Understanding Social Big data:
A literature review

Mahmoud HAOUAT

Phd student

Faculty of Legal, Economic and Social Sciences

Laboratory of Economics and Management of Organizations

Ibn Tofail University- Kenitra

Mahmoud.haouat@uit.ac.com

Mohammed QMICHCHOU
Professeur chercheur
Faculty of Legal, Economic and Social Sciences

Laboratory of Economics and Management of Organizations

Ibn Tofail University- Kenitra

qmichchou.mohammed@uit.ac.ma

***Abstract*—The advent and the popularity of both social media and ITC contributed to the birth of a new type of data called social data. The accumulation of this data gave birth to a new big data subgenre formally known as social big data. In the same way as big data, social big data is a new concept draw the attention of many studies. However, there is still a lack of consensus on its definition and also conceptualization. Hence the interest of our work, which aims to bridge the existing gap in the literature by providing a definition of the concept, its characteristics and also types. All of this was done based on a detailed analysis of previous works. In the same way as big data, it should be noted that the exploitation of social big data is very useful for several fields such as psychology, sociology, politics and business to achieve this firms therefore the second point addressed in this work, is how firms and users in general can extract value from all the social big data floating around them, to do so we put forward a series of actions and steps that falls under the name of social big data value chain, and also the main challenges encountered in these steps. With this paper we aim to foster future research activities around this concept.**

Keywords—Big data, Social big data, Social big data value, Digital human, Characteristics, Social media, Acquisition and preparation of data, Organization of data, Data Analysis, Data Exploitation.

I. INTRODUCTION

Since the start of the 21st century Big Data became one of the most discussed topics and terms in the world but at the same time it was the least understood.

Since its inception, the discussion around BD (Big Data) defined the concept using only qualitative terms. At that time BD was mostly defined as a larger than a certain number X of terabytes. This simple definition went through a lot of changes due to the many technological advances such as software tools, size of datasets, storage devices etc.… and also due to different fields and sectors where big data is used like business world, computer science etc.…

The rise of big data has coincided with two major events, the technological boom and the abundance of technological devices such as smartphones, computers etc. And also the birth and the development of a new communication tool called SOCIAL MEDIA. During the last two decades, social media completely reshaped not only the way people communicate but also the way they share and create information, these changes contributed to the birth of what researchers call “an always on society” where people constantly interact with each other thus creating an extensive amount of human generated diverse data called SBD (Social Big Data). According to Gandomi and Haider (2015) “such unstructured/semi unstructured yet semantically rich data has been argued to constitute 95% of all big data”. Given its important and quantity that cannot be ignored, many researchers and theorists tried to study and theorize this emerging concept.

Broadly speaking SBD refers to a large amount of data generated through the use of social media, ”the sheer volume and semantic richness of such data opens enormous possibilities for utilizing and analyzing it for personal, commercial as well as societal purposes” (Olshannikova et al, 2017).

II. MOTIVATION & METHODOLOGY

A. Motivation

The main goal of this literature review is to clarify and shed some light in the concept of social big data. To do so we will start by identifying the many ways SBD was defined and interpreted in the literature. We will also address the types of SBD and also the many benefits and challenges related to using SBD.

Based on what we said before, the value of this paper can be presented as follows:

— Firstly, we will examine the literature so we can bring clarity on big data and its characteristics.

— Secondly, we will aim to bring some clarity on various SBD concepts; we will also attempt to summarize the relation between SBD and many other fields. All of this will help us provide a synthesized definition of the concept and also identify its characteristics.

— Thirdly, we will focus on the various types of SBD.

— Finally we will address the way firms and users can create value des the collected data.

B. *Methodology*

Before stating which the mythology was followed in this work, it should be emphasized that the latter takes form of a literature review.

As we know, the literature review is considered the building block of almost all the academic research activities, regardless of the disciplines. It's usually used as way of collecting and synthesizing previous researches, and also a firm formation for advancing and facilitating any theory development. There are three different approaches to conduct a literature review: systemic, semi-systemic and integrative. Each of these approaches differs in their purpose, sample characteristics….

When it comes to our work we chose to follow the semi-systemic approach, this approach was designed is for "topics that have been conceptualized differently and studied by various groups of researchers within diverse disciplines" (Wong et al, 2013). Besides the aim of overviewing a topic, a semi-systemic review "often looks at how a research within a selected field has progressed over time or how a topic has developed across research traditions" (Snyder, 2019). Overall we choose this approach because it will help us understand all relevant research traditions that have implications for studied topic.

III. LITERATURE REVIEW:

A. *Big data:*

In this section, we will try to first present a list of popular definitions of big data, followed by a list of its essential characteristics. To do so we studied, reviewed and analyzed the related literature found on major databases.

- Definition:

While its ubiquitous big data as a concept has no formal and certain origin, despite its current popularity, there's no single unified definition. Fundamentally speaking, the term big data "applies to datasets that grows so large that they become awkward to work with using traditional datasets management systems" (Elgendy, Ebragal, 2014), this definition coincides with the one cited in the oxford dictionary. According to oxford dictionary big data is an "extremely large datasets that may be analyzed, computationally to reveal patterns, trends and associations especially relating to human behavior and interactions". This definition has since been reiterated by a number of other scholars, some of them even added new characteristics such as veracity, velocity, value etc...

When it comes to the literature, the quick and chaotic evolution of big data has contributed to the birth if several definitions, some of them focused on what big data is or represent, what others try to answer through it, what it does and even what kind of technology is needed to deal with it.

