International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Big Data Analytics and Its Applications

B. Manjulatha, Suresh Pabboju

Abstract*: *Big Data plays an important role in today's environment. As technology is rapidly growing, massive amount of data is being generated from various sources like social media, business organizations, healthcare sector, government sectors, educational institutions, iot applications through sensors and many more. It's a tremendous task to handle such large amount of data by using relational database management systems. This paper briefly describes about what are the various tools and techniques used to manage the data.

Keywords: Analytical tools, Big Data, Data mining algorithms

I. INTRODUCTION

Big Data [1] Analytics is the combination of Big Data and Analytics. Big Data is a term used for a collection of data sets that are large and complex, which is difficult to store and process using available database management tools or traditional data processing applications. Analytics involves studying past historical data to research potential trends, to analyze the effects of certain decisions or events, or to evaluate the performance of a given tool or scenario.

II. BIG DATA CHARACTERISTICS

The characteristics [2] that define Big Data are:

Fig .1. Big Data Characteristics

Volume - This describes the amount of data being transported and stored. The current challenge is to discover ways to most efficiently process the increasing amounts of data, which is predicted to grow 50 times by 2020, to 35 zettabytes.

Velocity - This describes the rate at which this data is generated. The data infrastructure must be able to immediately respond to the demands of applications accessing and streaming the data.

Manuscript published on November 30, 2019.
* Correspondence Author
B. Manjulatha*, Research Scholar, Osmania University, Hyderabad, India. Email: manjulathareddy86@gmail.com
Dr. Suresh Pabboju, Head of IT Department, CBIT, Hyderabad, India. Email: plpsuresh@gmail.com

Variety - This describes the type of data, which is rarely in a state that is perfectly ready for processing and analysis.

Veracity - This is the process of preventing inaccurate data from spoiling your data sets. For example, when people sign up for an online account, they often use a false contact information. Increased veracity in the collection of data reduces the amount of data cleaning that is required.

III. TYPES OF DATA

Big Data could be of three types:

- **Structured**
- **Semi-Structured**
- **Unstructured**

A. Structured Data

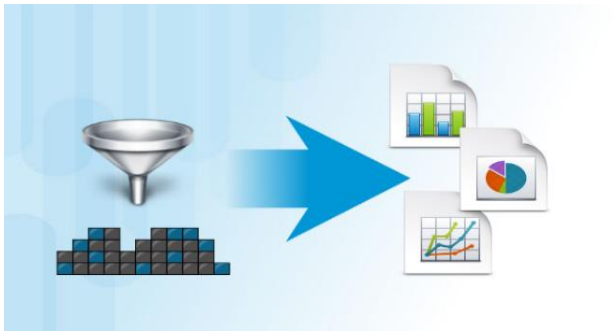


Fig .2. Structured Data

Structured data refers to data that is entered and maintained in fixed fields within a file or record. It is easily entered, classified, queried, and analyzed by a computer. This includes data found in relational databases and spreadsheets. The structure will force a certain format for entering the data to minimize errors and make it easier for a computer to interpret it.

B. Semi-Structured Data

It is a type of data which does not have a definite structure. It becomes a difficult task to analyze the semi structured data. Ex: XML files or JSON documents are examples of semi-structured data.

C. Unstructured Data

Fig .3. UnStructured Data

Unstructured data is a raw data which lacks the organization found in structured data.

It is not organized in Unstructured data lacks a set way of entering or grouping the data, and then analyzing the data. Examples of unstructured data include the content of photos, audio, video, web pages, blogs, books, journals, white papers, PowerPoint presentations, articles, email, wikis, word processing documents, and text in general.

IV. HISTORY BEHIND BIG DATA

In earlier days each and every data which has been collected from various sources are entered manually. It became a tedious task to maintain data in records as if any damage occurs, data can be manipulated etc. So, to overcome these difficulties a new technique was integrated i.e storing of data is done by relational database management systems. By using DBMS, data can be stored in a structured form i.e. in the form of tables (rows & columns). Retrieving of historical data was impossible as day to day transactions are done by modifying the data, even the previous data which has been stored earlier is also lost. As the data is growing rapidly, storing of data in DBMS becomes difficult. Therefore, Big Data came into existence.

