BIG DATA, STRUCTURE, CURRENT STATUS AND APPLICATIONS

Ilker Ali, Mr. Fehmi Skender , page 115-128

ABSTRACT

Formerly, Data in Big-Data is compiled from non-traditional sources, such as blogs, social media, emails, sensors, photos, videos, etc. Therefore they are usually unstructured and bulky. However, they have the promise to give enterprises deeper insight into their customers, partners and businesses. Such data can provide answers to questions that were not previously posted. Enterprises must learn to understand how best to use Big Data.

Nowadays big data has become a popular concept and it is interpreted as beginning of a new era. While a huge transformation occurs with the creation of big data concept, institution and organizations point of view and benefits gained from the data have changed and come to a different point. In this study, big data is conceptually analyzed and compared with structured and unstructured data. The usage areas of big data in the world and examples were given.

Key words: Big Data, Big Data Definitions, Big Data Development Model.

Mr. Ilker Ali PhD.
Faculty of Informatics,
International Vision University,
Gostivar - North Macedonia;
e-mail:
aybeyan@vizyon.edu.mk

Mr. Fehmi Skender PhD
Uluslararası Vizyon Üniversitesi
Gostivar / Kuzey Makedonya

e-mail:
Fehmi.skender@vizyon.edu.mk

UDK: 004.77:004.6

Date of received:
21.01.2019
Date of acceptance:
22.02.2019

Declaration of interest:
The authors reported no conflict of interest related to this article.

1 INTRODUCTION

In the global economy, almost all major organizations have begun to rely on feedback from their customers, business operations and ultimately the organization's internal processes to open up new opportunities for sustainable economic growth. In the process of detecting these observations, massive data sets are generated that are to be managed and manipulated by highly qualified data professionals.

Big Data is now a popular topic and is used to represent a huge volume of unstructured and structured data that is difficult to deal with just a relational database and techniques of traditional analysis to create a Big Data Analysis. In the most common enterprise scenarios, the data is too massive. Big Data has great potential to help organizations improve their operations and make better decisions (Jawell, 2014).

Today in the world online networking has turned into a basic flow of communication in everyday life of individuals. This flow of communication gives a gigantic measure of information called Big Data.

Big Data can encourage incredible pieces of knowledge; there may be a capacity to make sense as the main driver of issues and disappointments and further inconvenient behavior that affects the earnings of businesses to those.

Big Data allow connections to be found regarding the decision-making of business patterns, the nature of research, the legitimate references for connection, and controlling the simultaneous traffic conditions of internet traffic.

Working with Big Data has different means. It differs by relying on the abilities of the gathering that deal with the set and considering the applications they use. Big Data can deal with information packages that can hardly deal with conventional databases.

Some Of The Advantages Of Using Big Data In Marketing:

• Define the root causes of disasters, flaws in almost real time, which can save billions of dollars a year.

• Organizing a campaign in order to offer better services on available and past purchases to the buyer.

• Recalculate the target portfolios of risks in no more than a few minutes.

• Detection of customers that are important

Data

The concept on which our study is based is *data*. Therefore, some definitions of data should be considered here. Information, especially facts or numbers, collected to be examined and considered and used to help with making decisions (Dictionary, 2019).

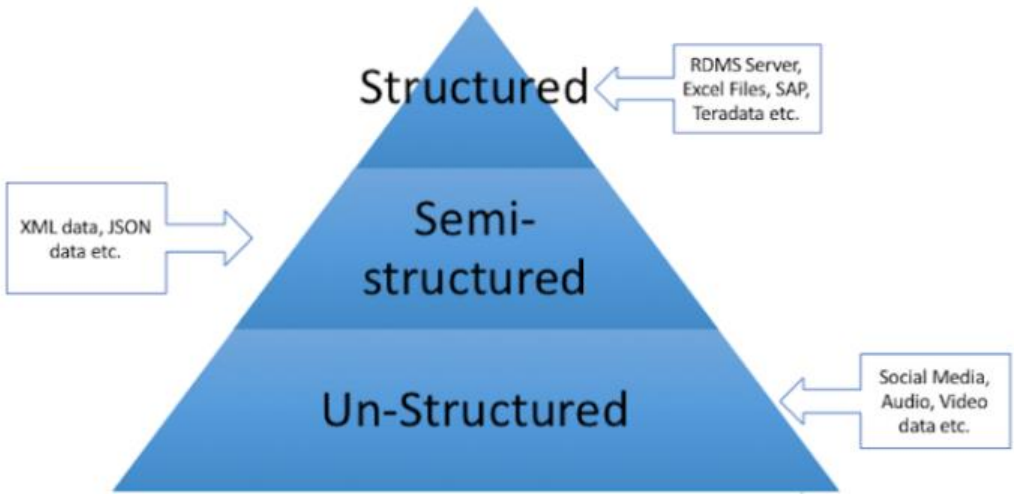
Data can be classified into groups as follows:
• Structured, Unstructured, Semi-structured
• Static, dynamic, flowing
• Secure / open, private / public
•Paid free
• Open government data
• Open data
• Big data.....

