International Research Journal in Global Engineering and Sciences. (IRJGES)
ISSN: 2456-172X | Vol. 3, No. 3, September - November, 2018 | Pages 69-76

A Review on Concepts, Applications, Challenges and Future Scope in Big Data

Gopinathan S, Assistant Professor, *Department of CSE, Agni College of Technology, Chennai, India*
S.Jayanthi, Assistant Professor, Department of CSE, Agni College Of Technology, Chennai, India

Abstract: Big Data is a storage system used to store huge amount of data. The data includes text, audio, video, images etc. The large amount of data is used in many applications such as Fraud detection, Telecommunications, Health and life sciences, E-commerce and customer service etc. Big data is most powerful for modern business which uses the intelligent automation. Machine learning is based on algorithm which can be learnt from data. Big data is supplied to analytical system of ML. Apart from machine learning big data is also plays a major role in artificial intelligence, deep learning an IoT. Big data challenges include data storage, data analysis, capturing data, visualization, querying, search, sharing, transfer, updating, information privacy and data source.

Keywords: 5 V's of big data, Machine learning, Deep learning, Data Visualization, HADOOP, Issues of big data, Applications of big data.

1. INTRODUCTION

Big data may include both structured and unstructured data for business use. It does not depends on the important of data. But it depends on what the organization will do with the collected data. These data can be analyzed in deeper and take good decision for business moves.

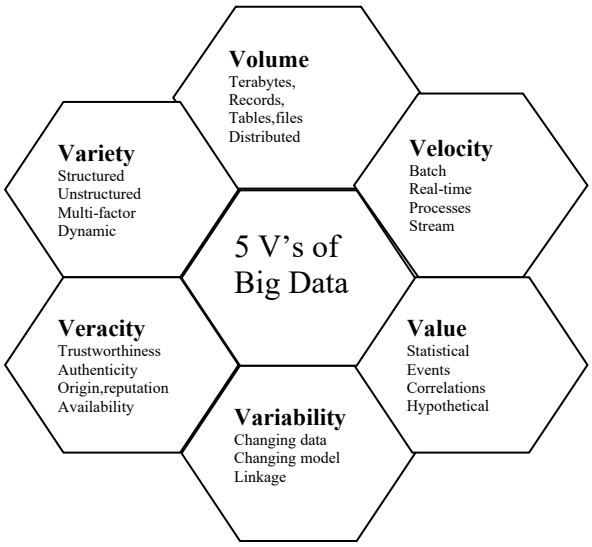


Image-1: 5V's of Big Data

1.1. Volume: It offers an increased amount of data. Because of the large volume of data distributed systems can be used to manage. In distributed system data can be stored in different locations and brought together by using any software when required. For example in face book there are 10 billion messages, 4.5 billion likes and 350 million pictures are uploaded every day. Analyzing this kind of data is really a very big challenge in engineering.

1.2. Velocity: It accelerates the data analysis process. Velocity deals with the speed of collecting and analyzing the data which can be generated already. Everyday data is increasing at every second. So the speed of transmission of data must be analyzed. It is done by the newer technology of big data while generating itself. Analyzing of the data happened before putting the data into the database.

1.3. Value: It is mainly for business growth. The value of the data is related to the cost of collecting and analyzing the data to guarantee that the data can be monetized. Link between the data and insights is not always mean that the data is value.

1.4. Veracity: It provides ultra-reliable data sets. Veracity also deals with noise and abnormality of data. In big data strategy the data should be keep in clean. Gathering loads and loads of data is not use if it is not a good quality.

1.5. Variety: It found new forms for investigation. Variety means different data types and categories of big data repository. Nowadays all the data are not structured in data table. Approximately 80% of data is unstructured. Latest technology of big data allows both structured and unstructured.

