BIG DATA FROM A TO Z

Elham Nazari[1], Marziyeh Afkanpour[1], Hamed Tabesh[1*]

[1]Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.

Article Info

Article type:
Short Communication

Article History:
Received: 2019-08-28
Revised: -
Accepted: 2019-09-20

**** Corresponding author:***
Hamed Tabesh
Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
Email: tabeshh@mums.ac.ir

A B S T R A C T

The rapid development of technology over the past 20 years has led to explosive data growth in various industries, including defense industries, healthcare. The analysis of generated Big Data has recently been addressed by many researchers, because today's Big Data analysis are one of the most important and most profitable areas of development in Data Science and companies that are able to extract valuable knowledge among the massive amount of data at logical time can earn significant advantages . Accordingly, in this survey, we investigate definition of the Big Data and the data sources. Also look at advantages, challenges, applications, analysis and platforms used in the Big Data.

Keywords:
Big Data, Data Mining, Challenge, Advantage, Application, Platform, Analysis

How to cite this paper
Nazari E, Afkanpour M, Tabesh H. Big Data from A to Z. Front Health Inform. 2019; 8(1): e20. DOI: 10.30699/fhi.v8i1.202

INTRODUCTION

Over the past 20 years, with the development of the Internet and advent of technology, the amount of data collected and stored digitally in a large volume is rapidly increasing in all industries [1, 2]. These data are known as Big Data [3, 4]. Big Data analysis refers to tools and methodologies that aim to convert a large amount of raw data into data about data for analysis purposes [5]. Analysis of this type of data has many benefits, such as cost reduction, information sharing, organizational competition, etc. Therefore it has become a hot topic that attracted the attention of many academics, researcher and governments [6]. Nowadays, Big Data analyzes have become one of the most important and profitable areas of development in Data Science. Management of this type of data is in the process of developing until able to extract useful information at the right time and applying available knowledge in the data to their purposes [7, 8]. Thereupon, due to the inevitable growth of data and the importance of Big Data analyzes, this survey peruse definition of the Big Data, its advantages, its applications, its challenges, its architecture and its platforms.

Big Data Definition

Big Data refers to data:

- that cannot be analyzed by old software due to the complexity and high volume [2, 9, 10].
- with volume, variety, and velocity properties [11, 12].
- of an Exabyte size (1018 bytes) or more [13].

Big Data Characteristics

The Big Data is defined by some characteristics; these characteristics are known as vs., which were initially identified with three attributes, and these features are increasing over time [14, 15].

1- Volume: Refers to the production of high-volume data.

2-Velocity: The data production rate is unpredictable.

3-Variety: It relates to the diversity of data and its various formats.

4-Veracity: It refers to bias, noise and abnormality in large data.

5-Viability: combine of the related information until a variety of predictions to be made in the future.

6-value: The descriptive feature of such massive data.

7-Viscosity: Refers to stability and resistance in Big Data flow.

8-Visualization: Refers to how present data to the user [16].

Moreover some studies also comment on other properties such as bellow:

- Validity: Correctness or accuracy of data used
- Volatility: Duration of Usefulness to the user
- Virality: Spreading Speed (rate at which the data is broadcast /spread by a user and received by

different users for their use)

- Variability: Data Differentiation
- Venue: Different Platform like personnel system and private & public cloud
- Vocabulary: Data Terminology likes data model and data structures
- Vagueness: concern the reality in information
- Verbosity: The redundancy of the information available at different sources
- Voluntariness: The will full availability of Big Data to be used according to the context
- Versatility: The ability of Big Data to be flexible enough to be used differently for different context [17, 18].

Big Data advantages

advantages of Big Data generally include better aimed marketing, more straight business insights, recognition of sales and market chances, automated decision making, definitions of customer behaviors, better planning and forecasting and identification consumer behavior [14].

Big Data applications

Big Data analysis generally is applied in Astronomy, atmospheric science, Genomics, Biogeochemical, biological science, physics, medical records, scientific research, natural disaster and resource management, military surveillance, financial services, social networks, web logs, Photography, search indexing, RFID(Radio-frequency identification), mobile phone, IOT(Internet Of Things), sensor network, education, transportation and telecommunication fields. [14, 19, 3].

Big Data Sources

These data are generated from online transactions, emails, videos, audio, images, click streams, logs, posts, search queries, sensors, mobile phones, and applications. These data are stored in databases and grow into massive volumes [3].

Big Data analysis

The steps to obtain valuable values from Big Data are as follows:

- Acquisition
- Information extraction and cleaning
- data integration
- modeling and analysis

And interpretation and deployment [20].
Some sources include the following stages:

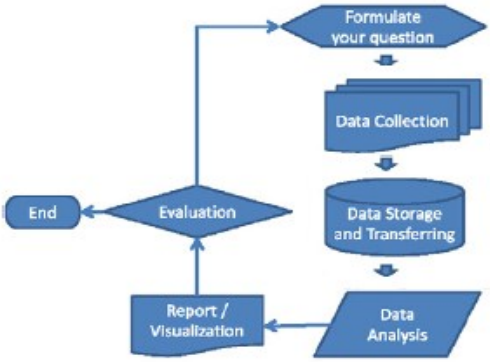

Fig 1: Stages of Big Data analysis [21]

In the Big Data analysis, the following techniques are usually used:

- Regression
- Correlation
- Classification
- Cluster analysis
- Factor analysis
- Statistical learning
- Data mining
- C4.5
- Association analysis
- K-means
- SVM(Support Vector Machine)
- Apriori
- EM(Expectation-Maximization)
- Naïve Bayes
- Cart and so on [3].

