International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-6, March 2021

A Survey on Some Big Data Applications Tools and Technologies

Nazia Tazeen, K.Sandhya Rani

Abstract: ***Big Data is a broad area that deals with enormous chunks of data sets. It is a word for enormous data sets having huge volume, more diverse structures of data originating from diverse sources are growing rapidly. Many data being generated because of fast data transmission between devices concerning different sectors like healthcare, science, media, business, entertainment and engineering. Data collection capacity and its storage is big concern. Apache Hadoop software is a store of accessible source programs to store big data and perform analytics and various other operations related to big data. Many organizations base their decisions by extracting knowledge from huge and complex data, because of this prime cause of decision making, Big Data has to be accurately classified and analyzed. In order to overcome the complex challenges encountered by Big Data, various Big Data tools and technologies have developed. Big Data Applications, tools and technologies used to handle it are briefly discussed in this paper.***

Keywords: Big Data, Veracity, Hadoop, structured data, unstructured data.

I. INTRODUCTION

The term Big Data implies huge data generating continuously in different formats which may be in structured format, unstructured format and semi-structured format. Big Data possess data complexity as data generated from online resources is in diverse formats like textual data, image data and audio data resulting in the format of multimedia data format and in order to process Big Data, it requires commanding technologies and innovative systems. Therefore, making use of traditional tools to process Big Data may not produce the desired results. In this paper a brief overview on different applications of Big Data, tools and know-hows are deliberated.

II. LITERATURE REVIEW

Big Data is defined as 3Vs by the experts and data scientists in this field [1]. Huge voluminous data is being generated ranging from millions of devices. It was approximately predicted that nearly 2.5 Exabyte's of data was generated per day in the year 2012 [2]. The data generated in terms of Exabyte's is replicating every 40 months. It was predicted that 4.4 zettabytes of data are being produced, replicated and used by an International Data Corporation in the year 2013. Big Data is being dwelling up for every two years and it was predicted that by the year 2015 digital data grew to 8 zettabytes [3]. In the present times, the data production has reached to 40 zettabytes as estimated and predicted well before and there is furthermore increase in the production of the data by year after year as more and more dives comes into picture with sharp increase in the technology [4]. The speed with which Big Data is generated from data repositories like Facebook, Twitter and Instagram is known as Velocity, one of the 3v's of Big Data. In present times, there is enormous growth in the transactional data and it is predicted that 2.5 petabytes of data are generated per hour from the Walmart retail chain. One of the best multimedia data used over the internet is YouTube, which generates Big Varieties of Big Data known as Variety, one of the 3v's in Big Data. These Big Data sets contains varied kinds of data belonging to different organizations, educational institutions, finance data, weather data and astronomical data etc. Furthermore, definition of Big Data has been expanded with addition of more v's like verification, validation, vision and value by some individuals to give more clear picture of Big Data Complexity. Analyzing Big Data is challenging because of diverse relationships among the data and therefore sound technologies to manage Big Data are need of the hour [5][6]. In order to process the huge generated data, organizations are developing effective tools and technologies so that Big Data can be processed well and proper decisions can be made from the Big Data. The methods utilized by various organizations to handle Big Data are known as Big Data Analytics, which comprise of dozens of algorithms, improved statistical techniques and applied analytics [7]. Last 20 years, Big Data production was increasing every day all over the world due to high usage of world wide web. Outmost production was yielded from social networking site, twitter which produces 2,77,000 tweets single minute, more than 2 million search queries are done on Google every minute, more than 100 million emails are received and sent through Gmail, 1000's of websites are created and lakhs of videos are viewed every second and also it was found data was generated interms of quintillion bytes [8]. More than 2.5 quintillion bytes of data was generated every passing day [9]. Tons of Big data is produced every second and to manage all these things, data extraction, transformation and loading has to be done in a proper manner as Big Data has become the center of trade, business and almost all the areas of life [10]. It was predicted by IBM that almost 2.5 exabytes of data is generated every day since 2003 and doubled, tripled and till date 2.7 zettabytes of data has been approximately generated [11].

