BIG DATA - Present Opportunities and Challenges

Andrei Marcel Paraschiv
Ștefan cel Mare University of Suceava
Str. Universității 13, Suceava 720229
Phone: 0230 216 147
andrei@paraschiv.ro

Mirela Danubianu
Ștefan cel Mare University of Suceava
Str. Universității 13, Suceava 720229
Phone: 0230 216 147
mdanub@eed.usv.ro

Abstract

Large data volumes having various types and moving at high speed is the main feature of the current environment and is closely related to the Big Data concept. As a source of valuable insights, these data volumes have become an important asset for organizations, helping to make more informed and high-quality decisions. A real problem is that such data sets cannot be managed and processed by traditional methods. This paper presents some general aspects regarding Big Data and Big Data analytics - as source of deep insights, and it highlights some of the general opportunities, challenges and the European Commission strategy for Big Data.

Keywords: Big Data, Big Data Features, Eu Strategy for Big Data Adoption, Opportunities, Challenges.

1. Introduction

A long time ago Sir Francis Bacon affirmed that "Knowledge is power". Paraphrasing this dictum, in the modern ages it has been demonstrated that information is power, so there can never be too much information as basis for decision-making.

The last decades have been marked by an explosive technological development accompanied by huge possibilities for data generation, collection and storage. In this context data has become a torrent that floods all areas of society.

The diversification of data sources (transactions, social networks, sensors, etc.) led to the emergence of new and complex data types whose storage and processing exceeds the possibilities of traditional databases.

This is the reason why the Big Data concept has emerged.

In this paper we intend to present the Big Data concept, some of its opportunities and challenges and the European Commission strategy for Big Data. Section 2 presents an overview of Data mining features and Section 3 briefly discuss aspects related to Big Data analytics. Section 4 shows some of the measures and actions taken at European Union level to implement Big Data projects. Section 5 briefly addresses the issue of current challenges. Finally, Section 6 points to some conclusions and future research directions.

2. Big Data

In fact, what does Big Data mean?

Big data is an evolving term that generally refers to a large volume of structured, semi-structured and/or unstructured data whose exploitation is beyond of the abilities of typical database management systems to collect, store, manipulate and analyze.

Moreover, Big Data implies the existence of a set of characteristics known as Big Data Vs. The main and most well-known characteristics are those related to the data *volume, variety*, and *velocity*.

These characteristics were first identified by Doug Laney in a 2001 Gartner report (Laney, 2001). In time, other features have been identified and defined so that the number of Vs has reached 42 for specific cases (Shafer, 2017).

Although the word "BIG" in the name may suggest a large *volume* as the primary attribute of the Big Data concept, the size of the dataset needs to be appreciated by context. There is no minimum data set limit to be used for categorization as Big Data. It is known that there are organizations whose activities involve working with gigabytes or terabytes of data, while others, such as social networks, collect and manipulate petabytes or exabytes of data. But in both cases, there may be requirements for complex data processing and analysis that are specific to Big Data applications (Danubianu & Barila, 2014).

The critical aspect for Big Data is *variety*, generated by associating data having different formats from different sources. Combining this data for further analysis is a great challenge.

Velocity is related to the speed at which information arrives and is analyzed and delivered. There are two ways for data moving through the components of an organization. First, integration and batch loading at predefined moments is currently used in data warehousing. Second, the real-time streaming of data is useful and used in areas such as complex event processing (CEP), text search and analysis, machine learning, etc. In order to assess the velocity needs for Big Data is mandatory to understand the business process and the requirements of end users. There are organizations for which obtaining real-time information is essential, and a delay of few seconds can make the difference between a good decision and a bad one. There are other organizations for which analysis should be done with great accuracy without strong constraints imposed on response times. So, it should be taken into account that for each cases, the correct information needs to be supplied at the right time (Danubianu & Barila, 2014).

IBM ("The Four V's of Big Data", 2019) proposes another feature for Big Data: *veracity*. This refers to deviations, noises and anomalies in data and, in this context, is equivalent to quality.

Starting from these four Vs and based on information we can obtain by a proper analysis of Big Data one can add the fifth V – *value*. Big Data analysis allows organizations to better understand the relationship between its evolution and the events from the environment. This can lead to enhanced decisions and to a stronger competitive position. Figure 1 gives a brief overview of IBM's vision of Big Data features.