Amongst the most cited definitions, we found the one included in the Gartner report published in 2001. However the Gartner report doesn't mention the term "big data", but it was still considered as an authority in the matter. According to Gartner big data can be defined based on its characteristics known as Vs. The report in question proposed a 3 fold definition encompassing the 3 Vs: Volume, Velocity and Variety. As reported by Gartner "big data is high volume, high velocity and high variety information assets that demand cost effective, innovative forms of information processing that enable enhanced insight, decision making and process automation". Similarly, (Schroeck et al, 2012) considers big data as "a combination of Volume, Variety, Velocity and Veracity that creates an opportunity for organizations to gain competitive advantage in today's digitized marketplace", In the same way Laney (2001) offers the following definition "big data represents voluminous, high-velocity and varied information resources that require innovative forms of processing ...». Recently, the 3Vs method developed by Laney (2001) has been extended to 3 other Vs, including the value cited by (Chen et al, 2014 and Oracle), veracity (Arun and Jabasheel, 2014) and variability / complexity (IAS Inc.). The addition of these new characteristics prompted the birth of new definitions based on the technological aspect of big data, according to Provost and Fawcett (2013) big data represents "data sets that are too large for traditional data processing systems and that therefore require new technologies with names like Hadoop, Hbase, Mapreduce etc…".

The attributes and the technological one were not the only ways to define and describe the concept in hand; some definitions consider big data in terms of crossing a certain thresholds. In his definition Dumbill (2013) conveyed the multidimensionality of big data, by stating that big data is "data too big, moves too fast or doesn't fit the structure of the database". The other aspect mentioned in the literature is big data's impact. Several definitions highlight the huge impact that big data have not only on firms but also on society in general. Boyd and Crowford (2012) proposed the following definition, big data is "a cultural, technological, and scholarly phenomenon", in the same way Mayer-Schonberger and Cuckier (2013) described big data in terms of three main shifts in the way of analyzing the data in hand, the analysis of the data will improve the way we understand the society and the firms in question, according to Mayer Schonberger and Cuckier (2013), the shifts in question include a high volume of messier data and correlation.

The many definitions stated before and the ones stated in table 1 confirm what we stated earlier and go along with the observation made by Bi and Cohen (2014) among other authors.

- ➢ Summary of the definitions:

For the rest of this work, we will offer a definition based on the aforementioned definitions as well as observations of the literature. "**Big data is a complex large data set characterized by high velocity and a variety of information; this said data requires new technological tools (for storage, analysis) to substrate value".** This definition highlights the concept of 5Vs; the omission of veracity was done on purpose since it's often attributed to a specific type of data characterized by uncertainties and inconsistencies or lack of precision, example: data collected from social media.

Authors	Definitions
Boyd and Crawford (2012)	A cultural, technological and scientific phenomenon that relies on the interplay of analytical technology and methodology.
Manyika et al (2011)	Refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
Zikopoulos et al (2013)	Big data contains four dimensions namely volume, variety, velocity and veracity.
Laney (2001)	Characterized by 3Vs theory: volume, variety, and velocity. Volume: with the generation and collection of data, data scale becomes increasingly big; Velocity: timeliness of big data, specifically data collection and analysis must be rapidly and timely conducted; Variety: the various types of data, which include semi-structured and unstructured, structured data

Table. 1: Some Big data definitions

- Big data characteristics:

The definition cited highlights the following characteristics: Volume, velocity, variety, value and complexity. These terms represent the characteristics of big data commonly known as 5Vs. However these dimensions are constantly developing.

— Volume:

Volume is the first dimension that we will discuss, it refers to the magnitude of data generated and collected (Chen et al, 2014), it is safe to say that the amount of data generated increases each day. That amount can exceed terabytes to reach petabytes, or even extabytes, and this growth is expected to continue over the next few years.

According to Hendler (2013) "the term volume originated to describe the amount of data held in large databases". This enormous amount of data is continuously created from multiple sources such as social media, clouds, business data and the Internet of Things which prompts us to ask the question what's big?, because what may be deemed big today may not meet the threshold in the future, since the storage capacities increase with time.

— Velocity:

Whereas volume refers to the size of data, velocity refers to the speed at which the data in question is generated and also the speed, at which it should be collected, analyzed and used. This was confirmed by Chen and Zhang (2014), they refer to velocity as "the speed characterizing incoming and outgoing data".

— Value:

This is the most important aspect of big data, and the purpose of big data technologies. It was introduced by Oracle as the defining attribute of big data. This view was also shared and expressed by IDC (International Data Corporation). According to them the big data architectures is designed to extract value from large volume of data.

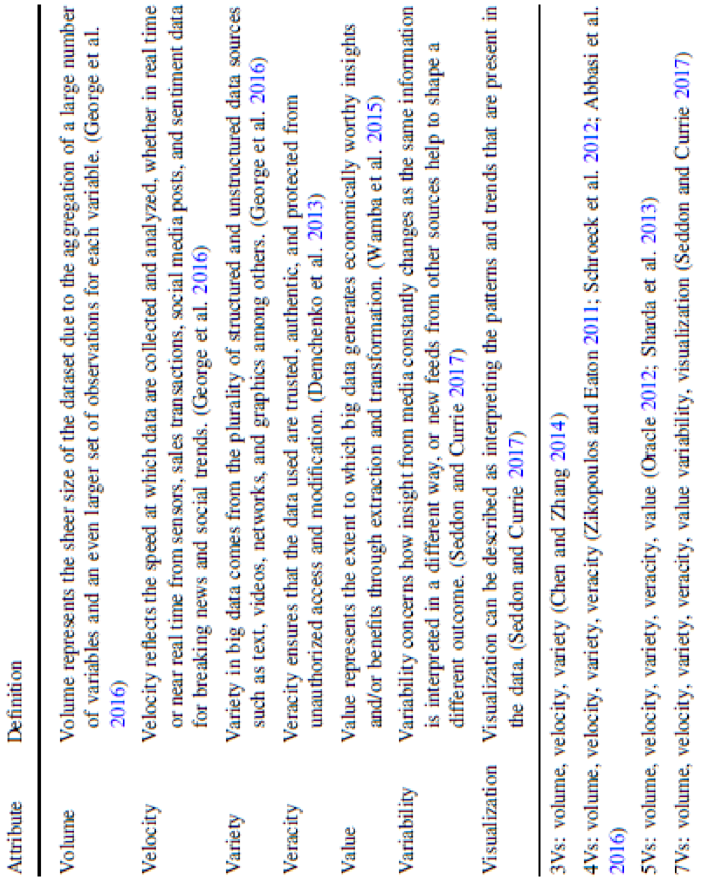
— Variety:

This attribute is an essential characteristic of big data, it refers to different data's structures (structured, semi structured and unstructured) and also the many sources it comes from for example smartphones, social networks, clouds, sensors among many others, and also the formats and types images, videos, texts, audios……

— Complexity:

This dimension was introduced by SAS Inc. It was added in order to extend the dimensional model of big data. It refers to the fact that big data is generated through a multitude of sources, which poses a critical challenge when processing data (the need to connect, cleanse and transform this data into useful information).

