A Review Paper on Big Data and HADOOP Architecture

[1]Sanket Gupta, [2]Nand Kishore Sharma, [3]Sonali Jain, [4]Sachin Sohra
Assistant Professor
Dept of Computer Science & Engineering,
Acropolis Technical Campus, Indore, M.P. India

***Abstract*— we are at a very high rate, on-command, on-demand entities, individuals and machines with data prolife ring live in the digital era. This data is known as Big Data due to its steep Volume, Variety, Velocity and Veracity. The mass of this data is unstructured, semi structured and it is heterogeneous in nature. In order for a cheap and efficient way to process vast amounts of data, parallelism is used. Big data is a collection of huge and multifaceted data sets that include the huge quantities of data, analysis of social data management capabilities, real-time data. The data is generated from various different sources and at different rates in the system can access. Hadoop and HDFS by Apache are commonly used for storing and managing Big Data. Hadoop is the main platform for processing Big Data, and solve the problem of making it useful for analysis purposes. Hadoop is an open source software project that capable the distributed processing of big data sets among clusters. It is designed to scale up from a single server to many machines, with a very high degree of fault tolerance. Map Reduce is widely been used for the efficient analysis of Big Data. In this research author review various aspects of big data and hadoop architecture.**

Keywords: **Hadoop, Big Data, MapReduce, Distributed System, Node, Cluster**

I. INTRODUCTION

A. Big Data: Definition

Big data is a term that refers to data sets or collection of data sets whose size (volume), complexity (variability), and rate of growth (velocity) Making them difficult to be captured, managed, processed or analyzed by traditional technologies and tools, such as RDBMS and desktop statistics or visualization packages, Within the time required to make them useful. Real Time analysis about Big data in now a days, facebook has 1490 million active user, Whatsapp has 800 million active user. Another example is flicker having feature of Unlimited photo uploads (50MB per photo), Unlimited video uploads, it also capable to show HD Video, Unlimited storage, Unlimited bandwidth. Flickr had a total of 87 million registered members and more than 3.5 million new images uploaded daily [2]. Below figure 1 shows the Big data analysis.

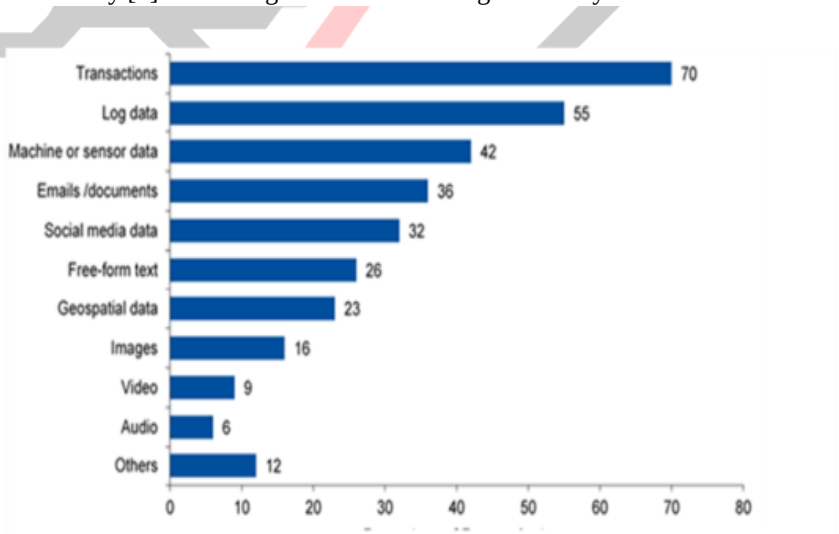

Fig.1 Big Data Analysis

- **Components of Big Data**

Volume of data: Volume shows the amount of data. It contains the large amount of data generated by organizations or individuals. Volume of data stored in enterprise database. It will grow from megabytes and gigabytes to petabytes and increase to zettabytes (10^{21}).

Variety of data: It refers types of data and sources where the data came. All data stored in enterprise repositories in the form of unstructured, semi structured, audio, video, XML etc.

