Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 1, No. 1, January : 2024

UNVEILING THE LANDSCAPE OF BIG DATA: CHALLENGES, TOOLS, APPLICATIONS, AND FUTURE HORIZONS – A SYSTEMATIC REVIEW

Prof. Bhavana A. Khivsara, Assistant Professor, Computer Engineering, SNJB's LKBJ , COE, Chandwad, Email-id: bhavana.khivsara@gmail.com
Dr. M. R. Sanghavi, Professor, Computer Engineering, SNJB's LSKBJ , COE, Chandwad
Dr. K.M. Sanghavi, Professor, Computer Engineering, SNJB's LSKBJ , COE, Chandwad

Abstract:
Big Data serves as the repository for vast daily data influx, encompassing text, audio, video, and images. Its applications span Fraud Detection, Telecommunications, Healthcare, E-commerce, and Customer Service. Leveraging intelligent automation, Big Data empowers businesses significantly. Machine Learning relies on data-learned algorithms, drawing from Big Data within analytical systems. Beyond Machine Learning, Big Data significantly influences Artificial Intelligence, Deep Learning, and IoT. Challenges encompass data storage, analysis, capture, visualization, querying, search, sharing, transfer, update, information privacy, and sourcing data.

Keywords: 6 V‘s of big data, Machine learning, Deep learning, Data Visualization, HADOOP, Issues of big data, Applications of Big Data.

I. INTRODUCTION

Big data is a large amount of data which may include structured, semi structure, Quazi structure and unstructured data for business use. It depends on what the organization will do with the collected data. These data can be analyzed in deeper and takes the decision for business moves and making business ahead.

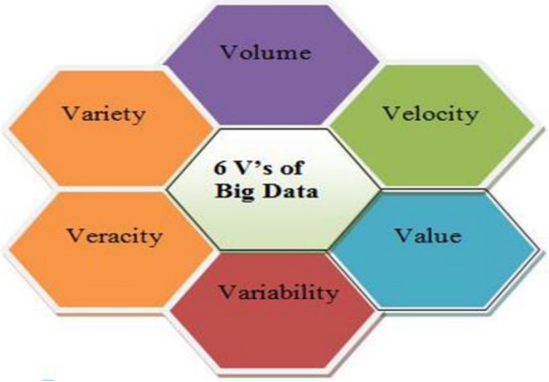

Figure-1: 6 V's of Big Data

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

b. 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 the data happened before putting the data into the database.

c. 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 does not always mean that the data is valuable.

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

e. Variability: how fast and up to what extent the structure of your data changes.

f. Variety: It found new forms for investigation. Variety means different data types and categories of big data repository. Nowadays all the data is not structured in a data table. Maximum of data is unstructured. Latest technology of big data allows both structured and unstructured.

II. CONCEPT

Most of the organizations are facing a big challenge to protect, finding patterns, analyzing the increasing volume of data which is available in variety. Big data analytics is used to analyze the large data set and also uncover the hidden patterns, rules, trends and connections among those data. Big data analytics used to support businesses to achieve more profit and also discover new revenue opportunities, improve customer service etc. Big data analytics deals with the challenges of unstructured and vast amounts of data. Hadoop is the best framework for big data analytics. It takes the incoming data and divides it into chunks for faster analysis. This technology is used to make better decisions in businesses.

Data science is a recent area which helps to collect, analyze, visualize, manage and preserve huge amounts 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 Learning, Deep Learning, Neutral Networks), all are algorithm based. These algorithmic methods are used on vast amounts of Data (Big Data) to produce desired patterns, results to predict recent trends and predictions. The complex analytical tasks are done faster on Big Data with the help of Machine Learning and Artificial Intelligence.

Machine learning is classified into supervised and unsupervised learning. In supervised learning, the model is trained first. 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, for example Buy computer or not buy computer. The regression algorithm is used where the output value is real or where prediction is required. In unsupervised learning, the information is neither classified nor labeled and allows the algorithm to act on the information without supervision or without training data. Unlike supervised learning no training and teacher is provided for learning.

Unsupervised learning is again classified into two algorithms. 1> Clustering and 2> Association.

Clustering is an algorithm where the similar kind of data is grouped and described very properly. Association is suitable when we have to find a pattern example is Market Basket Analysis.

