ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

A Study on Big Data Technologies

V. Srilakshmi[1.], V.Lakshmi Chetana[2.] , T.P.Ann Thabitha[3.]

Assistant Professor, Dept. of CSE, DVR & Dr HS MIC College of Technology, Kanchikacherla, AP, India[1]

Assistant Professor, Dept. of CSE, DVR & Dr HS MIC College of Technology, Kanchikacherla, AP, India[2]

Assistant Professor, Dept. of CSE, DVR & Dr HS MIC College of Technology, Kanchikacherla, AP, India[3]

ABSTRACT: The guarantee of information driven decision making is currently being perceived extensively, and it is creating eagerness for the idea of Big Data. Propels in data innovation and its widespread usage in the areas of business, health, engineering and scientific studies result in data/information impact. Knowledge discovery and decision making from such quickly developing voluminous data is a challenging task for data processing and organization, which is a rising pattern known as Big Data Computing. It demands large storage space and computing time for data curation and processing .This paper showcases about Big Data Dimensions, Hadoop Architecture, Hadoop Technologies, the programming model and architecture of Twister, an enhanced MapReduce runtime that supports Iterative MapReduce computations easy.

KEYWORDS: Big Data, MapReduce, Hadoop, Hadoop Ecosystem, Twister,HDFS

I. INTRODUCTION

We are immersed with a surge of data today. Data is being gathered at exceptional scale with the increase in the range of application areas. Decisions that beforehand depended on mystery, or on carefully developed models of reality, can now be made based on the data itself. Such Big Data analysis now drives about each part of our current society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences[2].
Consistently, we make 2.5 quintillion bytes of data — so much that 90% of the data on the planet today has been made in the most recent two years alone. This data originates from everywhere: sensors used to accumulate climate data, posts on social media sites, digital pictures and videos, purchase transaction records, and mobile GPS signals(global positioningsystem that gives location and time information in all weather conditions, anywhere on or close to the Earth) etc... This data is big data. Big data is a term, used to portray a huge volume of both structured and unstructured data that is so expansive and is hard to process using traditional database and software techniques. In most endeavor situations the volume of data is too big or it moves too quickly or it surpasses current processing limit. In spite of these issues, big data can possibly help organizations enhance operations and make quicker, more intelligent decisions. Big data is data that surpasses the processing capacity of conventional database systems. The data is too big, moves too quick, or doesn't fit the structures of your database architectures. To pick up worth from this data, we must choose analternative way to process it. Challenges include analysis, capture, data acquisition, search, sharing, storage, transfer and visualization and information privacy [2].

A) Big Data: Definition

Big Data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to capture, manage, process or analyze using traditional database and software techniques.

B) 6 V's of Big Data:

Big Data is characterized by the key dimensions - Volume, Velocity, and Variety.If the results of the Big Data is critical,then these additional 3 V's - Veracity, Value and Visibility are also considered to describe the nature of data.

Volume: The volume is how much data we have – what used to be measured in Gigabytes is now measured in Zettabytes or even Yottabytes .For example, YouTube users upload 48 hours of new video every minute of every day. By 2020, we will have created 40 Zeta Bytes of data which is 43 trillion Gigabytes.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Velocity: In this context, the speed at which the data is created and prepared to meet the requests and the challenges that lie in the way of development and advancement. One example is the real-time profiling of internet display adverts that are customized according to your usage pattern.
Variety: It implies diverse types of data to use for analysis, for example, structured like relational databases, semi structured like XML and unstructured like video, text.
Veracity: Veracity is all about making sure the data is accurate.
Visibility: Visibility is critical in today's world.Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spread sheets and reports.
Value: Value is the ultimate attribute of Big Data .Every organization after addressing all V's, wants to get value from the data.

