INTRODUCTION TO HADOOP ARCHITECTURE AND INSTALLATION ON UBUNTU

Anjali Deore[1], Bhagyashree More[2], Kaveri Sonawane[3], Jyoti Kharat[4]

[1]*Assistant Professor, Computer Engineering, Sandip Foundation Nashik, Maharashtra, India*
[2]*Assistant Professor, Computer Engineering, Wadia college of engineering Pune, Maharashtra, India*
[3]*Assistant Professor, Computer Engineering, Sandip Foundation Nashik, Maharashtra, India*
[4]*Assistant Professor, Computer Engineering, RSSOER, JSPM, NTC, Pune, Maharashtra, India*

Abstract

Currently the huge amount of data which is being generated through an online system, social networking sites, mobile which are not possible to manage with traditional methodologies and database system. Big Data' describes techniques to extract data from various sources, store, distribute on a different system for processing, manage, collect the result and analyze large-sized data sets with high-velocity. Structured, unstructured or semi-structured are belongs to bigdata, resulting in incapability of conventional data management methods. Hadoop is fundamental infrastructural software for storing and processing large data. It is an open source product under Apache and it is very popular. This paper illustrates the Hadoop architecture, HDFS architecture and installation procedure of Hadoop on a ubuntu system..

Keywords: *Hadoop, Bigdata, HDFS, MapReduce, Name Node, Data Node, Job Tracker.*

----------***----------

1. INTRODUCTION

1.1 Bigdata

Bigdata is used to describe a massive volume of both structured and unstructured data Structured data may include relational databases like MySQL. Unstructured data may include text files in .doc, .pdf formats as well as media files that are so large and it is difficult to process using traditional database and software technique.In 2010, a data analyst describes data in 3 V's:

a) **Volume:** Bigdata has a large volume which defines a size of data in Terabytes, Petabytes.Its size continues increasing, but not the percent of data that our tools can process.

b) **Variety:** Variety defines the Types of data. The data coming from various sources can be structured or unstructured with different types as text, sensor data, audio, video, graph, and more.

c) **Velocity:** Velocity is nothing but speed or rate at which data is moving around. Data is arriving continuously as streams of data, and we are interested in obtaining required information from it in real time

Nowadays, there are two more V's to define Bigdata:

d) **Veracity**: it describes the accuracy of data. Whenever any organization or applications require data, it gathers relevant data by analyzing and mining.

e) **Value:** The high volume of relevant fact as moving of different types of data compiles to create value.

User of Hadoop could be divided into two groups
1 Administrator
2 Users

Administrators are responsible for the task like installation monitor manage system tuning the system in short maintaining health of the software. Some of the areas where Hadoop is being used like social media Facebook is the second largest user of Hadoop after Yahoo then Amazon is a perfect example for retail Financial Services heavily used Hadoop, search tools Google was behind the Hadoop Technology as well Government and intelligence. Agencies any area where application dealing with big data is likely to use Hadoop.

Users are responsible for designing application importing and exporting data and working with tools.

In 2005 Hadoop was created by Doug Cutting and Michael J. Cafarella both used to work for Yahoo some projects related to Hadoop like hive, Hbase, oozie, pig, Mahout, scoop. In 2006 Yahoo was donated to Apache who is maintaining Hadoop and related tools

2. RELATED WORK

Vikram Phanindra [1] has explained the traditional database system that is RDBMS which is a structured database. Structure database store the data in specific schema or in structure like rows and columns but data is not always available in structured form so that we need unstructured database to handle big data. This paper has explained the areas where Hadoop is good for or where Hadoop can be used like large data set, scalable algorithm, log management.

Some times hadoop come across the failures and breakdown of name node, Hadoop provides efficient way of handling fault tolerance. This Paper discussed about the basic architecture of Hadoop and its components and also about

most important feature i.e. the fault tolerance mechanisms like data replication, checkpoint, heartbeat messages and recovery methods [3].