Manuscript received on March 22, 2021.
Revised Manuscript received on March 26, 2021.
Manuscript published on March 30, 2021.

Nazia Tazeen, Department of Computer Science & Engineering, Sri Padmavathi Mahila Viswa Vidyalayam, Tirupati, India. Email: naziabaseer@gmail.com

K.Sandhya Rani, Department of Computer Science & Engineering , Sri Padmavathi Mahila Viswa Vidyalayam, Tirupati, India. Email: sandhyaranikasireddy@yahoo.co.in

III. APPLICATIONS OF BIG DATA

The scope of Big Data is literally unlimited enveloping all the information generated by completely diverse devices and individual applications.

The following are some areas encompassed the big umbrella term holding immense knowledge. Some of Big Data applications are: Internet of things, Smart grid case, HealthCare, Political services and government monitoring, Transportation and logistics and weather forecasting, smart city [12].

A. Big data on Internet of things

Using shaky dial-up networks, at most 1 billion users were able to connect with internet in 1990s. With the advent of mobile phones in 20th century and latest technology features added into it, there was a possibility for over 2 billion users connected with internet to be in contact with friends around the world. The evolving trends in order to develop the Internet further, Internet of things came into existence. Accessing Physical objects by connecting to the Internet is IoT [13]. The embedded technology, which can interact with external environment like dishwashers and cars. By the vision 2020 the IoT may have the capability to link 10 multiples as more as 28X10^{9} devices to the internet from vehicles to armlets. Gartner has studied that the vast amount of revenue made from IoT, hopefully to surpass $300 Billion. The devices can have brought together different platforms like Bluetooth or any other technology in future. The brokers use a well-defined protocol to send required messages. Message Queue Telemetry Transport (MQTT) is the most accepted protocol, which is used. After receiving the data, the second procedure is to search out the better technology platform for hiring Internet of Things data. There are many companies, which are using Hadoop for data storage purposes. Also, Hive is next in demand after Hadoop and SQL databases are found to be more accurate for IoT data storage.

Security issues has also considered. The data, which has calculated from the Internet of Things devices supplies for mapping of device, inter connectivity. The various governments and companies use these mappings to increase the efficiency. The audiovisual data used in medicinal and growth contexts, which has been increasing day by day from IoT.

B. Healthcare

Due to advancement in big data analytics, there was seen tremendous increase in generation of health care data in various health sectors like mHealth, eHealth. Decision-making depends upon the quality of healthcare data**.** To predict best diagnosis of cancer patients or thyroid patients or to diagnose any disease, the researchers are mining the big data generated from hospitals. Patterns identified for different conditions based on patient's health record data, patient generated data, sensor data, imaging data and in other forms. Then, patient classified to which of tumor to which they classified so that best treatment given thereby reduction in the cost too. FDA uses big data technologies to enter huge loads of data to assess stimulate and medication.

C. Scientific Research

In order to "conduct, curate, and deliver data to association", a long term plan was initiated by the National Science Foundation which implemented the new approach for acquiring knowledge from data, expand new meets to education, build a new infrastructure.

D. Big data using Machine Learning

Various Machine Learning techniques are in use to mine the Big Data and extract patterns in order to make bright decisions pertaining to various fields like banking sector, financial sector and educational sector [14].

E. Weather Forecasting

Countrywide Oceanic world and to monitor Atmospheric Changes, Big Data technologies used to analyze and extract information in order to predict changes in weather. Apart from the above applications, big data generated from traffic database, media & Entertainment and tax compliance etc.

Traffic optimization, tax compliance and cyber security and intelligence are some of the fields where big data generated. By anticipating what the audience needs, organizing optimization and rising recovery.

