ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.6

Big Data: An Overview of Features, Tools, Techniques and Applications

M. Sowmya[1], N. Sravanthi[2]
Assistant Professor[1, 2]
Department of CSE
Aurora's Engineering College, Bhongir, India

Abstract:
Big Data is the huge amount of data that cannot be processed by making use of conventional methods of data processing. Due to extensive usage of many computing devices such as smart phones, laptops, wearable computing devices; the data processing over the internet has overreach more than the modern computers can handle. This paper provides a flying introduction to the Big data technology and its influence in the contemporary world. This paper addresses various properties and issues that need to be emphasized to present the full influence of big data. The tools used in big data technology are also explored in detail and Hadoop is the platform used in Big Data. It is an open-source framework that permits to store and process big data in a distributed environment across clusters of computers using simple programming models. Lastly, this paper also discuss about the applications of big data technology in detail.

Keywords-Big Data, Hadoop, HDFS, Horton works, Map Reduce.

I. INTRODUCTION

Big Data is a term used to describe collection of data that is huge in size and yet growing exponentially with time. The term Big Data is being progressively used practically extensively on the planet – online and offline. It is not associated to computers only, but part of almost all other technologies and fields of studies and businesses. Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. In short, such a data is so large and complex that none of the traditional data management tools can store it or process it efficiently. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.

II. FEATURES OF BIG DATA

Volume, Velocity, Value, Veracity and Variety (5 V's)

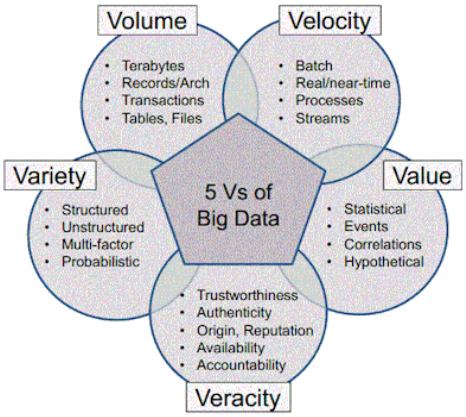

Figure .1. Features of Big Data

Volume:
Big data implies enormous volumes of data. It used to be employees created data. Now that data is generated by machines, networks and human interaction on systems like social media the volume of data to be analyzed is massive.

Value:
Value measures the usefulness of data in making decisions. User can run certain queries against the data stored and thus can deduct important results from the filter data obtained and rank it according to the dimensions they require.

Variety:
Variety refers to the many sources and types of data both structured and unstructured. We used to store data from sources like spreadsheets and databases. Now data comes in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. This variety of unstructured data creates problems for storage, mining and analyzing data.

Velocity:
Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. This real-time data can help researchers and businesses make valuable decisions that provide strategic competitive advantages and ROI if you are able to handle the velocity.

Veracity:
Big Data Veracity refers to the biases, noise and abnormality in data. Veracity in data analysis is the biggest challenge when compares to things like volume and velocity.

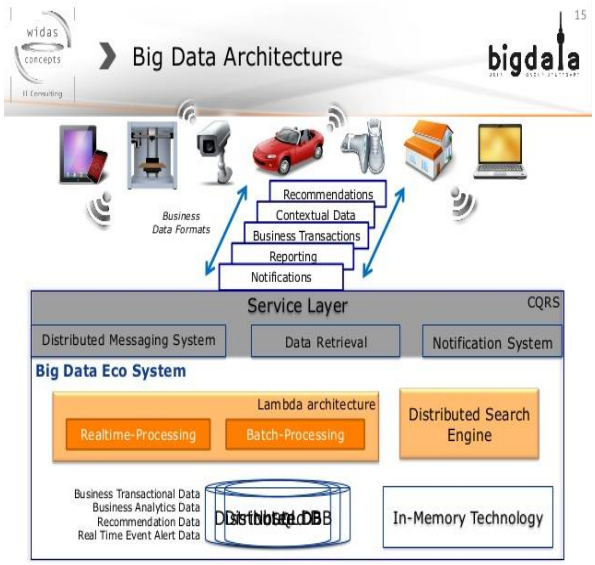

Figure .2. Architecture of Big Data

III. TOOLSUSED IN BIG DATA

1. Apache Hadoop: framework that can effectively store large amount of data in a cluster. This framework runs in parallel on a cluster and has an ability to allow us to process data across all nodes. Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster. This also replicates data in a cluster thus providing high availability.

