INTELLIGENCE ANALYSIS

BIG DATA AND ITS SECRETS: TYPES OF BIG DATA*

Alexandra ANGHEL*
Elena NOVĂCESCU*
Mădălina CUC*

Abstract:

The technological boom that characterized the last decades changes the rules in all societal domains, generating large volumes of data to be both processed and stored. The industry of communications developed new innovative communication channels and instruments (such as social media platforms and online applications that allow people to communicate across borders) that even though facilitated the communication process as a whole, generated several challenges for both network/platform administrators and cyber security agencies (the online environment becoming a new battlefield of the next generation wars, where data are the main weapons). In this context, one can conclude that mastering the art of managing large volumes of data it is a vital asset in understanding the architecture of the digital society – a reality of the 21st century. Therefore, the article aims to define, based on a process of literature review, the main characteristics of Big Data, a concept used to define the large quantity of data produced by any society nowadays. It will present the evolution of the data concept from small data to big data, the main types of big data and a common analysis framework for Big Data, all

* Acknowledgement: This work was possible with the financial support of the EEA Grants - Financial Mechanism 2014-2021, through the project THESEUS - *Connect the Disconnections – from Disparate Data to Insightful Analysis*, Contract Number 18-COP-0017. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. This paper was also used in materials developed for the THESEUS project and course.

* Junior researcher, "Mihai Viteazul" National Intelligence Academy THESEUS Project member. Email address: anghel.alexandra@animv.eu

* Junior researcher, "Mihai Viteazul" National Intelligence Academy, THESEUS Project member. Email address: elena.novacescu@animv.eu

* Lecturer, PhD, "Mihai Viteazul" National Intelligence Academy, THESEUS project member. Email address: cuc.madalina@animv.eu

in order to provide a better understanding of the Big Data concept and its application in current societies.

Keywords: *data, Big Data, analysis framework, small data, structured and unstructured data.*

Introduction

Given the emergence of new communication technologies and channels, such as social networks and platforms, mobile computing and online services, the data generated do not possess anymore a standard format or structure, and requires new specific models and instruments in order to be processed. Data now take different forms – from traditional text, images, videos to XML, weblogs, json, posts and so on – resulting in an increased number of new data types (Eberendu, 2016, p. 46), challenging the traditional analysis and process models and patterns.

As a consequence, the concept of Big Data became a common, extensively used term in the last years, with applicability in all vital domains of a society. However, in order to be able to analyse the relevance of Big Data or Small Data in an ecosystem based on the analysis of Big Data with Artificial Intelligence elements, we must start from the definition of the concept of "data".

In general, data are the representation of a phenomenon/event that occurs in reality in the form of information. The process of data production is based on the production of an event or phenomenon in the real world, is focused on information and uses the representation of the phenomenon modelled by the observer in various forms that define the notion of data. The need for representation arose from the need to make better decisions in the real environment where the impact of external and internal factors was decisive in the evolution of human society in its infancy (Bokulich & Parker, 2021).

In this context, this article aims to present the evolution of the Big Data concept, analysing the path from data to information, and then to digital data/information throughout an extensive literature review process. In addition, this paper will also focus on defining the different types of data that are characteristic for the Big Data domain, as well as identifying the key variables to be considered when analysing them.

Data versus information

Today, any organization, of any size, can have access to scalability, transparency, security compliance and reliability for computing and data systems, that in the past were accessible only to the largest companies. Starting with the industrial revolution, the business domain understood that technology can become a leverage when efficiently implemented and exploited so as to maximize the impact on the companies and their customers. Today, the success of businesses is defined by their speed in serving any customer, therefore it is essential to implement within their companies and equip their employees with state-of-the-art technology to ensure a better collaboration, a smarter communication channel, an efficient customer relations service, prolific human resources and so on (Cloud Backup Techniques & Tips, 2015). All these essential services made with the help of computing are "modern computing bulbs" and Big Data are the power plants that will start the "bulbs of tomorrow" (Johnson & Gueutal, 2011).