IV. BIG DATA TOOLS AND TECHNOLOGIES

This Big Data requires to be wisely handled and processed well to extract meaningful knowledge out of it. Along, with this, Big Data also needs to be stored properly, quality of data must be maintained and secured well and it should achieve scalability. In order to maintain all this, the companies are upgrading their infrastructure and have started implementing the Big Data technologies to predict much out of the heaps. Hadoop, Spark, SAP-HANA, High-Performance Cluster Computing (HPCC) etc. are various Big Data technologies in market, amongst these Hadoop is being used maximum. Hadoop uses the Google developed Map Reduce and an improved file system titled as Hadoop Distributed File System. All the technologies and tools which mining Big Data will be discussed in the subsequent section.

A. Hadoop

To perform big data storage in distributed systems and process it, Hadoop is the software frameworks written in java, one can access it freely as it is open source. Hadoop platform includes advanced declarative languages in order to write queries and handle big data. Hadoop is more preferred to accomplish and analyze enormous amount of formless logs and events. Hadoop is composed of many components but in Big Data two mostly used components are HDFS and Map Reduce. The other components provide complementary services and higher level of abstraction**.**

HDFS **is** used to store huge data files, which are high to store on a single machine typically in gigabyte to terabyte. It is a disseminated, accessible, reliable and manageable data base system written in java for Hadoop framework [15]. It maintains reliability by duplicating data through numerous hosts to facilitate parallel processing, for that, it split a file into blocks that will stored across multiple machines. The cluster of HDFS has master-slave relationship with single name-node and multiple data-node.

As illustrated in Figure1 below, Hadoop architecture is comprised of following components:

- Name Node, accomplishes the organization of file system namespace, controls access to files by clients, and decides the plotting of chunks to data Nodes.
- Data node, accomplishes data storing involved to the nodes that they run on save CRC codessend heartbeat to namenode.
- Accountable for merging fsImage and EditLog takes place by Secondary Namenode.
- Task Tracker
- HDFS

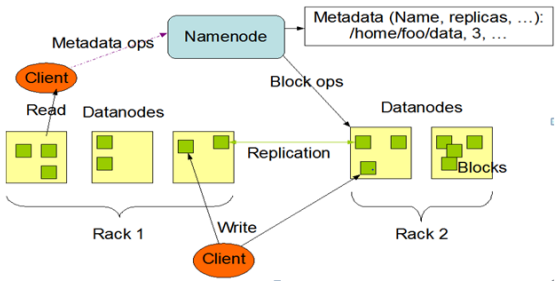

Figure1: HDFS Architecture [16]

B. Map Reduce

Map Reduce is a programming standard used to process huge bulk of data by partitioning the work into nodes by distributing the work into various independent nodes. A Map Reduce program corresponds to two jobs, A Map method which includes obtaining, filtering & sorting datasets and A Reduce method, which include finding out summaries and generate result. Map Reduce system arranges distributed servers, manage all communications, parallel data transfers, also provide redundancy and fault tolerance. Job tracker is accountable for allocating work to every task tracker, Work distribution management like eliminating and scheduling. Task Tracker is accountable for accomplishing work. Generally, task tracker launches another JVM to execute the job.

C. Cassandra and HBase

Both are open-source, non-relational, distributed DBMS written in java and supports in storage of structured data for bulky tables, and runs on top of HDFS. It is columnar data model with features like compression and in-memory operations.

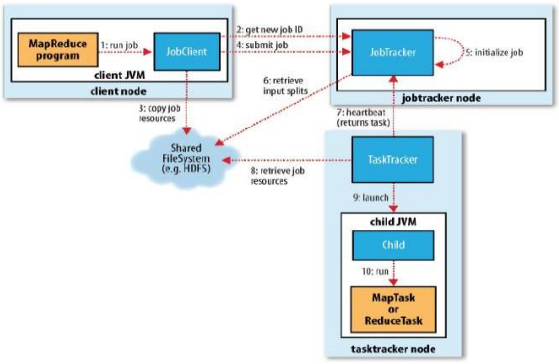

Figure2: Architecture of Map Reduce [16]

D. Hive

It is a warehouse infrastructure by Facebook, which provides data summarization, adhoc querying and analysis. HiveQL is to make powerful queries and get results in real time. The data after cleaning was supposed to be stored in some order and had to be in some shape (schema). The hive used for the same. However, the alternative was kept tackling the hive failure situation like if the data was coming from the multiple sources and complex transformations were to be done e.g. there were two datasets one from sensors and other one was the global survey data.

