ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)

AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad

Vol. 4, Special Issue 4, January 2017

Thrust Area in Data Science-Big Data and Data Analytics

Prof. Ashish N. Patil[1], Anagha Ajay Jadhav[2]

Professor, CSE, AGTI's DACOE, Karad, India[1]

Student, CSE, AGTI's DACOE, Karad, India[2]

Abstract: The dawn of big data has arisen. Big data may be defined as a term which indicates large sets of data. The challenges that occur due to the big data are its storage, management, capture, curation, sharing, analysis, security, handling, etc. The old or traditional methodologies are not much capable for the above purpose. Hence, it is necessary to know about the upcoming methods regarding the big data and the paper introduces the brief information about it. The concept of big data is becoming popular because of the constant increase need and creation of data. In general, big data is constantly moving data and its size has almost reached till yotta bytes. Certain tools used such a Hadoop, MongoDB, Cassendra and Apache spark to deal with big data are briefly presented in the paper. Also, the terms such as parameters, applications, storage and growth are described.

Keywords: Introduction to big data, parameters and tools of big data, Hadoop, HDFS, Map-Reduce, Apache Spark, MongoDB, Cassendra.

I. INTRODUCTION

Data can be described as set of values of different quantitative or qualitative variables. Data science can be termed as a field that is a cluster of all types of data [1]. In last two years, the creation of data is almost doubled as compared to its previous years. This data set can come from various sources such as industries, media, satellites, transaction records, etc. Also, the incoming, available or outgoing data can have various forms such as text, video, symbols, images, audio, e-mail etc. This all data when clustered together as large set of data is termed as big data. Big data is one of the upcoming generations of technologies. Big data can be looked upon as a term which doesn't have any specific quantity, so this term is mostly used when talking about zettabytes and yottabytes of data. The current estimated byte for digital information is yottabytes where 1 yottabyte= 10^{24} bytes.

Name	Symbol	Approximate Value for Reference	Actual Value
Byte			8 bits [Store one character]
Kilobyte	KB	About 10^3	2^{10} = 1,024 bytes
Megabyte	MB	About 10^6	2^{20} = 1,024 KB
Gigabyte	GB	About 10^9	2^{30} = 1,024 MB
Terabyte	TB	About 10^{12}	2^{40} = 1,024 GB
Petabyte	PB	About 10^{15}	2^{50} = 1,024 TB
Exabyte	EB	About 10^{18}	2^{60} = 1,024 PB
Zettabyte	ZB	About 10^{21}	2^{70} = 1,024 EB
Yottabyte	YB	About 10^{24}	2^{80} = 1,024 ZB

Fig.1: Storage capacity of data

Almost 500+ terabytes of data is generated per day on considering one single site such as facebook [2]. Hence a massive data is oriented every fraction of second and so it is essential to manage, store, secure, etc., the useful data and flush the unwanted data. For handling this all data which is termed as big data, it is necessary to create big data technologies, so that we get more accurate analysis. Also a concrete infrastructure is needed to store and process this huge volume of data[3].

Big data can be split into three forms:

1) Structured big data – The data which has semantic meaning is termed as structured data.
2) Unstructured big data – The data which has no latent meaning is termed as unstructured data.
3) Semi-structured big data – The format of this type of data between structured and unstructured format.

The major parameters to be considered for estimating big data are:

1) Volume – It is the quantity of data. Frequently, volume is the main parameter in deciding whether the data to be considered is the big data or not.
2) Variety – The big data can be there in various forms like texts, numbers, images, videos, audios, etc. which leads to different varieties of data.
3) Velocity – In case of big data, velocity is the speed of data flow from different sources.
4) Variability – The term refers to data whose meaning is changing constantly.
5) Veracity – Noise and abnormality present in data is known as veracity in data [4].

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)
AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad
Vol. 4, Special Issue 4, January 2017

To deal with this all, there are few challenges, which can be stated as follows:

- Curation
- Capturing data
- Storage
- Searching
- Sharing
- Presentation
- Analysis
- Transfer
- Management

II. APPLICATIONS OF BIG DATA [5]:

Banking: Commercial banks use big data analytics and visualization for high frequency trading and for predictive analytics.

Media and Entertainment: As per consumer requirement it is important for this field to understand real time patterns of big data and social media content. Amazon prime is one of great example that uses big data to provide better experience to consumers by providing kindle books, music, etc. [6].

Healthcare providers: Big data in healthcare is emerging as one of the promising field by providing different data types and speed at which it should be managed while reducing cost.

Education: A big challenge in education industry is to study big data from various sources, its challenges and to utilise it properly when needed and design various algorithms for its security, storage and management.

Government: Government agencies or some of public sectors such as FDA(Food and Drug Administration) are using big data to study and detect different patterns of food and drug related diseases.Also government agencies are using big data analysis to secure the country.

Insurance: The insurance industry is using big data for improving customer retention and for predicting behaviour of customer.