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

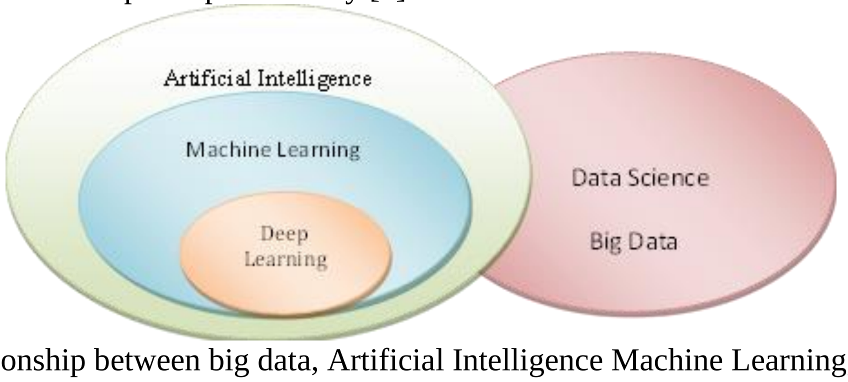

Figure- 2: Relationship between big data, Artificial Intelligence Machine Learning, Deep Learning

III. ADVANTAGES OF BIG DATA

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

3.1 Increased productivity: Big data analytics is used to increase productivity in business processes. Vendors are predicting the stock of any product through social media data and sentiment analysis, weather forecasts, web search patterns. Supply chain is one of the best big data analytics.[9]

3.2 Reduce costs:

Big data helps to reduce the cost. These big data analytics tools are used to automate the self- driving car. By implementing the new technologies the cost can be reduced.[9]

3.3 Improved customer service:

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

3.4 Fraud detection:

Fraud is a false representation of normal data. This 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]. Big data tools are helpful in the police department to catch the criminals and detect the activities of them.

IV. CHALLENGES IN BIG DATA

a. Problem in managing data: Information from various areas is exceptionally tremendous which requires more space to store and needs the board apparatuses to handle that information. To deal with the heterogeneous arrangement [8] of information a few instruments are utilized. It is a monotonous cycle. On the off chance that it isn't overseen appropriately, gives an inadmissible outcome. Numerous organizations had picked the business insight to deal with the enormous measure of information. Yet, it is hard to transform them from the customary working stage into the new stage. Subsequently still we need the trend setting innovation and instruments to deal with the present circumstance.[11][12]

b. Storage issues: For each business application or any sort of company's stockpiling of the enormous measure of huge information is significant issues. Typically huge information volumes are estimated as far as Exabyte. That is we need 25000 plate spaces to store the information. It is preposterous in a single framework. So we need to store the information in cloud [8]. Regardless of whether the information is put away on cloud it requires some investment to store from an assortment of information assortments and recovering from the cloud. This is a significant issue to store the massive information.[11][12]

c. Processing issues: Most of the organizations are moving to the online mode of processing to uplift their business and to improve the customer services. For this mode storage is required in zettabytes. This huge amount of knowledge processing remains a challenging task. A number of the organizations use MapReduce tool [8] which helps to try to execute for an extended time. It gives the result as accurate, but it's still slow processing.[11][12]

V. TOOLS AND SOFTWARE USED IN BIG DATA HADOOP (High Availability Distributed Object- Oriented Platform)

Hadoop is a framework for both structured and unstructured data in a distributed environment.

a. 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 the same block size except the last block. This file system is easy for data handling and storage of data.

b. Hive – Apache Hive empowers clients to handle information without unequivocally composing MapReduce code. Hive language, HiveQL (Hive Query Language), looks like Structured Query

Language (SQL) .A Hive table construction comprises lines and segments. The columns ordinarily relate to some record, exchange, or specific substance (for instance,
client) detail. The estimations of the comparing segments address the different credits or qualities for each line. Furthermore, a client may consider utilizing Hive if the client has insight with SQL and the information is now in HDFS. Hive isn't planned for continuous questioning

c. Hbase – HBase is on top of HDFS in Hadoop. HBase uses a key and value pair structure to store the data of an HBase table. Data is stored at the row, column, and version.

Each key consists of the following elements. Row length, Row Key, Column family length, Column family, Column qualifier, Version, Key type.

d. Pig – Apache Pig consists of a data flow language, Pig Latin, and environment to execute the Pig code. The important advantage of using Pig is to utilize the power of MapReduce in a distributed system. At the same time it also simplifies the tasks of developing and executing a MapReduce job.

e. Mahout- Hadoop is an open-source framework from Apache. Mahout used to store and process big data in a distributed environment across clusters of computers. Apache Mahout is mainly used for creating scalable machine learning algorithms. It implements popular machine learning techniques such as: Recommendation, Classification, and Clustering.

f. YARN: The technology used for job scheduling and resource management and one of the main components in Hadoop is called Yarn. Yarn stands for Yet Another Resource Negotiator though it is called Yarn by the developers. Yarn was previously called MapReduce2 and Nextgen MapReduce. This enables Hadoop to support different processing types. It runs interactive queries, streaming data and real time applications. Also it supports a broader range of different applications. Yarn combines a central resource manager with different containers. It can combine the resources dynamically to different applications and the operations are monitored well.

g. MapReduce: A software structure for distributed processing of huge databases 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) pairs. The reduce operation combines all the tuples based on the key and modifies the key value accordingly.

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

i. Apache Hadoop : It is a framework to store large amounts of data in a cluster using JAVA. It splits big data into chunks and distributes those chunks of data across nodes in clusters for further processing.

j. Non-Relational Database: It stores a massive set of data. Many organizations use non-relational databases to transmit structured data between web apps and the server. It is also known as NoSQL.

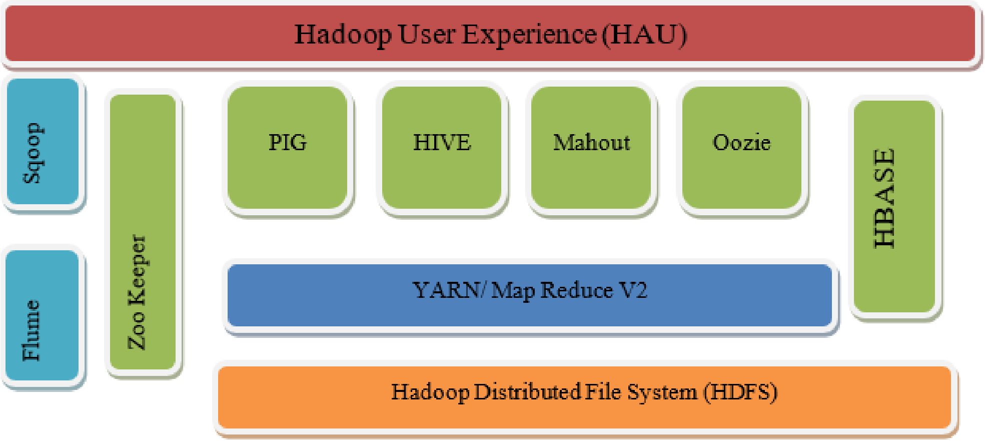
Hadoop User Experience (HAU)
Sqoop
Flume
Zoo Keeper
PIG
HIVE
Mahout
Oozie
HBASE
YARN/ Map Reduce V2
Hadoop Distributed File System (HDFS)

Figure- 3: Hadoop Ecosystem

VI. APPLICATIONS

Nowadays Big Data is used everywhere. It plays a major role in business, health and financial sectors.

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 software called computer graphics. 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. This kind of pictorial or graphical representation of a 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 offer best tools for visualization such as TIBCO, Qlink and Tableau software.

Tableau [4]: Tableau desktop is interactive data visualization software. This software uses the drag and drop of data to represent it visually. Programming skills are not required to use this software. It is a 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)

2 Big Data in Healthcare:

Big Data plays a vital role in all the areas of medicine and healthcare. In this paper we are mainly focusing on areas like: Image processing, genomics and signal processing. Medical images are used or the following processing:

1. Diagnosis of disease
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 system uses the singular physiological waveform of data. [5]

3 Big Data in Finance:

For a long period of time the historical data is stored in the larger data set. The financial process can be re-engineered with big data to manage volume of data. Big data in banking gives power to enhance the customer services, also 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.

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 a complex process and it was a challenging task to detect the fraud. Classification methods are used to find the fraud.[6]

5 Big Data and Sentiment Analysis:

Sentiment Analysis in big data is mainly used in social media such as WhatsApp, 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's opinions are very much important to make decisions. A Hadoop based environment is used to do sentiment analysis. The different opinions are collected from different users and the gathered information is stored in the 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 to 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:

VII CONCLUSION

This paper gives the literature survey of big data and big data analytics. It gives a brief introduction of big data, big data concepts, and its applications. We also discussed the Hadoop Ecosystem and How Hadoop is useful for Big Data. One of the major challenges is to manage big data with new innovations in the field of Hadoop is getting bigger. Latest technologies need to be carried out to exploit big data completely in the 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

[4] fraud detection." International Journal of Automation and Logistics 2.4 (2016): 332-348.

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

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

[7] 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.

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

[9] Wani, Mudasir Ahmad, and Suraiya Jabin. "Big Data: Issues, Challenges, and Techniques in Business Intelligence." Big

[10] Data Analytics. Springer, Singapore, 2018. 613-628.

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

[12] 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.

[13] Hariri, R.H., Fredericks, E.M. & Bowers, K.M. Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6, 44 (2019). https://doi.org/10.1186/s40537-019-0206-3

[14] C. Komalavalli and C. Laroiya, "Challenges in Big Data Analytics Techniques: A Survey," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 223-228, doi: 10.1109/CONFLUENCE.2019.8776932.