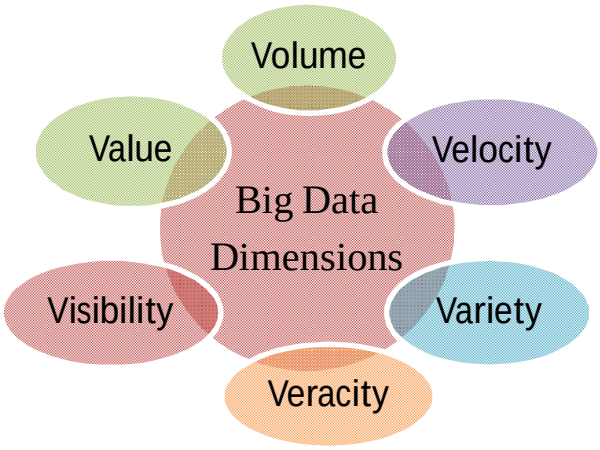

Figure 1: Big Data Dimensions

II.HADOOP: SOLUTION FOR BIG DATA PROCESSING

Hadoop is an open-source java based programming framework that supports the processing of data-intensive distributed applications.It was created by **Doug Cutting** and Mike Cafarella in 2005. It was originally developed to support distribution for the Nutch search engine project. **Doug**, who was working at Yahoo! at that time and is now Chief Architect of Cloudera, named the project after his son's toy elephant. One of the goals of Hadoop is to run applications on large number of clusters. The cluster is composed of a single master and multiple worker nodes [4].
Hadoop leverages the programming model of map/reduce. It is optimized for processing larger data sets. MapReduce is typically used to do distributed computing on clusters of computer. A cluster had many "nodes," where each node is a computer in a cluster. The goal of map reduce is to break huge data sets into smaller pieces, distribute those pieces to various slave or worker nodes in the cluster, and process the data in parallel. Hadoop leverages a distributed file system to store the data on various nodes.

A) HADOOP ARCHITECTURE

Apache Hadoop is an open source framework used to store and analyses big data which is present in Hadoop cluster. Hadoop always runs on a cluster means on homogeneous environment. Moreover homogeneous environment means all the systems which are present in cluster, their all components must be same in terms of RAM, CPU etc. Primarily Hadoop has two major components:
1. HDFS (Hadoop distributed File system)
2. Map Reduce

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

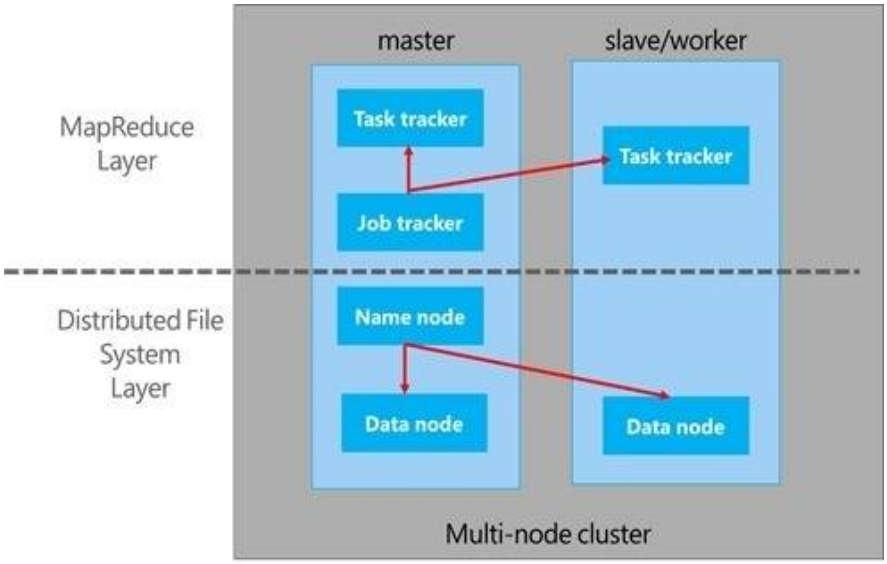

Figure 2:Hadoop Architecture

Mapper Function: It divides the problem into smaller sub-problems. A master node distributes the work to worker nodes. The worker node just does one thing and returns the work back to the master node.
Reducer Function: Once the master gets the work from the worker nodes, the reduce function takes over and combines all the work. By combining the work some answer and ultimately output is produced.

A) Hadoop Distributed File System

HDFS is a file system based on the master slave architecture. It breaks the large files into default 64 MB blocks and store them into large cluster in a very efficient way[4]. In this architecture there are three basic nodes in a Hadoop cluster, name node, data node and secondary name node.

