International Journal of Research in Engineering, Science and Management
Volume-1, Issue-11, November-2018
www.ijresm.com | ISSN (Online): 2581-5792

Big Data Techniques: Today and Tomorrow

Priyanka Khatana[1], Yashvardhan Soni[2]
[1]*M. Tech. Student, Dept. of Information Technology, Dronacharya College of Engineering, Gurugram, India*
[2]*Professor and HoD, Dept. of Information Technology, Dronacharya College of Engineering, Gurugram, India*

***Abstract*: This paper is written before doing any practical work on the above topic. It is based on the research done on the topic- "Big Data Techniques: Today and tomorrow". Big data refers to a large dataset which is not easy to handle or understand. We have given a brief description about the present scenario and the future scenario of big data. We are going to explain what big data is, how it came into existence, a brief about the technologies involved in data extraction. By the end of the paper we will get a basic knowledge about big data and its benefits and its use in our daily life also. This paper aims to analyze some of the different tools and techniques which can be applied to big data.**

***Keywords*: Organization products manufacturing, processing**

1. Introduction

Big Data describes the large amount of data that is very big and complex and exists in both structured and unstructured form. The building block upon which any organizations relay is Data. With every second data is being created and analyzed. Big data refers to the large amount of space which is needed by every organization. Now-a-days data has become cheaper to store, so organizations need to get as much value as possible. Furthermore big data is used everywhere now, banking, education, government, healthcare, manufacturing are some of these fields. It is very critical to understand that the primary value from big data do not comes from the raw data form, but from the processing and analysis of it, from the products, and services that emerge from analysis.

2. Big data

A. What is big data

Big Data is described as collection of data that is huge in size and, it is exponentially growing with time. Such data is very large and complex, none of the traditional data management tools are able to store it or process it efficiently. Although the term 'Big Data' is new but gathering and storing large amount of data for analyses is ages old. A very huge number of data is present but only 0.5% of the data has been analyzed yet. John Mashey was the scientist who coined this term and made it popular. As, big data is so complex it requires sets of tools and techniques to handle it.

B. Categories of big data

Big data can be categorized into three forms.

- Structured
- Unstructured
- Semi-structured

1) Structured

Any data which can be accessed, analyzed, stored and executed in a fixed form is known as the "structured data". In doing this, great success has been achieved. But now computers are facing problems in handling large amount of data. One example of such form is the data stored in relational databases.

2) Unstructured

Data which do not have a particular structure or it is present in raw form or any unknown form, is termed as the "unstructured data". Deriving information from unstructured form is data is a big task. One such example is a heterogeneous data source which contains videos, audios, text files.

3) Semi-structured

Data which is present in both the forms is termed as Semi-structured data.

C. Characteristics of big data

The most important V's of Big Data are:

1) Volume

As big data itself means dealing with da-ta which is enormous in size. Hence 'volume' is the uttermost important aspect of big data, as it defines the amount of data matters.

2) Variety

It tells about the source, type and nature of the data. It let us know whether the data is structured or unstructured as text, audio, video form of data needs additional preprocessing to derive meaning.

3) Velocity

It refers to the speed of the data by which it is received and acted upon. The real potential of data is determined by how fast the data is generated and meets the demand.

3. History of big data

The origin of large data sets was started in the year 1950 and 60's. In the year 1965 the first ever data center was opened in United States to store tax re-turns. Later in 1989 a British scientist invented the World Wide Web. 2005 was the year when people got to know that a huge amount of data is generated through mails, YouTube and other social media sites. This was the year when "HADOOP" came in.

HADOOP is an open source framework which is de-signed specially to store and analyze big data. It is created by 'Yahoo''. It made big data easier to work and cheaper to store the data. In 2009 the Indian government decided to take a photograph, fingerprint and iris scan of its 1.2 billion people. Its data is stored in the largest biometric database in the world. Big data has exponentially increased over the years. As with big data we have a large set of information and the company can retrieve useful information from that large set. That is the reason why more and more organizations are inclined towards big data. Also with the emergence of IOT and Cloud computing more number of devices are connected to inter-net. And also many new social media sites have been launched time to time. All these things produce new data every day. Machine learning has also opened new forms. However the large revolution is yet to come. The time is near when we will notice the changes as it's the era of Big Data now.

4. Big data today

A. *Importance*

Today Big data is on boom. It helps in the decision making process of the business. It helps in understanding the market forces. By understanding the market conditions, companies can make products according to it. With the help of big data analysis businessman can understand their customers and they get to know the demands and needs of the customer. Also, they can get timely feedbacks from the customer and can change their products according to their wishes. It also helps in controlling the social media sites and can also help in reducing the cost of products and makes it inexpensive.

B. *Tools for analyzing and storing*

Many tools are used for analyzing and storing data. Some of them are:

1) Apache hadoop

It is a java based free software framework. HDPS (Hadoop Distributed File System) is the storage system of Hadoop. Which splits data on every node in form of clusters.

2) NoSQL

Not Only SQL is used to handle the unstructured form of data. It gives a reliable and better performance to store massive amount of data.

3) Presto

It is a Query engine designed and developed by Facebook to handle Petabytes of data.