As we have already mentioned, data are **the representation in the form of information of a phenomenon that occurs in reality**. In comparison, the concept of information refers to a representation of reality, and at the same time to the result of the processes of reflection and projection, aspect that usually applies only for human intellect – throughout the medium of a structured set of symbols – not difficult to access by human senses and reason, but, in some situations also accessible for some devices, such as automatic calculation (computers). Information cannot be considered, therefore, neither as a specific content, nor as an agent, instruction, property, method or process, but information can rather be categorized as an independent category, characterized by an abstract and subtle immaterial existence, a category that is considered vital in the process of knowledge (Cimil & Plotnic, 2021). In this case, the information is the reflection after observing the phenomena in the real world around (Madden, 2000).

Therefore, when trying to differentiate between data and information, one should follow the following flow (as also described in the figure below):

1. Production of the phenomenon in the real world (followed by one or more observers);

2. Reflection (modelling) of the phenomenon in their consciousness (production of information; more observers, more representations, resulting in different types of information associated with each observer);
3. Symbolic representation of the record in the consciousness of each observer (production of the data associated with each observer) (Marchetti, 2018).

It is obvious that there can be several modelling models of the same phenomenon in the real world with several modes of representations generated by the existence of several observers with the consequence of generating more data types that will be the modelled representation of the same real phenomenon viewed from several perspectives (Marchetti, 2018).

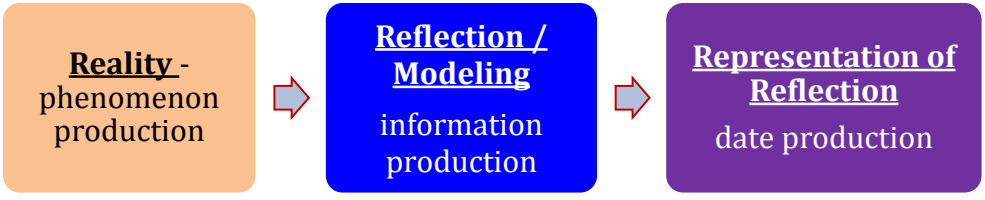

Figure 1: Genesis of the concept of data
(developed by the authors based on the conclusions of Marchetti, 2018)

The complexity of the relationship reality – information – data increases when the data are constituted in collections that must have material support. Thus, the event of storing data on material support becomes a reality, therefore information will now be generated about how to store previous information and these are called metainformation. Representing this as data defines metadata. Obviously, there can be several perspectives on the representation of metainformation, therefore metadata can be information about the location of data on the storage medium, and how it can be retrieved, how it can be updated or deleted (Sivarajah, Kamal, Irani, & Weerakkody, 2017).

After the appearance of writing and language, as well as the evolution of the medium on which the data were recorded, a biunivocal correspondence was found between the evolution of data and

information and the consumption of data support, respectively the storage capacity of these supports (stones, stone tablets, papyri, books, magnetic tapes, HDD etc.). The transition from information to data has been made since the beginning of their appearance, being transmitted along with data representing numbers (dimensions, days, months, years) and their reference and dependence (metadata) with reference to calendars, crop recipes agricultural, reserve management of ancient communities (Organisation for Economic Co-Operation and Development, 2007).

To resume, the concept of information can be defined in several ways, but whatever the meaning of the term, information will have a semantic character, contributing to the amount of knowledge of the recipient. Information always refers to objects, people, processes, phenomena, places, situations, conditions etc., so it has a very varied nature (economic, statistical, technical, scientific, administrative etc.). In contrast, data are the materialization, symbolic representation of information (through signs, letters, numbers, words etc.) in a conventional form (written, spoken, bright, graphic signs, drawings etc.), convenient for communication. Data have a (semantic) interpretation and is processed by humans directly through automated means.

Data are primary aspects of the surrounding reality, which are perceived through receivers, in various forms. For example, the data collected in a meteorological station come from various sources: measurements of atmospheric factors (temperature, humidity, atmospheric pressure) performed with specific instruments by the station staff, data collected by meteorological probes and satellites, etc. These data represent "raw material", unprocessed, without meaning and without obvious utility.

Types of data

In terms of types of data, the International standard ISO/IEC 11404: 2007 (E) presents three notions of types of data (International Organization for Standardization, 2007):

- the conceptual or abstract notion of a type of data, defined by its nominal values and properties;

- the structural notion of a type of data, characterized by the conceptual organization of the components and functionalities specific to some data types; and
- the notion of implementing a type of data, which identifies the type of data by defining the rules for representing the data type in a predefined environment.