According to Mikalef et al. (2018), various researchers focus on different aspects of big data, some authors introduced other characteristics such as veracity, visualization, variability.



Attribute	Definition
Volume	Volume represents the sheer size of the dataset due to the aggregation of a large number of variables and an even larger set of observations for each variable. (George et al. 2016)
Velocity	Velocity reflects the speed at which data are collected and analyzed, whether in real time or near real time from sensors, sales transactions, social media posts, and sentiment data for breaking news and social trends. (George et al. 2016)
Variety	Variety in big data comes from the plurality of structured and unstructured data sources such as text, videos, networks, and graphics among others. (George et al. 2016)
Veracity	Veracity ensures that the data used are trusted, authentic, and protected from unauthorized access and modification. (Demchenko et al. 2013)
Value	Value represents the extent to which big data generates economically worthy insights and/or benefits through extraction and transformation. (Wamba et al. 2015)
Variability	Variability concerns how insight from media constantly changes as the same information is interpreted in a different way, or new feeds from other sources help to shape a different outcome. (Seddon and Currie 2017)
Visualization	Visualization can be described as interpreting the patterns and trends that are present in the data. (Seddon and Currie 2017)

3Vs: volume, velocity, variety (Chen and Zhang 2014)

4Vs: volume, velocity, variety, veracity (Zikopoulos and Eaton 2011; Schroeck et al. 2012; Abbasi et al. 2016)

5Vs: volume, velocity, variety, veracity, value (Oracle 2012; Sharda et al. 2013)

7Vs: volume, velocity, variety, veracity, value variability, visualization (Seddon and Currie 2017)

Fig.1: The characteristics of big data, adopted from (Mikalef et al., 2018).

B. Social big data:

- Definition:

After defining big data in general, we will now move on to the main concept of our work, called social big data or social big data, as its name suggests, social big data represents the set of simple data and information collected from social media. When it comes to the literature there seems to be a lack of consensus on the general definition of social big data and associated terms. "In general, researchers focus on the analysis and use of social big data, having paid little attention to the concept and understanding of the associated phenomena" (Cambria et al, 2013). Faced with this dilemma it is therefore essential to describe and examine the literature in order to clarify this concept more and more. Therefore the goal of this section is to examine and review the existing literature on the concept by presenting and comparing the various definitions and approaches. While examining the literature, we noticed that social big data was defined and studied in several ways, some of these definitions were “simple” and others were considered “complex”. These differences were due to many factors:

- The rapid and constant development of social media.
- The enormous growth of data generated by social media and social networks, approximately 2.5 Exabyte of data is created and shared per second.
- The various fields of research: social media, social networks, social computing etc.…

As its name suggests, social big data represents the set of data and simple or raw information collected via social media. In the literature there seems to be a lack of consensus on the general definition of social big data and associated terms. When it comes to the literature the first definition of social big data is attributed to Lazer and al (2009). According to them social big data represents “data collected from social media platforms”. This definition was reiterated in different ways and by many authors such as Mukkamala and al (2014) among others. In a more detailed way Orgaz and al (2016) consider social big data as a “conjunction of two different concepts, social media and big data therefore SBD is a vast amount of data generated from multiple distributed sources but with an emphasis on social media”. Another school of thought proposes a definition by focusing on the components of social big data; Tang et al (2014) consider social big data as "the fusion or union of three large parts, user data, and relationships social and generated content”. Beyond the cited, definitions other theorists interpreted Social Big Data based on other concepts and approaches. For example, in his book “SBD mining”, Ishikawa considers Social Big Data as a science and describes it as “the science of analyzing physical real world data (heterogeneous data with implicit semantics such as science data, event data, and transportation data) and social data (social media data with explicit semantics) by relating them to each other”. In the same book Ishikawa also listed the characteristics of the concept. According to the latter since big data is characterized by 3 V’s: Volume, Variety and Velocity, Social Big Data shares the same characteristics but he also added a fourth one named Vagueness.

The second approach sees big data as a big part of social computing (social computing is the field of computing that focuses on the interaction between social behavior and computer systems). The most prominent scholars of this school of thought are Guellil and Boukhalfa (2015). Based on the work of other authors such as Mukkamala (2014) and Nguyen (2015), Guellil and Boukhalfa (2015) were able to conclude that SBD is a direct synonym of social media data. Guellil and Boukhalfa provided the following definition, “SBD might be interpreted as a synonym of SMD (Social Media Data) with qualities such as large volume, noisiness and dynamism”. In this context noisiness refers to the abundant spam found in the blogosphere as well as the existence of trivial posts on social networks. Even though Guellil and boukhalfa defined SBD, they did not propose any clear conceptualization of said concept. In the other hand, and following the same school of thought Mark coté (2014) attempted not only to define the concept but also to conceptualize it, by first of all distinguishing it from big data and second by defining it and showing its importance. According to Coté (2014) the difference between the two concepts resides in their sources, “big data is any data produced as the result of the quantification of the world that may include data from sensors, multiple industrial and domestic networks as well as financial markets, whereas BSD “comes from the mediated communicative practices of our everyday lives, whenever we go online, use our smartphone, use an app or make a purchase.” As for its importance Coté (2014) argues that SBD is not a novel concept, and has a huge impact for many reasons, processing a huge amount of data can provide valuable information.