Transportation: The transportation agencies relating to government use big data for traffic control, to predict traffic condition, route planning, etc. While private sector is also using big data for different technological enhancements to improve the transportation.

III. TOOLS TO DEAL WITH BIG DATA:

A) Hadoop:
Hadoop, commonly known as Apache Hadoop is termed as open source software which is written in java. Its core consists of two main units, one for storage and other for processing. The storage unit is called as HDFS(Hadoop Distributed File System) and processing unit is called as MapReduce. The fundamental approach that was done during the design of Hadoop was that hardware failures are common and should be handled by framework. Hadoop became popular because it has ability to store and process large amount of different types of data very quickly, especially from Internet of Things (IOT) and social-media. Hadoop provides good flexibility and scalability. Due to this we need not require to pre-process data before storing it, and also we can grow our system by adding nodes to handle big data. One of the major challenges before Hadoop is data security, and to overcome this issue Kerberos authentication protocol is one the major step taken[7].

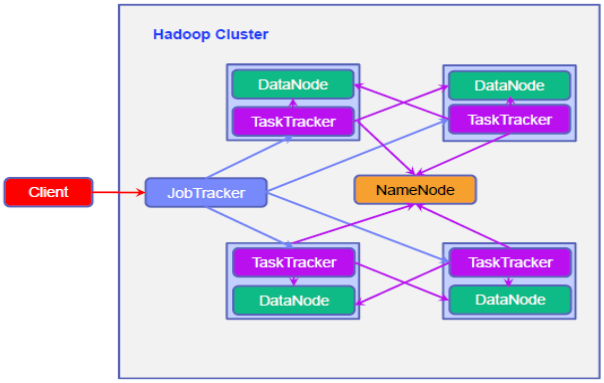

Fig 2: Hadoop Architecture

B) Cassandra:
The original authors of create this tool are Avinash Lakshman, Prashant Malik. The developer of this software is Apache software foundation. Like Hadoop, Apache Cassandra is a java based open source distributed database management system, providing a high value on performance.

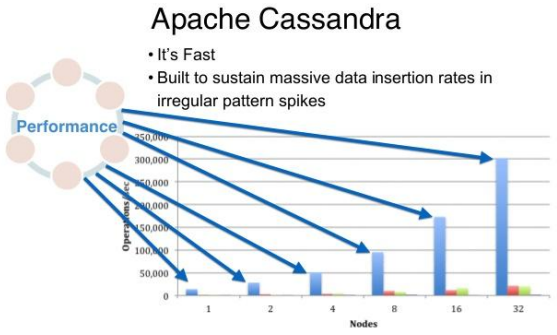

Fig 3: Overview of Cassandra

The authors initially developed Cassandra at facebook for facebook inbox search feature. In July 2008, facebook released Cassandra on Google code as an open source project [8].After research, Cassandra achieved highest throughput for maximum number of nodes. In Cassandra, every node has same role in computer cluster and hence occurrence of failure is much less. Main feature of Cassandra is it supports replication and is fault-tolerant.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)
AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad
Vol. 4, Special Issue 4, January 2017

With MapReduce support, Cassandra has Hadoop integration. Some of commercial companies that use Cassandra are DataStax, Impetus Technology and Instacluster, etc. [9].

C) MongoDB:

MongoDB is termed as document oriented cross-platform open-source data base program. MongoDB is written in three languages: C,C++ and Java. The major features of MongoDB are Ad hoc queries, Indexing, Replication, Load-balancing, File storage, Aggregation, Server-side JavaScript execution and Capped collection. The main units of MongoDB architecture are Programming Language accessibility, Management and graphical front-ends and Licensing.

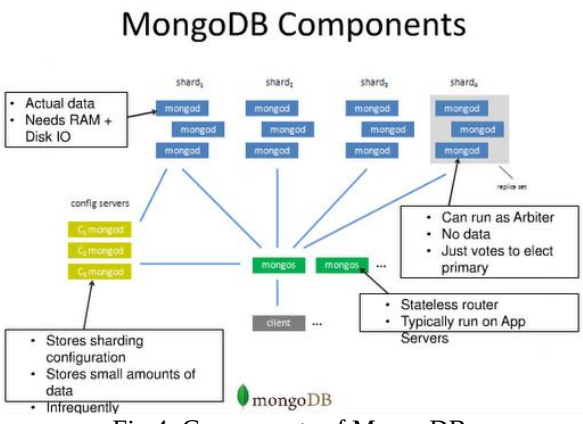

Fig 4: Components of MongoDB

A survey of January2017 claims that data has been stolen from tens of thousands of MongoDB installations. This bug may be due to MongoDB allows anyone to browse the databases, download them, write them and delete them[10].

D) Apache spark: Apache spark is open source software which provides clustered as computing format. This software was initially and originally developed at the University of California, Berkeley's AMP Lab, and was further donated to Apache Software Foundation.