This international standard includes an inventory of all data types, as well as a partial terminology for the different notions of implementation of data types, showing the use of this terminology in the definition process of data types, identifying thus common terms of implementation associated with data types and distinguishing them from conceptual notions. From the point of view of the physical representation of data in a computing system, at the processor level, a data is associated with one or more locations intended to store its current value and any information on its structure. Therefore, the list below comprises the main types of data that shape reality and are used in the Big Data ecosystem (International Organization for Standardization, 2007):

- Alphanumeric data – non-numeric data with values representing alphanumeric characters or strings. At the computer level, a character is represented by the code associated with it. The operations that can be performed with alphanumeric data are compaction, concatenation, comparison etc.
- Arithmetic data – numeric data with integer or real values. Due to the limited ability to represent numbers in the computer and due to the binary form of representation, the values of the arithmetic data correspond to a subset of real numbers, the axiomatic of real numbers is no longer fully observed and the calculations are approximate. In the computer the values of the arithmetic data are represented in fixed point or floating point. In programming languages, an arithmetic data with a constant value is presented in the form of a series of numbers and symbols, which indicate the sign, comma, exponent, etc. with a language syntax function.

- Data complex (time complex) – if falling on whose values are complex numbers (a + ib, i = sqrt (-1)). Every time the complex can be regarded as being made up of a structured imaginary part and the real one. There are programming languages that allow the direct performance of arithmetic operations with complex data.

Another classification model of data defines three main categories of data, as follows:

1. Structured data

Structured data is the type of data that is comprised by any database, defined as a collection of data that shapes a universe. This universe consists of several interacting objects, objects of the same type constituting an entity. An entity is an element of the real world. In addition, a model is an abstraction of a real-world system or process, a mathematical, formal description of the system. The model consists of a set of notations and terminology necessary to express ideas about the system or process described. The model is used either to study an existing system or to build a new system and it is obtained through a modelling process. If the purpose of the model is to study the modelled system, one can say that he/she performs an analysis process. The analysis process consists in studying the real-world system, identifying the representative features and retaining those characteristics relevant to the final goal. The rest of the features are ignored, only the relevant ones are included in the model (Castagna, 2021).

Structured data is, to summarize, data that adheres to a predefined data model where each record is constantly structured being efficiently described in a relational model. Structured data is easy to analyse because each data entry is conceptualized in a tabular format (as rows in a table), with each data item introduced in a singular cell (which form the columns and rows of a table). The most known examples of structured data are Excel files or SQL databases, built on a tabular format sortable rows and columns (Rambsy & Ossom-Williamson, 2021).

The collection of structured data can take the form of a dataset (records), multiple dataset (set of data organized according to certain criteria), data warehouse (a collection of multiple datasets between

which a series of links have been established that lead to a certain way of identifying and selecting the components) or metadata (the data about the previous data). The collection of data structured in the form of metadata is called "schema on write" and it assumes that the data (article values) are collected in the "schema" or structure developed before collection (Pickell, 2018).

2. Unstructured data

Unstructured data can be defined as information that does not follow a predefined data model or is not organized in a predefined way. The term unstructured data may seem to imply a complete randomization of the data collection without having an inherent form (Enterprise Big Data Framework, 2019). In fact, unstructured data incorporates a type of structure to a certain extent, in three ways:

- implicit but formally undefined structure;
- their structure is not useful for the targeted process;
- structure, called "semi-structured or even structured, but in unexpected or unforeseen ways" (Fritzner, 2017).

Unstructured information is usually text from web pages, but may also contain data such as general data, numbers, or events where the use of classic DBMS is no longer applicable (IBM Cloud Education, 2021). The Big Data analysis focused on extracting valuable information from unstructured data or from relationships between them but also on different methods that can more efficiently manage these types of data sets. Most eloquent examples of unstructured data include audio, video, or No-SQL databases (Rambsy & Ossom-Williamson, 2021).