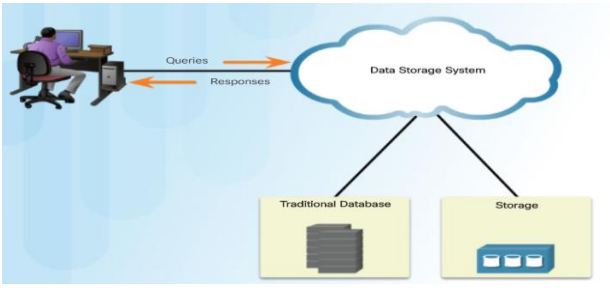


Fig .4. Traditional database system

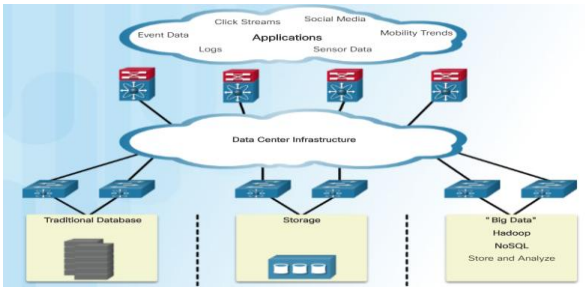
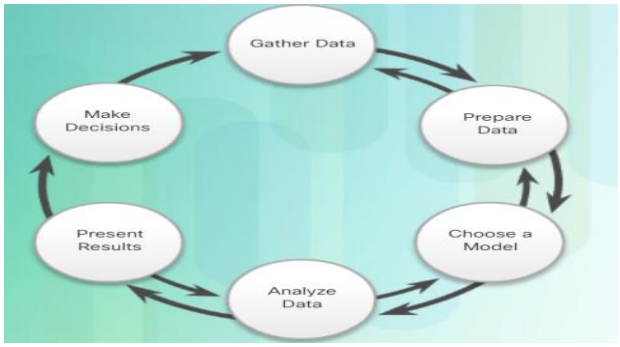


Fig .5. Big Data Infrastructure

A. Life Cycle of Data Analysis

It contains the following phases. Firstly, it gathers the data from various sources either it may be from flat files, documents, databases. After that data should be prepared for analysis. Pre-processing techniques are applied to remove all noisy data and finally make the data to be consistent. In the next phase choosing the correct algorithm is done. The results finally obtained will be helpful in making the decisions.



B. Big Data Analytics in Industry Verticals

Fig .6. Big Data in Educational Sector

Huge amount of data has been gathered from educational institutions [3] like from students, faculty, exam results, course details etc. Proper analysis of data should be done to improve the effectiveness and quality of the institution.

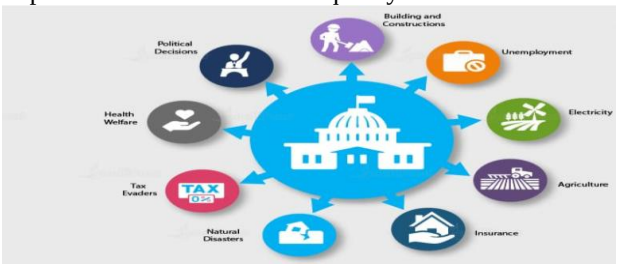


Fig .6. Big Data in Government Sector

As the government is offering variety of schemes to develop the society, a large amount of data will be generated. All the data which is stored is confidential data. Many data mining and security mechanisms are implemented to protect the data.

Fig .7. Big Data in Media and Entertainment

Many people are using various electronic gadgets for entertainment purpose. People are more addicted with social networks like facebook, Instagram, twitter etc. A huge amount is being generated through comments, likes, posts which has shared. Complexity increases and also security issues arises as huge amount of data is being stored.