Velocity of data: Velocity deals with the speed of data. It comes from various sources for data processing. For time-sensitive processes such as catching scam, Big data must be used not deal only for incoming data speed but also for speed of data flows.

Fig.2 Component of Big Data

B. Problem with Big Data Processing

- ***Heterogeneity and Incompleteness***

Now a day humans use more and more information, a great deal of heterogeneity is comfortably tolerated. In fact, the gradation and richness of natural language can provide valuable depth. However, machine analysis algorithms look forward to homogeneous data, and cannot recognize gradation. In consequence, data must be organized as a first step in data analysis carefully.

If we want Computer systems work most efficiently than arrange data in identical size and structure. It should be efficient in representation, easy in access and analyzed properly.

- ***Scale***

In the current technological trends data is come from various sources. So to maintain these huge amounts of data is very challenging task. So, for this purpose Data processing is used. But this technique becomes very complicated when data is in huge volume. Previously, this challenge was overcome by; increase the CPU speed and processor getting faster. But, now a day's data volume is scaling faster than compute resources and CPU speed are fixed.

- ***Privacy***

In this digital world, as the data exchange rate become faster, privacy is very necessary part of it. So, protect our data from unauthorized user. Security is needed in every field nowadays such as government sector, banking sector etc, where data is confidential.

II. SOLUTION OF BIG DATA PROCESSING

Hadoop is a framework which support programming and used to process a large data set in distributed environment. Bigdata requires processing on large amount of data that is supported by HADOOP. Hadoop was developed by Google's Map reduce that is a software framework where an application split into various parts. The existing Appache Hadoop ecosystem consists of the Hadoop Kernel, Map reduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper [4].

A. MapReduce

Map Reduce is used for processing on distributing system which is created by the Google in which divide and conquer method is used to split the large complex data into sub parts and process them. Map Reduce have two stages which are Map and Reduce.

B. MapReduce Components

- **Name Node** – It handle HDFS metadata, it doesn't allow deals with files directly.
- **Data Node** – It accumulate blocks of HDFS. It provides default replication level for each block.
- **Job Tracker** – It schedules, assign and observe job execution to Task Trackers.
- **Task Tracker** – It runs entire Map Reduce operations.

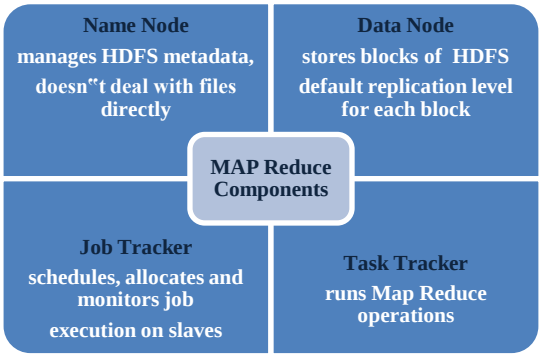

Fig.3 MapReduce Components

In below figure 4 input file are splits into number of block, each block required have one Map function for calculation. First record reader takes a block and converts into (key, value) pair and sends to Map function for further calculation. In second stapes only single reduce function are apply and generate output file. MapReduce divides workloads up into multiple tasks that can be executed in parallel.