For unstructured data, data collection is often followed by raw data storage. Unstructured data storage is done in "data lakes", which is an efficient way to store "all data" of an organization for further processing. The analyst will be able to focus on finding meaningful patterns in the data and not on the data itself. Unlike a hierarchical database in which data is stored in files and folders, Data Lake has a flat architecture. Each data element in a Data Lake receives a unique identifier and is labelled with a set of information stored in metadata. The data preparation processes then appear after storage and are managed by specific applications. This technique of storing raw data first and applying

a schema after interacting with the data is commonly referred to as a "read schema", as shown in the figure below (Pickell, 2018).

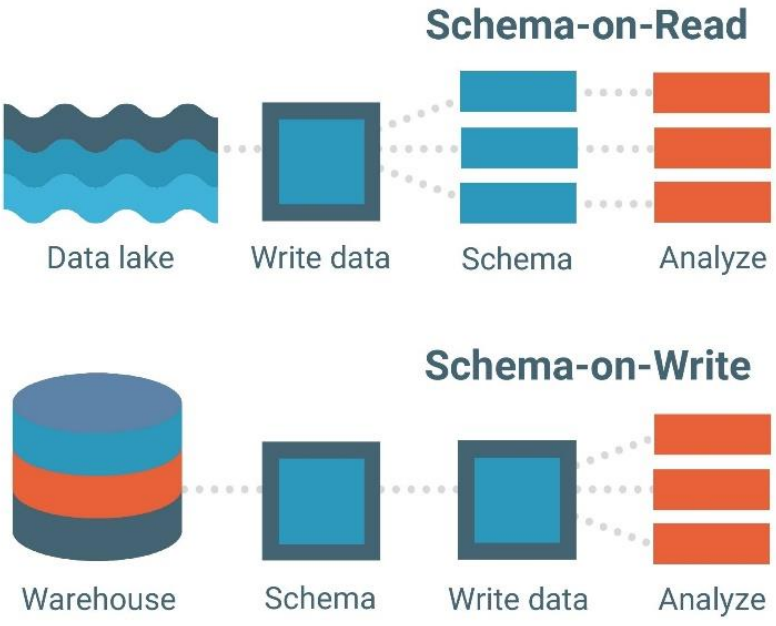

Figure 2: Structured vs. unstructured data (Pickell, 2018)

3. Semi-structured data

Semi-Structured Data is a "form of structured data that does not conform to the formal structure of data models associated with relational databases or other forms of data tables" (Kraus & Drass, 2020, p. 250-251), but still contains tags or other markers to separate semantic elements and to allow the hierarchy of records and fields in the data, known as the self-description structure. Semi-structured data are

composed of XML[1], JSON[2] or CSV[3] formats (Enterprise Big Data Framework, 2019).

Big data versus small data

With the emergence of the consumer society, companies and complex organizations appeared their information system, based on the analysis and interpretation of data and information from the internal and external environment. Initially these data were analogue (written tables, registers, nomenclatures, regulations, procedures for their use), and then with the advent of the first computers in the 1950s, the data appeared in digital format.

The first digital data was produced in 1937 at the International Telephone and Telegraph Co. in France by Alec H. Reeves (patent 1938) making possible the digital encryption of analogue voice transmission. The principle is used by Bell Labs (SIGSALY encryption) during WW2 in Winston Churchill's conversations with Franklin Roosevelt.

In comparison, the first digital storage medium was produced in 1939 by John V. Atanasoff and Clifford Berry, the first electronic computer with a drum storage device that used capacitors on its surface. The Atanasoff-Berry (ABC) computer had two drums to store 30 numbers on each drum, each number stored in 50 bits. As storage size on a current 16 GB USB flash fits the information of 85.3 million on such drums. In May 1955, IBM announced the first commercial magnetic disk (HDD), called RAMAC with a capacity of 5 million bytes[4] on 50 aluminium discs, each with a diameter of 61 cm, covered on both sides with magnetic iron oxide that weighed a ton. An ordinary 16 GB USB flash contained as

[1] XML (eXtensible Markup Language) XML is a model for storing unstructured and semi-structured data in native XML databases.

[2] JavaScript Object Notation or JSON is a format for representing and exchanging data between computer applications. It is a text format, intelligible to humans, used to represent objects and other data structures and is mainly used to transmit structured data over the network, the process being called serialization. It is used in GIS databases.

[3] Comma-separated values or CSV is the data format transmitted by most sensors and IoTs and lists comma-separated values.