2.1 Hadoop

Hadoop is a software framework where an application combination of various parts. Hadoop is an open source project. Hadoop was developed by Google's MapReduce [1]. It consists of many small sub projects which belong to the category of infrastructure for distributed computing. Hadoop is a cluster system. It can store structured and unstructured data because it is fundamentally a file system Hadoop is plug and play architecture. Any organization or application needs Hadoop because it processes large dataset.

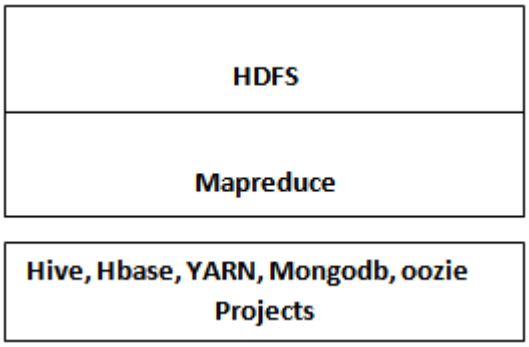

Fig -1: Hadoop Architecture

a) Hadoop Distributed File System (HDFS)

Apache Hadoop uses the Hadoop Distributed File System shown in Fig 2. It consists of a cluster of machines, and files are stored on them. It is highly fault tolerant and uses minimal cost hardware. It also provides file permissions and authentication and streaming access to system data. [2]

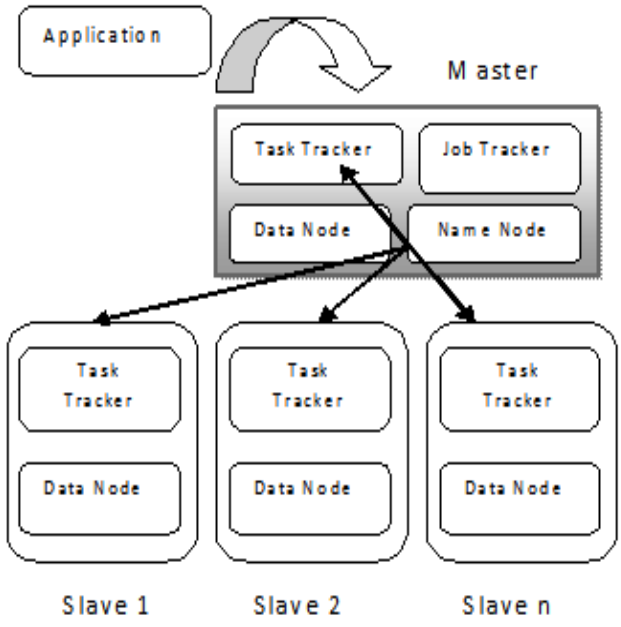

Fig 2: HDFS Architecture

Hadoop works on distributed model including numerous low cost computer institute of one powerful computer and Linux based system there are two types of system 1 master and another slaves.

That slave system will have two components of Hadoop.
1. Task trackers
2. Data node

Task Tracker

The job of task tracker component is to process the smaller piece of Task that has been given to particular node.

Data Node

The job of data node is managing the piece of data that has given to particular note. All these computers will be called slaves.

Difference between slaves and master is Master System will have two additional components
1. Task tracker
2. Data node
3. Job tracker
4. Name node

Jobtracker and task tracker is a part of high-level computing so they fall under Hadoop file system.The application that is running on Hadoop will contact the master node one of the attributes of Hadoop is batch processing. So application would assign are provided a task for Hadoop to perform and it's going in the queue. Once a task completed application will be informed hand result would give back to the application.

Job Tracker

Job tracker is running on master node and rule of job tracker is to break the higher and bigger task into smaller pieces and to send each to small computation to task tracker.

They will perform their smaller piece and send the result back to job tracker and it will combine the whole result together and send the final result back to the application.