It should be noted that other authors and researchers believe that social big data plays an essential role in the analysis of enormous of data generated via social media. Among the authors we find Bello-Orgaz (2016). The latter insinuate that SBD represents “a process and methods that are designed to provide sensitive and relevant knowledge to any user or company from social media data sources”. The collected data is characterized by their varied formats and content and also their very large size. Based on the work of Bello-Orgaz (2016), SBD incorporates three major different concept, big data (as a processing paradigm), social media (as the main source of data), and data analysis (as a method that will transform the collected data into relevant knowledge).

- Characteristics:

Since SBD is the result of the interaction between social media and big data, we can say that the two concepts share some characteristics (5Vs). However given the nature and the characteristics of social media, SBD is also characterized by Veracity and vagueness.

— Veracity:

This (recently added) feature does not yet have a uniform definition in the academic literature; each author uses / proposes a slightly different definition. The idea of veracity was used long before the birth of social big data and big data; it is deployed in several fields of research (biology, psychology, medicine, etc.). It alludes to the credibility of hypotheses and bibliographic resources used. We have to wait until 2012 that veracity was deployed in the context of social big data. Its first use is attributed to Snow (2012). The reasoning behind its incorporation was the complexity as well as the variety of the data collected and also the reliability of the sources.

When it comes to our research and social big data, there are several definitions of veracity. According to Orgaz et al (2016) it "refers to the accuracy and precision of information". This definition goes hand in hand with the one proposed by Bagiwa (2017), the latter considers veracity as "the disorder or reliability of the data, due to the large amount of data and its varied forms, the quality and precision are less controllable, for example Tweets or Facebook posts which are often characterized by abbreviations, typos and familiar speeches". The definitions do not end there, for example IBM and Microsoft have referred to "veracity as the fourth V, it represents the inherent unreliability of some data sources". For Gandomi and Haider (2015) "veracity refers to disorder and reliability of data", while for Storey and Song (2017) "veracity raises issues related to data quality". Other authors define veracity by citing associated dimensions such as precision, credibility, completeness, availability (Agrawal, 2012), consistency and accessibility (Corbellini et al (2017), integrity and authenticity (Denchenko et al 2013), reliability, authenticity, responsibility, availability (Mohan, 2016).

To conclude we can say that veracity refers to the accuracy, credibility and reliability of a set of collected data, it also refers to the integrity and objectivity as well as the reliability of the sources of these data especially when it's social media.

— Vagueness:

Vagueness is the new characteristic but also perhaps the trickiest one to define and address. In general vagueness refers to the indistinctness of existence in data. When it comes to big data vagueness represents the confusion over its meaning, nature, content availability, tools…. In the literature many authors tried to define vagueness as a characteristic such as Bonne (2014), Svetlana (2014), Panimabat et al (2017). When it comes to SBD, vagueness was introduced by Ishikawa (2015) in his book. Ishikawa (2015) maintains that vagueness "is a result of a combination of various types of data to be analyzed, which lead to inconsistency and deficiency. It also relates to the issues of privacy and data management as social data involves individuals' personal information. The proposed definition should not be mixed with the definition stated earlier that refers to the confusion over the meaning of big data.

When it comes to our work, we will focus on the definition made by Ishikawa.

- SBD types:

The rapid evolution of ICTs has completely transformed the role of users from a simple consumer to an active producer or mediator of information. This big change gave birth to a new type of users, "they are more in control of their profiles they can personalize/model the shared content according to their values, needs and preferences" (Cioffi-Revilla, 2013). These users are called "digital human».

The digital human represents the cornerstone of a society which balances between a physical world and a virtual world, during his interaction with the virtual world, the user gives birth to two types of data: data generated only by the machine and human generated data. Both of them can give users a fair share of social insight.

— Machine generated data:

This type data is considered the lifeblood of the Internet of Things. As its name suggests, machine generated data represents a type of data automatically generated by a computer process, application, or any other mechanism without any human intervention. The data generated by machine is characterized by being produced at very high rate and also by having no single form, type, format or metadata. According to Monash Research's Curt machine generated data can be defined as "data that was produced entirely by machines OR data that is more about observing humans than recording their choices". On the hand Abadi propose a more narrow definition of the concept, according to the author "machine generated data is data that is generated as a result of a decision of an independent computational agent or a measurement of an event that is not caused by a human action".

— Human generated data:

In general, the data generated by the human represents the content created following the interaction between the digital human (the user) and the social media, they exist in several forms created every day (email, audio file , video file, text, etc.). In the literature, human-generated data is divided into three subcategories: digital self-representation, Technology-mediated communication data and digital relationships data.

— Digital self-representation:

In what follows we will deal with the first type of social big data, even before defining this type, it should be noted that self-representation is not a new notion for human beings, because since its appearance humans have resorted to several methods in order to present themselves, starting with a simple painting in a cave, passing by well-detailed sculptures arriving at pictures / portrait photos. Today in the digital context, self-representation has retained its importance, it is considered the first step taken by the user (the digital human) in order to communicate and socialize with other users.

In the literature digital self-representation is defined in several ways, according to Warburton (2010) "digital self-representation constitutes a part of individual identity, and in an increasingly digitalized environment it represents an increasing part in sources of access and possible knowledge of an individual". For their part, Boyd and Heer (2006) consider digital self-representation as "the most frequent strategy in online participation and communication, this strategy is based on data such as virtual profiles, user-published content or a community".

— Technology-mediated communication data:

This type of social big data represents all the data generated from communications via social media, thanks to these communication platforms this type of data is easily created. In the literature there is a very minimal number of definitions of this type. According to, Olshannikova et al (2017) this type of data represents, "the data generated during two-way communication or during the collaborative creation of knowledge during the distribution of knowledge or information via social media platforms".

— Digital relationships data:

Social media platforms such as Facebook and Twitter offer their users the chance to create virtual communities based on already existing connections in the physical world or with other users (virtual connections), this characteristic has given rise to a type of data called digital relationships data, this type of data describes the links between and the implicit and explicit relationships between users of social media.