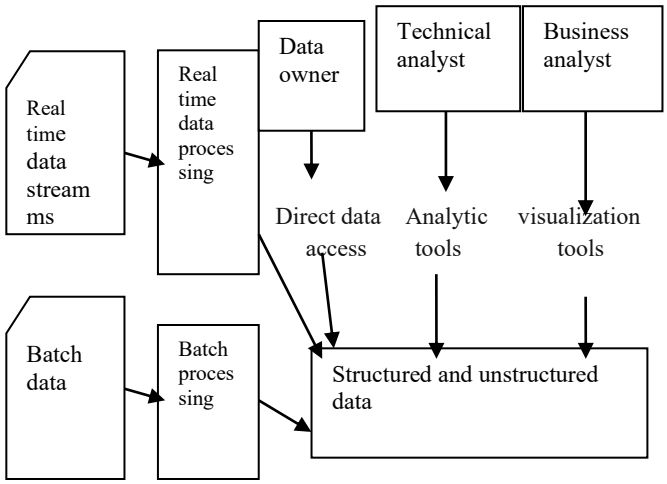


Image 2: Big data architecture

Big data means a large dataset which will grow continuously. But it is difficult to maintaining, process, visualize, analyze and store in existing database management system. So we need to do preprocessing and classified into structured and unstructured data. This big data technology is used in many applications. Most important is in business application.

2. CONCEPTS

Most of the organizations and industries are facing a big challenge to protecting and analyzing the increasing volume of data. Big data analytics is a strategy used to analyzing the large data set and also uncover the hidden patterns and connections among those data. Big data analytics used to support business to achieve more profit and also discovers the new revenue opportunities, improves the efficiency of customer service delivery etc. Big data analytics deals with the challenges of unstructured and vast data. Hadoop is a best framework for big data analytics. It takes the incoming data and divides it into cheaper disk. This technology is used to take better decisions in businesses.

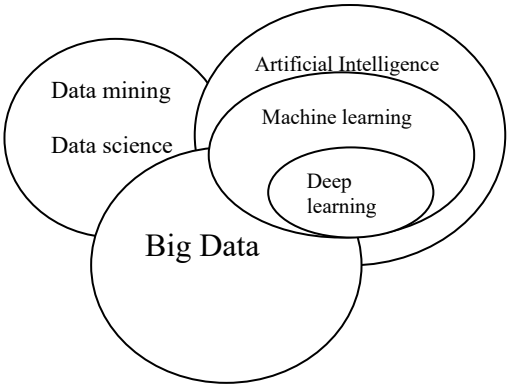


Image-3: Relationship between big data, ML, AI and DL.

Data science is a recent area which helps to collect, analyze, visualize, manage and preservation of huge data. Data mining is the process of examining the important patterns from the large data set.

Artificial Intelligence and its sub undergrowth (For example Machine Leaning, Deep Learning, Neutral Networks), all are algorithm based. These algorithmic methods are used on vast amount of Data (Big Data) to produce desired results and to find trends, patterns and predictions. Composite analytical tasks faster than human imagination are done on Big Data with the help of Machine Learning and Artificial Intelligence.

Machine learning classified as supervised and unsupervised learning. In supervised learning, training data includes both inputs and desired results. This kind of learning is fast and accurate. Supervised learning is classified into two different algorithms. One is classification and another one is regression. Classification algorithm is suitable where the output is categorized. The regression algorithm is used where the output value is real. In other terminology unsupervised learning, the information is neither classified nor labeled and allows the algorithm to act on the information without guidance. Unlike supervised learning no training and teacher is provided for learning. Unsupervised learning is again classified into two algorithms. One is clustering and another is association. Clustering is an algorithm where the related data grouped into several different types. Association is suitable when we describe the large portions of data.

Deep learning[1] is a subset of machine learning. It is uses multi-layered artificial neural networks to deliver the tasks such as speech recognition, object detection, language translation etc. Deep learning is helped to predictive analysis in most accurate manner. NVIDIA GUI-accelerated framework is best suitable to implement deep learning. This framework also provides the interfaces to the programming languages like c, c++ and python. TensorFlow and PyChart are other frameworks used for scientists, researchers to improve productivity[2].

3. ADVANTAGES OF BIG DATA

Big data is mainly used for human beings. It also used in science, technology and business.