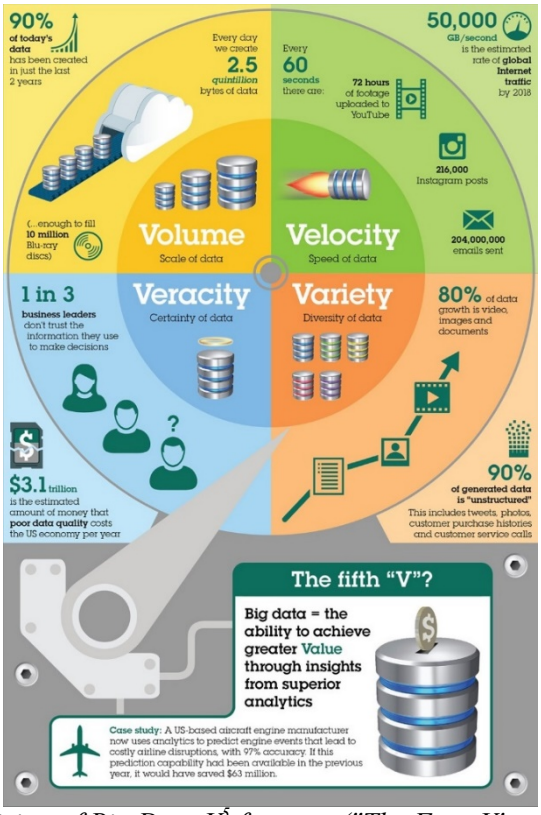

Figure 1. IBM's vision of Big Data V[5] features ("The Four V's of Big Data", 2019)

3. Big Data Analytics

Big Data analysis is a complex process consisting of several phases: data acquisition and capture, data cleaning, data integration, data aggregation, representation and visualization.

Big Data analytics refers analysis strategies aiming to discover models and connections that, otherwise, might be invisible. It helps organizations to harness their data and use it to identify new opportunities, and could provide valuable information to users. There are four types of Big Data analytics:

- descriptive analytics - combines raw data from multiple sources to consolidate past information without explaining why different events or phenomena occur. These analyzes are usually combined with other types of data analysis.
- diagnostic analytics - explains, based on historical data, the reasons or the context in which the events took place, finding dependencies and identifying patterns.
- prescriptive analytics focusing on models that aim to solve specific problems and indicate the actions that should be taken for that. The results are reflected in rules and recommendations for future steps.
- predictive analytics shows what may happen, based on descriptive and diagnostic analysis. It aims to detect trends, groups and exceptions, and to predict future trends being a valuable tool for forecasting.

Full information on a particular topic or scenario can be obtained by combining these types of analysis.

Figure 2 shows the relationship between the various data analytics types for a complete analysis cycle.

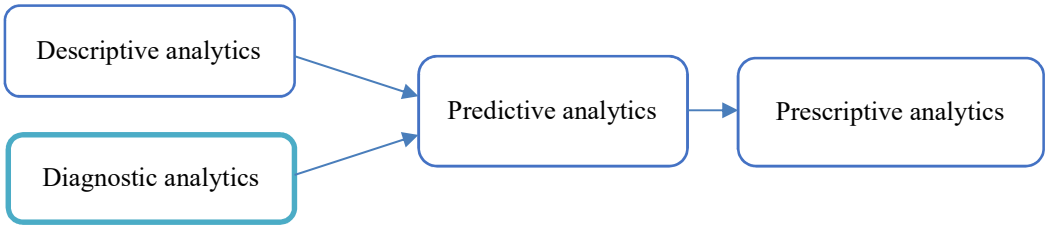

Figure 2. The relationship between Big Data analytics types

4. European vision concerning big data research, adoption and challenges

Like many other domains that benefit from the increased cooperation and synergies enabled by the European Union, Big Data has been identified as a key component for fostering innovation and enabling the data-driven economy. Among the first formal statements to recognize the importance of Big Data in the European context is the European Commission's Communication from 2014 "Towards a thriving data-driven economy" ("Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions", 2019). The Communication is the European Commission's answer to the conclusion of the European Council from October 2013 which was a call for EU action to provide the proper framework conditions for a single market for Big Data and Cloud Computing. The Communication acknowledged the current state-of-play of Big Data technologies and services in the global and European context, the disparities in the uptake speed between Europe and the USA, the great growth potential in multiple fields from health, energy, food security to smart cities and others, and proposed a series of application areas and actions to be considered in order for Europe to bridge the gap and be able to reap the benefits that Big Data technologies are foreseen to enable.

The three application areas that were identified are: *community building, developing framework conditions and regulatory issues*.

In the area of *community building*, the following actions were proposed:

- Creation of a European Public-Private partnership on Data
- Creation of open data incubators
- Development of a skills base
- Development of a Data market monitoring tool
- Identification of sectoral priorities for research and innovation

In the area of developing framework conditions, the following actions were proposed:

- Improving on the availability of data and interoperability (fostering Open Data policies, support innovation in data handling tools and methods, supporting new open standards)
- Enabling infrastructure for a data-driven economy (cloud computing, e-infrastructures, high performance computing, network, broadband, 5G, internet of things, public data infrastructures)

In the area of relevant regulatory issues, of priority were identified actions related to the personal data protection, consumer protection, data mining, security and ownership/transfer of data.