Name Node

It is running on the master node and it is responsible to keep an index of which data is residing on which data node. So when application contacts the name node it tells the application goes to this particular computer to get your data so it has all indexes.

Show the name node tells the application where the data is residing. It goes directly to that node and will get data directly from that node.

2.2 Fault Tolerance in Hadoop File system

The hardware failure is bound to happen. They will happen and good thing about hadoop is that it is keeping hardware

failures in mind. It has built in fault tolerance. By default Hadoop maintain 3 copies of each file and these copies are scattered along different computers, so this way when computer fails the system keeps on running data is available from different nodes and once you fix the failed node then Hadoop will take care of that and it will copy some other files to that node that is very important feature of Hadoop[3].

The fault tolerance is not limited to the disk failing at the one of that slave node or master node. it is also applicable to the task tracker services that running on slave computer it's any of computer fails or service fails jobtracker would detect failure and it will ask some other task tracker to perform the same job. If master node dies then it is single point of failure. Hadoop has taken care of that as well the tables that are maintained by the name node that has the entire index where the data is residing on which computer. All those tables are backup and backup copies are distributed over to different computers and Enterprise version of Hadoop computer also keeps two masters one has main master and one as backup master in case of Master dies.

Programmers do not have to worry about taking care of where is the file is located? How to manage failures? How to break computation into pieces? And they don't have to program for scaling. Programmer can now focus on writing skill free programs.

One of very important attribute of Hadoop is that it is highly scalable and Hadoop system would consist of 1 computer and it could go up to thousand of computers. All depends on requirement and changing needs.

b) MapReduce

Second characteristics of Hadoop are that it can process that data or it has framework to process the data that is called MapReduce. TheHadoop programs perform- represents two separate and distinct tasks Map Job and Reduce Job.

Mapping step accept input data set or unstructured data from application and start splitting each entity as keyword, after that it map separated data into <key, value> pair. Reduce job takes the output of the Map job i.e. the key value pairs and aggregates similar keys in separate file known as shuffling and pass generated output for reducing purpose to produce desired results. The input and output of the map and reduce jobs are stored in HDFS.

In Unique way suppose software which process large data, it could mean searching, counting keywords, aggregating. During processing data is stored at one place and software is installed at server where processing would be differently done and usually while processing data has to move at system where software is installed. Huge data set takes huge time to move data back and forth. What Hadoop does?
It moves processing software where the data is present so that processing across all the node it distribute the processing that called mapping to the data and it collect the result which is called reducing to bring the answer back.