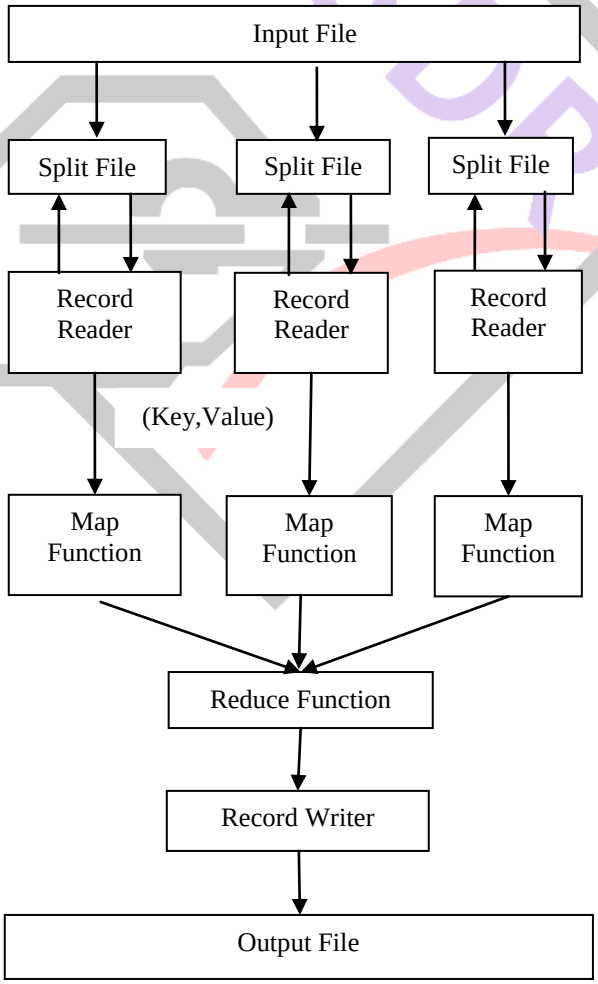

Fig.4 Map Reduce

C. MapReduce key attributes

- Resource Manager: to determine optimal computing operations, data locality and server resources are required.
- Optimized Scheduling: Here priority is the main concern. All jobs are completes according to prioritization.
- Flexibility: Here anyone can write their algorithm virtually any programming language.
- Resiliency & High Availability: Multiple jobs are run simultaneously. It ensures that if any job failure occurs it will restart automatically.
- Scale-out Architecture: To increase the processing power, increase the number of servers [3, 5].

D. HDFS

A fault tolerant storage system used by HADOOP called the Hadoop Distributed File System, or losing data. It builds clusters of machines and HDFS. HDFS is capable to store huge amounts of data, scale up incrementally and endure the failure of important parts of the storage without coordinates work among them.

Below figure 5 shows the architecture of HDFS. Clusters can be built with low configuration computers. If one fails, it continues to operate the cluster without losing data or interrupting work, by shifting load to the remaining machines in the cluster. It manages capacity on the cluster by splitting incoming files into small part, called "blocks," and storing each blocks across the servers. In general case, HDFS stores three duplicate copy of each file on three different servers. In case, a failure occurs in name node, HDFS does not support automatic recovery. But the secondary node is available for configuration [1].

