management of its code*.¹⁶⁷”. Instead of being an organization governed by a group, DAOs exploit a set of pre-defined rules written in a code and enforced by the network (nodes) which runs a shared software (blockchain)¹⁶⁸. To become a member of a DAO, users need to buy its cryptocurrency (or token). Usually, holding the token give the owner the right to vote, the decisional power will be determined by the amount of token possessed by the user¹⁶⁹. The first DAO was created and run on the Ethereum Blockchain. According to such definition, the concept of DAO can be very similar to DApp; However, there are some differences. A DAO can be seen as a fully autonomous DApp, but DApp is not necessarily a DAO. The key difference stands in automation, a DAO can be seen as an intelligent DApp, with a much higher level of automation¹⁷⁰.

To conclude this section on the blockchain technology some comments are worth mentioning. Although the aim of this thesis is to analyze and discuss different aspects of the

¹⁶² API is the acronym for Application Programming Interface, which is a software intermediary that allows two applications to talk to each other.

¹⁶³ <https://medium.datadriveninvestor.com/what-is-the-difference-between-smart-contracts-and-dapps-d252d88d32d3>

¹⁶⁴ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁶⁵ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁶⁶ Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

¹⁶⁷ <https://www.kraken.com/learn/what-is-decentralized-autonomous-organization-dao>

¹⁶⁸ <https://www.kraken.com/learn/what-is-decentralized-autonomous-organization-dao>

¹⁶⁹ <https://www.kraken.com/learn/what-is-decentralized-autonomous-organization-dao>

¹⁷⁰ <https://medium.com/swlh/whats-the-difference-between-dapp-idapp-and-dao-and-why-they-are-the-future-of-blockchain52758f50474e>

big data environment, this section helps us understand the topics that will be presented in the next. Blockchain technology is rapidly developing and evolving to disrupt and change more field that one can imagine. The scope of applications ranges from supply chain management, sanitary field, military operations, to big data applications. In the next sub chapter, we will see how blockchain technology can disrupt the management of big data and how the two technologies lead to natural improvement of the process. Moreover, we will exploit the knowledge acquired in this sub-chapter also in the fourth chapter, when we will discuss cryptocurrencies in a big-data-related case study (HUDI).

3.2 Integration of blockchain technology and big data

In this sub-chapter we will discuss the pros and cons of the integration between the blockchain technology and big data. We will first go through the main big data challenges, briefly mentioned in section 2.2.1, to better understand the deficient areas of big data. Then we will briefly analyze the main blockchain features and details that makes it a useful resource for big data applications. Finally, we will analyze some real-world cases/projects that exploit both big data and blockchain technology to implement new strategies.

3.2.1 Storage and security issues

When we refer to big data challenges we can mention a lot of features, issues, and peculiarities that could create some trouble to users and gatherers. As first feature of big data that can be seen as a “challenge” we find the storage solutions. As we already know, big data can prove to be very useful to corporations and other entities, especially when those entities based their whole business model on the exploitation of big data (see section 1.4). As a consequence to that, firms and other entities must have a place (virtual or physical) to safely store their data. Among the most popular solutions for data storage we find constant encryption, warehouse storage, and cloud. Constant encryptions consist in converting all available data to encrypted code, such code can be accessible and readable only to the user which possess the decryption key. Another solution is buying a physical warehouse storage for big data. Even if such solution can sound obsolete, there may are situations in which it can represent a reasonable choice. Big companies which possess huge amounts of data could not physically store it inside a warehouse, the volume and costs of such solution would jeopardize the utility. However, a small company may find useful to temporary store its data inside a physical warehouse which contains a storage center¹⁷¹. Finally, another possible solution for storing data can be the cloud. Cloud computing is a term that includes delivering hosted services among the internet, the type of clouds available are IaaS, PaaS, and SaaS. A SaaS is a distribution model that delivers software applications over the internet¹⁷². A PaaS is a cloud service that allows users to develop tools on the host infrastructures¹⁷³. Finally, an

¹⁷¹ <https://www.smartdatacollective.com/big-data-stored-managed/>

¹⁷² <https://searchcloudcomputing.techtarget.com/definition/cloud-computing>

¹⁷³ <https://searchcloudcomputing.techtarget.com/definition/cloud-computing>

IaaS (Infrastructure As A Service) is a type of cloud that offers storage and computing force to its users. We will focus only on IaaS-type of cloud. A cloud can either be public or private, a public cloud is accessible by everyone on the internet upon payment. On the other hand, a private cloud is accessible only by a pre-defined set of people with an access key. The key features of the cloud are its accessibility and flexibility, firstly, the cloud is accessible by any possible physical location in the planet, it is necessary just an internet connection. Then, it provides an insurance against a possible device disruption, that means that if a company stores its data into a physical warehouse storage and the warehouse burns down, that data is long gone. On the other hand, if the data were stored in the cloud, it could be easily downloadable and usable even after a warehouse fire.

Cloud storage solutions offer an interesting way of storing sensible data thanks to its low cost, high flexibility, and mobility. However, there are some features that can cause some problems with respect to clouds. More precisely, security is a primary concern when discussing cloud solutions, in fact, there are some sensible areas which often cause security issues:

Hacker attacks: this feature represents the most critical aspect of cloud security issues. When a corporation or a data gatherer exploits a public/private cloud to store its data, it must consider that it will be subject to possible cyberattacks. In a 2021 cloud data security report by Netwrix¹⁷⁴, it is showed that over half of the organizations (54%) that store customer data have had security issues linked to the cloud in 2020. As a consequence to that, 62% of interviewed plan to remove customers' sensible data from the cloud to improve their data security policy. It is vital for organizations to consider additional safety measures such as enhanced firewalls, and periodical security tests.

Misconfiguration: this feature is considered to be one of the main causes of cloud security issues. It consists in failing to properly configure the settings of the cloud to ensure a maximum safety level. Such configurations can concern making the cloud easily accessible to all users and maintaining a proper amount of access control at the same time¹⁷⁵. Misconfiguration can easily lead to the access in the cloud of unauthorized users which can cause serious data leaks.

¹⁷⁴ <https://www.netwrix.com/download/collaterals/2021%20Netwrix%20Cloud%20Data%20Security%20Report.pdf>

¹⁷⁵ <https://www.buchanan.com/cloud-computing-security-issues/>

Internal security: many instances of data leakage are caused by internal users which act with malicious intents. In fact, a 2020 report by Verizon¹⁷⁶ suggest that around 30% of data breaches was caused by internal actors¹⁷⁷.

Compliance with regulations: as we have seen in chapter 2, the world's lawmakers have engaged a progressive regulation of data-linked businesses. For instance, the European Union with the GDPR has increased the amount of security required by data gatherer to be compliant with the law. However, an interesting survey by Commvault¹⁷⁸ proved that only 12% of organizations properly understand how GDPR impact on their cloud businesses. This could imply that many operating cloud providers are not fully compliant with the law, causing possible concerns with respect to data security and privacy.

To recap, big data poses some challenges with respect to storage solutions and data security. The cloud does not assure full data integrity and a tamper-free solution, this suggest that, thanks to the feature of the blockchain, an integration of the two technologies can be fruitful.

3.2.2 A fruitful integration

Big data and blockchain technology have been the main characters of the digital revolution over the last decade. Both technologies present some unique features which makes them extremely useful and exploitable. In this section we will analyze why the integration between big data and blockchain can be a smart choice for overcoming some typical big data issues.

Firstly, we can consider the main features that makes the blockchain unique and useful:

Security: as we have seen in the first sub-chapter (3.1), the blockchain technology is popular for its safety and security. Thanks to its block-design structure and the decentralized architecture it is believed to provide an additional layer of security. The level of security of a blockchain depends on different factors, starting from the type of blockchain (private, public, consortium), and the purpose of the ledger. However, the fundamentals of the technology are common to all types of blockchain and represent a good starting point for implementing an enhanced security environment. If we consider a public blockchain with a PoW (Proof of Work) or PoS (Proof of Stake) consensus mechanism, the likelihood of

¹⁷⁶ <https://www.verizon.com/business/resources/reports/2020-data-breach-investigations-report.pdf>

¹⁷⁷ <https://www.buchanan.com/cloud-computing-security-issues/>

¹⁷⁸ <https://www.forbes.com/sites/forbestechcouncil/2018/07/05/four-trends-in-cloud-computing-cios-should-prepare-for-in-2019/?sh=6c4146c44dc2>

incurring in a security breach is not 0%. However, in order for a blockchain to be hacked, the simple majority of the validating nodes should be compromised. Therefore, for blockchains with a high number of nodes (high decentralization) it is possible, but highly unlikely the possibility of being hacked. The validation algorithm implemented in the blockchain ensures that the chain is non-manipulative and tamper-free.

Decentralization: the main feature of the blockchain technology is decentralization, which comes with numerous advantages. Firstly, since the data is not stored in a single spot, it makes it much more difficult for hackers to attack the system and access sensible data. Moreover, not having all the data under the control of a single entity comes with multiple pros. If the centralized entity goes through a software update, the system would not be accessible for a determined period of time. In a worst-case scenario, if the centralized entity happened to shut down for every reason, the data would remain inaccessible for anyone, causing a lot of damage to data gatherers¹⁷⁹. The blockchain environment is composed by a lot of nodes, each one of them holding the complete history of the blockchain (full nodes). theoretically if all nodes were shut down except for one, the blockchain would still be able to function properly. A decentralized environment ensures that all users can interact with each other without the need of a centralized party to oversee the interactions. Therefore, the blockchain is referred to as a peer-to-peer (Or P2P) system. As the name itself suggests, peer-to-peer systems are based on the transfer of information, currency, or other type of data directly from one user to another, with no need of an intermediary in order to carry out the transaction.

Transparency: another fundamental feature of the blockchain is transparency. In a public blockchain, all users are recognized through their public address. This implies that the identity of a person remains anonymous in order to ensure privacy. However, every user with access to a public blockchain can access all information with respect to a specific address. In other words, the identity is anonymous, but everyone can see each other's transactions in the blockchain. That is, if you knew the public address of a company, you could see how many crypto it possesses and all its previous transactions¹⁸⁰.

Immutability: this is one of the fundamental features of the blockchain, strictly connected to the aforementioned security feature. Immutability means that once information is added to the chain, it cannot be changed. This feature is ensured in different ways depending on the

¹⁷⁹ Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) "*Blockchain, Big data, and Machine learning: trends and applications*", CRC Press.

¹⁸⁰ Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) "*Blockchain, Big data, and Machine learning: trends and applications*", CRC Press.

blockchain we are into, for instance, in a PoW-based blockchain the SHA-256 hashing function ensures that each transaction is validated and tamper-free. In section 3.1.3 we suggested how the hashing process is a difficult task to be carried out moving forward, let alone performing it backwards in less than 10 minutes (in the Bitcoin blockchain).

All these features pave the way of the integration between blockchain technology and big data. More precisely blockchain can improve big data in several ways. The integration of such technologies can be more useful than one might think, while big data is used for prediction, blockchain technology is used for data validation and integrity. All this is carried out without the need of a third-party which act as a central authority. We will analyze the main areas of big data which can be improved by blockchain technology.