E. Pig

Provides a framework using high-level data flow language used to execute parallel computations. Pig has two major components. They are Pig Latin and Pig engine [17]. Pig supports a local mode for development purposes.

F. Karma sphere Studio and Analyst

Similar to Hadoop platform, works on Unstructured and structured data in a user-friendly manner and works on advanced business analytics. [17]

G. Zookeeper

It is a high-performance coordination service for distributed application that can store configuration information and have master-salve node [18].

H. Spark

Apache Spark open sourced tool that originally developed in AMP Lab at UC Berkley to run iterative algorithms. It provides in-memory analytics, which is faster than Hadoop (up to 100 times). It is highly compatible with Hadoop's Storage APIs. It can run on existing Hadoop Cluster Setup. Developers can write driver programs using multiple programming languages on Spark. The use of Spark is necessary because of machine learning algorithms that are iterative, and each iteration can improve the results [19].

I. Dryad

Dryad is a Scalable, user-friendly and scalable distributed computation programming model to hide the job distribution from users and a self-complete tool to pact with data-centric applications. Centralized job manager distributes workload to various computers developing directed acyclic graph and monitors execution. In case of a failure, dyrad updates graphs in order to provide robust computing framework [20].

J. Apache Mahout

Apache Mahout was started as a subproject of the Apache Lucene project in 2008. After some time, an open source project named Taste, which was developed for collaborative filtering, and it was absorbed into Mahout. Mahout is written in Java and provides scalable machine learning algorithms [21]. It offers profitable and scalable machine learning techniques for intelligent and extensive data analysis applications, which is known as Apache Mahout. Located and works on topmost of Hadoop with the Map/Reduce for data formats.

K. Jaspersoft BI Suite

Jaspersoft is capable to interface with diversities of databanks comprising MongoDB, Cassandra, Redis, Riak and Couch DB. Efficient for handling big data and is available as open source tool [22].

L. Pentaho Business Analytics

Processes and visualizes big data and available as an open source tool to the users, can accumulate data, store and make business decision implementing business analytics through Web interfaces. Many databases supported by this tool and also it offers several security and scalability features [22].

M. Sky tree Server

This tool offers sophisticated machine learning techniques and helps in improving data and handles different kinds of data on real-time basis [22].

N.Apache Kafka

To stream LinkedIn processes data and to meet real-time constraints using memory analytics, Apache Kafka was developed. It produces temporary solutions for both online and offline processing data and offers real-time computation.

O. IBM Infosphere

IBM Infosphere is a commercial mode of software comprising library of real-time data analytics. It is proficient in handling unlimited diverse data streams. It works on diverse data streams of emerging patterns, supports Mongo DB, DB, oracle on windows platform. After acknowledging the patterns, examination approved and if rejected, essential measure taken.

P. SAP HANA

To process real-time streaming data, SAPHANA tool is used. It uses advanced parallel architecture and algorithms for faster speed to process on block of data. It is a new tool for "In-memory computing".

Q. R

R is a statistical tool used for processing applications of big data and visualization of graphics. It provides numerous methods to process Big Data like time series, linear and non-linear modelling, classification and clustering etc. It offers data handling in a proper manner with security [23].

V. CONCLUSION

The Big Data is forthcoming technology for the data analysis in distinct ranges. Mostly used for business analytics, but progressively research is going on for its usage in distinct fields where the data is enormous. In recent years, big data analytics play a substantial role in various backgrounds including business watching, production development, healthcare applications, research, etc. Hadoop is used to store and manage huge data sets generated by big networking firms such as Google and Facebook. This paper gives brief overview of big data applications, technologies and its tools. A provisional study of distinct tools is carried out on which we can disciple unstructured data to structured data. We achieve acceptable results by current technologies, though few technologies are still under development to process Big Data in less time.