The analysis of this data offers a general and deep insight into social relationships and structure, as well as a better understanding of social phenomena, this information can be exploited to better know and target customers.

C. Unit Social big data management:

The proliferation of social big data allows the creation of added value in various areas, ranging from optimizing customer relation to significantly increasing margins. According to Stieglitz et al (2018) "social media data can be analyzed to gain insights into issues, trends, influential actors and other kind of information". In order to create and generate the value needed, companies and organizations must use and master several techniques and technologies essential for handling and extracting the coveted information from the data at their disposal. These techniques and technologies fall under what we call data management.

Basically data management "is seen as the gathering, processing, management of data producing new information for end users" (Emani et al, 2015; Krishnan, 2013). However, despite its benefits there's little to no research that focus on the management of SBD even though the latter has always been the order of the day for researchers and practitioners, this is not the case for big data management. The latter was always discussed and covered by many authors and researchers such as Karmasphere (2011), Mork and Miller (2013), Curry (2014), Gandomi and Haider (2015) etc..... Each of these authors proposed its own version of big data management process under the name "big data value chain". The proposed value chains have small differences (that can be overlooked) but also have several points such as the 4 big steps that should be followed: Acquisition, Organization, Analyze and Decision each of these steps contain many actions.

When it comes to SBD, and since it's considered as a subgenre of big data, we can say that the four steps cited before are also applicable here. However and considering its nature and characteristics it's only right to add more steps.

- Social Big Data value chain:

— Big data value chain:

The value chain was created thanks to the work of Porter (1985), "the value chain represents a series of activities that can lead to the creation of value", and based on this principle several authors, researchers and organizations have tried to apply the value chain in their respective fields. Rayport and Sviokla (1995) are considered among the first to adopt the value chain in another domain precisely to information systems. Based on the logic of Porter (1985) and Rayport and Sviokla (1995), Karmasphere (2011) a company specializing in the optimization of contact points proposed its own interpretation of the value chain in the field of data and big data, the merger between the two concepts has given birth to what is called - big data value chain -. The latter is considered by the European Commission as "the cornerstone of the economy of the future, the knowledge economy".

Karmasphere (2011) proposed a framework that can identify the key elements of BD value Chain, this framework is extremely focused on the processing of data already acquired and stored is split into four major stages called the 4As: acquisition, assembly, analysis and finally action. These steps can be segmented into:

- Data management: This category is made up of the following tasks, acquisition and assembly, it is considered as the basic element of the chain value, and during this step the data is retrieved / extracted from several sources and organized according to its types: structured, unstructured, semi-structured data.
- Data analysis: In turn, this category is made up of the following tasks: analysis and action. During this phase companies begin to extract value from the data collected, the extracted value will serve as a basis for decisions.

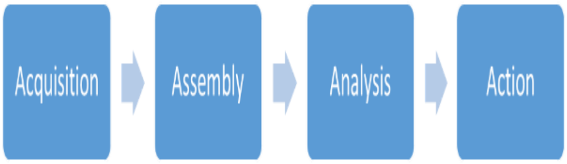


Fig.2: Value chain according to Karmasphere (2011)

Since this framework relegates the analysis, cleaning and filtering of data to the background, it is often judged by theorists to be too limited.

In the same way as Karmasphere (2011), several theorists and organizations have formulated their own versions of the big data value chain, according to Labrindis and Jagadish (2012) have proposed a model composed of five steps; these same steps constitute the two main sub-processes, data management and data analysis.

- Data management: involves the main things that enable the company to acquire, store and prepare data.
- Analysis: refers to the techniques used for the analysis and creation of knowledge and value.

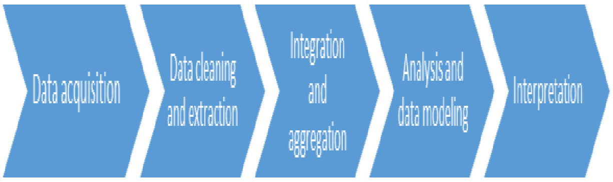

Data acquisition
Data cleaning and extraction
Integration and aggregation
Analysis and data modeling
Interpretation

Fig.3: Value chain according to Labrindis and Jagadish (2012).

Unlike the Karmasphere model, the model here emphasizes the phase of integration (or consolidation) and cleaning of data.

- Data cleaning: This is the operation of detecting, correcting or even deleting the various errors that exist on the data acquired and stored. This operation can be carried out using several technologies which will be mentioned below.

- Data integration: This is the operation that involves combining data from multiple sources (social media, internet of things, traditional data…), in order to provide users with a new idea and a benefit. It should be noted that non-compatible data is not used and is discarded.

Other theorists use the term pipeline to refer to the chain value, according to Gandomi and Haider (2015) the pipeline is made up of six major stages: data acquisition, data extraction and cleaning, data integration and aggregation. (These three steps represent the data management process), while the analysis process has the following two steps: data analysis and interpretation.

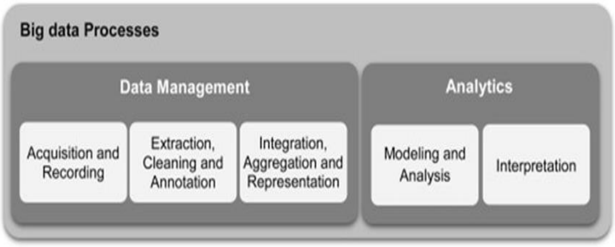

Big data Processes
Data Management
Analytics
Acquisition and Recording
Extraction, Cleaning and Annotation
Integration, Aggregation and Representation
Modeling and Analysis
Interpretation

Fig.4: Processes for extracting insights from big data according to Gandomi and Haider (2015)

.

- SBD value chain:

Based on the analysis of the frameworks mentioned above, it is obvious that certain steps are crucial to extracting value; these same steps will serve as a basis for us in the development of our value chain.