- **Name Node** is the master node which controls all the data nodes and it contains Meta data. It manages all the file operations like read, write etc.
- **Data nodes** are the slave nodes present in Hadoop cluster. All the file operations performed on these nodes and data is actually stored on these nodes as decided by name nodes.
- **Secondary name node** is the back up of name node. As name node is the master node, it becomes very important to take its backup. If the name node fails, secondary name node will be used as name node.

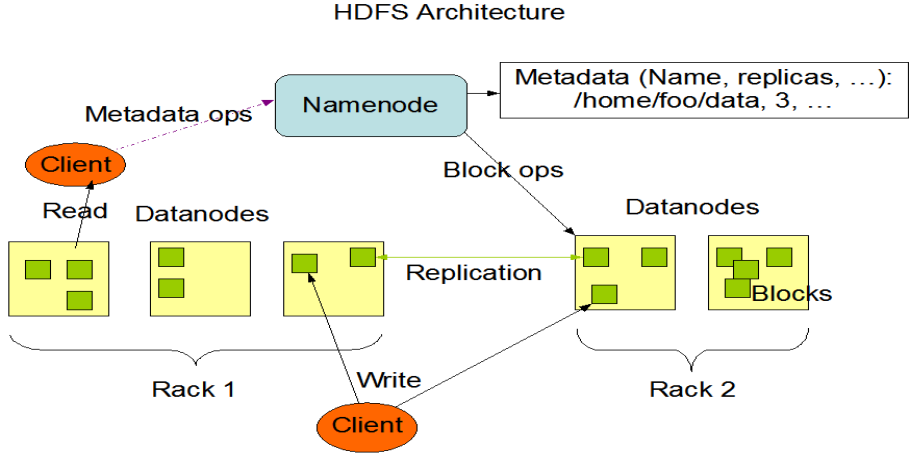

Figure 3:HDFS Architecture

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

B) Map Reduce

Map reduce is a core technology developed by Google and Hadoop implements it in an open source environment. It is a very important component of Hadoop and very helpful in dealing with big data. The basic meaning of map reduce is dividing the large task into smaller chunks and then deal with them accordingly [4].

The "map" in MapReduce

1. There is a master node and many slave nodes.
2. The master node takes the input, divides it into smaller sub-problems, and distributes the input to worker or slave nodes. Worker node may do this again in turn, leading to a multi-level tree structure.
3. The worker/slave nodes processes the data into a smaller problem, and passes the answer back to its master node.
4. Each mapping operation is independent of the others, all maps can be performed in parallel.

The "reduce" in MapReduce

1. The master node then collects the answers from the worker or slave
2. Nodes. It then aggregates the answers and creates the needed output, which is the answer to the problem it was originally trying to solve.
3. Reducers can also preform the reduction phase in parallel. That is how the system can process petabytes in a matter of hours.

Map Reduce has four core components: input, mapper, reducer, and output.

Input: Input means that data which gets for processing and it is divided into further smaller chunks which are further allocated to the mappers.

Mapper: Mappers are the individuals that are assigned with the smallest unit of work for some processing.
Reducer: Mappers output become input for the reducers to aggregate the data in form of final output.
Output Reducers' jobs are finally collected in the form of aggrgated output.

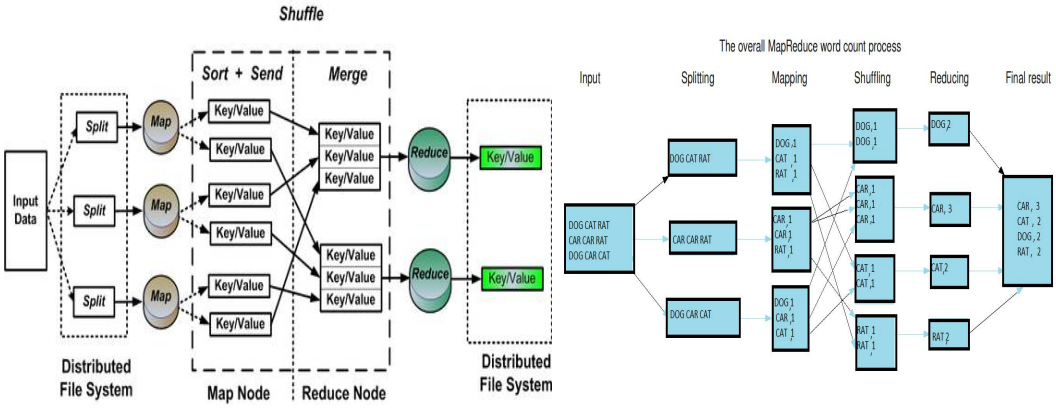

Figure 4:MapReduce Architecture and Word Count using MapReduce

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

III. HADOOP TECHNOLOGIES

Hadoop ecosystem has a wide range of technologies. Some of them are mentioned below [1]