3.1. Customer services:

Big data helps the customers to create predictive models for specific task. This kind of prediction is done by the data analysis. It helps the customers to understand the behavior of the entire process. Customer relationship management system is used to help the customers to contact with the enterprises.

3.2. Increased productivity: Big data analytics is used to increase the productivity in business processes. Vendors are predicting the stock of any product through social media data, weather forecasts and web search trends. Supply chain is one of the best big data analytics. Big data analytics is also improves the HR businesses.

3.3. Reduce costs:

Big data tools and automations are helped to reduce the cost**.** These big data analytics tools are used to automate the self-driving car with cameras, sensors, GPS and powerful computers. By implementing the new technologies the cost can be reduced.

3.4. Improved customer service:

Customer service is very much important for all the businesses and organizations. Because the customers feedback on service is taken to the common repository and later it should be analyzed that produces better decision or results. Customers can meet the product management team to improve the service better than others in market. If the organization does not responds for the customers service , they lose the customers which will affect the business.

3.5. Fraud detection:

Fraud is a false representation of normal data. These are frequently happens in financial industries. Anomalies can be detected easily by the machine learning techniques of big data. This technique helped in banking and credit card companies to spot the stolen cards easily. [3]Frauds can be detected easily in structured data. But it is a big challenge to detect in unstructured data which cannot follow any model. Another common use for big data analytics is fraud detection in the financial services industry. Deep learning and automated fraud detection is implemented for detecting the frauds in unstructured data.

3.6. Healthcare services: Big data analytics plays a major role in DNA analysis to find the new patterns for curing the diseases**.** It decodes the DNA strings within a minute. Hospital IT expertise who is familiar with the relational databases and SQL programming languages can only do these analysis.

3.7. Improved security: Big data tools are helped in police department to catch the criminals and detect the activities of them. The national security agency also uses the big data analytics to detect the terrorist who live with us. Some big data technique helps to detect the cyber-securities.

4. ISSUES IN BIG DATA

4.1. Problem in managing data: Data from different sectors are very huge which requires more space to store and need management tools to process those data. To manage the heterogeneous format[8] of data some tools are used. It is tedious process. If it does not managed properly, gives an unacceptable results. Many firms chosen the business intelligence to manage the huge amount of data. But this is difficult to change them from traditional working platform into the new platform. Therefore still we need the advanced technology and tools to manage this situation.

4.2. Storage issues:

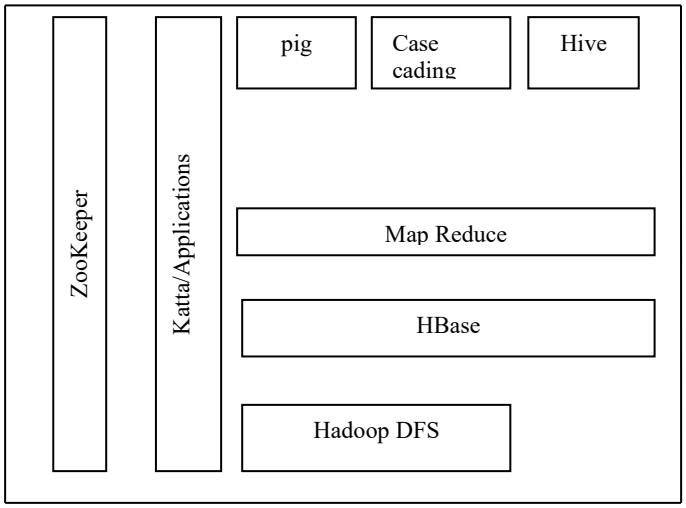
For every business applications or any kind of firms storage of the large amount of big data is major issues. Normally big data volumes are measured in terms of Exabyte. That is we need 25000 disk space to store the data. It is not possible in single system. So we need to store the data in cloud[8]. Even if the data is stored on cloud it take a long time to store from variety of data collections and retrieving from the cloud. This is a major issues to store the bulky data.