[4] Virtual Storage: 1Byte = 8bits, 1Kbyte (KB) = 1024bytes, 1MB = 1024KB, 1GB = 1024MB, 1TB = 1024GB, 1PB = 1024TB.

much information as 3200 such disks (weighing 3200 tons means 80 train cars loaded with 40t each or 4 trains with 20 cars each).

The Digital Revolution marked the beginning of the information age (Tucci, 2014). At the heart of this revolution is the mass production and widespread use of digital logic, transistors, integrated circuits and their derivative technologies, such as computers, microprocessors, digital cell phones and last but not least, the Internet (Debjani, 2014). These technological innovations have transformed traditional production and business techniques (Bojanova, 2014).

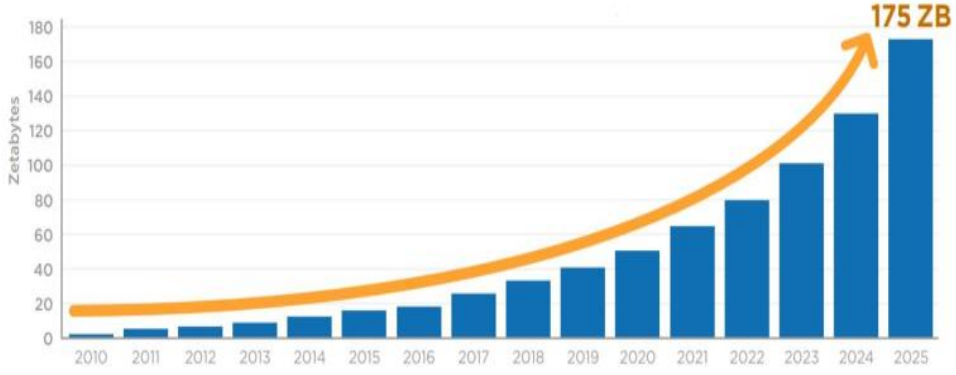

Figure 3: Evolution over time of digital data volume
(Reinsel, Gantz, & Rydning, 2018, p. 13)

The latest definitions (2021) of the concept of "Big Data presents the term as the field of knowledge that explores techniques, skills and technology to deduce valuable information from massive amounts of data. Big Data is considered information asset defined by a high level of volume, variety and velocity that require specific technological means and analytical methods to identify its value" (De Mauro, Greco, & Grimaldi, 2016). All in all, one can say that the concept of big data refers to massive data sets characterized by a varied and complex structure, which generate difficulties for companies and organizations in terms of storage, analysis and visualization (Sagiroglu & Sinanc, 2013, p. 42).

As far as small data is concerned, there are two approaches to this concept: (1) the first one refers not to the small size of the data (as the name would suggest), but to the inability of analysts to correctly assess the results if the data are too large and complex and there are not enough correlations leading to the intended purpose. Small data here defines the limits of the maximum size of data sets that analysts can directly evaluate and understand, beyond which the correctness of results and convergence towards the proposed goal can no longer be ensured (Kitchin & Lauriault, 2015); (2) the second approach defines Small Data as data that is "small" enough for human understanding, available in a format and volume that makes them accessible, informative and actionable (Sathyabama Institute of Science and Technology). Another definition enshrines Small DATA as small data sets capable enough to influence decisions today by allowing concrete action. In contrast to Big Data, the speed of ingestion of small data that reaches processing is constant and controlled, and the data flow is relatively slow, increasing both accessibility and processing speed. Usually Small Data is structured and located in the Data Base Management Systems (DBMS) and is easy to view (GeeksforGeeks, 2021).

Both types of data are relevant and vital for the processes of analysis and decision making, but in order to evaluate efficiently the efforts needed for the corroboration, correlation and analysis of different data sets it is important to acknowledge the difference between big data and small data. Therefore, there are a series of features that allow the differentiation between these two types of data, as follows (GeeksforGeeks, 2021):