Table:Hadoop Technologies

Hadoop Technologies	Description
Apache PIG	PIG is a scripting language which is used for writing programs for processing large data sets present in the Hadoop cluster. This language is known as PIG Latin. This language provides various operators using which programmers can develop their own functions for reading, writing, and processing data. Programmers need to write scripts to analyze data. All the scripts are converted to Map and Reduce tasks. PIG engine takes those scripts as input and convert them into map reduce jobs and then execute those jobs. It was created by Yahoo and now it is under Apache software foundation.
Apache HBase	HBase is non-relational or column oriented database which runs on the top of HDFS (Hadoop distributed file system).It also comes under Apache Software Foundation. It is open source and written in Java. Apache HBase allows reading and writing data on HDFS (Hadoop distributed file system) on the real-time scenario. HBase can deal with petabytes of data.
Apache Hive	Hive is SQL like language called HiveQL. It was developed by Facebook, but now it is owned by Apache Software Foundation and is used by many companies for data analysis. Moreover, it is a data warehouse infrastructure which provides all these functionalities. Hive allows querying of data from HDFS (Hadoop distributed file system) and these queries are converted into map reduce jobs.
Apache Sqoop	Sqoop is an application which helps in moving data in and out from any Relational database management system to Hadoop. So it is data management application built on the top of Hadoop by Apache Software Foundation.
Apache Flume	Flume is a data ingestion mechanism for collecting, aggregating and transferring large volumes of streaming data such as web log files, events etc.., from various sources to a centralized data store. Basic components of Apache Flume are source, channel, sink, agent, interceptor etc.
Apache Zookeeper	Zookeeper is open source project by Apache which provides distributed co-ordination service to manage large number of hosts. This framework was initially used by "Yahoo!" for accessing their applications in easy and robust way.

IV. ITERATIVE MAPREDUCE WITH TWISTER

Twister is a distributed in-memory MapReduce runtime optimized for iterative MapReduce computations. It reads data from local disks of the worker nodes and handles the intermediate data in distributed memory of the worker nodes. All communication and data transfers are performed via a publish /subscribe messaging infrastructure. It supports long running map/reduce tasks, which can be used in "configure once and use many times" approach. In addition it provides programming extensions to MapReduce with "broadcast" and "scatter" type data transfers. These improvements allow Twister to support iterative MapReduce computations highly efficiently compared to other MapReduce runtimes [3].
Twister provides the following features to support MapReduce computations.

1. Distinction on static and variable data
2. Configurable long running (cacheable) map/reduce tasks
3. Pub/sub messaging based communication/data transfers
4. Efficient support for Iterative MapReduce computations (extremely faster than Hadoop)
5. Combine phase to collect all reduce outputs
6. Data access via local disks
7. Lightweight (~5600 lines of Java code)
8. Support for typical MapReduce computations