The first field that would benefit from an integration between big data and blockchain technology is data integrity. We can define data integrity by exploiting the definition provided in the Harvard Business School website, “*Data integrity is the accuracy, completeness, and quality of data as it’s maintained over time and across formats.*”¹⁸¹. From this definition we can grasp that data integrity refers to the quality, reliability, truthfulness, and consistency of data. Blockchain can improve this feature by its very definition, thanks to its tamper manner nature, it can be proved that data has not been modified in the past. Blockchain can be used for proving the source of data. Each single transaction/action carried out on data is written in an immutable way in the history of the blockchain and it cannot be deleted. Two main challenges to data integrity are¹⁸²:

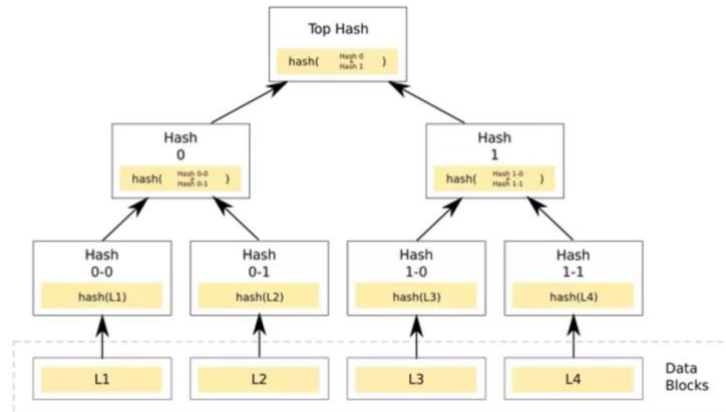
- Proving data source: proving where and when the data has been acquired and who gave the permission to acquire the data.
- Proving data authenticity: proving that the data is authentic, and it is not modified, or in case it has been modified, proving that the changes were legitimate.

Merkle trees (see section 3.1.3) are a fundamental element that can ensure data integrity. Thanks to the Merkle tree data structure, hashes of child nodes are combined into the parent node’s header and this technique is repeated iteratively until a final node (or root) is reached. The final node is unique and act like a fingerprint for the entire tree¹⁸³.

¹⁸¹ <https://online.hbs.edu/blog/post/what-is-data-integrity>

¹⁸² <https://www.zdnet.com/article/three-data-integrity-challenges-blockchain-can-help-solve/>

¹⁸³ <https://medium.com/@hmishfer17/blockchain-data-integrity-e70e17cac086>



184

Image 13 shows how 4 different data transactions are validated. Each data block is individually hashed (Hash 0-0; Hash 0-1; Hash 1-0; Hash 1-1), then each couple of hash is hashed and stored into a parent node (Hash 0 and Hash 1), finally the two parent nodes are combined into the final node which provides a timestamp and a nonce which is used to generate the block header. Therefore, Merkle trees provide the hash-based architecture or blockchain which ensures data integrity.

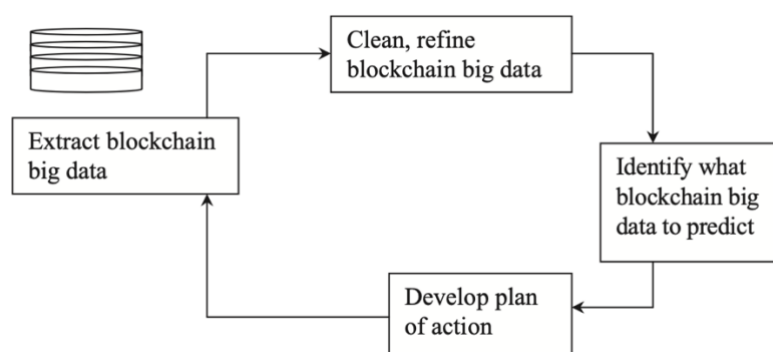
Another field that can be improved thanks to the integration between big data and blockchain is security¹⁸⁵. As we already know, the blockchain technology is known for being safe and immutable. Such features come in handy when we are considering the management of large amounts of sensitive data. Each user who tries to change, erase, or modify existing data would face the computational power of the entire network. Such user would be easily noticeable and banished from the network. The consensus algorithm principle ensures the well behavior of the validating nodes, guaranteeing a high level of data security. Moreover, as already discussed in the previous sub-chapter, to change the blockchain rules it would be necessary that the simple majority of the validating network agree to that. Of course, the greater number of nodes, the more unlikely will be such possibility.

Another field that would benefit from the integration of the two technologies is predictive analysis. Blockchain data can be exploited to analyze and retrieve some precious insight from large data sets (as in the normal process). Such insights then, can be used to forecast future events, trends or preferences referring to a specific market area. The real advantage of exploiting blockchain data to do so, is that such data is already structured and ready to be

¹⁸⁴ Image 13: <https://medium.com/@hmishfer17/blockchain-data-integrity-e70e17cac086>

¹⁸⁵ Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) “Blockchain, Big data, and Machine learning: trends and applications”, CRC Press.

processed. Moreover, thanks to the distributed network which provides huge computational power, data scientists can implement intensive predictive analysis in less time¹⁸⁶. Here we present a graph depicting the process:



187

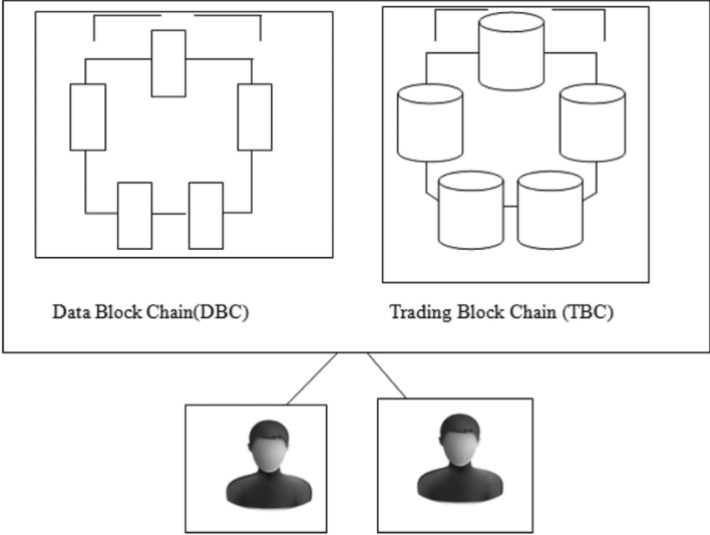
Real time data analysis is another field that would benefit from the integration between big data and blockchain. As we already know, the blockchain technology allows for the (almost) instant exchange of tokens such as cryptocurrencies and other assets. Especially in the banking and finance sector such technology has raised quite the interest. The possibility of exchanging huge amounts of money (for instance) without any delay in time, no transaction costs and with almost no concerning regarding the security is revolutionary. Such features are useful also when we consider the data analysis framework, since transactions are carried out almost instantaneously, also real time data analysis is possible. Organizations could track, process, and exploit real time transaction data to better implement fast decision making such as blocking suspicious transactions, suggesting different purchase solutions to customers, etc. The integration between big data and blockchain will also lead to a decrease in cost for storing and processing data. Moreover, thanks to the decentralized network architecture, companies could process data much more quickly and in a more efficient way. Finally, the access to data could be facilitate by making people part of the blockchain.

To conclude this analysis of the possible field of improvement thanks to the integration between big data and blockchain we shall talk about management of data sharing. Whenever data is stored inside the blockchain network and accessible to all data scientist, the level of coordination would increase exponentially. The data that has already been processed would

¹⁸⁶ Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) “Blockchain, Big data, and Machine learning: trends and applications”, CRC Press.

¹⁸⁷ Image 14: Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) “Blockchain, Big data, and Machine learning: trends and applications”, CRC Press.

not be processed again erroneously. Moreover, the blockchain would allow to share the data with a determined network of analysts without incurring in the risk of data breached and leaks. The management of data sharing is composed of two types, Data Blockchain (DBC) and Trading Blockchain (TBC). To better explain the process, we present a graph depicting the architecture:



188

The two blockchains operate in an independent manner, the Data Blockchain (DBC) stores the actual user data and the trading Blockchain (TBC) stores only relevant information which is useful for executing the transaction. Both blockchains exchange information between each other and data authenticity is achieved via cryptographic techniques¹⁸⁹.

3.2.3 Real-life case studies of integration

In this section we will discuss some real-life case studies of integration between the big data environment and the blockchain technology.

Storj: Storj is a decentralized cloud storage founded by Shawn Wilkinson, James Prestwich, Jim Lowry, John Quinn, and Tome Boshevski in 2014¹⁹⁰. Storj provides a service similar to those of the most popular cloud providers like Amazon Web Services and Google Cloud. The

¹⁸⁸ Image 15: Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) “Blockchain, Big data, and Machine learning: trends and applications”, CRC Press.

¹⁸⁹ Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) “Blockchain, Big data, and Machine learning: trends and applications”, CRC Press.

¹⁹⁰ <https://www.storj.io>

peculiarity of Stroj is that it offers a blockchain-based infrastructure to allow customers to store their data in. When customers use Stroj to store their personal/company data, the information is divided in more than 80 pieces and encrypted with an end-to-end encryption technology. Moreover, since the cloud is blockchain based, the fragments of the data are distributed among a vast number of nodes active in the world. Currently, the number of available nodes in the Stroj platform is 13'112¹⁹¹, this allows the company to have available space for 6,6 Petabytes overall. When exploiting Stroj for storing personal data, customers can benefit from the decentralized and encrypted nature of the service. Since the data is not stored in one place but it is divided into 80+ nodes, this makes it almost impossible for a hacker to access the whole data. Moreover, each single piece of data is encrypted for ensuring the maximum amount of reliability. Stroj enforces a strong privacy policy, each file is encrypted before being uploaded in order to allow access only to the owner of the data and the people the data is shared with. For what extent the performance of the service, Stroj claims to offer faster download rates than other centralized cloud provider, thanks to the decentralized structure that provides additional computational power. According to a study of VentureBeat¹⁹², exploiting a blockchain-based cloud service could reduce the storage costs of data of 90%.

Filecoin: Filecoin is a project launched by Protocol labs in 2017, whose founder is Juan Benet (current CEO). Filecoin is similar in purpose to Stroj, in fact it offers a blockchain-based cloud storage service to privates and companies. The mission of Filecoin is to exploit the world's unused storage space to provide a worldwide decentralized cloud service in a transparent way. The environment of Filecoin is composed by three actors¹⁹³: users, storage miners, and retrieval miners. Users are the clients who wish to store their data in a safe and affordable way, the data is encrypted and divided into various nodes to ensure maximum decentralization and security of the data. Storage miners are individuals that hold and store user's data in their hard disk/drive and in exchange they are compensated with a cryptocurrency named FIL. Retrieval miners on the other hand, are individuals which help users retrieve their data among the numerous nodes of the network. As storage miners, also retrieval miners are compensated for their work with the cryptocurrency FIL¹⁹⁴. FIL is the

¹⁹¹ Measured on February 2nd

¹⁹² Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) "*Blockchain, Big data, and Machine learning: trends and applications*", CRC Press.

¹⁹³ <https://filecoin.io/store/>

¹⁹⁴ Topic regarding cryptocurrencies will be discussed in next chapter.

native asset (cryptocurrency) of the Filecoin ecosystem¹⁹⁵. One of the main goals of Filecoin, is to reduce the risks of attacks and leakages of user's data while providing a cheap and much more affordable service. In fact, decentralization paved the way to progressive costs distribution and economies of scale. Such features will ease the transition from an oligopoly (currently few big players manage the cloud service worldwide) to a widely distributed and decentralized environment.

Omnilytics: Omnilytics is a Malaysian company founded in 2017 by Kendrick Wong, Sylvia Yin, and Nikolai Prettnner¹⁹⁶. It operates in the business of blockchain-based big data analytics, it exploits artificial intelligence and machine learning to obtain real-time data on trends, preferences, orders, etc. Omnilytics supports companies to exploit real-time data analytics for gaining a competitive advantage over their competitors. Thanks to the large amount of data points, and to the decentralized infrastructure of the blockchain they can gain real-time data analytics and provide customers (companies) with strategic advisory regarding pricing of the products, positioning, trends on products. Omnilytics focused its business model on the fashion and clothing area but provides insights also to other businesses¹⁹⁷.