Big Data platform

Hadoop is the most common platform for storing and analyzing of Big Data in view of its scalability characteristics. The main components of the Hadoop platform are:

1) The Hadoop Distributed File System (HDFS), which is used to store data between clusters of systems.
2) The resource management layer, YARN (Yet another Resource Negotiator) is the new model of distributed work and put jobs among the cluster.
3) Map Reduce is a distributed programming and processing model of Big Data [7].
4) Common libraries used in different parts of the Hadoop that are also Blur, Solar: Warehouse documents

- Hbase: NOSQL database with random access
- Cassandra: Key-value storage
- Giraph: Graph based database

- AMBARI: Manage and monitor a Hadoop cluster

Oozie: A workflow scheduler for managing complex mu used elsewhere [22, 23].

Some of the important tools of Hadoop are listed in the following:

- AVRO: Serialization of information
- Hive: Data interaction
- Ltiparty tasks of Hadoop.
- Pig: High-level data streaming language for data processing
- Mahout: A set of scalable machine learning algorithms that runs on the Hadoop.

In Table1 can be see Mahout Map Reduce Algorithm [22-26]. In the Table2, we introduce and comparing the Hadoop, Spark and Flink platforms. [7, 27-29].

Table 1: Mahout map reduce algorithm

Mahout Algorithm
k-means clustering / fuzzy k-means
Latent Dirichlet Assignment
Singular Value Decomposition
Logistics- regression- based classifier
Complementary naïve Bayes classifier
Random forest decision tree –based classifier
Collaborative filtering

Table 2: Differences between platforms

The differences	Hadoop	Spark	Flink
Processing method	Batch processing	stream base	Stream based - Batch processing
Speed	Slow in complex analysis, weak in interactive and online computing	The higher the speed, especially in the Iterative and Online processes	
Fault Tolerance	High	Recovering Missing Data Sections - High	Very High
Flexibility	No	Yes	No
Supports a variety of data models	No	Yes	No
Cashes data set in memory to reduce latency	No	Yes	No
Simplicity	Yes	No	No
Programming language	Java	R, Java, Python, Scala	Java
Others	Variable share-custom-partition-local memory	User code optimization	User code optimization

Considering the advantages of spark, Mllib is introduced: MLlib, a Machine learning tool that is Used for Spark.In Table3 can be see MLLIB Algorithm [25].

Table 3: MLLIB algorithm

MLLIB Algorithm
Linear SVM and Logistic Regression
Classification and Regression Tree
k-means clustering
Suggested through squares at least periodically
Simple polynomial Bayesians
Basic statistics
Feature extraction and conversion
Dimension reduction

Big Data challenges

There is not enough knowledge about which data to use for the purpose. There is not appropriate IT infrastructure. Also, there is no enough knowledge about which algorithm is pertinent and what tools are be fitting for analysis.

Another challenge is the high diversity of data and scalability. Missing data and statistical uncertainty and fuzziness are another challenge. The issue of security, privacy and trust is another problem. Also cost is another challenge. The low quality of these data affects analyzes. [11, 20, 14, 30-32].

CONCLUSION

Today, with the growing data production in all industries, Big Data analysis have been considered. These analyzes have numerous applications in traffic management, astronomy and so on. At the same time, there exists many challenges such as the lack of data with proper quality and unaware use of the appropriate method and platform that should be considered. In view of the specific features of this type of data, it is suggested that future studies explore methods, tools, and suitable platforms. Also,

discover more challenges that these analysis confront to them and then to examine. Finally, to take advantage of the capabilities of these analyzes, provide solutions to the challenges.

AUTHOR'S CONTRIBUTION

All the authors approved the final version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE

No financial interests related to the material of this.

