iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VII **Month of publication:** July 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue VII, July 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Big Data Analysis and Its Tools – A Review

Shobhit Srivastava[1], S.Venkatesan[2,] S.Amutha[3]
Department of Computer Science and Engineering
Dayananda Sagar College of Engineeirng, Bangalore, India

Abstract—The term Big Data is blooming very much these days, in this new era of large data sets, it has became a challenge for the organization for its storage and processing. The massive volume of data which is also referred to as metadata has very wide applications in our industry , this unstructured data have lead to development of new and advanced tools , as the traditional tools of structured data was unable to do so.This has led to the development of tools like Hadoop , Hive , Pig and the newly introduced Apache Spark which has been able to processes this unstructured data very swiftly. This paper presents an overview of Big Data its applications and the tools like Hadoop and Apache Spark.
Keywords— Hadoop, MapReduce, Hive, Pig, HBase, Spark

I. INTRODUCTION

Big data is the super abundance of data arising from various industries of which the major contributor is the internet. Companies like Facebook, Google, Twitter, Instagram, and You Tube are generating huge amounts of data that can be close to zeta bytes per day. The concept of big data is not very old but earlier the huge amount of data that was produced by the companies was very difficult to store. The major obstacles were the cost of storage and the cost of purchasing the data. These days this is not a huge problem. Facebook users send an average of 31.25 million messages and view 2.77 million videos every minute. Google alone processes 40,000 search results per second. Every minute 300 hours of video is uploaded to You Tube. Scientists, doctors, and government analysts require huge amount of data for their research and studies.

The file sharing websites and the video sharing websites such as You Tube have recorded a huge rise in data. According to IBM, 90% of the world's data was created in last 2 years alone. Big data is considered to be of high volume, high variety, and generated with high velocity. These are known as the three V's of big data. High volume suggests that the data generated today is of massive volume and expected to increase several folds in the near future, hence storing and processing the data is a real challenge for organizations. Big data is of high variety as majority of the data is unstructured and could be in the form of text, binary, audio, video and various other formats. Big data is generated with high velocity and is produced in real time growing so rapidly that traditional software tools are incapable of handling data generated at such velocity.

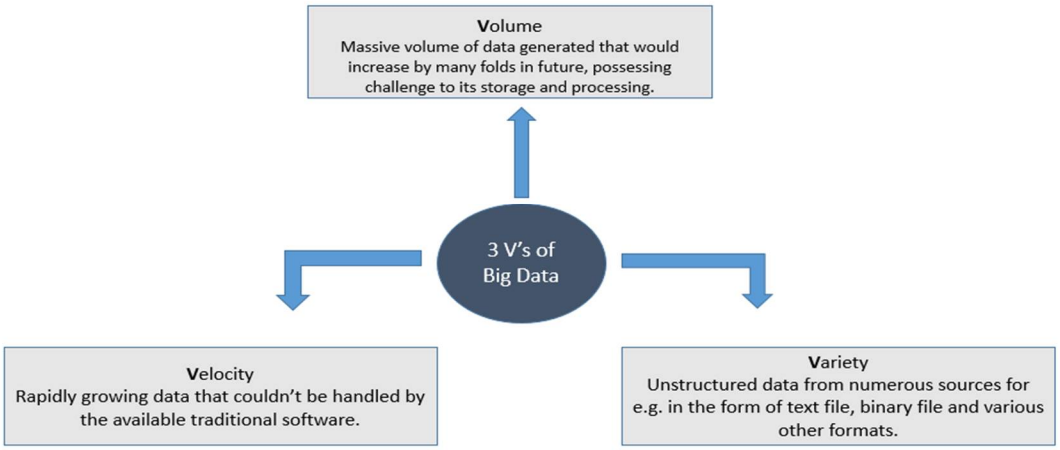


Fig 1. The 3 V's of Big Data

II. TYPES OF DATA GENERATED

The data generated can generally be categorized into three different types: structured, semi structured and unstructured data.

Volume 4 Issue VII, July 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Structured data is data that is highly organized and stored in the form of table. The collection of such data is called a schema and can be processed using relational database systems such as SQL or MySQL. Unstructured data is the largest component of data in big data. More than 90% of the data generated are unstructured. These kind of data does not follow any proper data model and consists of text, binary, audio, video and several other file types. Semi structured data is a hybrid of both structured and unstructured data. It does not possess a proper data model to store the data. The most common semi structured data formats are XML and JSON files.