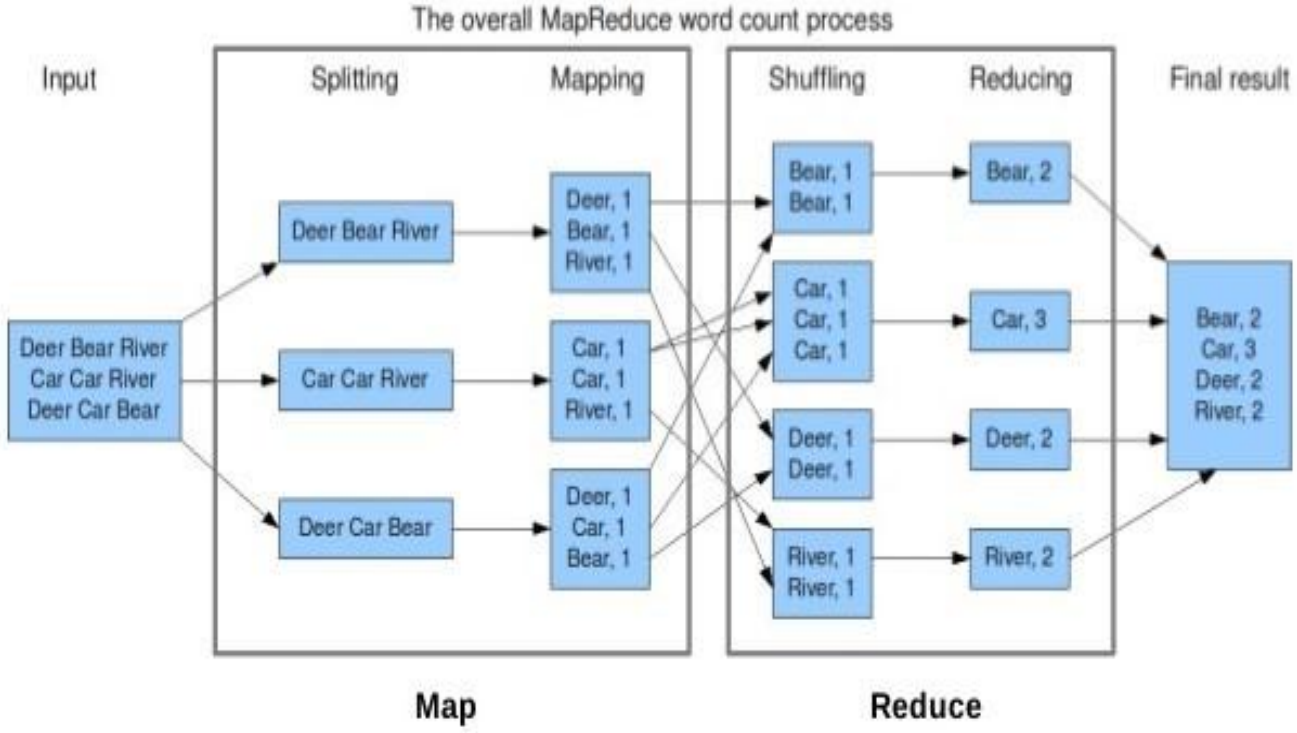

Fig-3: Map Reduce Architecture with word count example

2.3 Challenges in Big Data Analytics

1. Speed and Data Quality: In today's' competitive environment, every domain required relevant data for business decision and development, they want data quickly and accurate.

2. Storage and Network Requirement: Now a day's data is generated in all the areas in huge amount which is unable to handle and store. Many organisations are using cloud as choice to fulfil the storage requirement of Big data [4].But when it comes to uploading the huge amount of data, network bandwidth becomes insufficient and it leads to bottleneck and effect on bandwidth.

3. Talent Gap: Enough number of experts are not available for Big data implementation, as it is an emerging technology. At the same time, it does not only require the technology acquaintance but one should have analytical and interpretive skills too [4]. Many companies are hiring Big data consultants to train their employees.

4. Data Integration: Integrating disparate data from various heterogeneous sources and unstructured which is an open challenge in big Data analytics [5].

5. Selection of Right Project: Identifying the right business problem is critical for success and business needs to get involved as quickly as possible.

6. Budget: Traditional enterprise data servers are not capable to process Big data and additional hardware required for storage purpose [6]. Additional IT investments like to purchase high performance computing (HPC) servers, GPU units and analytical servers.

7. IT Knows -How: Due to the sheer volume of data, HPC and parallel processing are used in Big data in which several processing threads run concurrently on data. To facilitate parallel processing, storage strategies are also required to be changed.

8. Data Cleanup: Data cleaning is a process of eliminating inaccurate, redundant data and incomplete from the input data source. It should be the first step of any Big data project. Generally, datasets contain high level of redundancy which should be eliminated to reduce the overall cost of the project [7].

9. Data Confidentiality: Security is a main feature to focus during storing and analysing bigdata. A very lower granularity data which can drive any business such as transactional data which is confidential is used by the big data projects [7].

3. INSTALLATION OF HADOOP ON UBUNTU

Installation of hadoop is nothing but starting hadoop server. Hadoop server is suppose to start at every execution of hadoop program .Basic and minimum configuration for hadoop is ubuntu operating system, 4 GB Random Access Memory, 500 GB Hard Disk.