REFERENCES

1. Nambiar R, Bhardwaj R, Sethi A, Vargheese R. A look at challenges and opportunities of Big Data analytics in healthcare. International Conference on Big Data. IEEE; 2013.
2. Sagiroglu S, Sinanc D. Big Data: A review. International Conference on Collaboration Technologies and Systems. IEEE; 2013.
3. Chen M, Mao S, Liu Y. Big Data: A survey. Mobile Networks and Applications. 2014; 19(2): 171-209.
4. Murdoch TB, Detsky AS. The inevitable application of Big Data to health care. JAMA. 2013; 309(13): 1351-2. PMID: 23549579 DOI: 10.1001/jama.2013.393 [PubMed]
5. Mayer-Schönberger V, Cukier K. Big Data for development: Challenges & opportunities. Houghton Mifflin Harcourt; 2013.
6. Duggal PS, Paul S. Big Data analysis: Challenges and solutions. International Conference on Cloud, Big Data and Trust. 2013.
7. Ramírez-Gallego S, Fernández A, García S, Chen M, Herrera F. Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce. Information Fusion. 2018; 42: 51-61.
8. Jin X, Wah BW, Cheng X, Wang Y. Significance and challenges of Big Data research. Big Data Research. 2015; 2(2): 59-64.
9. Liyanage H, deLusignan S, Liaw S, Kuziemsky C, Mold F, Krause P, et al. Big Data usage patterns in the health care domain: A use case driven approach applied to the assessment of vaccination benefits and risks. Yearb Med Inform. 2014; 9: 27-35. PMID: 25123718 DOI: 10.15265/IY-2014-0016 [PubMed]
10. Emani CK, Cullot N, Nicolle C. Understandable Big Data: A survey. Computer Science Review. 2015; 17: 70-81.
11. Mehta N, Pandit A. Concurrence of Big Data analytics and healthcare: A systematic review. Int J Med Inform. 2018; 114: 57-65. PMID: 29673604 DOI: 10.1016/j.ijmedinf.2018.03.013 [PubMed]
12. deMauro A, Greco M, Grimaldi M. What is Big Data? A consensual definition and a review of key research topics. AIP Conference Proceedings. AIP; 2015.
13. O'Driscoll A, Daugelaite J, Sleator RD. Big Data, Hadoop and cloud computing in genomics. J Biomed Inform. 2013; 46(5): 774-81. PMID: 23872175 DOI: 10.1016/j.jbi.2013.07.001 [PubMed]
14. Zhang Q, Yang LT, Chen Z, Li P. A survey on deep learning for Big Data. Information Fusion. 2018; 42: 146-57.
15. Bello-Orgaz G, Jung JJ, Camacho D. Social Big Data: Recent achievements and new challenges. Information Fusion. 2016; 28: 45-59.
16. Manogaran G, Lopez D, Thota C, Abbas KM, Pyne S, Sundarasekar R. Big Data analytics in healthcare Internet of things. In: Qudrat-Ullah H, Tsasis P. (eds) Innovative healthcare systems for the 21st century. Understanding Complex Systems. Springer, Cham; 2017.
17. Arockia Panimalar S, Varnekha Shree S, Veneshia Kathrine A. The 17 V's of Big Data. International Research Journal of Engineering and Technology. 2017; 4(9): 329-33.
18. Shafer T. The 42 V's of Big Data and data science [Internet]. 2017 [cited: 1 Jul 2019]. Available from: https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html.
19. Archenaa J, Anita EM. A survey of Big Data analytics in healthcare and government. Procedia Computer Science. 2015; 50: 408-13.
20. Jagadish H, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, et al. Big Data and its technical challenges. Communications of the ACM. 2014; 57(7): 86-94.
21. Huang T, Lan L, Fang X, An P, Min J, Wang F. Promises and challenges of Big Data computing in health sciences. Big Data Research. 2015; 2(1): 2-11.
22. Sinha S. What is a Hadoop ecosystem? [Internet]. 2017 [cited: 1 Jul 2019]. Available from: https://www.quora.com/What-is-a-Hadoop-ecosystem.
23. Dean J, Ghemawat S. Map reduce: Simplified data processing on large clusters. Communications of the ACM. 1958; 51(1): 107-13.
24. Raghupathi W, Raghupathi V. Big Data analytics in healthcare: Promise and potential. Health Inf Sci Syst. 2014; 2: 3. PMID: 25825667 DOI: 10.1186/2047-2501-2-3 [PubMed]
25. Sitto K, Presser M. Field guide to Hadoop: An introduction to Hadoop, its ecosystem, and aligned technologies. O'Reilly Media Inc.; 2015.
26. Kumar VN, Shindgikar P. Modern Big Data processing with Hadoop: Expert techniques for architecting end-to-end Big Data solutions to get valuable insights. Packt Publishing; 2018.
27. Mangai UG, Samanta S, Das S, Chowdhury PR. A survey of decision fusion and feature fusion strategies for pattern classification. IETE Technical Review. 2010; 27(4): 293-307.

28. Ponti MPJ. Combining classifiers: From the creation of ensembles to the decision fusion. 24th Conference on Graphics, Patterns and Images Tutorials. IEEE; 2011.

29. Ferranti A, Marcelloni F, Segatori A, Antonelli M, Ducange P. A distributed approach to multi-objective evolutionary generation of fuzzy rule-based classifiers from Big Data. Information Sciences. 2017; 415: 319-40.

30. Fan J, Han F, and Liu H. Challenges of Big Data analysis. National Science Review. 2014; 1(2): 293-314.

31. Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. J Integr Bioinform. 2018;15(3): 1-5. PMID: 29746254 DOI: 10.1515/jib-2017-0030 [PubMed]

32. Bossé É, Solaiman B. Information fusion and analytics for Big Data and IoT. Artech House; 2016.