4.3. Processing issues:

Most of the organizations are moving to the online mode of processing to boost their business or customer services. For this mode storage is required in zettabyte. This huge amount of data processing is still challenging task. Some of the organizations use MapReduce tool[8] which helps to do the batch processing for a long time. It gives the result as accurate, but it is still slow processing.

5. PLATFORM, TOOLS AND SOFTWARE USED IN BIG DATA HADOOP(High Availablity Distributed Object-Oriented Platform)

Hadoop is a framework for both formatted(structured) and unformatted(unstructured) data in distributed servers.



Characteristic of Hadoop Platform Stack – HDFS + Hive + HBase + Pig

Hadoop Distributed File System(HDFS) is a file management framework for data distribution and storage of the system. In this storage system the files are stored sequentially with same block size except the last block. This file system is easy for data handling and storage of data.

Hive – It is a data warehouse tool which will converts the SQL to MapReduce jobs. This system was implemented by face book initially. Query language used in hive is called Hive Query Language(HQL) which is similar to standard SQL. It deals with queries of data, storage of large data set and analysis of data models.

Hbase – Hbase is a column-oriented Database Management System that works on first layer of Hadoop. It does not support Structured Query Language(SQL). Hbase applications are developed by java. Hbase wont support relational database.

Pig – Apache pig is a high level language platform which is developed to execute SQL on large data set. PigLiatin is a language used for the application development. Pig is a procedural programming language used to simplify the query to access the huge database present in hadoop and MapReduce.

YARN: Yet Another Resource Negotiator and MapReduce. It is present in resource management layer of hadoop.

MapReduce: A software structure for distributed processing of huge database on computing clusters. MapReduce is a core component of hadoop. MapReduce performed in two different operations. One is map operations which will convert the set of data into another set of data in which elements are broken up into key/value(tuple) pair. The reduce operation combines all the tuples based on the key and modifies the key value accordingly.

Spark: It is an open source parallel processing framework Which is faster in memory operations. Spark is another big data processing engine which have the capabilities of machine learning environment with 100 times faster than Hadoop.

Kafka: A distributed processing framework used to stream data in Hadoop.

Apache Hadoop : It is a free framework to store large amount of data in a cluster using JAVA. It splits big data and distributes those data across nodes in cluster.

Microsoft HDInsight : It is a solution for big data from Microsoft. It gives high availability with low cost.

Non Relational Database: It stores massive set of data. Many organizations uses non-relational database to replace XML and transmit structured data between web app and the server.

Advanced tools to use in Big Data analytics

Each tool is aimed at specialists of different level, so it can be a real challenge for you to choose the suitable one for any application. Some kinds of tools are listed below.

Dundas BI

This tool is used to represent a web solution specifically for e-commerce Big Data analytics. All users can customize its platform according to their needs. Special tools for analytics and unlimited number of metrics can be considered to work with it. Likewise, you sync this tool to other similar tools to make use of all their capabilities. It can be used in a cloud as a web version, or it can be saved in our desktop as an app. It is an open source platform that is a great idea for programmers, and it has many pricing offers in which you can choose the most appropriate one.

Pentaho

This one platform also offers users with Big data analytics, reports drawing up, and predictive analytics in which we can adjust according to our needs as well. A large number of different databases and Hadoop platform can be connected and integrated with Pentaho. Pentaho platform uses a visualized analysis which will collect data from other resources and draws them on it. you can use this platform easily during the trial period to make a final choice - is it good for you or not.

Oracle Advanced Analytics

Oracle can boast the development of a large number of different analytical tools. The most fashionable solution is Oracle R Enterprise. This Oracle R Enterprise is working along with Oracle Databases. Using this mixture, you can work with predictive analytics which allows you to analyze the market and goods that you are interested in. Apart from this developers can use tools provided by Oracle to make their software more advanced and smarter.

Even though it is a customizable solution, it is rather difficult to use it without relevant knowledge in software development, so take some decision to find a good software developer that will help you to implement everything in a high-quality .