2. Microsoft HDInsight by Apache Hadoop which is available as a ser**:** It is a Big Data solution from Microsoft powered vice in the cloud. HDInsight uses Windows Azure Blob storage as the default file system. This also provides high availability with low cost.

A.NoSQL: Stands for Not Only SQL. While the traditional SQL can be effectively used to handle large amount of structured data, we need NoSQL to handle unstructured Apache Hadoop is a java based free software data. NoSQL databases store unstructured data with no particular schema. Each row can have its own set of column values. NoSQL gives better performance in storing massive amount of data. There are many open-source NoSQL DBs available to analyse big Data.

B.Hive: This is a distributed data management for Hadoop. This supports SQL-like query option HiveSQLin short HSQL to access big data. This can be primarily used for Data mining purpose. This runs on top of Hadoop.

C.Sqoop: This is a tool that connects Hadoop with various relational databases to transfer data. This can be effectively used to transfer structured data to Hadoop or Hive.

D.PolyBase: This works on top of SQL Server 2012 Parallel Data Warehouse (PDW) and is used to access data stored in PDW. PDW is a data warehousing appliance built for processing any volume of relational data and provides integration with Hadoop allowing us to access non-relational data as well.

E.Big data in EXCEL:As many people are comfortable in doing analysis in EXCEL, a popular tool from Microsoft, you can also connect data stored in Hadoop using EXCEL 2013. Horton works, which is primarily working in providing Enterprise Apache Hadoop, provides an option to access big data stored in their Hadoop platform using EXCEL 2013. You can use Power View feature of EXCEL 2013 to easily summarise the data.

F.Facebook has developed and recently open-sourced its Query engine (SQL-on-Hadoop) named presto which is built to handle petabytes of data. Unlike Hive, Presto does not depend on MapReduce technique and can quickly retrieve data.

IV.TECHNOLOGIES FOR BIG DATA HANDLING

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business. To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in real-time and can protect data privacy and security. There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology.

A. Operational Big Data

This includes systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored. No SQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.

B. Analytical Big Data

This includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data. MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low-end machines. The Big Data handling techniques and tools include Hadoop, Map Reduce, and Big Table. Out of these, Hadoop is one of the most widely used technologies.

Hadoop

Hadoop is an Apache open source framework which is written in java. High volumes of data, in any structure, are processed by Hadoop. Hadoop allows distributed storage and distributed processing for very large data sets. The main components of Hadoop are:

1. Hadoop distributed file system (HDFS)
2. MapReduce

The architecture of Hadoop is shown in the figure. Hadoop has three layers. The two major layers are MapReduce and HDFS.

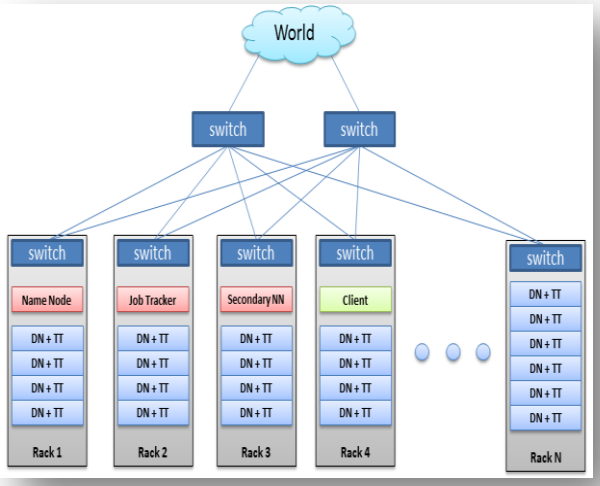

Figure.3. Hadoop Architecture

HDFS The Hadoop Distributed File System is a scalable and reliable distributed storage system that aggregates the storage of every node in a Hadoop cluster into a single global file system. HDFS stores individual files in large blocks, allowing it to efficiently store very large or numerous files across multiple machines and access individual chunks of data in parallel, without needing to read the entire file into a single computer's memory. Reliability is achieved by replicating the data across multiple hosts, with each block of data being stored, by default, on three separate computers. If an individual node fails, the data remains available and an additional copy of any blocks it holds may be made on new machines to protect against future failures.