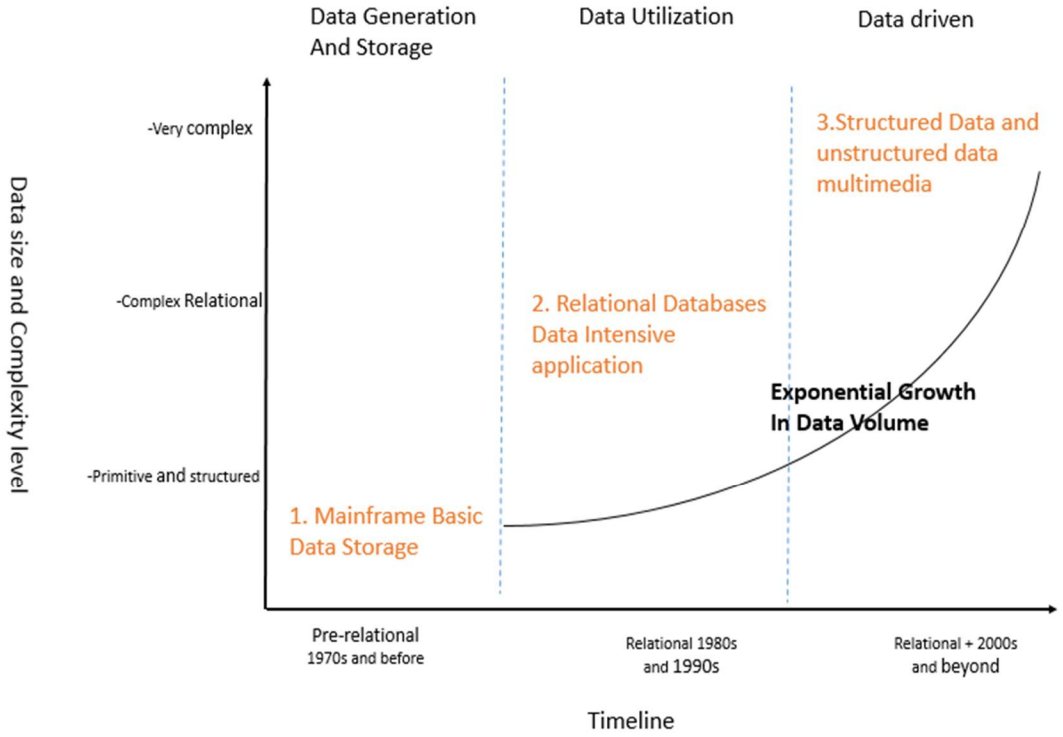


Fig. 2.Evolution of Data

III. SOURCES AND APPLICATIONS OF BIG DATA

Social and digital media has observed an enormous growth in the last five years. The reason behind this growth is that content is being generated for each person according to his or her taste. Big data plays a key role in the user specific content which enables the companies to provide better customer experience.

A. IT companies can leverage several sources of data such as application data, machine data, and social data.
B. The large number of wireless sensor networks which consists of several sensor nodes transmit a huge amount of data.
C. In field of medical sciences big data is been a revolution. The large amount of medical data generated helps in the predictive analysis of any disease like tumor in the human body.

IV. CHALLENGES FACED BY BIG DATA

Due to the huge growth in the rate of data explosion, big data faces various challenges. Heterogeneity refers to the property of big data that as the large volume of data generated increases with great velocity, the data contains both structured and unstructured data. This requires development of analytical tool which can process both types of data. The second challenge is of storage. The amount of data generated is so huge that we need massive storage capabilities to store this data. It is also found that already the data centers storing have grown to be the size equivalent to 6000 football fields. Another challenge is of combining the accumulated data**.** The large volume of data is stored in distributed machines and big data analytics needs tools that can process these data independently as well as merge the data analysis together. The challenges mentioned above can be addressed using the big data tools Apache Hadoop and Apace Spark.

Volume 4 Issue VII, July 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

V. HADOOP

Hadoop is framework developed by the Apache foundation to solve the storage and distributed processing problems of big data. Hadoop is architected using the Hadoop Common libraries that contains the java libraries required by other modules of Hadoop. The Yarn module acts like a cluster resource manager for Hadoop and separates the management and processing functions of Hadoop and provides a platform for other processing tools. The Hadoop Distributed File System well known as HDFS enables interaction among the nodes. The Map-Reduce framework is then used to solve the problem of big data analysis using the distributed file systems.

The input files in HDFS splits a huge file into chunks of blocks with each block having a definite memory allocation(usually 128mb).The HDFS stores usually replicates 3 copies of each block into different nodes hence the replication factor is referred to as 3,the nodes in which the blocks are stored are called as data nodes. HDFS works on a Master-Slave architecture, where the master node is called the name node and the data nodes acts like slaves. In case of any failure of any one of the data nodes, the master node asks the other data node having the copy of failed data node block to perform the block assigned to it.

A. MapReduce

Map reduce function separate the data into two functions which are:

1) *Mapper:* It accepts the input data and transform into data that exists in key/value pairs, the intermediate value is then given as an input to the reducer.
2) *Reducer:* The reducer accepts the key and the values corresponding to the key which is merged together to produce a smaller result which is then written to the HDFS.