Cloudera

This platform is aimed at high-skilled specialists in the development, so it is not a good solution for people who are not good in development. Cloudera is integrated with Hadoop platform for processing data. But it is an open source so all developers can use it easily to implement all tasks which they need. It is very good tool for working with Big Data, and also it provides a high level of security.

Vertica

This platform is mandatory for e-commerce field. It is one of the best solutions in the market among other tools. It is very easy to use, everybody can work easily with it. There are 3 solutions to Vertica that distributes for Big Data analytics. First one is a desktop solution that requires the Internet connection and it offers you a number of tools to make an advanced analysis of your e-commerce market and so on. The second one is a cloud based solution that can be commenced with the help of Microsoft Azure.

6. APPLICATIONS

Now a days Big Data is used in everywhere. It takes a major role in business, health and financial sectors.

6.1. Data Visualization

Visualization is the process of creating visual images which is used in complex applications like detective agency, police enquiry and health issues. It is done by using the software called computer graphics. Two basics representations to view the data : 1. Tables 2. Graphs . In tables users need to concentrate on specific values And its precision. It has multiple data sets with different values. But in graphs the message is represented using different shapes. The relationship among multiple values can be identified very easily. It maintains a large data set. This data visualization is applied in big data to represent data patterns and insights of data. These kind of pictorial or graphical representation of large data set is easy for decision makers to make a good decision. By this Big Data visualization a scientist can visualize the data efficiently. It improves the return on investments in business. Many software vendors offering best tools for visualization such as TIBCO, Qlink and Tableau software.

6.1.1. TIBCO

This software is used to access, analyze and visualize the large data sets. These data can be accessed from all the places through the remote access. It can have the capabilities of the report generation and converted into presentation for meetings or sharing information to colleges and friends.

Major four interfaces are present in TIBCO spot fire

1. Visualization

It is a key to analyzing data in spot fire. The types of visualizations are as follows

Tables: cross Tables, Graphical Tables, Summary Tables

Charts: Bar Charts, Line Charts, Combination Charts, Pie Charts, Map Charts

Plots: Scatter Plots , 3D Scatter Plots , Parallel Coordinate Plots, Box Plots

Maps: Tree maps, Heat Map

2. Text area

This area is used to type the users opinion by seeing the different visualization. Text areas can also used to allow to keep controls, filters and actions to view particular data.

3. Filters

By using the filters the data seen can be reduced to view the interested and important data. There are several types of filter forms available such as check box, drop down menu, sliders etc.

4. Details On-Demand

This window shows the exact values of particular row. The values may be numerical or textual. This can be done by clicking an item in visualization, dragging an item around the mouse and marking many items by clicking.

6.1.2. Tableau

Tableau [4]desktop is an interactive data visualization software. This software use the drag and drop of data to represent it visually. Programming skills are not required to use this software. It is very easy and fastest tool and also free for students.

Sheets in Tableau dashboard: Bubble, Tree Map, Line Graph.

1. Bubble : partner(text), Measure(Trade value-size), Filter.
2. Tree Map: Commodity description (Text), Measure(Commodity value share-Size), Trade Flow, Filter
3. Line Graph: Measure(Trade value-Size)

6.1.3. Qlik view

Qlik software is used to find the important visualization. It is one of the fastest business Intelligence in data analytics. This software is best suitable for particular customized setup.

6.2 Big Data in Healthcare:

Big Data plays a major role in all the areas of medicine. But in this paper three categories are explained. Image processing, signal processing and genomics. For each and every activity in hospital the images are very much important to spot the diseases. Medical images are used for the following processing:

1. Diagnosis
2. Computed Tomography(CT)
3. Magnetic Resonance Imaging(MRI)

Signal processing is another technology used for high-resolution acquisition and the multitude of monitors is connected to the patient. Now the healthcare systems uses the singular physiological waveform of data. [5]

6.3. Big Data and the World of Finance:

For long period of time the larger data set contains historical data. The financial process can be re-engineered with big data to manage growing volume of data. Big data in banking is very much useful to enhance the customer services, increases the revenue and business engagement. Currently big data technologies are combined with financial services to improve the efficiency of services to the customers. Securing the storage of data in banking and financial institutions is a challenging process. Security should also extend for online banking and electronic communications of sensitive information. Dell's shareplex connector for hadoop is used to improve the security in these firms.