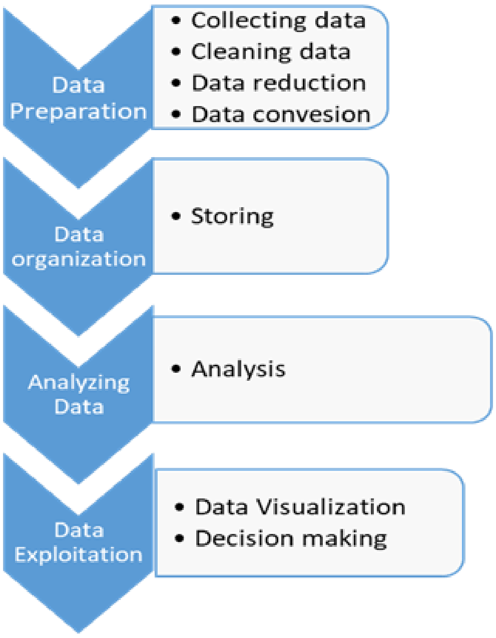


Fig.5. Social Big Data value chain:

— Acquisition and preparation of data:

The preparation of the data represents the first phase of our value chain, it is considered particularly crucial for the analysis of SBD and the creation of value, given the nature and characteristics of this type of data. The objective of this phase is to collect and acquire high speed data from different social media sources by using APIs and WebCrawler. One the data has been collected, firms can now process raw datasets, improve its quality, verify its veracity and limit the disclosure of private and sensitive information as much as possible. Once these tasks are completed the firm can now store the data.

— Organization:

This phase represents the second phase of the value chain; it's mostly composed of a single major task “storing”. The purposes if this phase is to store, manage and organize data in an efficient way (even when it's subjected to a heavy load of data) in order to facilitate the next phase. Organization is considered one of the most complex aspects of the social big data world, because unlike other elements of the process, this step does not only rely on software but it also requires a significant infrastructure. The latter should be able to deal with various data formats and store it in the right location and also must be able to parse them and extract the actual information like named entities, relation between them, etc.

It should be noted, that a storage system is said to be efficient when it takes into account in one hand, one of the factors proposed by Brewer (2000): Consistency, Availability and Partition of Tolerance. And on the other hand when it offers users simultaneity the chance to store an unlimited amount of data and manage a high rate of random access to said data.

— Analyze:

During this phase, the collected data is subjected to several techniques in order to extract information deemed necessary to facilitate decision making.

The collected SBD data in itself has no value and requires deep analysis in order to extract and acquire information from the collected data hence the critical status of this step. Some refers to this phase as the element that bridges the gap between data and knowledge.

In the literature, it is defined in several ways. According to Davenport and Harris (2007) it represents "the intensive use of data, quantitative statistical analysis, explanatory and predictive models as well as factual management to guide decisions and actions to be taken". On the hand for (Gudivada et al, 2016), it represents “any actionable information that results from computational analysis of data using mathematical and statistical methods. Data analytics is an interdisciplinary domain encompassing mathematics, statistics, and computer science”. From these definitions it is evident that this phase is interdisciplinary and multidimensional, it involves the use of disciplines such as statistics, operational research, information systems to name a few examples.

SBDA is generally categorized into five major categories that differ from each other in complexity, value and techniques used: method that describes what happened, method that describes why it happened, method that describes what will happen, method that facilitates obtaining the desired results and finally the one that describes the best action or decision to be taken. These methods are respectively named as follows: Descriptive, diagnostic, predictive, prescriptive and finally cognitive analysis.

Analytics	Définitions	Technics used
Descriptive	Based on the exploitation of historical data in order to identify patterns and create management reports based on past behavior. (Assunçao et al, 2015)	▪ Standard reports and dashboards ▪ AD / HOC reporting ▪ OLAP ▪ Line graph
Diagnostic	An application of data analytics to investigate the causes and effects of situations. (Fleckenstein et al, 2018).	▪ Correlations ▪ Drill Down
Predictive	Use Data to identify Past patterns to predict the future (Bagiwa, 2017).	▪ Data mining ▪ Text mining ▪ Web / media mining
Perspective	Uses data and mathematics to determine a set of high-value alternative actions given a complex set of objectives, requirements, with the goal of improving business performance (Damirkan and Delen, 2013).	▪ Machine Learning ▪ Neural networks ▪ Prescriptive dashboards
Cognitive	"The natural evolution of both data mining and visual analytics. It removes humans from the loop and is completely Automated, it combines the computing and cognitive science approaches" (Gudivada et al, 2016).	▪ Deep Learning ▪ Computer learning systems ▪ Pattern recognition

Table. 2: 5 Types of Analytics

— Decision:

It represents the last phase of our value chain, during this phase the user interprets the results found during the previous phase, then it relies on these interpretations in order to make the appropriate decisions. It is generally made up of two tasks: data visualization then interpretation / decision making.

- Visualization:

Broadly speaking, it can be described as the science of analytical reasoning facilitated by static or interactive visual interfaces, to be precise it represents an iterative process that involves information gathering, data preprocessing, knowledge representation. In the literature visualization is defined by several authors, according to Batrinca and Treleaven (2015) "it is the visual representation of data via diagrams in order to communicate information in a clear, efficient and compact way", according to Nasser and Tariq (2015) "visualization reveals so-called hidden patterns and patterns as well as unknown correlations to improve decision-making".

In general, graphs and dashboards are the most used techniques (for decades) to synthesize data in a coherent, compact and understandable format. However, as the volume of data increases, traditional visualization techniques cannot handle this huge volume, which has given rise to a new field of advanced visualization that uses interactive methods to represent thousands or even millions of points.

- Decision making or interpretation:

Decision-making is the last link in the value chain, this step consists in determining the necessary actions to be taken based on the gathered and visualized results, it includes an interpretation and a critical evolution of the reports, diagrams tables, while taking into account the limitations, the validity and the reliability of the methods used. In the literature it is considered to be the phase with the greatest value in the chain.

- Challenges related to the use of SBD value chain:

In order to make the best use of the value chain, companies face several challenges. These challenges can be categorized according to the phases of the chain as well as the tasks to be accomplished.