HDFS Architecture

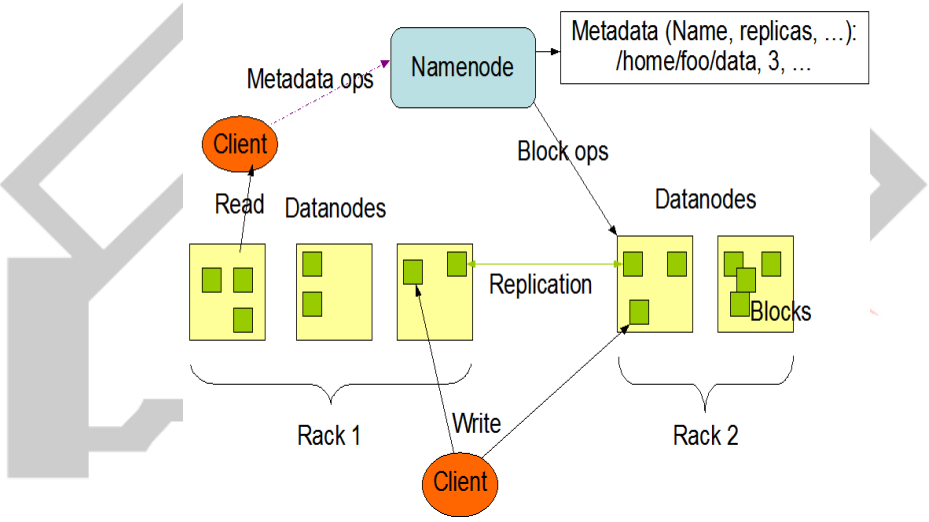

Fig.5 HDFS Architecture

E. HDFS key attributes

- High Availability: Provides mission-critical workflows and applications.
- Fault Tolerance: if any failure occurs, it will automatically and flawlessly recover.
- Scale-Out Architecture: To increase capacity we can add more servers.
- Flexible Access: It provides open frameworks for serialization and file system mounts.
- Load Balancing: For maximum efficiency and utilization, data should be intelligently positioned.
- Tunable Replication: To provide data protection, multiple copies of each file is used and it will increase computational performance [10]. Other components of Hadoop are discussed below [2].

  - **HBase**: It is open source, Non-relational, distributed database system. It operates on the top of HDFS. It can give as the input and output for the MapReduce. It is written in Java.

- **Pig**: Pig is high-level platform where the MapReduce programs are produced which is used with Hadoop. It is a high level data processing system where the data sets are analyzed that occurs in high level language.
- **Hive**: For SQL interface and relational model, Data warehousing application is used. Its infrastructure is built on the top of Hadoop. It support for summarization, query and analysis.
- **Sqoop**: For transferring data between relational databases and Hadoop, Sqoop is used. It works on command line interface.
- **Avro**: It is mainly used in Apache Hadoop. It provides data serialization system and data exchange service. This functionality can be used together as well as independently.
- **Oozie**: Oozie is based on java. It supports web application which runs on java servlet. It maintains the entire task performed on hadoop. To store definition of Workflow (collection of actions) database is used.
- **Chukwa**: It is data collection and analysis framework. To process and analyze the large amount of logs Chukwa is used. It is situated on the top of the HDFS and MapReduce framework.
- **Flume**: For streaming of data from multiple sources a high level architecture is used which is called as Flume.
- **Zookeeper**: it is a centralized service. It offers distributed synchronization and group services and preserves the configuration information etc [6].

3. CONCLUSIONS

In current scenario data is generated very high speed. Now a day's data is generated from various devices in the form of structure and unstructured data. This paper expresses the problems and solution of big data along with 3Vs.we has express of problem related big data like privacy, Incompleteness etc. Here describes the HDFC architecture with all components and key attributes. We have discussed how HDFS save multiple copies of block data in different- different cluster. At the end we have explain the working of map reduce. This paper gives the complete exposure of big data processing.

REFERENCES

[1] Harshawardhan S. Bhosale & Prof. Devendra P. Gadekar "A Review Paper on BigData ana Hadoop", IJCR,Volume 4,Issue 10, Oct 2014.
[2] Shilpa & Manjit Kaur "Big Data and Methodology- A Review", IJARCSS, Volume 4, Issue 10, Oct 2013.
[3] E.Sivaraman, Dr.R.Manickachezian" High Performance and Fault Tolerant Distributed File System for Big Data Storage and Processing using Hadoop", International Conference on Intelligent Computing Applications, 2014.
[4] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data- solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19, 2013.
[5] Neil Raden, "Big Data Analytics Architecture - Putting All Your Eggs in Three Baskets", 2012.
[6] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.
[7] Sameer Agarwal, Barzan MozafariX, Aurojit Panda, Henry Milner, Samuel MaddenX, Ion Stoica "BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data" Copyright © 2013ì ACM 978-1-4503-1994 2/13/04.
[8] Yingyi Bu _ Bill Howe _ Magdalena Balazinska _ Michael D. Ernst "The HaLoop Approach to Large-Scale Iterative Data Analysis" VLDB 2010 paper "HaLoop: Efficient Iterative Data Processing on Large Clusters.
[9] Shadi Ibrahim, Hai Jin _ Lu Lu "Handling Partitioning Skew in MapReduce using LEEN" ACM 51, 107–113, 2008.
[10] Neil Raden, "Big Data Analytics Architecture - Putting All Your Eggs in Three Baskets", 2012.