Provenance: Provenance is a company founded by Jessi Baker in 2013¹⁹⁸ and it operates in the sector of transparent communication for businesses. More precisely Provenance exploits a blockchain infrastructure to follow step-by-step the data relating to some specific product. The idea of the CEO is that, due to progressive globalization, customers have been detached to the world of production. People do not know the impact that a product can have on the environment and on producers¹⁹⁹. Thanks to Provenance's blockchain-based technology producers can provide verified data as the product passes through the various steps of the supply chain. By so doing, the final customer will be able to acknowledge the origin of the product, the various steps that it went through, the people concerned with the production, the impact of the ecosystem, etc.

¹⁹⁵ <https://blog.coinlist.co/filecoin-why-its-a-big-deal/>

¹⁹⁶ <https://pitchbook.com/profiles/company/433136-26#overview>

¹⁹⁷ <https://omnilytics.co/omnilytics-overview>

¹⁹⁸ <https://www.crunchbase.com/organization/provenance>

¹⁹⁹ <https://www.provenance.org>

3.2.4 GDPR regulation on blockchain-based data processing

In chapter 2 we analyzed the main implications and principles of the data privacy regulatory environment for both the European Union and the United States. In analyzing the features and implication we focused on the main “traditional” cases of data gathering. As a consequence to that, we skipped the implication of the regulations on entities that employs blockchain-based technologies to gather and process big data. Now that we analyzed and properly understood the functioning and advantages of the integration between blockchain and big data, we can briefly analyze the implications of the regulatory environment.

As we already know from the previous sections, blockchain enhances big data in multiple ways. Among the main advantages we find data integrity, increased security, predictive and real-time analysis, and better data sharing. All such features make the process of gathering and processing big data a more fair, secure, and transparent process. To this end, it should be natural thinking that the GDPR regulation does not concerns with imposing additional security measures to this process, seen its reliability. However, there are some features of the blockchain environment that cause some issues when discussing regulations.

First²⁰⁰, the GDPR is based on the individuation of the so-called data controller (see section 2.2.2). Such controllers are the subject which are responsible for the management and well-processing of the data. The data controller is the person/entity whom data subject can address to enforce their rights under the EU protection law. However, in a blockchain-based environment the ownership of the data is distributed over a decentralized network of nodes. In such instance, the GDPR regulation may struggle to locate and address some responsible entities to enforce its principles.

Second²⁰¹, the GDPR assumes that the data which has been previously collected can be modified, updated, and in some cases, erased from the controller database (see articles 16 and 17 of GDPR²⁰² or section 2.2.2). However, due to the very nature of the blockchain and its consensus mechanism, it is almost impossible to modify some previously added data (modifying a previous block, see section 3.1.3). These features cause some issues that may arise from the application of the GDPR regulation to blockchain-based technology.

²⁰⁰ Michèle Finck, (2019), “*Blockchain and the General Data Protection Regulation*”, European Parliamentary Research Service (EPRS)

²⁰¹ Michèle Finck, (2019), “*Blockchain and the General Data Protection Regulation*”, European Parliamentary Research Service (EPRS)

²⁰² (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

One important section of the GDPR regulation is the territorial scope (Art. 3 GDPR²⁰³), such article says that the GDPR's rules apply to every data controller located in a country inside the EU. Moreover, even if the controller is located outside the EU, GDPR applies if the data subject is located inside the EU. This underlines that the GDPR had doubtlessly a broad territorial scope, as a consequence, many (if not all) of the blockchain-based data processing will fall under the application of the GDPR rules²⁰⁴.

Another fundamental part of GDPR's rules is the definition of personal data. Depending on whether the data collected and processed is identified as personal data, there will/will not be an application of the GDPR regulation. To this extent, Article 4(1) defines personal data:

“Any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person²⁰⁵”

We can understand how according to GDPR, personal data is data that directly or indirectly relates to an identified or identifiable natural person. However, with the new regulation has been introduced another concept, pseudonymization. Article 4(5) states:

“Processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person²⁰⁶”

Pseudonymization is a key concept when discussing blockchain-related data analysis, that is, the GDPR do not recognize pseudonymization as a form of anonymization. This implies that

²⁰³ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

²⁰⁴ Michèle Finck, (2019), “Blockchain and the General Data Protection Regulation”, European Parliamentary Research Service (EPRS)

²⁰⁵ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

²⁰⁶ (GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

the public keys, often used in the blockchain environment to identify users, are considered a pseudonymization of personal data and fall in the scope of application of GDPR rules²⁰⁷.

Finally, recital 26 of GDPR states that, it is possible to manipulate personal data in a manner removing the reasonable likelihood of identifying a data subject through such data²⁰⁸. Whenever such process is carried out successfully, personal data can be considered anonymous data and can therefore evade the GDPR's rules.

The possible points of friction between GDPR and blockchain technology are numerous, this section is not aimed at providing a rigorous analysis of the subject. However, it is worth mentioning few conclusive points to discuss how in practice the regulation applies to decentralized environments.

As we have seen, one of the first issues is the individuation of the data controller, the territorial scope of application, and the definition of data processing and personal data. Anonymity and pseudonymity are key concept when discussing blockchain-based data systems because it captures the nuances of the blockchain technology such as public keys, private keys, and encryption. To conclude the overview of the application of GDPR's rules to blockchain data systems we can cite the conclusions of the aforementioned study:

“The study has concluded that it can be easier for private and permissioned blockchains to comply with these legal requirements as opposed to private and permissionless blockchains. It has, however, also been stressed that the compatibility of these instruments with the Regulation can only ever be assessed on a case-by-case basis... Indeed, the key takeaway from this study should be that it is impossible to state that blockchains are, as a whole, either completely compliant or non-compliant with the GDPR. Rather, while numerous important points of tension have been highlighted and ultimately each concrete use case needs to be examined on the basis of a detailed case-by-case analysis.”²⁰⁹

²⁰⁷ Michèle Finck, (2019), “Blockchain and the General Data Protection Regulation”, European Parliamentary Research Service (EPRS)

²⁰⁸ Michèle Finck, (2019), “Blockchain and the General Data Protection Regulation”, European Parliamentary Research Service (EPRS)

²⁰⁹ Michèle Finck, (2019), “Blockchain and the General Data Protection Regulation”, European Parliamentary Research Service (EPRS)

Chapter 4: Data monetization, HUDI case study

4.1 The data monetization process

In chapter 1 we underlined how important big data can be for companies and entities such as governments and associations. We analyzed the popular McKinsey Global paper²¹⁰ which pointed out some typical big data features that allow companies to gain relevant advantage over competitors. Moreover, we analyzed some case studies in which big data allowed some companies to gain enormous success, the data giants. The aim of this sub-chapter is to further analyze the process of data monetization from both the business and customer perspective.

4.1.1 Business data monetization

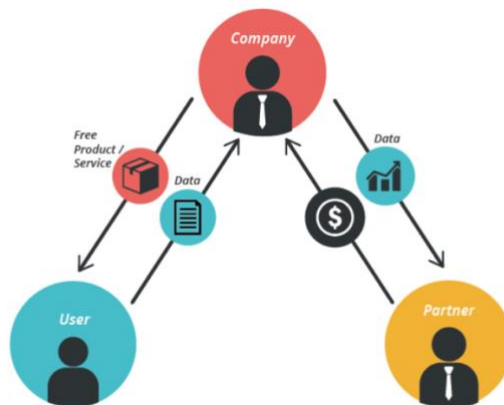
The world's economy comprehends different types of businesses with different business models. Each one has its own peculiarities, strengths, and weaknesses. One common factor to all types of business models is that they can be enhanced through the use of big data analytics. A report by Ernst Young²¹¹ provides some interesting insights on the process of data monetization for businesses. Figures show that the data monetization market for businesses is growing rapidly all over the world. Estimates suggest that the e-commerce and retail sector alone, will experience a 36% compound annual growth²¹² from 2016 to 2023 in the data monetization process. The process of data monetization can be implemented on different levels, firstly, companies can exploit big data to enhance the internal decision-making process. Such process can lead to monetization through different steps, enhancing the marketing function by profiling and tailoring ads and post-sale customer service. Another area that can lead to internal data monetization is the purchase/supply chain management improvement. Other than monetizing their data internally, companies have another option, which consist in selling customer's data to third parties. To better explain the process, we will present a graphical example:

²¹⁰ McKinsey Global Institute, (2011), "*Big data: The next frontier for innovation, competition, and productivity*"

²¹¹ https://www.ey.com/en_us/consumer-products-retail/retailers-can-use-data-as-an-alternative-profit-source

²¹² https://www.ey.com/en_us/consumer-products-retail/retailers-can-use-data-as-an-alternative-profit-source

User Data Monetization



213

The process starts from the company, which gathers customer's data and offers a discount or a free product/service in exchange. The company can then collect data from all customers (in accordance with GDPR or US privacy regulations) and sell it to third parties. Such third-party companies have built their businesses around buying and selling data and they are the one's mentioned in section 1.3.2 (data selling market). Of course, there are different types of data to be sold, for instance structured data, unstructured data, or insights (see chapter 1). In each type of transaction, the purpose of the sale is the same, generating an alternative revenue stream for sustaining the business model. Such process is also known as data monetization. There is a popular quote commonly referred to a tweet by Tim O'Reilly²¹⁴ which states: *"If you're not paying for it, you're not the customer; you're the product being sold."* Such quote consistently confirms that nowadays businesses have the chance of monetizing data in different ways to improve and sustain their business model. It is sufficient thinking to case studies of sub-chapter 1.4 (Data giants) to confirm such statement.

²¹³ Image 16: <https://bmttoolbox.net/patterns/customer-data-monetization/>

²¹⁴ <https://www.oreilly.com/tim/bio.html>

4.1.2 Customer data monetization

As we approach an always more digitalized economy, the number of companies that seek to collect, exploit, and sell customer's data increase day after day. However, the companies are not the only entities which are focusing more and more into the big data ecosystem. More precisely, the data subjects (the owner of the data) are showing an increasing awareness over the usage of their personal data and the purposes of the collection. Some statistics suggest that: *"86% of US citizens have attempted to somehow remove or decrease their digital footprint online²¹⁵"* or *"As many as 79% of Americans on the web worry about companies infringing their online privacy²¹⁶"*. Such statistics suggest an increasing level of awareness among data subjects. Moreover, it is interesting that customers are more aware of how their data is used and how much companies make out of it. To this extent, a practice is becoming more popular among users, data licensing. *"Data licensing is a form of personal data monetization where individuals permit companies to access specific data sources social streams, bank data, browser data, etc. in exchange for financial compensation.²¹⁷"* Which such process, users can retain a sort of copyright on their data and allow specific companies to exploit it and therefore monetize their data. More precisely, the process of data licensing implies different steps and necessary considerations:

- Individuation of the license: in this phase is identified the owner of the data and it is checked if the data is protected by an IP. Whenever an owner is found, the data is addressed to the subject so that he becomes the legal owner. In this phase is also determined the rights of usage of data for each party (the licensee, party who receives the license and the licensor, the party who gives the license)²¹⁸.
- Crucial factors in licensing: the safety and protection of the license depends on the manner of aggregation of the data and the kind of data collected. Among various types of licensed data, we find market data, search engine data, financial data, map data, and business data²¹⁹.
- Consideration on licensing: in the European environment both the licensee and the licensor should carefully consider compliance with GDPR's rules. More precisely,

²¹⁵ <https://dataprot.net/statistics/internet-privacy-statistics/>

²¹⁶ <https://dataprot.net/statistics/internet-privacy-statistics/>

²¹⁷ <https://www.invisibly.com/learn-blog/data-monetization>

²¹⁸ <https://www.infoclutch.com/infographic/what-is-data-licensing>

²¹⁹ <https://www.infoclutch.com/infographic/what-is-data-licensing>

the crucial factor is the data ownership which constitute the basis for accessing, controlling, modifying, and deleting one's data²²⁰.