6.4. Big Data in Fraud Detection:

Today people are using credit cards for shopping and bill payment[10]. They spent the amount based on the card limit and then they paid it later through the bank. If a card is stolen and used by some other persons than the transaction shows abnormal expenditure, this is called fraudulent transaction. Identification of fraud detection is complex process and it was a challenge task to detect the fraud. Classification methods are used to find the fraud.[6]

6.5. Big Data and Sentiment Analysis:

Sentiment Analysis in big data is mainly used in social media such as watsapp, facebook and other social media. The major purpose of sentiment analysis is to decide the user's attitude and moods. The opinion is expressed in positive or negative emotions. Other people opinions are very much important to take decisions. Hadoop based environment is used to do the sentiment analysis. The different opinions are collected from different users and the gathered information is stored in HDFS environment. The data are classified based on sentence level. Some machine learning techniques and algorithms are used to find whether the sentiment is positive, negative or neutral. The sentiment analysis is also referred as Natural Language Processing(NLP).

IBM developed IBM Social Media Analytics [7] that captures structured and unstructured data from social media networks to develop a common understanding of opinions, attitudes and trends. It has the following structure:

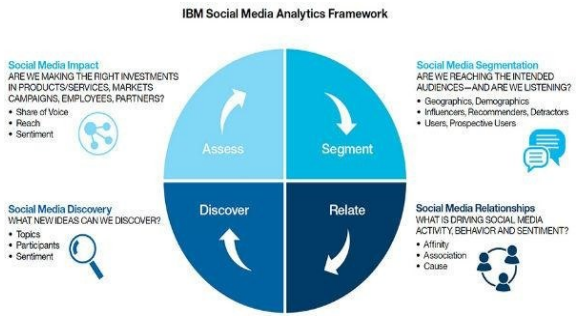


Image-4: IBM's Social Media Analytics [7] framework

VIII. CONCLUSION

This paper gives the literature survey of big data and analytics. It details about the introduction of big data, concepts, advantages, disadvantages and applications. We also discussed about big data with hadoop environment. The one of the major challenge is to manage the big data with new innovations in the field of Hadoop is getting bigger. Revolutionary technologies need to be carried out to exploit the big data completely in future. Another challenge in big data is to provide greater security among Social Media Networks.

REFERENCES

[1] Mao, Feng, et al. "Small Boxes Big Data: A Deep Learning Approach to Optimize Variable Sized Bin Packing." *arXiv preprint arXiv:1702.04415* (2017).

[2]Reference: Available from: https://developer.nvidia.com/deep-learning/

[3] Sharma, Vikash, Bhavna Pandey, and Vipin Kumar. "Importance of big data in financial fraud detection." *International Journal of Automation and Logistics* 2.4 (2016): 332-348.

[4] Reference: Available from: https://www.tableau.com//

[5] Belle, Ashwin, et al. "Big data analytics in healthcare." *BioMed research international* 2015 (2015).

[6] Kamaruddin, Sk, and Vadlamani Ravi. "Credit card fraud detection using big data analytics: use of PSOAANN based one-class classification." *Proceedings of the International Conference on Informatics and Analytics*. ACM, 2016.

[7] Reference: Available from http://www- 01.ibm.com/software/analytics/solutions/customer- analytics/social-media-analytics/

[8] Wani, Mudasir Ahmad, and Suraiya Jabin. "Big Data: Issues, Challenges, and Techniques in Business Intelligence." *Big Data Analytics*. Springer, Singapore, 2018. 613-628.

[9] Satyanarayana, L. "A Survey on Challenges and Advantages in Big Data." *International Journal of Computer Science and Technology* 6.2 (2015): 115-119.

[10] Sharma, Vikash, Bhavna Pandey, and Vipin Kumar. "Importance of big data in financial fraud detection." *International Journal of Automation and Logistics* 2.4 (2016): 332-348.