C. *Applications*

1) Internet of things (IOT)

Data extracted from IOT devices provides a map-ping of device inter-connectivity which helps the industries and government to increase its efficiency.

2) Manufacturing

With the help of big data we can forecast the output, increase energy efficiency, check the product and its quality and many more other things can be done.

3) Healthcare

With Big data analytics, personalized medicines and prescriptive analytics have improved the healthcare. Researchers are mining the data to see what other treatments can be found in the field. Government sectors, Cyber Security and Intelligence, Scientific Research etc. are a few more applications of Big Data.

D. *Challenges*

Hadoop software is not easy to manage. Many professionals find it difficult to handle. There is a lack of talent to work on big data. There is a shortage of Data scientists. Everyday new form of data is being produced. Data quality and storing the data is a big concern. Data security is also a very big challenge in big data

5. Big data future

A. *Importance*

It is said that the data in future will be double of the present data. More and more industries will start using the concept of big data by realizing its power and benefits. As more and more companies will adopt big data analytics, new technologies will be provided in the future to determine the exact output. With the help of new tools and technologies the percentage of accurate prediction will increase. Machine learning will be a top strategic trend in future. New algorithms will also emerge in the future. With the emergence of new data and technologies, the needs for data scientist will also increase and hence chances for new jobs in this field will also in-crease.

B. *Tools for analyzing and storing*

1) Terracotta In-Genius

In future, it will be a new invention as Terracotta In-Genius speeds up data analysis by moving into RAM For example it will be helpful in trying to stop credit card frauds.

2) Medalogix

It is a healthcare risk assessment tool. It will help in determining the risk rate of a patient of readmitting in the hospital. It will be done by going through the patients past treatment history and its records.

3) SAS Text Miner

It can look for trends in a large text and can predict issues.

Let's say a company is facing a problem with product, this tool will look into the complain posted by the users and will look into the trends. Then it will predict the correct output for that product. So, that it can fulfill the customer's demands and needs. This will be the role of the SAS Text Miner.

C. *Applications*

By 2020 more and more companies will move to-wards big data. Banking and securities, communications, Media and services, Government sectors, Engineering firms, Education, Healthcare providers, Insurance, Manufacturing and natural resources, retail, transportation, Utilities, wholesale Trade will

use Big data in all possible quantities.

D. Challenges

The most important challenge that users will face in the future is Scalability. It is predicted that the number of users will increase exponentially and with that data handling will become difficult. Only professionals can handle such big amount of data. Thus gaining expert knowledge will be a challenge. Security will be the most important challenge. As se-curing such huge amount of data in not easy. Also the quality of data should be maintained. It is quite inexpensive today but will be expensive in the future.

6. Techniques used for analyzing

A. A/B Testing

Also known as Bucket testing and Split testing. In this technique, variety of test groups are com-pared with a control group in order to conclude what changes will improve the given objective variable. For example, what color, text font, layout will be suitable for an e-commerce website so that it attracts more auditions.

B. Association rule learning

This technique is used to determine some correlations between the variables in large databases. It was first used in major supermarket chain to determine correlations between its products and the customers. It is also used to uncover new relationships by examining biological data.

C. Classification tree analysis

As the name specifies, it classifies the category of the new data or the document. It is used to categories organisms into their respective bunch. It requires a lot of training.

D. Machine learning

As we all know that machine learning is based on learn from data concept. The machine learns from the “training data set” and acts according to the given conditions and produces output. It is used to differentiate between spam and non-spam mails or messages. It is helpful in determining the probability cases. It learns user’s predilections and guides the user ac-cording to the information given.

E. Social network analysis

This technique was first used by telecommunication industry. It is used to see how people interact and tie bonds with people outside their zone. It helps in tying relationships between individual. It helps in understanding the social domain of an in-dividual.

7. Conclusion

This paper can be seen as a complete summary of Big Data. As the amount of data generated is increasing day by day at a very high rate, it has become a necessity to develop techniques that can handle such a huge amount of data. Big data is helping us in doing so. The analyzing techniques Big Data provides us with today also need some modifications in them to make them even more compatible in future. This is a field which needs continuous research and modifications to work efficiently in future also. With time, more new techniques are also needed to be developed.

References

[1] Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11-19 (2010).
[2] Asur, S., Huberman, B.A.: predicting the Future with Social Media. In: ACM International Conference on Web Intelli-gence and Intelligent Agent Technology, vol. 1, pp. 492-499 (2010).
[3] Bakshi, K.: Considerations for Big Data: Architecture and Ap-proaches. In: Proceedimgs of the IEEE Aerospace Confer-ence, pp. 1-7 (2012).
[4] Cebr: Data equity, Unlocking the value of big data in: SAS Reports, pp. 1-44 (2012).
[5] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceed-ings of the ACM VLDB Endowment 2(2), 1481-1492 (2009).
[6] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM International Workshop on Data Warehousing and OLAP, pp. 101-104 (2011).
[7] Econimist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1-24 (2012).
[8] Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013).
[9] EMC: Data Science and Big Data Analytics. In: EMC Educa-tion Services, pp. 1-508 (2012).
[10] He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space Efficient Data Placement Structure in Map Reduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199-1208 (2011).
[11] Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261-272 (2011).