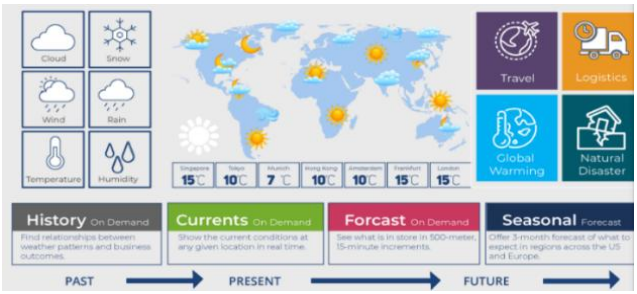


Fig .8. Big Data in Weather Patterns

With the help of sensors which are embedded in iot devices collects the weather data and predicts the weather conditions and also gives alerts to protect from natural disasters.

Fig .9. Big Data in Transportation

Big Data is used in transportation to make our lives easier.

Fig .10. Big Data in Banking Sector

As banking sector produces lots of confidential data of the users, there are chances of stealing that information by unauthorized persons/hackers. Big Data helps in detecting and eradicate those issues by applying some techniques.

Fig .11. Big Data in Traffic control

Smart cities came into existence to overcome the difficulties of traffic congestion.

Smart City can have the following:

1) Real Time Information on Transport, Population and Waste
2) No Traffic Congestion
3) Access to Energy-Saving Resources
4) Clean and Fresh Air
5) Water Shortage or Power Outages

V. CHALLENGES WITH BIG DATA

Few challenges [4] which come along with Big Data:

1) Complexity in managing the data which is stored
2) Maintaining integrity, security and privacy
3) Shortage of skilled people
4) Quality of data

VI. BIG DATA FRAMEWORKS

The functions of Big Data include privacy, data storage, capturing data, data analysis, searching, sharing, visualization, querying, updating, transfers, and information security. The best Big Data Frameworks are as follows:

Apache Hadoop

Apache Hadoop [5] is important as it can store vast amount of both structured and unstructured data and also protects the data processing from hardware failure. Big data tools which are associated with Hadoop are Apache Flume, Apache HBase, Apache Hive, Apache Pig etc.

Apache Storm

Apache Storm is a stream processing framework and is simple, can be used with any programming language.

Apache Spark

Apache Spark is a general purpose and lightning fast cluster computing system. It is faster than Hadoop.

Cassandra

It is a database that provides a mechanism to store and retrieve data other than the tabular relations used in relational databases.

Flink

Apache Flink is an open-source stream processing Big data tool. It is distributed, high-performing, always-available, and accurate data streaming applications.

Cloudera

Cloudera is the fastest, easiest and highly secure modern big data platform. It allows anyone to get any data across any environment within single, scalable platform.

VII. CONCLUSION

There is a strong impact in almost every sector and industry today. In this paper, we have briefly reviewed about the challenges that big data faces. Finally, the goal is to integrate with the methods and techniques proposed to handle and to protect such massive amounts of data.

REFERENCES

1. V. Maria Antoniate Martin, Big Data and Its Challenges, International Journal of Scientific Research in Computer Science, Engineering and Information Technology,2018.
2. \Laney, D.: 3D data management: controlling data volume, velocity and variety. Appl. Deliv. Strateg. File, 949 (2001).
3. Jens Baum, Applications of Big Data analytics and Related Technologies in Maintenance—Literature-Based Research, 2018.
4. D. P. Acharjya, A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, International Journal of Advanced Computer Science and Applications,2016.
5. Ms. Komal, A Review Paper on Big Data Analytics Tools, International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES) ,2018.

AUTHORS PROFILE

B. Manjulatha is a Research Scholar in Computer Science and Engineering of Osmania University. Currently working as an Assistant Professor in VBIT, Ghatkesar, Hyderabad.

Dr. Suresh Pabboju currently working as a Professor & Head, Dept. Of IT, IQAC Coordinator, Alumini Affairs Coordinator, CBIT, Hyderabad.