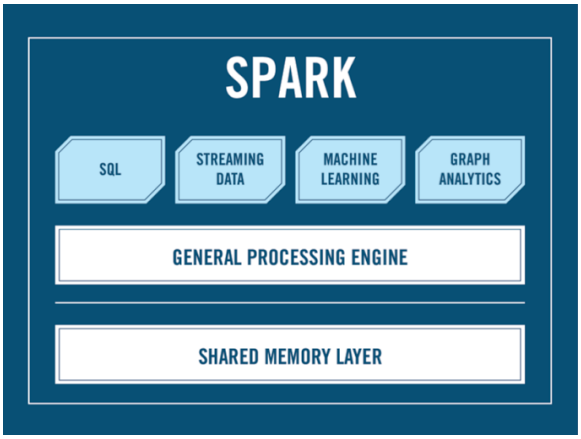

Fig 5: Apache Spark units

Spark provides data-parallies and fault-tolerance. One of the differences between MapReduce and Apache spark is that programs of MapReduce read input data from disk, while those of Apache Spark read input data from distributed shared memory. Apache Spark needs a distributed storage system and cluster manager. For distributed storage, spark can interface with Cassandra, Amazons3, HDFS or some sort of custom solution can be implemented. In case of cluster management, Standalone, Hadoop YARN Apache Mesos is supported by spark [11].

IV. CONCLUSION

Big data is not a fixed quantity.Technologies regarding big data are on high demand. Big data is emerging as one of the thrust field in Computer Engg.. Major V's of the big data are volume, velocity, variety, variability and veracity. Big data has many applications in different industries such as banking, media and entertainment,healthcare providers, insurance, transportation, etc. Some of the tools on big data are Hadoop, Cassandra, MongoDB and Apache Spark. These tools are all open sources and very much in use.

V. FUTURE SCOPE

Like other technologies, big data is influencing many industries.The study of big data is not a destination, but a journey. Some of the predictions done for the future of big data are as follows:

- Visualization tools for big data with more accuracy and efficiency will grow[12].
- The sources of rich media such as videos, audios, images, etc. will grow rapidly and will emerge as a key driver for big data analytics.
- Most of the customers by the year of 2019, will interact with cognitive computing based services.
- It is estimated by different surveys that data will grow upto 203 billion till the year 2020, than the current data which is 130 billion.
- Issues of security will arise and hence there will be need for different algorithms for security purpose.
- After studying the current data analysis, by the year 2019, large organizations will purchase 100% of external data.
- The domain of Internet of Things is growing rapidly and so is the the growth of data. Hence to deal with this, new advanced technologies need to be invented.
- Companies will become more aware about data accuracy.
- The era of big data is not just defining the generation of storage needs, but a new dictionary which has still not existed completely.

ACKNOWLEDGEMENT

We would like to express our sincere gratitute to our **Prof. Ashish Patil** for helping and guiding us throughout the

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

National Conference on Innovative Applications and Research in Computer Science and Engineering (NCIARCSE-2017)

AGTI's Dr. Daulatrao Aher College Engineering, Vidyanagar Extension, Karad

Vol. 4, Special Issue 4, January 2017

paper. We also take opportunity to thank Prof. Sayali Shinde and everyone who supported and encouraged us.

REFERENCES

[1] Data Science- Available at https://www.coursera.org/specializations/jhu-data-science
[2] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: a survey", Tsai et al. Journal of Big Data, A Springer open journal 2015.
[3] Feng Xia; Wei Wang; Teshome Megersa Bekele; Huan Liu, "Big Scholarly Data: A Survey", Volume: PP, Issue 99, January 2017.
[4] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, "Survey Paper on Big Data", Vol.5 (6)-2014
[5] Applications of big data available at https://www.simplilearn.com/big-data-applications-in-industries-article
[6] "Improving Media & Entertainment Performance with Big Data", Oracle Enterprise Architecture White Paper, February 2015.
[7] Apache HDFS. Available at http://hadoop.apache.org/hdfs
[8] ^Hamilton, James (July 12, 2008). "Facebook Releases Cassandra as Open Source". Retrieved 2009-06-04.
[9] Apache Cassandra. Available at https://Cassandra.apache.org/
[10] Apache MongoDB- Available at https://en.wikipedia.org/wiki/MongoDB
[11] Apache Spark. Available at https://en.wikipedia.org/wiki/Apache_Spark
[12] Yixian Zheng; Wenchao Wu; Yuanzhe Chen; Huamin Qu; Lional M.Ni,"Visual Analytics in Urban Computing: An Overview", Volume 2, Issue 3, July 2016.

BIOGRAPHIES

Prof. Ashish N. Patil has master degree in Computer Engineering and Research Scholar from BVDUCOE, Bharati Vidyapeeth Deemed University, Pune. He has 6.7 years of experience in teaching field. Currently working as Assistant Professor at AGTI's DACOE, Karad. His area of interest is Distributed System and Big Data Analytics

Ms. Anagha A. Jadhav Student of CSE Department, AGTI's DACOE, Karad