**BIG DATA FROM A TO Z**

Elham Nazari¹, Marziyeh Afkanpour¹, Hamed Tabesh¹[16]

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

---

**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

**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 below:

- 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

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 [17, 18].

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
• Ltparty 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

<table>
<thead>
<tr>
<th>Mahout Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means clustering / fuzzy k-means</td>
</tr>
<tr>
<td>Latent Dirichlet Assignment</td>
</tr>
<tr>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>Logistics- regression- based classifier</td>
</tr>
<tr>
<td>Complementary naïve Bayes classifier</td>
</tr>
<tr>
<td>Random forest decision tree –based classifier</td>
</tr>
<tr>
<td>Collaborative filtering</td>
</tr>
</tbody>
</table>

Table 2: Differences between platforms

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

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

<table>
<thead>
<tr>
<th>MLLIB Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear SVM and Logistic Regression</td>
</tr>
<tr>
<td>Classification and Regression Tree</td>
</tr>
<tr>
<td>k-means clustering</td>
</tr>
<tr>
<td>Suggested through squares at least periodically</td>
</tr>
<tr>
<td>Simple polynomial Bayesians</td>
</tr>
<tr>
<td>Basic statistics</td>
</tr>
<tr>
<td>Feature extraction and conversion</td>
</tr>
<tr>
<td>Dimension reduction</td>
</tr>
</tbody>
</table>

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**