Analyzing data can be structured or unstructured

Structured data: The data inside the changed constraints, records, or documents are known as organized information. Safe from the way in which organized information - even in large quantities - can be entered, separated, challenged and terminated in a basic and clear way, this information is best served by the traditional database.

Unstructured data: Data coming from different sources, for example, emails, text documents, videos, photos, etc.

According to some estimates that by 2020, the digital universe will contain more than 40 zettabytes of data. That's 40,000,000,000,000,000,000,000. They also estimate that 90 percent of what we call "Big Data" is unstructured data. And this big data can be handled with the help of applications like Hadoop.



Semi-structured: Data somewhere between structured and unstructured data such as It is not organized in a complex manner that makes sophisticated access and analysis possible; however, it may have information associated with it, such as metadata tagging, that allows elements contained to be addressed. And we can understand more clearly about semi-structured Here's an example: A Word document is generally considered to be unstructured data. However, you can add metadata tags in the form of keywords and other metadata that represent the document content and make it easier for that document to be found when people search for those terms — the data is now semi-structured. (Wigmore, 2019)

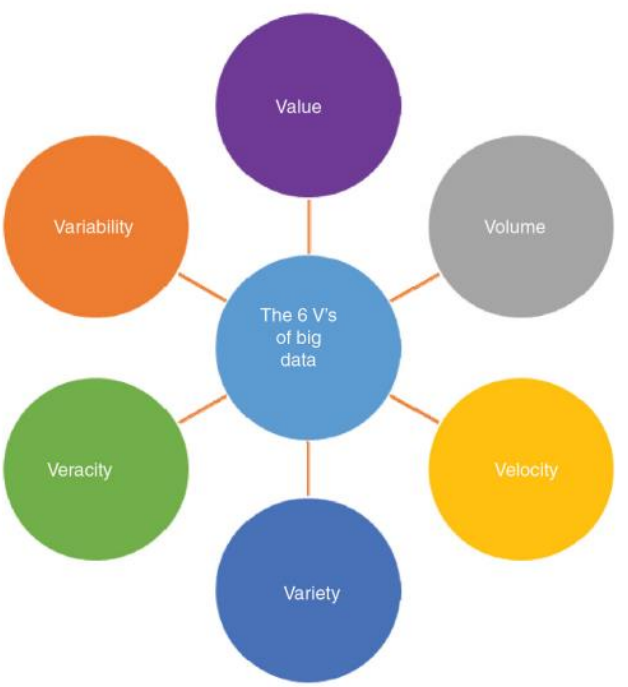
Getting data from unstructured data through data mining. Data mining. Based on the semantic structure of Natural Languages; research technique aimed at finding and uncovering " Data mining can also be defined as the discovery of information from data. In data mining, it is aimed to extract information from large amounts of data by using automated and partial automated methods. uses algorithms from various disciplines such as statistics, artificial intelligence, computer science. It combines all the technologies that can analyze database information to find useful information in the data and to find possible, meaningful and useful relationships (Monino, 2016).

WHAT IS BIG DATA - BIG DATA?

The term Big Data is used and defined first by (Laney, The Importance of 'Big-Data': A Definition, 2012). Since then, the number of articles and reports has appeared on this topic. Some of them are included in the references.

When the database becomes so large that it is not possible to analyze, process, and visualize this data set using simple database tools then the database becomes Big Data. (Kumar, 2016)

Big data can be characterized by the following 6V data: volume, value, velocity, variety, veracity and variability (Ristevski, 2018)



Database vs. big data

Today, the social world generates huge data collections. As a result, the Big-Data analysis has become an important device for businesses that hope to exploit large-scale valuable data on benefits and competitive advantage. While the Big Data lives up to a great deal of excitement, there are certain situations where workloads on a traditional database can be a better arrangement (Dijks, 2013).

Will the implementation of Big Data be economical?