Looking at the current situation, it can be observed that starting from the Digital Single Market [6], one of the nine overarching priorities of the current European Commission, multiple initiatives, platforms and support structures that were devised in order to foster growth and further the creation of value chains related to Big Data have been put in place. All of the application areas and action points identified in the European Commission's Communication have been implemented and many have reached a very advanced level of maturity. Of particular note, having a clear multiplier effect on the growth of the Big Data ecosystem in Europe, have been two of the community building application area action points: the development of a Data market monitoring tool and the creation of a European Public-Private partnership on Data.

The European Data Market Monitoring Tool (http://datalandscape.eu/)

As, in order to improve, you first need to measure where you are, in 2013 the European Commission financed an initiative to assess the current situation of the EU Data Market and Data Economy. The result of the initiative was a European Data Market Study and the European Data Market Monitoring Tool (European Data Market SMART 2013/0063) that was used to gather facts and figures on the size and trends of the EU Data Market and Data Economy. In order to provide a holistic view, the tool measured a number of indicators covering several areas of interest. The following indicators where measured (covering all 28 EU member states and, in addition, for Brazil, Japan and the United states):

- Indicator 1.1 Number of data professionals;
- Indicator 1.2 Employment share of data professionals;
- Indicator 1.3 Intensity share of data professionals;
- Indicator 2.1: Number of data supplier companies;
- Indicator 2.2: Share of data supplier companies;
- Indicator 2.3: Number of data user companies;
- Indicator 2.4: Share of data user companies;
- Indicator 3.1: Revenues of data companies;
- Indicator 3.2: Share of data companies' revenues;
- Indicator 4: Value of the Data Market;
- Indicator 5.1: Value of the Data Economy;
- Indicator 5.2: Incidence of the Data Economy;

• Indicator 6: Data professionals' skills gaps.

Initially the tool was used to assess the period 2013-2015 with a follow-up update covering 2016-2017. The studies provided three future scenarios (for 2020 and 2025, respectively) that would be used to inform decision makers in EU and guide future funding priorities:

• The Baseline scenario
• The High Growth scenario
• The Challenge scenario

The European Public Private Partnership (PPP) on Big Data Value ("Big Data Value Public-Private Partnership - Digital Single Market - European Commission")

The *European Public Private Partnership (PPP) on Big Data Value* is a partnership between the European Commission and the Big Data Value Association-BDVA. The BDVA represents some of the most relevant industry players (including SMEs), researchers and academia, and was created with the aim to establish a functional *D*ata Market and Data Economy in Europe**.** The partnership is funded and implemented through calls for proposals from the 'Leadership in enabling and industrial technologies' area ('Industrial Leadership' pillar), under the Horizon 2020 Research and Innovation programme**.**

The funding priorities are based on the Strategic Research and Innovation Agenda (SRIA) ("European Big Data Value Strategic Research and Innovation Agenda", 2017) which *defines the overall goals, main technical and nontechnical priorities, and a research and innovation roadmap for the European Public Private Partnership (PPP) on Big Data Value.* There are four types of mechanisms to support the implementation of the strategy: **I-Space** – acting as incubators for new businesses and creation of competence, **Lighthouse projects -** to support large scale innovation projects, **Technical projects** - to support specific technical priorities, **Cooperation and coordination projects** to facilitate cooperation and information exchange.

5. Big Data challenges

As we have noted in Section 3, Big Data analysis process goes through many phases, and each of these phases can raise its own challenges. Generally speaking, these challenges can be classified into three major categories: data, process and management challenges (Zicari, 2013).

Data challenges relate to the characteristics of the data itself, such as volume, variety, velocity, veracity, volatility, etc.

The second group of challenges concerns the process. Specifically, there is the question of how data is captured, how it is integrated and transformed and which is the appropriate way to data analyze and to deliver results.

The third category is the management challenges that cover all confidentiality, security, governance and ethical issues ("Big Data Challenges and Opportunity").

The most obvious challenge associated with BIG DATA is the need to store and analyze huge volumes of data. There are reports ("The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things", 2014) which estimates that every two years the volume of generated data is doubling and requires appropriate storage systems. Most of these data are semi-structured or unstructured. They are not found in databases, making it difficult their search and analyze. Solving these problems requires the use of appropriate technologies. For example, converging or hyper-converging infrastructures may be storage solutions. Additionally, technologies such as compression, deduplication and automated tiered storage can reduce the space and costs associated with large volumes of data.

Another problem is the data velocity. Addressing this issue has led to the development of a new generation of data extraction, transformation and load (ETL) tools, as well as to the deploying of analytical tools that considerably reduce the time needed for reporting. There are software with real-time analysis capabilities that allow for an immediate response to environmental events.