Feature	Small Data	Big Data
Technology	Based on traditional technology	Based on modern technology
Collection	Generally, collection is conducted in an organized format, all the obtained data being inserted into a database	Comparatively, the collection of big data is conducted through pipelines that use queues such as AWS Kinesis or Google Pub/Sub in order to balance high-speed data
Volume	Comprises data contained between tens or hundreds of Gigabytes	The volume of this type of data registers more than Terabytes
Analysis Areas	Data marts (for analysts)	Clusters (for data scientists) and data marts (for analysts)
Quality	Given the fact that data is less collected in a controlled format, small data contains less noise	In general, the quality of data cannot be guaranteed
Processing	For processing small data it requires batch-oriented processing pipelines	For processing big data, it can be used both stream and batch processing pipelines
Database	Format: SQL	Format: NoSQL
Velocity	Data aggregation is slow, determining a regulated, constant flow of data	Data is produced at extremely high speeds, determining large volumes of data aggregation in short time intervals
Structure	Small data is structured in a tabular format, following a fixed schema (relational structure)	Big data consists of a variety of data sets, including tabular data, text, media files (audio, images, video), logs, JSON etc. (non relational structure)
Scalability	Small data are, in general, vertically scaled	Big data are mostly horizontally scaled, generating more versatility at a lower cost
Query Language	only SQL	Python, R, Java, SQL
Hardware	One single server	Requires more than one server

Feature	Small Data	Big Data
Value	Analysis and reporting, Business Intelligence	Complex data mining techniques and methods for pattern finding, recommendation, prediction and so on
Optimization	Manually optimization of data (human powered)	Data optimization based on machine learning techniques
Storage	Small data requires storage systems within different companies/enterprises, local servers etc.	Usually, big data requires distributed storage systems available on cloud or uses external file systems
People	Categories: data analyst, database administrator and data engineer	Categories: data analyst, data scientist, database administrator and data engineer
Security	Security protocols for small data include, but is not limited to: user privileges, data encryption, hashing etc.	Securing protocols for big data systems are complex and include, without being limited to: data encryption, strong access control protocols, cluster network isolation etc.
Nomenclature	Database, data mart, data warehouse	Data Lake
Infrastructure	Preponderently vertically scalable hardware and predictable resource allocation	More agile infrastructure with horizontally scalable hardware

Table 1. Big data vs. small data (GeeksforGeeks, 2021)

Having these aspects in mind, one can say that even though the concept of big data known an unprecedented exposure and growth, small data will remain a vital part of the research landscape in the near future. There is not expected a paradigm shift in which studies using small data will replace those employing big data, but small and big data will work together in a complementary manner (Sawyer, 2008).

Big Data and the 10Vs

Most definitions of the concept of big data highlights the size of data in storage, which is an important aspect, but there are also other valuable features specific for big data, such as velocity and variety. Thus, for a better understanding, the term of big data can be analysed through the lenses of the three Vs (volume, variety, and velocity), lenses that can also help building a comprehensive definition, busting the myth that big data refers only to the volume of data (Russom, 2011, p. 6):

- **Volume:** this variable is considered to be the primarily attribute of Big Data, referring to the size of the data sets to be analysed and processed, which are larger than petabytes and exabytes. The large volume of data requires processing technologies that are distinct and different from traditional storage and processing capabilities. In other words, this means that Big Data datasets are too large to be processed with a regular laptop or desktop processor (Russom, 2011, p. 6-7);
- **Velocity:** speed is a measure of the rate of data flow. Traditionally, high-speed data transfer systems have been described as streaming data. The collection of big data in real time is not a process specific for current times, as many companies have been collecting and corroborating clickstream data from Web sites for years, using streaming data in order to adapt their advertising strategies to the expectations and different profiles of Web visitors (Russom, 2011, p. 7-8);
- **Variety:** this variable refers to data from multiple repositories, domains, or types. The variety of data in several domains was addressed by identifying features that would allow the alignment of data sets and their merging into a data warehouse. Although volume and speed allow for faster and more cost-effective analysis, the variety of data allows analytical results that have never been possible before. The Big Data variety consists of storing and correlating data of various types: text, numerical data, time series, video, audio, images, tabular data (databases), hierarchical data,

> documents, XMLs, e-mails, blogs, instant messages, click streams, as well as all the different log files produced by any system with an embedded computer (Trnka, 2014, p. 144).