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

9. Tools to manage data

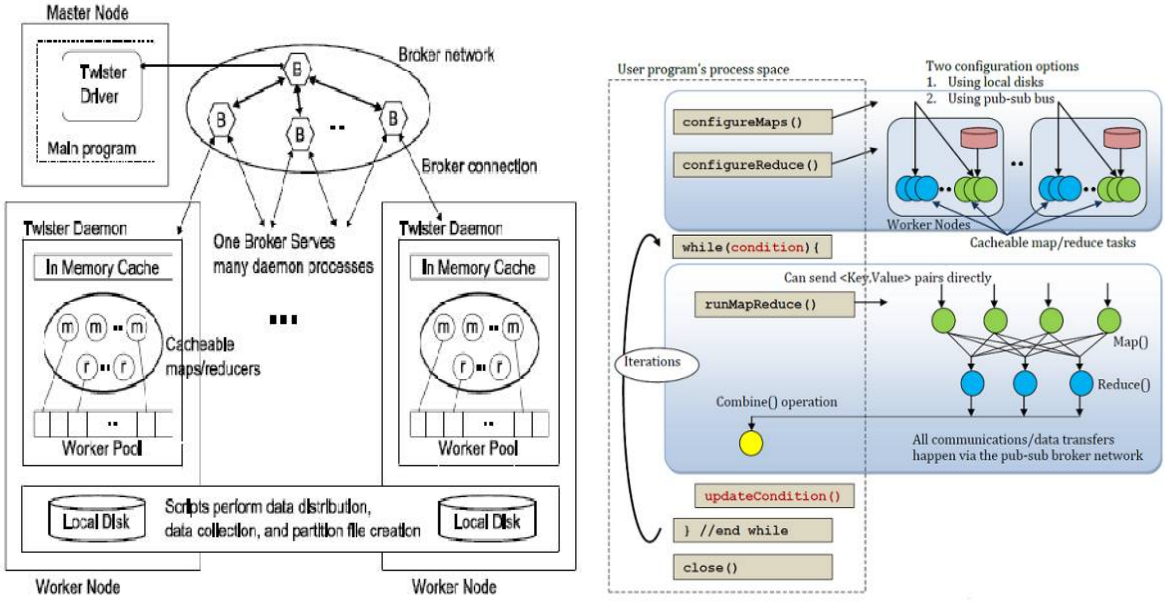

Figure 5: Twister Architecture and Programming model

V.CONCLUSION

We are in the era of Big Data where the world is capable of generating data in terabytes and petabytes every day. In 2012 2.5 Exabyte data was generated on each day. Social networking business such as Facebook, Twitter, Tumblr, Google have completely changed the view of data. This data has a great impact on the business. It is very helpful to make the business intelligent if used properly by extracting information from it. This paper gives a brief overview of Big Data and Hadoop technologies.However, big data technologies are still in the stage of development, facing opportunities and challenges. Today lot of platforms and challenging situations are there ahead of this world to make new technologies for both processing and storage.

REFERENCES

1. Jaskaran Singh, Varun Singla "Big Data: Tools and Technologies in Big Data". In International Journal of Computer Applications (0975 – 8887) Volume 112 – No 15, February 2015.
2. Raghavendra Kune, Pramod Kumar Konugurthi, Arun Agarwal, Raghavendra Rao Chillarige, and Rajkumar Buyya "The Anatomy of Big Data Computing".
3. Jaliya Ekanayake, Hui Li, Bingjing Zhang, Thilina Gunarathne, Seung-Hee Bae, Judy Qiu, Geoffrey Fox, Twister: A Runtime for Iterative MapReduce," The First International Workshop on MapReduce and its Applications (MAPREDUCE'10) - HPDC2010.
4. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar "A Review Paper on Big Data and Hadoop". In International Journal of Scientific and Research Publications, Volume 4, Issue 10, October 2014 ISSN 2250-3153.
5. Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}
6. Umasri.M.L, Shyamalagowri.D ,Suresh Kumar.S "Mining Big Data:- Current status and forecast to the future" Volume 4, Issue 1, January 2014 ISSN: 2277 128X
7. Albert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) 15–20 DEC 2012
8. Shadi Ibrahim⋆ _ Hai Jin _ Lu Lu "Handling Partitioning Skew in MapReduce using *LEEN*" ACM 51 (2008) 107–113
9. Sanjeev Dhawan1, Sanjay Rathee2 "Big Data Analytics using Hadoop Components like Pig and Hive" AIJRSTEM ISSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-3629
10. J. Dean, S. Ghemawat, MapReduce: Simplified data processing on large cluster, Communications of the ACM, 51(1) (2008) 107-113.
11. Jaskaran Singh , Varun Singla "Big Data: Tools and Technologies in Big Data "International Journal of Computer Applications (0975 – 8887), International Journal of Computer Applications (0975 – 8887).
12. Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money. "Big Data: Issues and Challenges Moving Forward". In IEEE, 46th Hawaii International Conference on System Sciences, pages 995-1004, 2013.