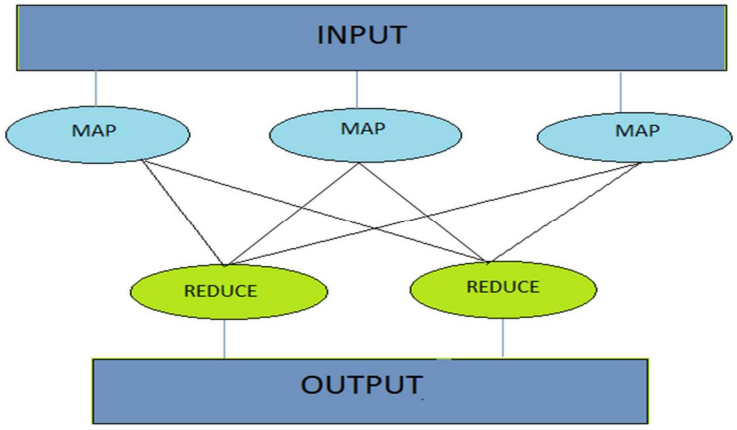


Fig.3. MapReduce

EXAMPLE: (input) <k1, v1> -> map -> <k2, v2> -> combine -> <k2, v2> -> reduce -> <k3, v3> (output)

B. Other components of Hadoop

1) *HBase :* It is an database management system ,which works on top of HDFS and provides a real time data analysis with quick read and write feature
2) *Pig:* Pig is a platform which is used for the analysis of large data sets which occurs in a high level language.
3) *Flume:* It a distributed and flexible service which is responsible for movement of large volume of data into HDFS.
4) *Sqoop:* It is an Apache project mainly used for transfer of data from relational databases such as MySQL to HDFS.
5) *Storm:* It is real time distributed system used for the processing of unbounded streams of data it is fast, scalable and reliable.
6) *Oozie:* It is a workflow scheduler system used to manage MapReduce jobs, it is a collection of nodes in directed acyclic graphs.
7) *ZooKeeper:* It is a fast, reliable and fault tolerant distributed coordination service which provides group services, synchronization services etc.

VI. APACHE SPARK

Spark is an open source, very powerful engine used for rapid processing of huge data sets on a large scale. It is similar to MapReduce and extends the functions of MapReduce very efficiently so that large workloads are managed very easily. It was

developed in 2009 in AMP Lab of UC Berkeley. Though both Hadoop and Spark are the frameworks used for big data the functionality of both the tools are quite different. Hadoop is a distributive framework which distributes huge data into different nodes within the cluster, while apache is data processing tool which processes the distributed data. Spark is also way faster than MapReduce because of the processing methods Map Reduce processes the data in steps while spark processes the data as a whole.

A. Features of apache spark

1) *Speed:* Spark runs 100 times faster Hadoop Map Reduce programs in memory. It also supports DAG execution engine for cyclic data flows and in memory processing.
2) *Easier:* The programs can be written with ease and it provides great support of platforms to write programs like java, python, Scala and R.
3) *Generic:* In support to Map Reduce programs, it provides an extensive support of platforms such as SQL queries, machine Learning algorithms and graphs.
4) *Integration:* Spark can run on Hadoop CRM Yarn and can read any existing Hadoop files.

B. Apache spark tools

Spark has a great support of tools which are SparkSQL, Spark Streaming, GraphX, and MLlib.

Apache Spark Tools			
Spark SQL	Spark Streaming	MLlib	GraphX
SparkSQL is used for querying structured data as Resilient Distributed Data(RDD) with APIs supported for Java,Python,R.	Spark Streaming hold the capability to read the data from sources like HDFS,Twitter,Kafta and use it for stream processing in Java or Scala.	MLlib comprises the library for various algorithms of Machine Learning like Regression,Clustering, Classification	GraphX is used to view data as both graph and collection and join the graphs using RDDs