In addition to these three V's there are also other seven variables that can be considered when analysing big data, developing an extended V's framework, as follows:

- **Value:** data value refers to data usefulness in decision making and is one of the most significant factors in Big Data, because it has direct impact on business profits. Data value refers to data usefulness in decision making and is one of the most significant factors in Big Data, because it has direct impact on business profits (Khan, et al., 2018, p. 54);
- **Veracity:** this attribute focuses on the quality and accuracy of data and defines how data can be trusted when important decision needs to be made regarding the collected data. In the context of increasing the volume, speed and variety of data, the veracity (confidence in data) decreases. Accuracy is the characteristic that shows the degree of trust or distrust of the data in the need to store and process the real data, considering deviations as well as the information noise in the stored data (Khan, et al., 2018, p. 54);
- **Visualization:** visualization is an important step in any scientific data application to allow human understanding of data, analysis or results. Present data visualization tools face technical challenges as a consequence of the limitations specific to the current technological developments in terms of memory and reduced scalability, functionality and response time. It is true to say that for traversing a billion data points one cannot rely only on traditional graphs, needing various supplementary ways of representing data, such as data clustering, or tree maps, parallel coordinates, pie charts, cones etc. (Ashfaque Aatqb, 2020, p. 10);
- **Variability:** this variable refers to inconsistent data flow (in terms of data set, and not in the data format/structure and/or volume that affects its processing) and can be shown at times.

Variability in data volumes implies the need to expand or reduce virtualized resources to efficiently manage the additional processing task, one of the advantageous capabilities of cloud computing (Katal, Wazid, & Goudar, 2013);

- **Volatility:** refers to the life duration - for how long-time data is valid and for how long time it should be stored. Volatility answers the question: How old must the data be in order to be considered irrelevant, historical or useless? For how long should the data be stored? Before the apparition of the concept of Big Data, organizations have tended to store data indefinitely, but in the Big Data ecosystem these policies are no longer applicable (Firican, 2017);
- **Viability:** it refers to the capacity of Big Data to be live and active forever, and able for developing, and to produce more data when need. In simple words, viability focuses on identifying those features, as well as the relationship between them, which allow Big Data to remain available and further develop in order to extend its relevance and applicability (Khan, et al., 2018, p. 55);
- **Validity:** although the collected data may have a high degree of veracity (the exact representation of the real-world processes that created them), there are times when the data are no longer valid for the required hypothesis. Similar to veracity, validity also refers to the level of accuracy and correctness of the data in connection with the desired outcome. According to Forbes, an estimated 60% of data researchers spend time cleaning up data before they can do any analysis (Firican, 2017).

Conclusions

Big data represents a matrix of data sets which is constantly growing in volume as a consequence of the technological boom that created new means of communication – as factors continuously generating data and information. There is no common definition unanimously accepted for the concept of Big Data, but one can consider

that this term refers both to the techniques, instruments and methods used to collect and process the multitude of data produced worldwide, as well as to data sets that score high levels of volume, variety and velocity.

Even though the big data technological spectrum created various opportunities in terms of collection and analysis processes, it also brought new challenges and issues for analysist and system administrators, who were forced to adapt their mechanisms and methods to the realities of the 21st century. As a consequence, in order to facilitate the transition from traditional data analysis processes to digital and Big Data ones, researchers and scientist developed a general framework for defining big data, based on ten main features characterizing this type of data: among volume, variety and velocity, the framework also identifies value, veracity, visualization, variability, volatility, viability and validity as key components of the big data ecosystem.

Reference:

1. Ashfaque Aatqb, J. (2020, February). *The 10 Vs of Big Data.* Retrieved from ResearchGate: https://www.researchgate.net/publication/339107749_The_10_Vs_of_Big_Data

2. Bojanova, I. (2014). The Digital Revolution: What's on the Horizon? *IT Professional*, 8–12.

3. Bokulich, A., & Parker, W. (2021). Data models, representation and adequacy-for-purpose. *European Journal for Philosophy of Science*.

4. Castagna, R. (2021, April 15). *Structured vs. unstructured data: The key differences*. Retrieved from TechTarget: https://www.techtarget.com/whatis/feature/Structured-vs-unstructured-data-The-key-differences

5. Cimil, D., & Plotnic, O. (2021). European Union view on Personal Data in Intellectual Property Rights. *Eastern European Journal of Regional Studies*, 92-106.