Economical efficiency is always a concern for companies that hope to adopt new technologies. When considering the implementation of Big Data, companies need to do their job to ensure that the benefits derived from the deployment of Big Data exceed the costs. All the things that are being considered, Big Data have a number of things that draw out all the stops to make the implementation more financially exhausted by the companies they can realize. One thing, Big-data saves money by plugging open source with ware servers.

Why big data is needed

Companies and industries are increasingly aware that data analysis is increasingly becoming a vital factor to be competitive, to discover new insights and personalize services (Ahmed Oussous, 2017).

The organization needs a Big Data and Analysis strategy for three reasons (Waddell, 2014):

To create smarter, independent organizations

Today, the number of Google searches for housing and real estate, starting from a quarter, and then to the next, ends with a prediction more precisely.

To equip the organization

As most organizations agree, it is essentially unrealistic to conduct the conversations they once had with the customers. There are many dialogues ranging from different sources.

Application of big data and some examples

There are four areas of application for Big Data according to (McGuire, 2012):

- **As organizations create more transactional data.**

They can gather more detailed information about the performance of everything, from the inventory of objects to frail days, and therefore exhibit variability and performance support. In fact, some leading companies use their ability to collect and analyze Large-scale data for direct controlled experiments in order to make better management decisions. Големи-податоци поддржува потесна сегментација на клиенти.

- **Big Data analysis can improve predictions, minimize risks**.
- For example, firms use the information obtained from machine sensors embedded in products to create an innovative maintenance process. Many people consider Big Data as an extraordinary trendy expression. (White, 2012).

Google MapReduce

MapReduce is a programming model as well as the application associated with the processing and creation of large data sets (Dean and Ghemawat 1). Google has used the MapReduce programming model for many different purposes. Google MapReduce links its success to a variety of reasons. First, the use of this model is easy for programmers who do not have parallel and distributed system experience. Second, a wide variety of problems, such as MapReduce calculations can be expressed easily. For example, MapReduce can be used to generate data, for Google's search engine service, for ranking, for data mining, for machine learning, and for many other systems. Third, Google has developed a scalable application of MapReduce to large machine clusters that contain thousands of machines. This application enables the effective use of machine resources and is therefore suitable for large numerical problems encountered on Google (Dean Jeffrey, 2016)

HADOOP

Apache has developed a Hadoop library that is 100% open source and develops a fundamentally new way of processing data. Hadoop lists parallel processing of large amounts of data and can scale unlimited.

Hadoop is Google MapReduce's biggest competitor. It is very popular today in processing big data and has become the symbol of the change that comes with big data. It divides large data into smaller clusters and shares them to other machines. It assumes that the data is not clean and organized, that is, the data is too large to be cleaned before processing. While basic data analysis requires a process called ETL to move the data to the place where it will be analyzed, Hadoop recognizes that the amount of data instead of this process is very large, so it cannot be moved and should be analyzed where it is. The outputs of Hadoop are not as precise as relational databases. But it is much faster than relational databases in many areas where precise answers are not required. That only 5% of all

digital data is structural, it is useful to use these programs to use the remaining 95% non-structural data, such as web pages and videos. Today, Hadoop has become indispensable for many institutions and organizations that develop and use big data technology.

Some examples of Big Data

Giving examples of some of the successful applications of big data in different areas of the world can make the subject easier to understand.

• **Sentimental motives: rules of addiction**

They introduce a new paradigm of the analysis of sensitive concept levels that combines semantics, practical skills for knowledge judgment and machine learning to improve the accuracy of tasks, for example, polarity detection. By allowing feelings to penetrate from concept to concept based on the dependence of the info-sentence. (Poria, 2014)

- **Walmart**

In 2004, Walmart and Teradata's digital analysts examined huge databases including data on which customers bought which product, total costs, what else was available in shopping carts, times of day and even situations. In conducting this review, the company noticed that not only flashlight sales were increasing before a hurricane but also sales of Pop-Tarts, a sweet American flakes. Afterwards, it increased sales significantly by storing Pop-Tarts next to hurricane supplies at the front of the store for customers quickly entering and exiting. In the past, a central employee had to be born before data collection and testing of ideas, but since today Walmart has such data and better tools, it has been able to generate correlations much faster and inexpensively and use them in company operations, and today it (Schönberger Viktor Mayer, 2013).

- **Time Corporation: epochs, opinions and changes**

(Popescu, 2014) suggest to explore diachronic phenomena using large churches of chronologically ordered languages.

- **Analysis of future communities**

(Jung, 2014) see the research group as a social network, where communication occurs through academic work.

• **Meta-level Sentiment Models**

As different dimensions of feelings, for example, subjectivity, polarity, intensity, and emotions, complement each other in specific scenarios (Bravo-Marquez, 2014).

• PoliTwi: Early detection of new political topics

(Rill, 2014) is a system designed to discover new political points in Twitter earlier than other standard information channels. In addition, the authors checked their determination through Google Trends to notice that the themes appeared earlier in Twitter.

• **Extract relevant knowledge to detect sarcasm**

A system is needed that can manage some sort of knowledge to interpret the emotional language used. The results of the paper show that the task of detecting sarcasm is beneficial by incorporating etymological and semantic sources of information (Justo, 2014).

• **Big Data analysis of news and social content**

The analysis of media content is central in the social sciences. This process provides opportunities for conducting mass surveys, real-time monitoring and modeling the system-wide level of the global media system. This study describes how the analysis of Twitter content can reveal changes in the disposition throughout the population, as political relations between US leaders can be extracted from the huge data name (Flaounas, 2012).

- **World Cup 2014 Brazil**

As is known, in 2014, FIFA World Cup champion Germany (Brazilian Federal Government). In the tournament, SAP and the German Football Federation (DFB) collaborated in an innovative way to transform big data into smart decisions to improve player performance in the cup. Working on the SAP HANA platform, this solution is designed to facilitate analysis of training, preparations and tournaments, and to improve player-team performance. Oliver Bierhoff states that 10 players in 10 minutes produce data from more than 7 million data points. With this solution, the data of this magnitude could be analyzed in the training and preparation of the next match (SAP SE). These analyzes played a major role in bringing the trophy to Germany. This solution, which has a share in the success of the

football world, is an important development both in terms of the diversity and development of the fields of use of big data and in terms of the sports world. In this example, it is possible to evaluate the importance of the data collected in the match to the transformation of information and to obtain important information by using this information in decision processes.

• Big data analysis using a Naive Bayes classifier

A typical method of obtaining valuable information is to get an attitude or a conjecture from a message. Therefore, machine learning technologies have the ability to learn from a set of training data to anticipate or strengthen the decision-making process with relatively high accuracy (Liu, 2013).

• Distributed Analysis of Real Life

The Big Data trend has forced remote data systems to have constant fast data (Rahnama, 2014). In recent years, real-time analysis of current data has been established in a new field of research, which aims to answer questions about what is happening - now with a negligible delay. The real challenge with real-time data processing is that it is impossible to store data cases and therefore use online analytical algorithms. To perform real-time analytics, pre-processing data should be executed in such a way that only a brief summary of the stream is stored in the main memory. In addition, due to rapid arrival, the average processing time for each data instance should be in such a way that the approximation of the cases is not lost without being trapped. Lastly, it should provide high analytical measures for accuracy. Sentinel is a distributed system written in Java that aims to solve this challenge by implementing the process of learning and processing in a distributed structure. Sentinel is based on the Apache Storm's top, a distributed figurative platform. Sentinel also uses storage space to keep the data flow summary and saves the summary in the data collection structure.

CONCLUSION

The importance of big data is increasing day by day. Although the works and services of various states, institutions and organizations are available on the internet, they have not been evaluated in a good way. States, institutions and organizations have realized that by processing large data they can provide great benefits for themselves. Therefore, this issue is given great importance today. If institutions and organizations do not pay attention to this technology, it is impossible for them to progress. According to the results of the research, the following things can be listed as follows:

Universities and other science-related institutions and organizations need to support the development of technology and applications related to big data. For those who want to evaluate big data in line with their professions, the necessary environment should be provided for them to receive the necessary training and this issue should be included in the training programs. Practical seminars and in-service trainings on the subject of big data and tools should be organized. In the trainings to be provided, the results obtained from the processing of big data should be put forward in a concrete manner, and thus, the people who will receive training should be able to realize the importance of the subject more realistically.

BIBLIOGRAPHY

Ahmed Oussous, F.-Z. B. (2017). Big Data Technologies: A Survey. *Journal of King Saud University - Computer and Information Sciences*.

Apache. (н.д.). *Welcome to Apache Hadoop*. Повратено од http://hadoop.apache.org/

Bravo-Marquez, F. M. (2014). Meta-level sentiment models for big social data analysis, doi:10.1016/j.knosys.2014.05.016, Volume 69. *Knowledge-Based Systems*, 89-99.

Collins, E. (2014). Big-Data in the Public Cloud. *Cloud Computing, IEEE (Volume: 1, Issue: 2), ISSN: 2325-6095*, 13-15.

Dean Jeffrey, S. G. (2016). MapReduce: Simplified Data Processing on Large Clusters. *Google, Inc. Web*.

Dictionary, C. A. (2019). *cambridge.org*. Повратено од dictionary.cambridge.org: https://dictionary.cambridge.org/dictionary/english/data

Dijks, J. (2013). *Oracle Big-Data for Enterprise.* Повратено од http://www.oracle.com/us/products/database/Големи-податоци-for-enterprise-51913.pdf

FigureEight. (2018). *Figure Eight, Airline Twitter Sentiment*. Повратено од https://www.figure-eight.com/data-for-everyone/

Flaounas, I. S.-W. (2012). *Big-Data Analysis of News and Social Media Content.* Повратено од http://www.see-a-pattern.org/sites/default/files/Big%20Data%20Analysis%20of%20News%20and%20Social%20Media%20Content.pdf

Hortonworks, & Community, H. (2015). *Hortonworks Community, Open Enterprise Hadoop*. Повратено од http://hortonworks.com

Jawell, D. R. (2014). Performance and Capacity Implications for Big Data. *Retrieved from: http://www.redbooks.ibm.com/redpapers/pdfs/redp5070.pdf*.

Jung, S. a. (2014). Knowledge-Based Systems. *Article:Analyzing future communities in growing citation networks, doi:10.1016/j.knosys.2014.04.022, Volume 69*, 34-44.

Justo, R. T. (2014). Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web, doi:10.1016/j.knosys.2014.05.021. *Knowledge-Based Systems*, 124-133.

Kumar, A. (2016). A Big Data MapReduce Framework for Fault Diagnosis in Cloud-based Manufacturing. *Loughborough University Institutional Repository*.

Laney, D. (2001). Повратено од 3D Data Management: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Laney, D. (2012). *The Importance of 'Big-Data': A Definition.* Повратено од http://www.gartner.com/resId=2057415

Liu, B. E. (2013). Scalable sentiment classification for Big-Data analysis using Naïve Bayes Classifier. *Big-Data, 2013 IEEE International Conference on, INSPEC Accession Number: 13999322, Conference Location: Silicon Valley, DOI:10.1016/BigData.2013.6691740, Publisher: IEEE*, 99-104.

M. Dhavapriya, N. Y. (2016). Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table. *International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 1*.

McGuire, T. J. (2012). *Why Big-Data is the new competitive advantage*. Повратено од http://iveybusinessjournal.com/ competitive-advantage/

Monino, J.-L. (2016). Big Data, Open Data and Data Development. *3. London: ISTE Ltd*.

Namrata Singh, S. A. (2016). A Performance Analysis of High-Level MapReduce Query Languages in Big Data. *Springer Science+Business Media Singapore, Proceedings of the International Congress*.

Oracle. (2018). *Oracle Virtual Box*. Повратено од https://www.virtualbox.org/wiki/Downloads

Popescu, O. a. (2014). Knowledge-Based Systems. *Timecorpora: Epochs, opinions and changes, Volume 69, doi:10.1016/j.knosys.2014.04.029*, 3-13.

Poria, S. E. (2014). Sentic patterns: Dependency-based rules for concept-level sentiment analysis, j.knosys.2014.05.005. *Knowledge-Based Systems*, 45-63.

Rahnama, A. H. (2014). Distributed Real-Time Sentiment Analysis for Big-Data Social Streams. *Control, Decision and Information Technologies (CoDIT)*.

Rill, S. D. (2014). Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, doi:10.1016/j.knosys.2014.04.022, Volume 69. *PoliTwi*, 23-34.

Ristevski, B. (2018). Big Data Analytics in Medicine and Healthcare. *Journal of Integrative Bioinformatics*. doi:https://doi.org/10.1515/jib-2017-0030

Rupinder Singh, P. J. (2016). Analyzing performance of Apache Tez and MapReduce with hadoop multinode cluster on Amazon cloud.

Schönberger Viktor Mayer, K. C. (2013). *Big Data - A Revolution That Will Transform Our Way of Living, Working and Thinking.*

Waddell, T. (2014). *3 Reasons You Need a Big-Data and Analytics Strategy*. Повратено од http://blogs.adobe.com/Big-Data-analytics-strategy/

White, C. (2012). *What Is Big-Data and Why Do We Need It?* Повратено од http://www.technologytransfer.eu/article/98/2012/1/What_Is_Big_Data_and_Why_Do_We_Need_It_.html

Wigmore, I. (2019). Повратено од WhatIs.com: https://whatis.techtarget.com/definition/structured-data