— Phase 1 : data acquisition and data preparation:

In the quest for data acquisition and preparation, companies face several challenges. Each of these challenges is directly related to the task to accomplish.

- Data collection:

In this first task, companies face primarily technological challenges among them we find:

✓ Difficulty in collecting, acquiring, deciphering data related to experiences, due to the irregularity as well as to the diversity of the language used, the latter often contains an informal language (sarcasm, acronyms, spelling mistakes), this content is often ambiguous and subject to human interpretation, not algorithms.

✓ Difficulty in deciphering opinions/feelings. In order to overcome this obstacle, it is necessary to establish lists of pseudo codes/terms one which designates positive feelings/opinions while the other designates negative feelings/opinions.

✓ Despite their necessity, APIs prove to be quite limited in terms of space and the number of units allowed, for example YouTube data API sets a limit of 30,000 units/users, while the total quota per day is set at 50,000,000 units, for its part Twitter allows only 15 requests per minute. These limits prevent a large amount of data from being obtained.

- Data cleaning:

When it comes to data cleaning firms face diverse set of challenges some of them were mentioned in the literature, according to (Freitas and Curry, 2016) "the central challenge of cleaning models is to manage the long data trail and improve the extensibility of data curation». In addition to this challenge, practitioners distinguish many others such as the high costs of curation projects, the time constraint this is directly related to the volume and variety of data collected as well as to security.

- Data reduction and data conversion:

The most obvious challenge is that of technology, companies must at all costs update IT and technological services in order to cope with the complexity of the task and the complexity of the data in their hands this in order to avoid any corruption that can affect data. The other challenge that companies face is directly related to redundancy elimination methods, deleting non-duplicate data, in order to prevent this from happening, it is recommended to create duplicate media.

Another concern is data security due to open source programs, adding to this the loss of the original data format, which in some cases poses problems of compliance with legal constraints.

— Phase 2 : data organization:

Despite the immense technological progress the world has experienced, companies face several challenges:

- Privacy and Security:

Data storage is no longer the main challenge faced by companies following the birth of Cloud computing. The main threat encountered is data privacy and security; if ever the systems are compromised personal data can be disclosed. Therefore, it is of utmost importance to secure and protect data against threats.

- Integrity check:

This challenge is directly linked to the use of the Cloud. When businesses use this option, they lose full or partial control of the data. In this case the data is outsourced and always remains at risk. Verification is of utmost importance, this task can be carried out by companies or by third parties.

- Cost:

The cost of this operation is always high, whether the companies use disks or systems. It should also be noted that the transfer of data to the Cloud as well as their hosting remains very expensive due to the excessive volume.

— Phase 3 data analysis:

In order to apply this operation correctly, companies face certain challenges, according to Ahmed and Ji (2013), companies face the following challenges: Budget and investment costs, also the data availability. Other authors cite data security and confidentiality as the main challenge faced by companies, as well as the talent gap (even if the field of big data is growing; there is still a lack of experts who can perform these tasks quickly, efficiently and correctly).

— Phase 4 data exploitation:

Companies face several challenges when carrying out this step:

- The difficulty in making decisions, since data is collected from social media, the latter are constantly developing. In this case the platforms must be constantly monitored in order to have a major margin of time to take the necessary measures, it is necessary to use real time big data analytics.
- Overestimation of the accuracy of the analysis.
- The use of erroneous data.

IV. CONCLUSION:

The goal of this work was to introduce and study social big data as a separated concept in all its multidisciplinary and multidimensional nature, to accomplish this task we decided to divide our work in two chapters. In the first chapter we defined social big data based on an in depth literature search, while in the second we tried to tackle social big data management.

When it comes to the first chapter, our literature overview showed us the existence of two different ways of treating and studying social big data. The first stream of thought tries to define the concept in its own rights, while the other focuses more on the analysis of social big data and the concepts related to it. These differences led to a lack of consensus and a little bit of vagueness. In order to remedy this problem we tried to define the concept in question based on the other definitions and also our own observations. Based on the proposed definition we outlined social big data's characteristics. Finally, to close the chapter we proposed a brief classification of the concept.

As for the second chapter, it was entirely devoted to social big data management; in this chapter we presented the concept of social big data chain value. This chain value covers both the entire data lifecycle, and also how companies and users can create value from the collected data. In general, the chain value is made up of four main phases each of which play a different and specific role in the value creation process: data preparation (discovery of data), data organization (discovery and preservation of data), analysis (creation of value), and exploitation (realization of value). During the same chapter we defined each phase and identify every challenge the users might face. It should be noted that these challenges vary from one organization to another.

In summary with this paper we aim to draw researchers to this concept in the hopes of developing a better conceptualization and understanding of social big data as a concept.

REFERENCES

[1] Abadi Daniel, 2010. Machine vs. human generated data. (http://dbmsmusings.blogspot.com/2010/12/machine-vs-human-generated-data.html).

[2] Ahmed, Zafor and Ji, Shaobo, Business Analytics: Current State & Challenges. CONF-IRM 2013 Proceedings. 12. 2013.

[3] K. Arum, L. Jabasheela, Big Data: Review, Classification and Analysis Survey, International Journal of Innovative Research in Information Security, Vol. 4, N°3, p.17-23, 2014.

[4] M.D. Assunção, R.N. Calheiros, S. Bianchi, M.A.S. Netto, R. Buyya, Big Data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, Vol. 79, p.3-15, 2015.

[5] L.I. Bagiwa, Big Data: Concepts, Approaches and Challenges, International Journal of Computer Networks and Communications Security, Vol. 5, N°8, p.181–187, 2017.

[6] B. Batrinca, P.C. Treleaven, Social media analytics: a survey of techniques, tools and platforms, AI & SOCIETY, Vol. 30, N°1, p.89-116, 2014.

[7] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: Recent achievements and new challenges, Information Fusion, Vol. 28, p.45-49.