Basic steps of hadoop installation has been explained below:

1. Download HADOOP.TAR.GZ file from apache site and put at specific location.

2. Open terminal

3. Extract downloaded file using
   >>tar –xzf HADOOP.TAR.GZ

4. Set path where hadoop folder has extracted and enter into bin folder of hadoop as given below.
   Ex: >> cd /usr/local/hadoop/bin

5. Format Name node of hadoop with
   >> ./hadoop namenode –format

   Above command delete currently available data from namenode and shutdown the namenode.

6. Start again namenode with
   >>./start-all.sh

   This command newly starts all components/services of hadoop like namenode , job tracker, data node, Task Tracker

7. >>$JAVA_HOME/bin/jps
   This command shows list of started services. There are 5 services to be start to execute any hadoop application.

8. To crosscheck, whether hadoop processes have started or not, go to internet browser and run below url in address bar.
   a. Localhost:50070
      It shows status of Name node.
   b. Localhost:50030
      It shows status of JobTracker.
   c. Localhost:50060
      It shows status of TaskTracker..

4. CONCLUSION

Traditional and structured database system is unable to store all type of data. Massive amount of data move across the internet with high speed is nothing but bigdata. This paper introduces Bigdata and characteristics. Hadoop is a platform to handle bigdata and it has 2 important module HDFS and MapReduce, Various challenges associated with it that has been also explained. The paper also highlights the steps of hadoop installation in short.

REFERENCES

[1] S. Vikram Phaneendra1, E. Madhusudhana Reddy2 "Big Data - Solutions for RDBMS Problems – A Survey- Vol. 2, Issue 9, September 2013

[2] "Research Paper on Big Data and Hadoop " IJCST Vol. 7, Iss ue 4, Oct - Dec 2016- ISSN : 0976-8491 | ISSN : 2229-4333.
[3] "Big Data Processing with Hadoop : A Review" Volume: 04 Issue: 02 | Feb -2017- e-ISSN: 2395 - 0056 p-ISSN: 2395-0072.
[4] Cowsalya and S.R. Mugunthan "hadoop architecture and fault tolerance based hadoop clusters in geographically distributed data center" - VOL. 10, NO. 7, APRIL 2015
[5] Katal, Avita, Mohammad Wazid, and R. H. Goudar. "Big data: issues, challenges, tools and good practices" Contemporary Computing (IC3),2013 Sixth International Conference on. IEEE, 2013.
[6] Cuzzocrea, Alfredo, Il-Yeol Song, and Karen C. Davis. "Analytics over large-scale multidimensional data: the big datarevolution!." Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP. ACM, 2011.
[7] Mary shacklett "10 roadblocks to implementing Big Data analytics" http://www.techrepublic.com/blog/10-things/10-roadblocks-toimplementing big-data-analytics
[8] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: A survey." Mobile Networks and Applications 19.2 (2014): 171-209.

BIOGRAPHIES

Prof. Anjali Deore working as Assistant Professor in SITRC ,Sandip Foundation, Nashik, India. She has pursued ME in Computer Science and engineering in 2014. Area of interest is data mining, database, internet security.

Email: anjalideore13@gmail.com

Prof. Bhagyashree More working as Assistant Professor in Wadia college of engineering Pune, India. She has pursued ME in Computer Science and engineering in 2015. Area of interest is data mining, database, internet security, Image Processing

Prof. Kaveri Sonawane working as Assistant Professor in SITRC ,Sandip Foundation, Nashik,India. She has pursued ME in Computer Science and engineering in 2016. Area of interest is data mining, database, internet security,cloud computing.

Email: kaverisonawane28@gmail.com

Prof. Jyoti Kharat working as Assistant Professor in RSSOER, JSPM, NTC, Pune, India, She has pursued ME in Computer Science and engineering in 2013. Area of interest is data mining, database, internet security.

Email: kharat.jyo@gmail.com