Of course, each type of data has a different value according to the type of data subject and the buyer of the data²²¹. To this extent, an interesting study by MacKeeper and YouGov²²² investigated how much personal data is worth. This graph represents a summary of the study:

Demographic		Cost for Data Per Person	Percentage of Population
Sex Assignment	Male	\$0.15	48.59%
	Female	\$0.14	51.41%
Age	Age 18-24	\$0.36	11.92%
	Age 55+	\$0.05	32.33%
Ethnicity	Middle Eastern	\$0.62	1.21%
	Hispanic	\$0.01	8.09%
Family Annual Income	\$40,000-\$49,999	\$0.02	4.94%
	\$120,000-\$149,999	\$0.33	1.84%

223

As can be easily noticed data gatherers are willing to pay more for personal data referring to data subjects under the age of 24. Moreover, Middle Eastern people seems to attract buyer which are willing to pay more for their data. If instead of considering single data points on individuals we consider full health records, the valuation changes dramatically:

Record Type	Average Price
Health Care Record	\$250.15
Payment Card Details	\$5.40
Banking Records	\$4.12
Access Credentials	\$0.95
Social Security Number	\$0.53
Credit Record	\$0.31
Basic PII	\$0.03

224

²²⁰ <https://www.infoclutch.com/infographic/what-is-data-licensing>

²²¹ <https://www.invisibly.com/learn-blog/data-monetization>

²²² <https://mackeeper.com/blog/most-desired-data/>

²²³ Image 17: <https://www.invisibly.com/learn-blog/how-much-is-data-worth>

²²⁴ Image 18: <https://www.invisibly.com/learn-blog/how-much-is-data-worth>

We can see how the price for a data point changes from around \$0,03 up to \$250 for a full health care record.

Data licensing is one way of monetizing customer's data, however, there are other ways and companies that are investing in such field. In the next sections we will analyze a case study of a project which aims to create a full environment where users can freely control and monetize their data. To fully understand the topics that will be presented in the next case study, is necessary to make an introduction to cryptocurrencies. The topics in the next sessions will concern data monetization, blockchain, cryptocurrencies and privacy.

4.2 The cryptocurrencies environment

This chapter focuses on the monetization of personal data, more precisely, we will analyze a case study of a project that includes such topics. Among the fundamental elements of the project, we find cryptocurrencies, vital to the functioning of the data environment. To this extent, we will exploit this section to provide a clear overview of the crypto environment in order to be able to understand the next topics.

4.2.1 Cryptocurrencies 101

To begin this introduction to cryptocurrencies, we can start from the definition of crypto and what is the difference with respect to traditional money. Over the internet we find different definitions of cryptocurrency, however, the following is considered the most accurate: “A *cryptocurrency is a digital or virtual currency that is secured by cryptography, which makes it nearly impossible to counterfeit or double-spend. Many cryptocurrencies are decentralized networks based on blockchain technology*²²⁵”. We can notice from the definition that crypto are defined as digital currency, secured by cryptography. Cryptography is the science of safe and private communication which we discussed already in chapter 3. Through cryptography, the data is encrypted and is readable only from the intended recipient of the message which possess the decryption key. Another feature which characterizes almost all cryptocurrencies is blockchain technology, this allow crypto to be fully decentralized, safe, and tamper-proof. To be more specific, cryptocurrencies are a type of digital asset which rely on the blockchain technology to operate. It is crucial to understand the difference between blockchain and crypto (see sect 3.1.3).

Cryptocurrencies are different from “traditional” money in a number of ways, to better appreciate the differences, we provide a brief definition of “FIAT” money. Fiat money comprehend all kinds of money that are made legal tender by a government decree of fiat²²⁶. To further simplify the concept, fiat money comprehends all kind of government-backed

²²⁵ <https://www.investopedia.com/terms/c/cryptocurrency.asp>

²²⁶ <https://www.britannica.com/topic/flat-money>

currencies such as euro, dollar, pound, etc. Fiat money can be in the form of paper or metallic money (cash), or digital fiat money (banking payments, bank account, etc.).

Now that we understood the difference between fiat money and cryptocurrencies, we can further analyze the world of digital money by making some important distinction. The first distinction that has to be made is between a cryptocurrency and a token. Both cryptocurrencies and tokens are a type of digital asset, which represents a broader class. Both type of assets exploit cryptography to safely execute transactions, and that is why they are also called crypto-currencies and crypto-tokens. The key distinction between cryptocurrencies and tokens stands in the blockchain they are built on. More precisely, cryptocurrencies are built on their own blockchain, like bitcoin which operates on the Bitcoin blockchain, and ether which operates on the Ethereum blockchain. On the other hand, tokens are just like cryptocurrencies, but they are built on an already existing blockchain. Examples of tokens are those of the class ERC-20 which are built on top of the Ethereum blockchain. To understand properly the cryptocurrencies environment, we must analyze the principal class of digital assets and their properties. We will start by analyzing tokens and the different purposes that they can serve depending on the type of crypto assets that they represent. Outside the crypto environment tokens represent “*something serving to represent or indicate some fact, event, feeling, etc.*”²²⁷. Real-life examples of tokens are, for instance, a gym membership card which represents the right of the holder of the card, to train in a specific gym. Therefore, tokens serve the purpose of representing something (a right, a proof, a value) to someone. If we come back to the crypto environment, we find that, essentially, tokens serve the same purpose. We will start by analyzing the main classes of tokens:

- Security tokens: security tokens are a type of digital assets which derive their value from an external security that can be traded²²⁸. To this extent, security tokens are much like a financial instrument such as derivatives, bonds, or equities. Security tokens can give the holder the right to cash interest, control partnership in a company or simply speculate on the value of an asset. Due to their similarity with financial instruments, security tokens need to be compliant with national regulations that control securities. Security tokens are similar to utility tokens, the difference between them is that security tokens qualify as investments, while utility tokens do not. In order to be qualified as investment instruments, security tokens must pass the “Howey

²²⁷ <https://www.dictionary.com/browse/token>

²²⁸ <https://www.blockchain-council.org/blockchain/security-tokens-vs-utility-tokens-a-concise-guide/>

test”. The Howey test was created by a U.S. Supreme Court case²²⁹ and determine whether a token qualifies as “investment contract”. Therefore, a security token is a utility token which passed the Howey test. Security tokens are issued via STO (Security Token Offering)²³⁰ which is similar to an IPO for tokens. Whenever a user buys and possess a security token, he is entitled to receive interest, dividends, capital gains, and any other type of profit/loss deriving from the contract.

- Utility tokens: utility tokens are a type of tokens which serves some pre-defined purposes inside a certain ecosystem. Usually, utility tokens are used to provide its holder with a specific product or service. However, since utility tokens did not pass the Howey test, they are not classified as investment instruments and therefore, are not subject to national regulations²³¹. Utility tokens can serve different purposes depending on the entity that issued them. Among the most popular use cases for utility tokens, we find:
 - Provide the owner with the right to utilize or own a product or service. Moreover, it can also be used for voting purposes inside a decentralized ecosystem. An example of utility token is FIL the native token of the Filecoin decentralized storage (see section 3.2.3)²³².
 - Provide the owner with decentralized storage²³³;
 - It can provide the owner of the token with rewards for executing specific tasks²³⁴;

Utility tokens serve different purposes; however, one must consider that such instruments are not regulated and can lead to huge losses if used without knowledge.

- Fungible tokens: fungible tokens are tied to the old concept of economics of fungibility and non-uniqueness of a coin, token, etc. Fungible tokens are all types of tokens which are divisible and non-unique²³⁵. We can take as an example fiat money such as the dollar, one dollar is worth one dollar, no matter who possess it. By the same reasoning, also bitcoin can be seen as a fungible token. That is because, one bitcoin is always worth one bitcoin, and everyone who possess one, has access to the

²²⁹ <https://www.investopedia.com/terms/h/howey-test.asp>

²³⁰ <https://www.blockchain-council.org/blockchain/security-tokens-vs-utility-tokens-a-concise-guide/>

²³¹ <https://www.sofi.com/learn/content/what-is-a-utility-token/>

²³² <https://www.blockchain-council.org/blockchain/security-tokens-vs-utility-tokens-a-concise-guide/>

²³³ <https://www.blockchain-council.org/blockchain/security-tokens-vs-utility-tokens-a-concise-guide/>

²³⁴ <https://www.blockchain-council.org/blockchain/security-tokens-vs-utility-tokens-a-concise-guide/>

²³⁵ <https://cointelegraph.com/nonfungible-tokens-for-beginners/fungible-vs-nonfungible-tokens-what-is-the-difference>

same level of wealth as everybody else²³⁶. Non-uniqueness and fungibility is a feature which affects almost all type of currencies in the world.

- Non-fungible tokens (NFTs): NFTs are a type of cryptographic asset with unique identification codes and metadata²³⁷. As the name suggests, these types of tokens are non-fungible and unique, as opposed to fungible ones. The blockchain technology allowed for increasingly simpler methods for proving the ownership of an item (digital or not) and transferring it to other parties without the need of a central authority. To this extent, NFTs provide a useful tool for exchanging unique data which can assume different formats (music, images, videos, etc.). The array of applications of NFTs increased a lot during recent years, together with the popularity of such tokens. The hype and interest that grew around NFTs during the years caused the market size of such instruments to reach over \$41 billions²³⁸ in 2021.

NFTs sector is increasing rapidly thanks to the huge amount of real-world application that such instruments allow. Firstly, thanks to the unique identification code, cryptography, and blockchain technology, NFTs are not replicable²³⁹. Such features enable them to conserve their non-fungibility trait, which allow them to be useful in real-life scenarios. One of the main use cases of NFTs is representation of real-world items like artwork²⁴⁰. To this extent, NFTs are the most popular instrument for selling and owning pieces of artwork from different artists. For instance, a popular NFT collection called “Bored Ape Yacht Club” comprehend 10’000 NFTs which are images of stylized apes with peculiar features. Most of them are being sold for huge amounts and the most expensive “Bored Ape” NFT has been sold for approximately \$2.3 million²⁴¹. NFTs are also being employed in the so-called “tokenization” of real-world assets such as real estate²⁴². By so doing, a person can be identified as the legal owner of a real estate without the need of a notary or other intermediaries.

Finally, NFTs can also be used to represent individual’s identity, property rights and other unique features²⁴³.

²³⁶ <https://cointelegraph.com/nonfungible-tokens-for-beginners/fungible-vs-nonfungible-tokens-what-is-the-difference>

²³⁷ <https://www.investopedia.com/non-fungible-tokens-nft-5115211>

²³⁸ <https://markets.businessinsider.com/news/currencies/nft-market-41-billion-nearing-fine-art-market-size-2022-1>

²³⁹ <https://www.investopedia.com/non-fungible-tokens-nft-5115211>

²⁴⁰ <https://www.investopedia.com/non-fungible-tokens-nft-5115211>

²⁴¹ <https://www.prestigeonline.com/hk/pursuits/art-culture/most-popular-nft-projects/>

²⁴² <https://www.investopedia.com/non-fungible-tokens-nft-5115211>

²⁴³ <https://www.investopedia.com/non-fungible-tokens-nft-5115211>

- **Currency tokens:** currency tokens are a type of cryptoasset whose purpose is to serve as a currency²⁴⁴. Since the beginning of the expansion of the blockchain technology, currency tokens were the unique, and most popular form of tokens. Depending on whether the token runs on his own blockchain, it can be called “cryptocurrency” or “currency token”. Some of the most popular cryptocurrencies can, in fact, be labeled as currency tokens. If we consider both bitcoin and ether, they were initially created for payment purposes in a fast, reliable, and decentralized way. This means that their main and original goal was to serve as a digital currency; therefore, they are currency tokens.
- **Governance tokens:** governance tokens are a type of token designed to allow blockchain users to vote and gather power in a decentralized ecosystem²⁴⁵. We can imagine a governance tokens as an evolution of a utility token. Instead of simply allowing the holder to benefit from a product or service, governance tokens empower their holder with the voting right. Such voting right can be used to decide to patch or update an existing decentralized application (see section 3.1.7), or to review and change and existing voting protocol.

To this point, it comes naturally the fact that “tokens” is a broad class which encompass strictly cryptocurrencies. Cryptocurrencies are a type of currency token which operates on its own blockchain and serve monetary purposes such as payments. Nonetheless, the terms “token” and “cryptocurrency” are frequently interchanged in the popular jargon.

4.2.2 ICO, IEO, IDO

To conclude this overview on cryptocurrencies we shall discuss about the different way of issuance of such instruments. Historically, when a company decided to go public, an IPO (Initial Public Offering) was necessary. After the IPO the shares of the company were released in the market and the price resulted from the intersection between offer and demand.

²⁴⁴ <https://www.twobirds.com/en/news/articles/2019/global/ico-legal-classification-of-tokens-3>

²⁴⁵ <https://www.bairesdev.com/blog/governance-tokens-they-threaten-blockchain/>

By the same reasoning, to launch a new token, it necessary to go through an ICO, STO, IEO, or IDO, depending on the case.

Nowadays, whenever companies need some funding to expand or start a business, they can choose among many more options with respect to 20 years ago. There are traditional banking loans, VCs funds, equity issuance, crowdfunding, or tokenization. Tokenization is growing popular in recent years thanks to the development of the blockchain technology and the common interest that gathered around cryptocurrencies. Companies now have a new option to collect money from the public through tokenization. Tokenization consists in the issuance of tokens that are sold to the public in exchange for cryptocurrencies or fiat money. In such a way, people can invest money and in return they do not receive a share, but a token. This concept is similar to the one of crowdfunding, whereas in certain instances (reward crowdfunding) investors are awarded with something different from a share. The tokens awarded to investors can be used to speculate on the price, to receive a product/service, or to obtain a security, depending on the type of token. While in an equity issuance there are few ways of selling new shares to the public (IPO, SEO, private sale), in a tokenization there are more.

The Initial Coin Offering (ICO) is the first type of token issuance that became popular in 2017 alongside with the growth of interest towards cryptocurrencies. In an ICO, a company which desires to raise funding attempt to issue a new token. Such token will be sold to interested parties in pre-sale and afterwards will be traded on some exchange. The peculiarity of the ICO is the massive lack of regulation from governmental authorities. That is because, if tokens are not linked to a real security (security tokens), they are not considered an investment and therefore are exempt from regulation. This means that everyone can issue a new security, it is just necessary a white paper, to be published in the company's website²⁴⁶. The white paper is much like a company's business plan with a description of the token to be issued, a roadmap, and some clarifications. The lack of regulation and the extreme simplicity of issuance allowed ICO to gain enormous popularity in 2017. Even though ICO have been around since 2013, in 2017 it is estimated that the market for ICO amounted to \$680 million²⁴⁷. While ICOs allowed astonishing returns to certain investors, the lack of regulation and the extreme simplicity of issuance attracted a lot of fraudulent ICOs during the years. In fact, many projects promised unreasonable returns based on fake information and fraudulent

²⁴⁶ <https://phemex.com/blogs/what-is-a-dex-ido>

²⁴⁷ <https://news.crunchbase.com/news/2017s-ico-market-grew-nearly-100x-q1-q4/>

intentions. A lot of projects turned out to be scams and the issuers withdrew all liquidity right after the ICO, leaving the investors with huge losses. Due to such reasons, ICO popularity dropped dramatically, leaving the floor to STOs.

STOs are Security Token Offerings, and as the name suggests, these are initial offerings of security tokens (discussed in the previous section). Security tokens are backed by real world securities such as shares or debt. Therefore, they are considered investment vehicles by the SEC and are subject to substantial regulation²⁴⁸. For these reasons, STOs are considered to be much safer than ICOs, but also much more expensive. The first STOs began in 2017, but the real boom started after the decline of ICOs in 2018. A report by PwC shows that in 2018 a total of 28 security token offerings were concluded for a total sum of \$442 million²⁴⁹. Although STOs are considered more safer than ICOs, the cost and difficulty of compliance made these instruments obsolete in a short time.

IEOs are Initial Exchange Offerings and are much similar to ICOs. The difference between ICOs and IEOs is that the latter are executed on centralized exchanges such as Binance²⁵⁰. IEOs gained popularity in 2019, after the rise of the aforementioned STOs²⁵¹. The advantage of IEOs is that they preserve the simplicity of execution of ICOs and offer additional security measures due to the controls that the exchange carries on the projects before launching them. Among the advantages of IEOs we find increased credibility thanks to the reputation of the hosting exchange, increased reach thanks to the exploitation of the exchange's existing user base, and liquidity guaranteed by the exchange²⁵².

Finally, the last option to issue new tokens which has gained relevant popularity in recent periods is Initial Dex Offerings, where Dex stands for "decentralized". In an IDO, the procedure is quite similar to an IEO, but the listing is performed on a decentralized exchange such as Uniswap²⁵³ or Pancakeswap²⁵⁴. Among the advantages of IDOs, we find:

- Immediate liquidity provided post-sale;
- Non need to directly deal with a project's smart contracts²⁵⁵;
- Affordability and accessibility;

²⁴⁸ Robert Lui, Wilson Cheung, (2020), "Security token offerings: The next phase of financial market evolution?", Deloitte and King & Wood Mallesons.

²⁴⁹ Steve Davies, Daniel Diemers, Henri Arslanian, Günther Dobrauz, Lukas Wohlgemuth, Axel von Perfall, Henrik Olsson, John Shipman, Pierre-Edouard Wahl; (2019), "STO report, a strategic perspective", PwC and Cryptovalley.

²⁵⁰ <https://www.binance.com/en>

²⁵¹ <https://www.gemini.com/cryptopedia/ieo-crypto-ido-crypto-initial-exchange-offering#section-initial-dex-offering-ido>

²⁵² <https://www.gemini.com/cryptopedia/ieo-crypto-ido-crypto-initial-exchange-offering#section-initial-dex-offering-ido>

²⁵³ <https://uniswap.org>

²⁵⁴ <https://pancakeswap.finance>

²⁵⁵ <https://academy.binance.com/en/articles/what-is-an-ido-initial-dex-offering>

To recap, companies that want to issue new tokens to get funding have different options. Depending on the aim and goal of the issue, companies can choose among ICOs, STOs, IEOs, and IDOs. To recap the differences between such instruments, we present a graph:

	IDO	IEO	ICO
Vetting process	DEX vets the project	CEX vets the project	No vetting process as the project runs the sale themselves
Fundraising	DEX handles investors' funds	CEX handles investors' funds	The project handles investors' funds
Smart contracts	DEX creates and runs smart contracts	CEX creates and runs smart contracts	The project creates and runs smart contracts
Token listing	Liquidity pools open on the DEX	Exchange lists the token	The project has to find an exchange to list on

256

As can be noticed IDOs allow for more advantages compared to ICOs and IEOs if we consider the costs of such operations. STOs are not included in the graph because these instruments are comparable with proper investment issuance and are too costly and difficult to comply with. To conclude, we state that every token has its own peculiarity and can have a better issuance method based on the goal to be reached. Recently, utility, currency, and other types of tokens have been issued mainly through IDOs thanks to their flexibility and security.

²⁵⁶ Image 19: <https://academy.binance.com/en/articles/what-is-an-ido-initial-dex-offering>

4.3 HUDI – Human Data Income, data monetization

To write and update this chapter, I had the opportunity to interview HUDI's CEO, Francesco Ballarani. Some of the information available in this chapter is the result of Francesco's comments and answers to my questions.

HUDI is an English project launched in 2017 with the aim of creating a data ecosystem where users can monetize and control their data. HUDI has been founded by Francesco Ballarani (CEO), Gianluigi Ballarani (CMO), and Andrea Silvi (CTO). The idea behind HUDI was born when Francesco and Gianluigi were working in Hotlead²⁵⁷, a consultancy marketing firm founded by them in 2013. The role of Francesco and Gianluigi was concerned with online advertising, finding new clients for customers, in general, lead generation online²⁵⁸. Therefore, they both used to work with data in an extensive way and were already familiar with the topic. The core idea behind HUDI was creating an ecosystem which linked big data management with monetization. That is because, the vast majority of online advertisement is based on the extensive exploitation of user's data. However, the data subject very rarely profits from the exploitation of his own data. Everyday data giants like Facebook, Google, and Apple (see chapter 1) collect and exploit intensively people's data and use it to create tailored online advertisement services which represents the lion's share of their revenue's streams. Users benefit from this process only by not paying the service (Facebook, Google, Instagram, etc.). Given the relevance of big data nowadays, HUDI aims at providing data subjects with control over their data and to repay them whenever they wish to share them. HUDI's goal is not to collect people's data in shady ways or by leveraging free services/products, the goal is to create an ecosystem which act as a databank. In such databank, users will be able to choose which type of data to share, which one not to, and to profit thanks to their data. To achieve such goal, the founders believed that the data ecosystem could profit from an integration with the crypto environment. The experience in crypto began for Francesco Ballarani in 2016 when he was in Toronto, a city where the popularity of crypto developed early. By combining big data management with cryptocurrencies, HUDI is able to provide users with a safe and reliable way of monetizing their data. That is because, payments made through electronic fiat money would erode all the profit due to transaction fees. Moreover, by exploiting a utility token, HUDI's users can benefit from the increase in price

²⁵⁷ <https://hotlead.it/chi-siamo/>

²⁵⁸ Source: Francesco Ballarani's interview.

that will occur in case of increase of demand of the token itself. Of course, HUDI token can be hold as a long-term investment, or can be converted into fiat money through a decentralized exchange (more on HUDI token in next sections).

4.3.1 HUDI ecosystem

As stated before, HUDI is a project which aims at creating an ecosystem which enables users to control their data and to profit from it. The key issue that HUDI is trying to address is that the big data market is worth billions of dollars a year and the owners of such data do not get nothing in return for it. User's data is the fuel that keeps supplying the data-driven economy we are progressively switching into. And yet, users keep providing their data with little or no control over them and with no economic return. Therefore, to recap the main issues that the HUDI project wants to address:

- Limited access to data owners and small data providers: data owners have limited or no control over the management of their personal data. Moreover, small companies struggle to access the data market because they lack resources and popularity²⁵⁹.
- Limited transparency and control for data owners: in most of the cases the owners of the data (data subjects) have limited or no control over the management of their personal information. According to GDPR's requirements, data subjects should be able to access, manage and eventually delete their own data²⁶⁰.
- Absence of monetization for data subjects: apart from benefits such as discounts or free access to services, data subjects do not get compensated for the trade and sale of their data²⁶¹.

HUDI aims to address such problems by creating a data ecosystem where the three main actors of the data environment can interact and profit together (data owners, data providers, and data buyers²⁶²). Data owners/subjects will be able to actively insert their own data or to

²⁵⁹ HUDI team. "Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.", December (2021).

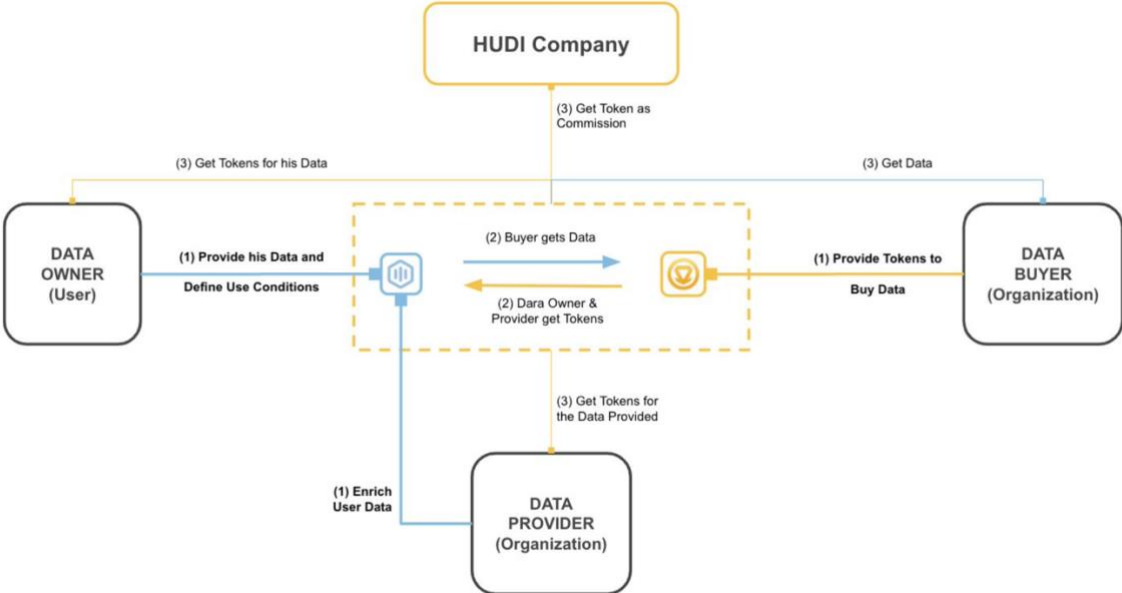
²⁶⁰ HUDI team. "Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.", December (2021).

²⁶¹ HUDI team. "Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.", December (2021).

²⁶² HUDI team. "Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.", December (2021).

transfer their data from other websites/databanks. By doing so, they will have full control over which type of information to share and with whom. Moreover, they will be able to profit every time their data gets traded/used by third parties. Data providers will be able to gain an additional profit while executing their usual business activities (data trading). Data buyers will be able to access an increasingly more complete databank with quality data and a transparent privacy policy.

Although HUDI has been founded in 2017 and has operated for 5 years, it is a continuously developing ecosystem, which has changed through time. As for now, the working mechanism of the HUDI ecosystem could be represented as follows:



263

Users (data owners) provide their data to the HUDI platform through the web application²⁶⁴, they choose which type of data to share through the active monetization section. In the active monetization section users can provide basic demographical data such as name, age, status, etc. Moreover, some surveys are made available to users and based on the eligibility conditions, users can answer and earn HUDI tokens. Another option available to users who wish to actively add data to the HUDI platform is data takeout. In a data takeout users can

²⁶³ Image 20: HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁶⁴ <https://humandataincome.com>

export their data from another databank (Facebook, Instagram, etc.) and transfer it to HUDI platform²⁶⁵.

In addition to the active monetization section, a passive monetization opportunity will be soon available on the HUDI web application. Such passive monetization section will allow users to receive tailored ads on different platforms (social media, etc.) based on the information that they provided to HUDI. Data buyers can join the HUDI ecosystem and buy user's data by providing HUDI tokens (which will be used to compensate users for their data). Also data providers can exploit the HUDI platform, they can transfer their data from other sources and upload it on HUDI to gain an additional profit. In all this process, HUDI take a percentage of the data transactions as a fee for sustaining the environment. However, the users/providers get around 50% of each transaction which concerns their data. Thanks to this environment users can choose to earn from their own data in two ways, by spontaneously providing data in the active monetization section, or by passively earning each time their data gets traded to third parties. In both ways they get compensated with the HUDI token, which can be held as an "investment" opportunity or can be converted in fiat money²⁶⁶.

4.3.2 Roadmap and products

As we stated before, HUDI is a project launched in 2017 that is been operating for 5 years and continued to develop to improve its architecture. According to the HUDI's roadmap²⁶⁷, the first two years have been focused on testing and developing both the concept of HUDI and the first infrastructures. In January 2019 it started the first development of the HUDI app and browser extensions.

²⁶⁵ Source: Francesco Ballarani's interview.

²⁶⁶ Source: Francesco Ballarani's interview.

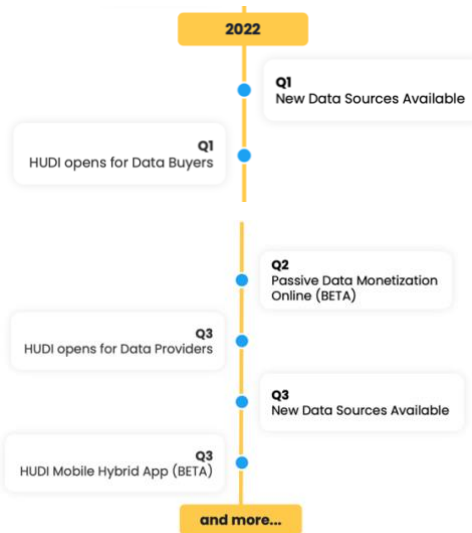
²⁶⁷ HUDI team. "Litpaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.", December (2021).



268

2020 can be considered the year of development, from June 2020 to December, different betas were released, HUDI browser beta IOS, HUDI browser beta Android, HUDI wallet beta IOS, etc.

In 2021, in addition to other app/browsers extension development, an important achievement has been done, the official IDO of the HUDI token on Pancakeswap on the 15 September 2021. For what extent the near future, 2022 will be a year which will focus on the business development and therefore all that directly concerns the data.



269

270

²⁶⁸ Image 21: HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁶⁹ Image 22: HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁷⁰ Image 23: HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

2022 will be the year of introduction of new data sources, such as the surveys section added in early 2022. HUDI will also be available for data buyers and data providers to complete the experience and increase the turnover of the data in the HUDI platform.

For what extent the products, HUDI currently let users and buyers/providers use the web platform “HUDI data exchange”²⁷¹. Such web application is cloud based and allows to execute different operations:

- Collect data from multiple sources and monetize it²⁷²;
- Enrich the current customer base with new data²⁷³;
- Data buyers can launch surveys and market research on highly profiled and consenting users²⁷⁴;
- Data buyers can acquire new profiled users leads and customers²⁷⁵.

In addition to the web application, a cloud-based app for users is expected to be launched, but in this respect, a comment is worth mentioning. During my interview to the company’s CEO, he explained to me that the focus now is to complete and fully integrate the data environment for users, buyers, and providers. Francesco explained that the optimal type of product (mobile app, desktop app, browser extension, etc.) depend on the market positioning of the project. Each product has its strengths and weaknesses, mobile apps are highly centralized and require massive workloads just for bug fixing and updates, browser extensions can be problematic based on the type of browser or device is used to connect to the internet, etc. Therefore, HUDI has already developed different types of products which support different types of technologies, those products will be released when there will be a reasonable opportunity to do so. In the meantime, the web application is running and providing all necessary services to all stakeholders in the HUDI data environment.

²⁷¹ <https://humandataincome.com>

²⁷² HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁷³ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁷⁴ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁷⁵ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

4.3.3 Security and GDPR compliance

Operating in a data-related business implies a lot of regulation and laws to be compliant with. To this extent, HUDI terms of services²⁷⁶ are in accordance with Articles 13 and 14 of GDPR, and the whole data management process is fully compliant with the regulation. In the HUDI website²⁷⁷ we can find all relevant specification with regard with terms of use and compliance with regulation. More precisely, are specified the conditions of use, user's rights and duties, types of data that will be processed, etc.

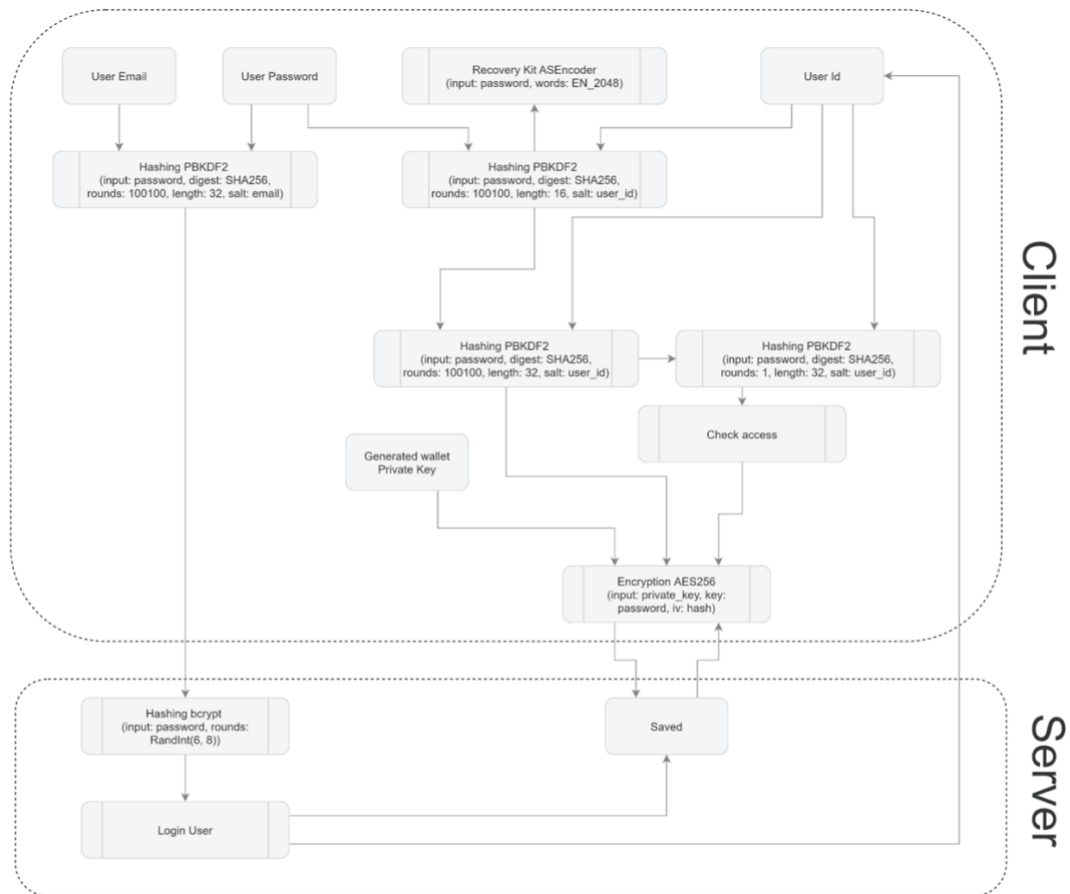
For what extent the practical acquisition and storage of personal data HUDI employs a “local-only²⁷⁸” encryption method, also known as “host-proof hosting”. Such method ensures that all sensitive data is encrypted before syncing with HUDI's servers, and only the user's local device possess the decryption key necessary to decrypt the data. By so doing the data owner will be the only one able to decrypt and access all data. In addition to that, data is transferred to HUDI as a Base64 encoded stripe of encrypted data²⁷⁹. HUDI employs the Advanced Encryption Standard (AES) in Chiper Block Chaining (CBC) with a 256-bit key generated from each user master password. Here follows a graphical representation of HUDI's security procedure:

²⁷⁶ <https://humandataincome.com/policy/user-terms>

²⁷⁷ <https://humandataincome.com/policy/privacy-policy>

²⁷⁸ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁷⁹ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).



280

4.3.4 HUDI token overview

As we already mentioned in the previous sections, the HUDI token is a fundamental element in the data environment. In fact, thanks to the HUDI token is possible to compensate users for their data in a safe and transparent way. To better understand the utility that such token apportos to the environment, we can define it and list its features.

HUDI is a BEP-20 utility token, BEP-20 is a standard which allows users to release fungible tokens on the Binance Smart Chain Blockchain²⁸¹. We can think of BEP-20 standard as a blueprint for tokens which defines how they can be spent, and who can spend them²⁸².

Being a utility token (see section 4.2.1), HUDI empowers the functioning of the whole data ecosystem, providing the basis for data monetization and absence of transaction fees. When

²⁸⁰ Image 24: <https://asset.humandataincome.com/docs/litepaper.pdf>


²⁸¹ The Binance Smart Chain Blockchain is a blockchain that gained enormous popularity in recent years.

²⁸² <https://academy.binance.com/en/glossary/bep-20>

user’s data gets traded or simply uploaded, users get HUDI tokens which can be exchanged for product or services, kept as an investment, or converted into fiat money²⁸³.

The blockchain on which the HUDI token operates is the Binance Smart Chain (BSC), created in 2020. We can think of BSC as a parallel blockchain to Binance Chain, the original one. Binance Chain was created in 2019 and its goal was to support fast and decentralized trading²⁸⁴. Due to technical features, the Binance Chain is not designed to efficiently support smart contracts; Therefore, Binance Smart Chain has been created to support smart contracts and Ethereum Virtual Machine (EVM)²⁸⁵.

For what extent the token governance, HUDI’s goal is to power the whole ecosystem through a DAO (see section 3.1.7) directly controlled by HUDI’s stakeholders²⁸⁶. The HUDI token has been issued through an IDO the 15th of September 2021 on Pancakeswap, a decentralized exchange which runs on the Binance Smart Chain²⁸⁷. The token specifications are the following:

Details	
Name	HUDI
Symbol	
Supply	69420.000,80085
Emission	No new Tokens will ever be created

288

If we analyze more precisely the token distribution features, we can find on the Litepaper²⁸⁹ the followings:

- “Seed 6.9M (10%) tokens are allocated for the private sale, with a 3-month lockup period starting from the IDO event. Afterwards, a linear vesting schedule is applied, unlocking the 5% each month²⁹⁰”. This means that after the IDO, starts a 3-months-

²⁸³ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).
²⁸⁴ <https://academy.binance.com/it/articles/an-introduction-to-binance-smart-chain-bsc>
²⁸⁵ <https://academy.binance.com/en/articles/an-introduction-to-binance-smart-chain-bsc>
²⁸⁶ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).
²⁸⁷ <https://humandataincome.com/#crypto>
²⁸⁸ Image 25: <https://humandataincome.com/#crypto>
²⁸⁹ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).
²⁹⁰ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

period called “lockup period” during which the token cannot be sold for price stabilization purposes. The unlocking procedure consists in a 5% each month on the amount of HUDI token possessed.

- “*Pre-Sale 3.47M (5%) tokens are allocated for the pre-sale event, with a 3-month lockup period starting from the IDO event. Afterwards, a 10-MONTHS linear vesting schedule is applied, unlocking the 10% each month.*²⁹¹”. Each token can be sold before it gets listed to allow early adopters benefit from the issuance, such procedure is called “pre-sale”. 5% of HUDI tokens have been sold in pre-sale.

To conclude, a comment is to be made, the HUDI utility token serve an essential purpose inside the HUDI data ecosystem. It allows the remuneration of users in exchange for their attention and data, and it allows the HUDI project to sustain itself as a form of financing. Although HUDI is a token, and therefore can be erroneously identified as a cryptocurrency, its main goal is not to allow for price speculation. In fact, price speculation can cause high price fluctuation, increasing the token volatility and finally causing a decreasing interest in holding the token as an investment security. The main goal of the HUDI token is a relative price stability which can allow users to be compensated for their data in a reliable and economical way.

4.3.5 DeFi functionalities

To conclude the overview of the HUDI environment, we shall discuss about the DeFi functionalities which are and will be available on the HUDI web application. Although the DeFi environment is less concerned with big data, we think that a brief analysis of this last section will help us complete the discussion of HUDI data ecosystem.

DeFi stands for Decentralized Finance, and it is a term that gained enormous importance over the last years. DeFi is used to indicate all technologies and knowledge which links finance to a decentralized network, a blockchain. The main goal of Decentralized Finance is to build a

²⁹¹ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

payment and banking network that will not require the presence of a central authority for operating²⁹². Among the advantages that Defi could offer to its users we find:

- Deletion of the banking and transaction fees applied by financial institutions²⁹³;
- The possibility of holding your money in your own digital wallet without asking for permission to access the funds²⁹⁴;
- Almost instant transactions of funds from one party to another²⁹⁵.

The HUDI DeFi ecosystem allow users to earn a passive income while staking their tokens. Thanks to such functions, HUDI users have the opportunity of gaining a passive income while holding tokens and supporting the HUDI's ecosystem²⁹⁶. To better understand the opportunities offered by the HUDI DeFi ecosystem we can analyze the products. Similar to the mining activity in the Proof-of-Stake consensus mechanism (The one adopted in the Ethereum blockchain), users can “pool” their coins/tokens and get rewarded for it. There are different way of pooling depending on the purpose of the pool:

- Staking pool: users will decide an amount of their tokens that will be “locked” for a specified period of time. Thanks to such operation, users will be able to earn a passive income on their tokens, similar to the Proof-of-Stake consensus mechanism. The main goal of the staking pool is to increase the TVL (Total Locked Value) which will in turn, increase the credibility of the token²⁹⁷.
- Liquidity pool: users will allocate a specified amount of their token which will be redirected to an external liquidity pool. The tokens will be balanced in pools with other major tokens like BNB, ETH and will provide liquidity to the market. Also, in this type of pooling the user will receive an interest commensurate to the amount pooled²⁹⁸.
- Mini liquidity pool: users will be able to put 50% of a major token (BNB, ETH) in a pool and HUDI will borrow the remaining 50% in HUDI, so it will be able to put it

²⁹² <https://www.investopedia.com/decentralized-finance-defi-5113835>

²⁹³ <https://www.investopedia.com/decentralized-finance-defi-5113835>

²⁹⁴ <https://www.investopedia.com/decentralized-finance-defi-5113835>

²⁹⁵ <https://www.investopedia.com/decentralized-finance-defi-5113835>

²⁹⁶ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁹⁷ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

²⁹⁸ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

in a liquidity pool. When the user exits the pool, he will get 50% in the original token and 50% in HUDI token, so that the pool will remain balanced²⁹⁹.

²⁹⁹ HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

Conclusions

Throughout this thesis we analyzed and discussed different aspects of the big data environment. The main research procedure has been presenting a topic, discussing hypothetical implications (both opportunities and issues), and draw conclusions. The main goal of this research work is to demonstrate how crucial and important the role of big data is, inside the digital economy.

We started by analyzing how the big data environment is composed and how important data is to corporations. Throughout the analysis of the first chapter, we demonstrated how companies can profit from customers' data. Moreover, we analyzed the key areas that can be improved by big data management and understood how companies can access the market of data. Finally, we showed how big data allowed some companies to create whole new business models based on the absence of fees to access the service and based on the extensive exploitation of data. Such companies are called "Data Giants" and are now among the biggest and most powerful corporations in the world. This suggest that, thanks to big data management, some companies can achieve full customer satisfaction, provide quality services, and successful diversification. Those who do not adapt to the new dynamic environment of the digital economy are meant to lose market shares and progressively run out of business.

In the second chapter we discussed some issues that may arise when dealing with big data management. Research work proved that most individuals do not fully understand the importance of their data and are willing to give it away in exchange for a little discount on a product/service. This work focused on a case study in particular, which covers some of the most critical aspects of big data, the "Cambridge Analytica" case study. Such analysis discussed how easily users' data can be obtained and traded illegally, and how powerful are the instruments and techniques which exploit big amounts of sensitive data. In the case study the improper use of data lead to manipulated governmental election and fraudulent actions. Successively, we analyzed how such issues can be prevented by the adoption of the privacy regulations concerning data treatment in both the USA and European Union. More precisely, we analyzed the impact and consequences of the adoption of the GDPR's rules and the local USA's state regulation. Moreover, given the complexity of the privacy laws, we discussed about the economical effort made by businesses to adapt and be compliant with said

regulations. Such topics suggest that big data can be a powerful instrument and when used in improper ways it can cause severe damages to both peoples and the economy. To this extent, privacy regulations aim at minimizing that risk by ensuring that the entities which collect, and process data are law compliant. After having analyzed both the European and American regulatory environment, we feel to underline how the European Union's regulation (GDPR) appears complex but effective. On the other hand, the USA's fragmented regulatory environment causes some users to be less protected than others just because of their location inside the federal state. To this end, it can be stated that users inside the European Union are more protected than American users with respect to data privacy.

The third chapter discussed about the opportunities provided by the integration of big data and blockchain technology. More precisely, the first part was aimed at introducing and explaining what the blockchain is and how it operates. By doing so, we discussed the process of mining, how new blocks are added to the chain, and the different types of blockchains. Moreover, we discussed how the blockchain is mainly used nowadays, and what instruments (smart contracts, DApps, etc.) can be useful. After having introduced the blockchain technology, we presented different big data features that would benefit from the integration with such technology. Transparency, security, integrity and decentralization are characteristics that distinguish decentralized big data solutions such as storages and services. We can argue that big data management can benefit immensely from the integration with the blockchain technology. That is, a decentralized infrastructure solves the typical big data issues that arise when considering data integrity and reliability. Moreover, it has been proved that blockchain technology can reduce storage costs and increase transaction speed.

The fourth and last chapter discussed about data monetization, it considered the opportunities available to both businesses and individuals. Moreover, it showed how much personal data can be worth, depending on ethnicity, nature of the data, age, gender, etc. The core of the fourth chapter aimed at presenting a project whose goal is creating a data ecosystem (HUDI project). To allow beginner readers to understand the topics in the chapter we introduced the cryptocurrencies environment and discussed about the main topics. We introduced the main types of token and crypto and the process of issuance of such assets. Successively, we analyzed the HUDI ecosystem and its functioning, starting from the databank to the HUDI utility token. Finally, we discussed about the past and future steps to be taken in order to achieve a full data decentralization, leaving the data subject with complete ownership over his data. To this extent, the HUDI project is not yet definitively positioned inside its market

share and has not yet released the full resources to the public. However, given the importance that topics such as big data and cryptocurrencies are gaining, the forecast is to expect a massive increase in subscriptions in the near future. Big data is a priceless market, and the opportunity to gain and control over your data will attract a relevant audience. That said, this project is an example of how big data is flexible and can benefit from the integration with other technologies such as blockchain and cryptocurrencies.

Big data has driven the digital revolution for the past ten years and will continue to do so. There are still some areas that can be improved and that will further disrupt the way big data is managed and used. To this extent, it is becoming obvious that the companies that will dominate the market in the future, are those who are investing in big data management right now.

Bibliography

(GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Anupam Chander, Meaza Abraham, Sandeep Chandy, Yuan Fang, Dayoung Park, Isabel Yu, (2021) “*Cost of compliance and enforcement of data protection regulation*”. World bank paper in collaboration with Macroeconomics, trade, and investments global practice. Policy research working paper 9594.

Bernard Marr, (2016), “*Big Data in practice*”, Wiley.

Deepa N, Quoc-Viet Pham, Dinh C. Nguyen, Sweta Bhattacharya, B. Prabadevi, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, Fang Fang, Pubudu N. Pathirana, (2021), “*A Survey on Blockchain for Big Data*”.

Doug Laney, (2001), “*3D Data Management: Controlling Data Volume, Velocity and Variety*”, META group research.

Horst Treiblmaier, Roman Beck, (2019), *Business transformation through blockchain*, Volume II, Palgrave Macmillan.

HUDI team. “*Litepaper 3.0 - The #1 DeFi Data Ecosystem empowering People and Organizations to Collect, Enrich and Trade their Data for a Profit in Crypto.*”, December (2021).

Imran Bashir, (2020), “*Mastering blockchain: A deep dive into distributed ledgers, consensus protocols, smart contracts, DApps, cryptocurrencies, Ethereum, and more*”. Third edition, Packt Birmingham, Mumbai.

Janice C. Sipior, Burke T. Ward, Ruben A. Mendoza, (2011), “*Online Privacy Concerns Associated with Cookies, Flash Cookies, and Web Beacons*” *Journal of Internet Commerce*.

John E. Grable, CFP Angela C. Lyons, (2018), “*An introduction to big data*”, *Journal of financial services professional*.

Legal TechNews, (May 14, 2018), *A brief history of blockchain*, ALM publications.

M.u.S.A, “*Data privacy and protection fundamentals*”, Hellenic open University

McKinsey Global Institute, (2011), “*Big data: The next frontier for innovation, competition, and productivity*”

Michèle Finck, (2019), “*Blockchain and the General Data Protection Regulation*”, European Parliamentary Research Service (EPRS)

Neeraj Kumar, N. Gayathri, Md. Arafatur Rahman, B. Balamurugan. (2020) “*Blockchain, Big data, and Machine learning: trends and applications*”, CRC Press.

Robert Lui, Wilson Cheung, (2020), “*Security token offerings: The next phase of financial market evolution?*”, Deloitte and King & Wood Mallesons.

Satoshi Nakamoto, (2008), *Bitcoin, a peer-to-peer electronic cash system*, www.bitcoin.org.

Steve Davies, Daniel Diemers, Henri Arslanian, Günther Dobrauz, Lukas Wohlgemuth, Axel von Perfall, Henrik Olsson, John Shipman, Pierre-Edouard Wahl; (2019), “*STO report, a strategic perspective*”, PwC and Cryptovalley.

Thomas Erl, Wajid Khattak, Paul Buhler, (2015), “*Big data fundamentals*”, Prentice Halls.

Sitography

<https://about.google/our-story/>
<https://academy.binance.com/en/articles/what-is-an-ido-initial-dex-offering>
<https://academy.binance.com/en/glossary/bep-20>
<https://academy.binance.com/it/articles/an-introduction-to-binance-smart-chain-bsc>
<https://asset.humandataincome.com/docs/litepaper.pdf>
<https://bernardmarr.com/where-can-you-buy-big-data-here-are-the-biggest-consumer-data-brokers/>
<https://bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/1285>
<https://blog.coinlist.co/filecoin-why-its-a-big-deal/>
<https://bmtoolbox.net/patterns/customer-data-monetization/>
<https://cointelegraph.com/nonfungible-tokens-for-beginners/fungible-vs-nonfungible-tokens-what-is-the-difference>
<https://datafloq.com/read/big-data-history/239>
<https://filecoin.io/store/>
<https://firstmonday.org/ojs/index.php/fm/article/view/548>
<https://firstmonday.org/ojs/index.php/fm/article/view/548/469>
<https://gadgets.ndtv.com/cryptocurrency/features/what-is-a-blockchain-node-how-does-cryptocurrency-work-2515427>
<https://gdpr.eu/what-is-gdpr/>
<https://growthrocks.com/blog/big-five-tech-companies-acquisitions/>
<https://guides.loc.gov/this-month-in-business-history/april/apple-computers-founded>
<https://hotlead.it/chi-siamo/>
<https://humandataincome.com>
<https://humandataincome.com/#crypto>
<https://irishtechnews.ie/5-ways-big-data-gets-misused/>
<https://markets.businessinsider.com/news/currencies/nft-market-41-billion-nearing-fine-art-market-size-2022-1>
<https://martech.org/83-percent-of-consumers-now-aware-of-marketers-tracking-their-locations-study/>
<https://medium.com/@blairmarshall/how-do-miners-validate-transactions-c01b05f36231>
<https://medium.com/@hmishfer17/blockchain-data-integrity-e70e17cac086>
<https://medium.com/coinmonks/blockchain-what-is-a-node-or-masternode-and-what-does-it-do-4d9a4200938f>
<https://medium.com/coinmonks/public-vs-private-blockchain-in-a-nutshell-c9fe284fa39f>
<https://medium.com/swlh/whats-the-difference-between-dapp-idapp-and-dao-and-why-they-are-the-future-of-blockchain52758f50474e>
<https://medium.datadriveninvestor.com/what-is-the-difference-between-smart-contracts-and-dapps-d252d88d32d3>
<https://news.crunchbase.com/news/2017s-ico-market-grew-nearly-100x-q1-q4/>

<https://oag.ca.gov/privacy/ccpa>
<https://omnilytics.co/omnilytics-overview>
<https://online.hbs.edu/blog/post/what-is-data-integrity>
<https://pancakeswap.finance>
<https://phemex.com/blogs/what-is-a-dex-ido>
<https://pitchbook.com/profiles/company/433136-26#overview>
<https://pro.bloomberglaw.com/brief/what-is-the-vcdpa/>
<https://qz.com/2000350/the-inventor-of-the-digital-cookie-has-some-regrets/>
<https://retinacromatica.it/capire-l-uso-dei-big-data-cambridge-analytica/>
<https://searchcloudcomputing.techtarget.com/definition/cloud-computing>
<https://support.apple.com/it-it/HT208944>
[https://techjury.net/blog/big-data-statistics/#gref Connectiva Systems, \(2021\),](https://techjury.net/blog/big-data-statistics/#gref Connectiva Systems, (2021),)
[https://techjury.net/blog/big-data-statistics/#gref Forbes, \(2021\),](https://techjury.net/blog/big-data-statistics/#gref Forbes, (2021),)
[https://techjury.net/blog/big-data-statistics/#gref IBM, \(2021\),](https://techjury.net/blog/big-data-statistics/#gref IBM, (2021),)
[https://techjury.net/blog/big-data-statistics/#gref Internet live stats, \(2021\),](https://techjury.net/blog/big-data-statistics/#gref Internet live stats, (2021),)
<https://uniswap.org>
<https://usercentrics.com/knowledge-hub/childrens-online-protection-act-coppa/>
<https://www.analyticssteps.com/blogs/how-apple-uses-ai-and-big-data>
<https://www.bairesdev.com/blog/governance-tokens-they-threaten-blockchain/>
<https://www.binance.com/en>
<https://www.blockchain-council.org/blockchain/security-tokens-vs-utility-tokens-a-concise-guide/>
<https://www.blockchain.com/charts/difficulty>
<https://www.britannica.com/topic/flat-money>
<https://www.buchanan.com/cloud-computing-security-issues/>
<https://www.businessnewsdaily.com/10625-businesses-collecting-data.html>
<https://www.cdc.gov/phlp/publications/topic/hipaa.html>
<https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html>
<https://www.cnbc.com/2021/08/31/kid-siblings-earn-thousands-per-month-mining-crypto-like-bitcoin-eth.html>
<https://www.concordia.net/about/>
<https://www.crunchbase.com/organization/provenance>
<https://www.dataversity.net/brief-history-big-data/#>
<https://www.devteam.space/blog/10-uses-for-smart-contracts/>
<https://www.dictionary.com/browse/token>
https://www.ey.com/en_us/consumer-products-retail/retailers-can-use-data-as-an-alternative-profit-source
<https://www.forbes.com/profile/robert-mercier/>
<https://www.forbes.com/sites/forbestechcouncil/2018/07/05/four-trends-in-cloud-computing-cios-should-prepare-for-in-2019/?sh=6c4146c44dc2>
<https://www.gemini.com/cryptopedia/ieo-crypto-ido-crypto-initial-exchange-offering#section-initial-dex-offering-ido>

<https://www.igi-global.com/dictionary/has-bitcoin-achieved-the-characteristics-of-money/59928>
<https://www.ilpost.it/2018/03/19/facebook-cambridge-analytica/>
<https://www.infoclutch.com/infographic/what-is-data-licensing>
<https://www.investopedia.com/tech/how-does-bitcoin-mining-work/>
<https://www.investopedia.com/terms/c/consensus-mechanism-cryptocurrency.asp>
<https://www.investopedia.com/terms/c/cryptocurrency.asp>
<https://www.investopedia.com/terms/h/howey-test.asp>
<https://www.investopedia.com/terms/m/merkle-tree.asp>
<https://www.investopedia.com/terms/p/proof-stake-pos.asp>
<https://www.kaspersky.com/resource-center/definitions/cookies>
<https://www.kraken.com/learn/what-is-decentralized-autonomous-organization-dao>
<https://www.natlawreview.com/article/and-now-there-are-three-colorado-privacy-act>
<https://www.netwrix.com/download/collaterals/2021%20Netwrix%20Cloud%20Data%20Security%20Report.pdf>
<https://www.ntt.com/en/about-us/hp/webbeacon.html>
<https://www.oreilly.com/tim/bio.html>
<https://www.profolus.com/topics/pos-advantages-and-disadvantages-of-proof-of-stake/>
<https://www.provenance.org>
<https://www.sdxcentral.com/security/definitions/what-are-the-data-privacy-fundamentals/>
<https://www.seba.swiss/research/Classification-and-importance-of-nodes-in-a-blockchain-network> Seba Bank, 2020, The bridge
https://www.sec.gov/Archives/edgar/data/1652044/000165204421000010/goog-20201231.htm#id55be7992b374e1a9a2bc48887d4bb3f_106
<https://www.selecthub.com/big-data-analytics/types-of-big-data-analytics/>
<https://www.smartdatacollective.com/big-data-stored-managed/>
<https://www.sofi.com/learn/content/what-is-a-utility-token/>
<https://www.statista.com/statistics/254266/global-big-data-market-forecast/>
<https://www.statista.com/statistics/254266/global-big-data-market-forecast/>
<https://www.storj.io>
<https://www.theguardian.com/news/2018/mar/18/what-is-cambridge-analytica-firm-at-centre-of-facebook-data-breach>
<https://www.tokenex.com/blog/data-privacy-vs-security>
<https://www.twobirds.com/en/news/articles/2019/global/ico-legal-classification-of-tokens-3>
<https://www.verizon.com/business/resources/reports/2020-data-breach-investigations-report.pdf>
<https://www.youtube.com/watch?v=mpbeOCKZFfQ>
<https://www.youtube.com/watch?v=n8Dd5aVXLcC>
<https://www.zdnet.com/article/three-data-integrity-challenges-blockchain-can-help-solve/>