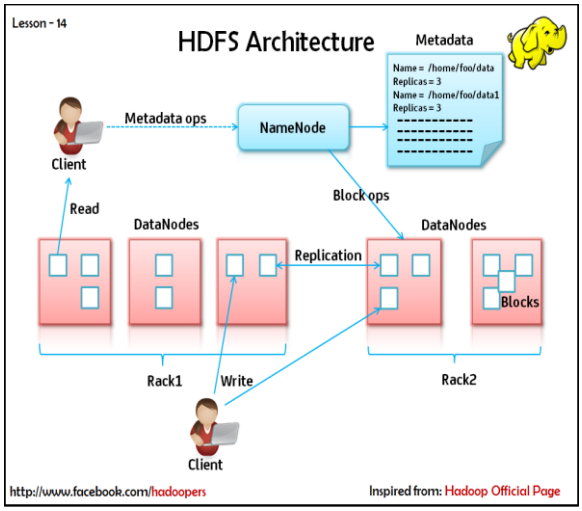

Figure .4. Architecture for Hdfs

This approach allows HDFS to dependably store massive amounts of data. For instance, in late 2012, the Apache Hadoop clusters at Yahoo had grown to hold over 350 petabytes (PB) of data across 40,000+ servers. Once data has been loaded into HDFS, we can begin to process it with MapReduce. **Map Reduce** Map Reduce is the programming model that allows Hadoop to efficiently process large amounts of data. MapReduce breaks large data processing problems into multiple steps, namely a set of Maps and Reduces, that can each be worked on at the same time (in parallel) on multiple computers. MapReduce is designed to work with of HDFS. Apache Hadoop automatically optimizes the execution of MapReduce programs so that a given Map or Reduce step is run on the HDFS node that contains locally the blocks of data required to complete the step.Map Reduce has proven itself in its ability to allow data processing problems that once required many hours to complete on very expensive computers to be written as programs that run in minutes on a handful of rather inexpensive machines. And, while MapReduce can require a shift in thinking on the part of developers, many problems not traditionally solved using the method are easily expressed as MapReduce programs.

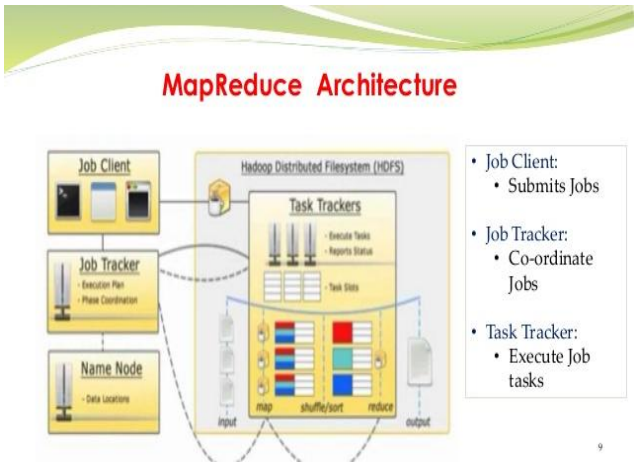

Figure.5. Architecture of Map Reduce

Hadoop **MapReduce** is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The term MapReduce actually refers to the following two different tasks that Hadoop programs perform:

- **The Map Task:** This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples (key/value pairs).
- **The Reduce Task:** This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.

Typically both the input and the output are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

The MapReduce framework consists of a single master **JobTracker** and one slave **TaskTracker** per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing thefailed tasks. The slaves Task Tracker execute the tasks as directed by the master and provide task-status information to the master periodically.

The JobTracker is a single point of failure for the Hadoop MapReduce service which means if Job Tracker goes down, all running jobs are halted.

V. APPLICATIONS OF BIG DATA

As per the market strategy, companies who miss big data opportunities of today will miss the next frontier of innovation, competition, and productivity. Big Data tools and Technologies help the companies to interpret the huge amount of data very

faster which helps to boost production efficiency and also to develop new data-driven products and services. So, Big data applications are creating a new era in every industry. The below figure shows examples of Big Data applications in different fields.