[8] D. Boyd, J. Heer, Profiles as Conversation: Networked Identity Performance on Friendster, Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), Vol. 3, 2006.

[9] D. Boyd, K. Crawford, CRITICAL QUESTIONS FOR BIG DATA, Information, Communication & Society, Vol. 15, N°5, p.662-679, 2012.

[10] L. Burkholdee, “PHILOSOPHY AND THE COMPUTER. ROUTLEDGE”. [S.l.], 1992.

[11] E. Cambria, D. Rajagopal, D. Olsher, D. Das, Big Social Data Analysis, Big Data Computing, p. 401-414, 2013.

[12] C.L. Philip Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, Vol. 275, p.314-347, 2014.

[13] M. Chen, S. Mao, Y. Liu, Big Data: A Survey, Mobile Networks and Applications, Volume 19, N°2, p.171-209, 2014.

[14] C. Cioffi-Revilla, Introduction to Computational Social Science. Springer International Publishing, 2017.

[15] M. Coté, Data Motility: The Materiality of Big Social Data, Cultural Studies Review, Vol. 20, N°1, 2014.

[16] T.H. Davenport, J.G. Harris, “Competing on Analytics: The New Science of Winning”, Harvard Business Review Press, Boston, 2007.

[17] Dbms2.com. 2021. Examples and definition of machine-generated data | DBMS 2: DataBase Management System Services. [online] Available at: <http://www.dbms2.com/2010/12/30/examples-and-definition-of-machine-generated-data/> [Accessed 17 October 2021].

[18] D. Delen, H. Demirkan, Data, information and analytics as services, Decision Support Systems, Vol. 55, N°1, p.359-363, 2013.

[19] E. Dumbill, Making Sense of Big Data, Big Data, Vol. 1, N°1, p.1-2, 2013.

[20] N. Elgendy, A. Elragal, Big Data Analytics: A Literature Review Paper, Advances in Data Mining. Applications and Theoretical Aspects, p.214-227, 2014.

[21] C.K. Emani, N. Cullot, C. Nicolle, Understandable Big Data: A survey, Computer Science Review, Vol. 17, p.70-81, 2015.

[22] A. Freitas, E. Curry, Big Data Curation, New Horizons for a Data-Driven Economy, p.87-118, 2016.

[23] M. Fleckenstein, L. Fellows, Modern Data Strategy, 1st ed., Cham, Switzerland, p.133, 2018.

[24] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, Vol. 35, N°2, p.137-144, 2015.

[25] V.N. Gudivada, M.T. Irfan, E. Fathi, D.L. Rao, Cognitive Analytics: Going Beyond Big Data Analytics and Machine Learning, Handbook of Statistics, p.169-205, 2016.

[26] I. Guellil, K. Boukhalfa, Social big data mining: A survey focused on opinion mining and sentiments analysis, 12th International Symposium on Programming and Systems (ISPS), 2015.

[27] J. Hendler, Broad Data: Exploring the Emerging Web of Data, Big Data, Vol. 2, N°1, p.18-20, 2013.

[28] P. Ishikawa, “Social big data mining”, CRC Press, Boca Raton, Florida, 2015.

[29] D Laney, 3D Data Management: Controlling Data Volume, Velocity, and Variety, Gartner, file No.949, 6 February 2001.

[30] D. Lazer and al, Computational Social Science, Science, Vol. 323, N°5915, p.721-723, 2009.

[31] J. Manyika, J. Chui, M. Brown, et al, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011.

[32] V. Mayer-Schönberger, K. Cukier, “Big Data: A Revolution that Will Transform how We Live, Work, and Think”, Houghton Mifflin Harcourt, Boston, 2013.

[33] P. Mikalef, I. Pappas, J. Krogstie, M. Giannakos, Big data analytics capabilities: a systematic literature review and research agenda, Information Systems and e-Business Management, Vol. 16, N°3, p.547-578, 2017.

[34] R.R, Mukkamala, A. Hussain, R. Vatrapu, Fuzzy-Set Based Sentiment Analysis of Big Social Data, 2014 IEEE 18th International Enterprise Distributed Object Computing Conference, 2014.

[35] T. Nasser, RS. Tariq, Big Data Challenges, Computer Engineering & Information Technology, Vol. 04, N°03, 2015.

[36] D.T. Nguyen, D. Hwang, J.J. Jung, Time-Frequency Social Data Analytics for Understanding Social Big Data, Intelligent Distributed Computing VIII, p.223-228, 2015.

[37] E. Olshannikova et al, Conceptualizing Big Social Data, Journal of Big Data, Vol. 4, N°1, 2017.

[38] M.E. Porter, “Competitive advantage”, Free Press, New York, 1985.

[39] F. Provost, T. Fawcett, Data Science and its Relationship to Big Data and Data-Driven Decision Making, Big Data, Vol. 1, N°1, p.51-59, 2013.

[40] M. Schroeck, R. Shockley, J. Smart, et al, Analytics: The realworld use of big data. New York, NY: IBM Institute for Business Value, Said Business School, 2012.

[41] H. Snyder, Literature review as a research methodology: An overview and guidelines, Journal of Business Research, Vol. 104; p.333-339, 2019.

[42] S. Stieglitz et al, Social media analytics – Challenges in topic discovery, data collection, and data preparation, International Journal of Information Management, Vol. 39, p.156-168, 2018.

[43] V. Storey, I.Y. Song, Big data technologies and Management: What conceptual modeling can do, Data & Knowledge Engineering, Vol. 108, p.50-67, 2017.

[44] J. Tang, Y. Chang, H. Liu, Mining social media with social theories, ACM SIGKDD Explorations Newsletter, Vol. 15, N°2, p.20-29, 2014.

[45] S. Warburton, “Digital Identity Matters”, London: King’s College London, 2010.

[46] Wong, G., Greenhalgh, T., Westhorp, G. et al. RAMESES publication standards: meta-narrative reviews. BMC Med 11, 20, 2013.

[47] P. Zikopoulos, R.B. Melnyk, “Harness the power of big data”, New York, 2013.