Last but not least, the issue of data quality, which is essential in generating information appropriate to decision-making, must be mentioned. In Big Data, due to the fact that data is less structured and comes from multiple sources, data quality may suffer. In this case, it is necessary to apply a data quality control process to develop quality metrics, evaluate data quality, repair erroneous data, and estimate a cost-benefit compromise associated to a better data quality.

Another important challenge for Big Data is its security. In the absence of adequate security mechanisms, sensitive and / or confidential information may be deliberately or not transmitted to third parties. Insufficient data protection can lead to financial losses, damage to a company's reputation, and may determine users' resistance to Big Data adoption (Harvey, 2017). This problem is addressed in most cases by using authentication and access authorization mechanisms, data encryption and data segregation. A real help can be a powerful security management protocol combined with security solutions such as intrusion prevention and detection systems.

With the maturity of Big Data technologies, extensive personal data collection becomes a serious concern for individuals, businesses or governments. For example, sensors, which are an important data source for Big Data, can provide significant amounts of data related to the location and movements of individuals, their health conditions, or their buying preferences. All this can generate major concerns about keeping their privacy. Privacy can sometimes be a hindrance to improving service quality or cost savings, so a balance should be found between using personal data and confidentiality concerns.

Despite the promised benefits, the actual use of Big Data is limited. This is due to the concerns expressed by many managers about the large investments associated with the implementation and use of Big Data analyzes. For many such projects, the definition of problems and objectives is unclear, and the use of emerging technologies leads to increased risk of failure associated with the impossibility of recovering the made investments.

6. Conclusions

It's obvious that we live in Big Data era. Although the opportunities related to the technological evolution and Big Data adoption policies and strategies have allowed important steps in the development of the field, there are still a number of different challenges that require finding the right solutions.

This paper briefly outlined some aspects of the European Union vision of Big Data adoption and, moreover, highlighted the types of challenges still to be solved. We are proposing that in the near future we will conduct a detailed study of possible solutions for these challenges.

References

Big Data Value Public-Private Partnership - Digital Single Market - European Commission. Retrieved from https://ec.europa.eu/digital-single-market/en/big-data-value-public-private-partnership

Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions. (2019). Retrieved from http://ec.europa.eu/newsroom/dae/document.cfm?action=display&doc_id=6210

Danubianu M., Barila A. (2014). Big Data vs. Data Mining for Social Media Analytics, International Conference on Social Media in Academia - Research and Teaching – SMART2014

Digital single market. (2019). Retrieved from https://ec.europa.eu/commission/priorities/digital-single-market_en.

European Big Data Value Strategic Research and Innovation Agenda. (2017). Retrieved from http://www.bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf

Harvey, C. (2017). Big Data Challenges. Retrieved from https://www.datamation.com/big-data/big-data-challenges.html

http://datalandscape.eu/
http://datalandscape.eu/sites/default/files/report/EDM_D2.1_1stReport-FactsFigures_revised_21.03.2018.pdf
https://sites.google.com/a/open-evidence.com/download/repository/SMART20130063_Final%20Report_030417_2.pdf?attredirects=0&d=1
Laney, D. (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, 6.
Shafer, T. (2017). The 42 V's of Big Data and Data Science. Retrieved from https://www.elderresearch.com/blog/42-v-of-big-data
The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. (2014). Retrieved from https://www.emc.com/leadership/digital-universe/2014iview/index.htm
The Four V's of Big Data. (2019). Retrieved from https://www.ibmbigdatahub.com/infographic/four-vs-big-data
Zicari. R.V., (2013). Big Data: Challenges and Opportunities. ODBMS.org; Big Data Challenges and Opportunity. Retrieved from https://www.qubole.com/resources/big-data-challenges

Andrei Marcel Paraschiv (b. June 28, 1982) received his BSc in European Studies (2005), MSc in European Studies (2007) from the Romanian-American University. Currently he is an IT Service Manager and Team Leader in the European Chemicals Agency (ECHA) in Helsinki, Finland. Previously he has been working as the Head of ICT in the Ministry for Research and Innovation in Bucharest, Romania. His current research interests include different aspects of Big Data applied in Project and Service Management.

Mirela Danubianu (b. July 13, 1961) has obtained the B.S. and M.S. degree in Computer Science from University of Craiova in 1985, and the PhD. degree in Computer Science in 2006 from “Stefan cel Mare“ University of Suceava. She has also obtained the B.E. degree in Economics from University of Craiova in 2001. Currently, she is Associate Professor and Head of the Computers Department at “Stefan cel Mare University” of Suceava. She is the author/co-author of 5 books, 7 chapters and more than 100 papers which have been published in journals and presented at different conferences. Her current research interests include databases theory and implementation, modern data architectures, data analytics, application of Data Science in economics, education and healthcare.