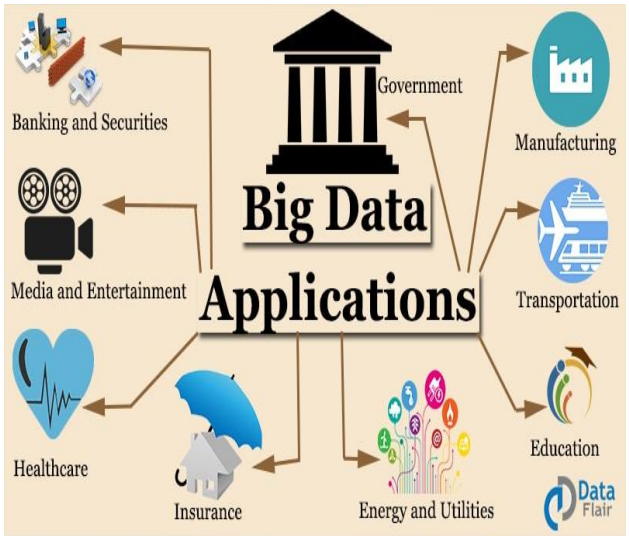

Figure.6. Applications of Big Data

VI. FUTURE SCOPE

The Future of Hadoop has "crossed the chasm" from a framework for early adopters, developers and technology enthusiasts to a strategic data platform embraced by innovative CTOs and CIOs across mainstream enterprises. People who want to improve the performance of their companies and unlock new business opportunities, realize that including Apache Hadoop as a deeply integrated supplement to their current data architecture offers the fastest path to reaching their goals while maximizing their existing investments,. Going forward, Horton works and the Apache Hadoop community will continue to focus on increasing the ease with which enterprises deploy and use Hadoop, and on increasing the platform's interoperability with the broader data ecosystem. This includes making certain it is reliable and stable and more importantly, ready for all and any enterprise workloads.

VII. CONCLUSION

As there are huge volumes of data that are produced every day, so such large size of data it becomes very challenging to achieve effective processing using the existing traditional techniques. Big data is data that exceeds the processing capacity of conventional database systems. In this paper fundamental concepts about Big Data are presented. These concepts include Big Data characteristics, tools, techniques and applications for handling big data.

VIII. REFERENCES

[1]. MrigankMridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N "Analysis of Bigdata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014" 27

[2]. www.Wikibon.org

[3]. A.Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices." Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.]

[4]. Golfarelli, M., &Rizzi, S. (2009). Data warehouse design: modern principles and methodologies. Columbus: McGraw-Hill

[5]. https://www.progress.com

[6]. M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014

[7]. Apache Hadoop (2013). HDFS Architecture Guide [Online]. Available: https://hadoop. apache.org/docs/r1.2.1/hdfs_design.ht

[8]. Almeida, F., and Calistru, C, "The Main Challenges and Issues of Big Data Management", International Journal of Research Studies in Computing, 2(1), 2013, pp. 11-20.

[9]. Apache Hadoop, http://hadoop.apache.org/

[10]. Amrit pal, PinkiAggrawal, Kunal Jain, Sanjay Aggrawal *"A Performance Analysis of MapReduce Task with Large Number of Files Dataset in Big Data using Hadoop"* Forth International Conference on Communication Systems and Network Technologies, 2014.

[11]. A.Ghoting, P. Kambadur, E. P. D.Pednault, and R. Kannan. Nimble: a toolkit for the implementation of parallel data mining and machine learning algorithms on mapreduce. In SIGKDD, pages 334–342, 2011.

[12]. W. Lang and J. M. Patel. Energy management for mapreduce clusters. PVLDB, 3(1):129–139, 2010.

[13]. Rahm, E., & Hai Do, H. (2000). Data cleaning: problems and current approaches. Bulletin of the Technical Committee on Data Engineering, 23(4), 3-13.),

[14]. Apache Hadoop (2013). HDFS Architecture Guide [Online]. Available: https://hadoop. apache.org/ docs/r1.2.1/ hdfs_ design.htm

[15]. K. E. Martin, "Ethical Issues in the Big Data Industry." *MIS Quarterly Executive*, vol. 14, 2015. pp. 67-85.