REFERENCES

1. B. Furht and F. Villanustre, "Introduction to big data. In: Big Data Technoogies: Springer International Publishing, Chambers, pp 3–11, 2016.
2. 2, A. McAfee, E. Brynjolfsson, "Big Data: The Management Revolution," Harvard Business Review, pp.60–68, 2012.
3. Rajaraman, "Big data analytics," Resonance vol.21, pp.695–716, 2016.
4. Kune, P.K. Konugurthi, P.K. Agarwal, A. Chillarige, R. Buyya, "The anatomy of big data computing," Software: Practice and Experience 46(1), pp.79–105, 2016.
5. C.K. Emani, N. Cullot, C. Nicolle, "Understandable big data: a survey," Computer Science Review, vol.17, pp.70–81, 2015.
6. Gandomi, M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," IJIM, vol.35, pp.137–144, 2015.
7. Iqbal, F. Doctor, B. More, S. Mahmud, U. Yousuf, "Big Data Analytics: Computational Intelligence Techniques and Application areas,' IJIM, 2016.
8. https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/
9. Shilpa, M. Kaur, "International Journal of Advanced Research in Computer Science and Software Engineering," vol.3(10), October, pp. 991-995, 2013.
10. Sagiroglu, D. Sinanc, "Big Data: A Review," pp.20-24, May 2013.
11. N. Mangla, R.K. Khola, "Application Based Route Optimization," IOSR Journal of Engineering, vo.2(8), pp.78-82, 2012.
12. https://www.edureka.co/blog/big-data-applications-revolutionizing-various-domains
13. https://www.simplilearn.com/how-big-data-powering-internet-of-things-iot-revolution-article
14. M Bishop, "Pattern Recognition," Machine Learning, 128, 1-58, 2006.
15. https://mapr.com/products/apache-hadoop/
16. Mehta and N. Mangla, "A survey paper on Big Data Analytics using Map Reduce and Hive on Hadoop Framework," IJRAET Volume.4, Issue 2, NCRISTM-2016.
17. Maheshwari, "Big Data," McGraw Hill Education India private limited, second edition.
18. J. Wang, W. Liu, S.kumar and S. Chang, "Learning to Hash for indexing Big Data A survey," arXiv:1509.05472v1 [cs.LG]
19. https://www.qubole.com/products/qubole-data-service/apache-spark-service/
20. M. Israd, M. Budiu, Y. Yu, A. Birrell, D. Fitterly, "Dryad: Distributed Data-parallel Programs from Sequential Building Blocks," Proc. of 2007 Eurosys conf.
21. Gupta, "Learning Apache Mahout Classification," Packt publication, UK, 2015.
22. M. Chen, S. Mao, Y. Liu, "Big Data: A Survey," Springer Science Business Media New York, Mobile Network Applications pp.171-209, 2014.
23. Chambers, J.: Bell Laboratories: What is R? The R Foundation. http://www.r-project.org/. Accessed 5 Aug 2018

AUTHORS PROFILE

Mrs. Nazia Tazeen, completed B.tech, M.Tech in Computer Science and Engineering, at JNTU Hyderabad. I am pursuing my PhD in Computer Science and Engineering at Sri Padmavathi Mahila Viswa Vidyalayam, Tirupati, India. My research area is Big Data Analytics, Data Mining, Cloud computing.

Prof. K. Sandhya Rani, graduated M.Tech in Computer Science from IIT Kharagpur, in 1992 and Ph.D. in Computer Science from Sri Padmavati Mahila Visvavidyalayam, Tirupathi in 2003. She is having total 33 years Work Experience in various fields. She has worked as the Rector of Sri Padmavati Mahila Visvavidyalayam, Tirupathi. Research area: Big Data Analytics, Pattern Recognition, Artificial Neural Networks and Privacy Preserving Data Mining