6. Cloud Backup Techniques & Tips. (2015, April 21). *Changes to Computer Thinking – Stephen Fry Explains Cloud Computing*. Retrieved from Cloud Backup Techniques & Tips: https://www.corelnet.com/changes-computer-thinking-stephen-fry-explains-cloud-computing/

7. De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 122-135.

8. Debjani, R. (2014). Cinema in the Age of Digital Revolution. *International Journal of Interdisciplinary and Multidisciplinary Studies*, 107 -111.

9. Eberendu, A. C. (2016). Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology (IJCTT)*, 46-50.

10. Enterprise Big Data Framework. (2019, January 19). *Data Types: Structured vs. Unstructured Data*. Retrieved from Enterprise Big Data Framework: https://www.bigdataframework.org/data-types-structured-vs-unstructured-data

11. Firican, G. (2017, February 8). *The 10 Vs of Big Data*. Retrieved from TDWI: https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

12. Fritzner, J. E. (2017, June). *Automated Information Extraction in Natural Language.* Retrieved from Norwegian University of Science and Technology Open: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2454576/17567_FULLTEXT.pdf?sequence=1

13. GeeksforGeeks. (2021, September 29). *Difference Between Small Data and Big Data*. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/difference-between-small-data-and-big-data/

14. IBM Cloud Education. (2021, June 29). *Structured vs. Unstructured Data: What's the Difference?* Retrieved from IBM: https://www.ibm.com/cloud/blog/structured-vs-unstructured-data

15. International Organization for Standardization. (2007). *ISO/IEC 11404:2007 Information technology – General-Purpose Datatypes (GPD).* International Organization for Standardization.

16. Johnson, R. D., & Gueutal, H. G. (2011). *Transforming HR through technology. The Use of E-HR and HRIS in organizations.* Alexandria, VA 22314: SHRM Foundation's Effective Practice Guidelines Series.

17. Katal, A., Wazid, M., & Goudar, R. (2013). Big data: issues, challenges, tools and good practice. *Contemporary Computing (IC3)*, 404–409.

18. Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. A., & Salehian, S. (2018). The 10 Vs, Issues and Challenges of Big Data. *ICBDE '18: Proceedings of the 2018 International Conference on Big Data and Education*, (pp. 52–56).

19. Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal 80*, 463-475.

20. Kraus, M. A., & Drass, M. (2020). Artificial intelligence for structural glass engineering applications - overview, case studies and future potentials. *Glass Structures & Engineering*, 247-285.

21. Madden, A. (2000). A definition of information. *Aslib Proceedings.*, 343-349.

22. Marchetti, G. (2018). Consciousness: a unique way of processing information. *Cogn Process*, 435-464.

23. Organisation for Economic Co-Operation and Development. (2007). *Data and Metadata Reporting and Presentation Handbook.* Organisation for Economic Co-Operation and Development.

24. Pickell, D. (2018, November 30). *What Is a Data Lake and Why Is It Essential for Big Data?* Retrieved from G2 Business Software Reviews: https://www.g2.com/articles/what-is-a-data-lake

25. Rambsy, K., & Ossom-Williamson, P. (2021, November 8). *The Data Notebook*. Retrieved from Mavs Open Press: https://uta.pressbooks.pub/datanotebook/

26. Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World. From Edge to Core.* IDC.

27. Russom, P. (2011). *Big Data Analytics. TDWI Best Practices Report.* TDWI Research.

28. Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, (pp. 42-47).

29. Sathyabama Institute of Science and Technology. (n.d.). *Unit 1 – Big Data – SIT1606.* Retrieved from Sathyabama School of Computing: https://sist.sathyabama.ac.in/sist_coursematerial/uploads/SIT1606.pdf

30. Sawyer, S. (2008). Data wealth, data poverty, science and cyberinfrastructure. *Prometheus: Critical Studies in Innovation*, 355-371.

31. Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 263-286.

32. Trnka, A. (2014). Big Data Analysis. *European Journal of Science and Theology*, 143-148.

33. Tucci, L. (2014, March). *Information Age*. Retrieved from TechTarget: https://www.techtarget.com/searchcio/definition/Information-Ag

34. Wonderflow. (2019, April 01). *What is small data (in just 4 minutes)*. Retrieved from Wonderflow: https://www.wonderflow.ai/blog/what-is-small-data