Table 1. Tools used in Apache Spark

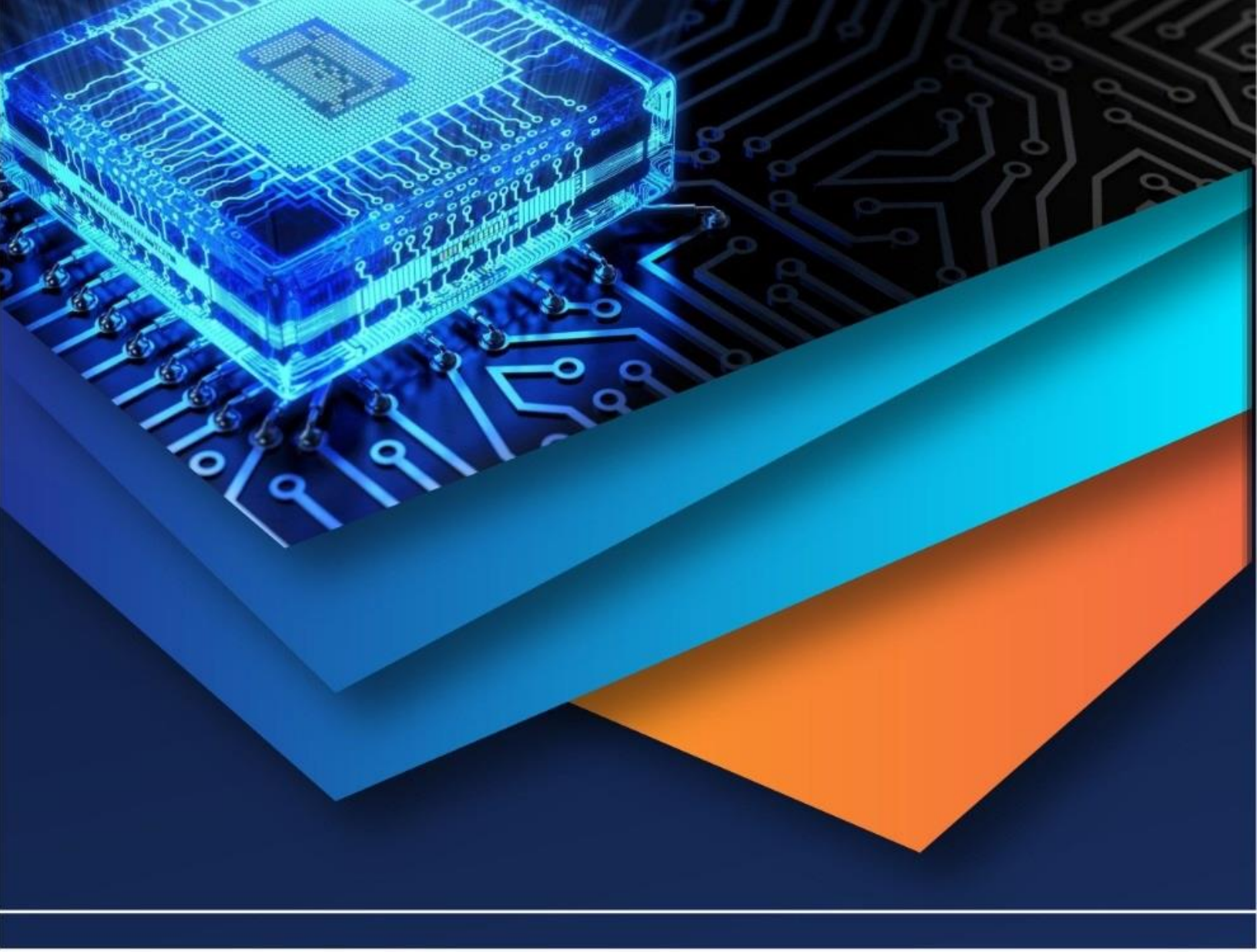
VII. CONCLUSION

In this review paper the huge challenges in front of Big Data analytics and related tool is reviewed and how Hadoop, Map Reduce embedded in processing data and also the application of Apache Spark which provided a better computational speed than other tools.

But we just cannot rely upon the current tools be used because the data that will be generated in future would be more than double as of now , there is a need of continuous enhancement to meet the challenges especially storage and processing needs.

REFERENCES

[1] Shilpa, Manjit Kaur, "BIG Data and Methodology-A review, ijarcsse, Volume 3, Issue 10, October 2013.
[2] Jeffrey Dean and Sanjay Ghemawat ," MapReduce: Simplifed Data Processing on Large Clusters, Google Inc.
[3] Hortonworks ,"Apache Hadoop The *Big Data* Refinery Whitepaper"
[4] Aditya B. Patel, Manashvi Birla, Ushma Nair ,(6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce"
[5] Han Hu, Yongyang Nen, Tat Seng Chua, Xuelong Li," Towards Scalable System for Big Data Analytics: A Technology Tutorial", IEEE Access, Volume 2, Page No 653, June 2014.
[6] Big Data And Hadoop: A Review Paper by Rahul Beakta CSE Deptt., Baddi University of Emerging Sciences & Technology, Baddi, India.
[7] Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A ReviewPaper on Big Data and Hadoop" in International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014.
[8] D.Fisher, R.Deline, M.Czerwinski and S. Drucker,"Interaction with big data analytics", Volume 19, No.3, May 2012.
[9] Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013),"Big Data: A Review"
[10] http://www.tutorialspoint.com/Hadoop
[11] http://www.en.wikipedia.com/Hadoop
[12] http://www.apache.org



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY