Analytical Survey On Big Data

Neha M. Yadav, Asst. Prof. Pushpanjali M. Chouragade

Abstract—Big data is the term for extremely large data sets that may be analyzed computationally to reveal patterns, especially relating to human behavior and interactions.. This information can be utilized to get distinctive results and forecasts utilizing diverse sorts of examination.. Data sets grow in size in part because they are increasingly being gathered by low price information-sensing mobile devices, aerial (remote sensing). Big data is hard to work with utilizing most social database frameworks and desktop measurements and perception bundles, requiring rather "hugely parallel programming running on tens, hundreds, or even a huge number of servers". Big data usually works with such large size sets that are above the limit of commonly used software tools to capture, curate, manage, and process data within a capable time of event occuring time. There is a set or a bunch of technologies and techniques that require some new forms of parts to uncover large values that are not accessible to view from large datasets that are distinct, complicated, and of a large scale which is called Big Data. Big data environment is used to acquire, organize and analyze the various types of data. There is an observation about Map Reduce framework is a framework that generates data in a large amount that lies between extremes within a particular time period and space. Therefore, as well as the tasks finishes there is need of throwing the data that is present in a large amount, because MapReduce is not able to put that data into the work .

Index Terms—Big data, Hadoop, HDFS, MapReduce, Pig, Hive, Hbase.

1 INTRODUCTION

In today's world , 2.5 quintillion bytes of data is created by us — so much that 90% of the data. The amount of data comes from everywhere: to collect information about climate, sensors are used and posts to social media sites, digital pictures and videos, purchase transaction records, etc. This huge amount of the data is known as "Big data"[14]. Big data is a stock phrase that have become of no sense through endless repitition, or catch-phrase, that works to describe a large volume of both structure.
In most enterprise scenarios there is large size data or the data that moves very fast or sometimes it may cross the present capacity of processing. There is nothing comes before big data that comes into picture no.[3].

Big data is a term that evolves and describes any large amount of definite structured, semi definite structure and not in a definite structure data that has the potential to be mined for information. Although big data does not refers any particular quantity, so this term might be used when discussing about petabytes and Exabyte's of the data. Big data is an all-part of a broader term for large collection of the data sets .

When dealing with massive datasets, lots of difficulties has been faced by the organizations in creating, manipulating, and managing big data.

Various challenges has been included that may include capture, privacy violation, curation, analysis, visualization, storage, transfer, etc. Thus the massive amount of data is known as Big data [14].
Scientists regularly finds or notices the limitations due to massive data sets in many areas, including weather forecasting, branch of genetics i.e, genomics, connectomics, complex physics, biology and environmental research. The limitations also affect Internet search, finance, etc.
The world's technological per-capita capacity to every day 2.5 Exabyte's (2.5×1018) of data were created. Determining who should own big data initiatives is the challenge for large enterprises. The challenge for large enterprises is that to straddle the entire organization. Big data industry analyst Doug Laney (currently with Gartner) articulated the now mainstream definition of big data as the three Vs of big data: velocity, volume, variety [18].

Big data can be characterized by well known 3Vs: the extreme volume of data, the wide variety of types of data and the velocity of processing a data. Since big data does not refer to any specific quantity.

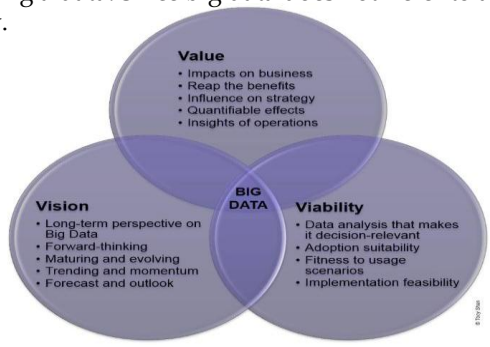

Figure 1:Three V's

Volume: Numerous elements add to the increment in information volume. Exchange based information put away as the years progressed. Unstructured information spilling in from online networking. Expanding measures of sensor and

machine-to-machine information being gathered. Previously, unnecessary information volume was a stockpiling issue. In

any case, with diminishing stockpiling expenses, different issues develop, including how to decide importance inside of extensive information volumes and how to utilize investigation to make esteem from significant data.

Velocity: Information is spilling in at exceptional speed and must be managed in a convenient way. RFID, sensors and savvy metering are driving the need to manage downpours of information in close continuous. Responding rapidly enough to manage information speed is a test for most organizations.

Variety: Information today comes in a wide range of arrangements. Organized, numeric information in customary databases. Data has been made from the line-of-business applications. Unstructured content records, email, video, sound, stock ticker information and money related exchanges. Overseeing, combining and representing distinctive assortments of information is something numerous associations still hook with.

2. Hadoop:

Hadoop is an open-source programming structure written in Java for circulated stockpiling and appropriated handling of vast information sets on PC bunches fabricated from item equipment. In Hadoop all modules are designed with a assumption that hardware failures (of individual machines, or racks of machines) handled automatically in software by the framework. Hadoop was created by Doug Cutting and Mike Cafarella in 2005.The base Apache Hadoop framework is composed of the following modules:

- ***Hadoop Common* –** It includes libraries and services needed by other Hadoop modules;
- ***Hadoop Distributed File System (HDFS)* –** a distributed file-system stores data on commodity machines that provides sum of bandwidth (aggregate), across the cluster;
- ***Hadoop YARN* –** a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications; and
- ***Hadoop MapReduce* –** a programming model for large scale data processing.

The hadoop system is consisting of several functioning system, called nodes.

- ***Master Node*:** Hadoop system generally consists of several master node instances, which are the main controlling node in the system. Several instances of master nodes are included so that even if one of the master node fail to service other master nodes can handle the work without directly failing the system. Master node has three major constituent parts- *Job Tracker*, *Task Tracker*, *Name Node*. *JobTracker* interacts with client applications and distributes MapReduce tasks to particular nodes within a cluster.Task tracker on the other hand receives tasks (like *Map()*, *Shuffle(), Reduce()* etc.) assigned by

JobTracker (HDFS) and also keep track of file data that is kept within the cluster. Client applications contact *Name Node*s when they need to locate a file, or add, copy, or delete a file.

- ***Data Node:*** The *Data Node* stores data in the Hadoop Distributed File System, and it replicates data among the clusters of the system. Data Nodes interact with client applications when the Name Node has supplied the Data Node's address.

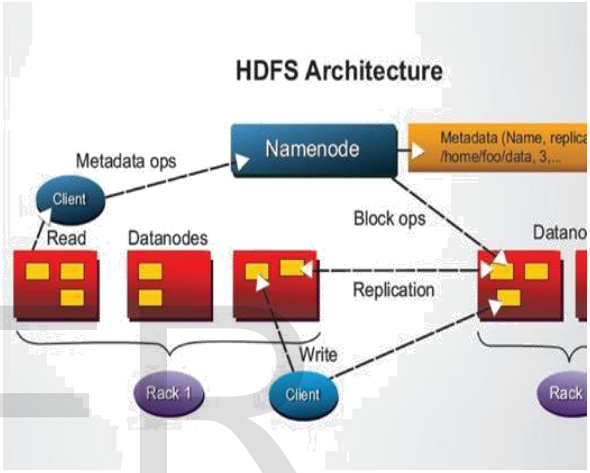

Figure 2: HDFS Architecture

- ***Worker node:*** Generally one Hadoop system includes dozens or even hundreds of *worker nodes*, that provides processing power. Each worker node includes a *Data Node* as well as a *Task Tracker*. Using the *Task Tracker* it tracks the job it's assigned to y *Job Tracker* of the controller *Master Node*.

Hadoop Architecture

Hadoop system consists of three layers. These logical Hierarchies together implement the MapReduce operation. The three layers are- *Application layer/end user access layer, Workload Management layer, Data layer.*

- **Application layer/end user access layer:** This layer provides a programming framework. It serves as the point of contact for applications to interact with Hadoop. These applications may be internally written solutions, or third party tools. To build any one of these applications, programming interfaces such as *Java,* a specialized, higher-level MapReduce language calld as PIG, or *Hive* (a specialized, SQLbased MapReduce language).

- **MapReduce workload management layer :**

It is commonly known as JobTracker. A single Hadoop environment will likely need to sustain multiple workloads. The only way that one can realistically support this variety is to document and then carefully plan for each type of workload.

For each workload, you need to know many details:

a. Number of Users

b. Volumes of Data

c. Data types

d. Processing windows (e.g., a few seconds, minutes or hours)

e. Anticipated network traffic

f. Applications that will consume Hadoop results.

Scheduling itself is frequently performed by software developed from the Apache Oozie project. This layer is the most critical for guaranteeing enterprise-grade Hadoop performance and reliability.

Distributed parallel file systems/data layer:

This layer is responsible for the actual storage of information. Along with HDFS, this layer may also consist of commercial and other third-party implementations. These include IBM's GPFS, the MapR filesystem from MapR Technologies, Kosmix's Cloud Store, and Amazon's Simple Storage Service (S3).

The inventors of the HDFS made a series of important design decisions:

- **Files stored in the form of blocks:** These are much larger than most file systems, with a default of 128 MB.
- **Through replication, Reliability is achieved:** Each block is replicated across two or more Data Nodes; the default value is three.
- **A single master Name Node coordinates access and metadata:** This simplifies and centralizes management.
- **No data caching:** It's not worth it given the large data sets and sequential scans.
- **There's a familiar interface with a customizable**

API: This lets you simplify the problem and focus on distributed applications, rather than performing low-level data manipulation.

4 CONCLUSION

Hadoop MapReduce which is an open source software framework. It breaks larger data into smaller chunks and handles scheduling. It is reliable and fault tolerant. Then these broken small chunks of data get solvable in parallel.

FUTURE ENHANCEMENT

Usually it is observed that the MapReduce framework generates a large amount of intermediate data. Such abundant information MapReduce is unable to utilize them. Therefore, we propose Dacha, a data-aware cache framework for big-data applications then its tasks submit their intermediate results to the cache manager. A novel cache description scheme and a cache request and reply protocol is designed.

REFERENCES

[1] Dhole Poonam B, Gunjal Baisa L, "Survey Paper on Traditional Hadoop and Pipelined Map Reduce"

[2] International Journal of Computational Engineering Research Vol, 03,Issue 12

[3] Nilam Kadale, U. A. Mande, "Survey of Task Scheduling Method for Mapreduce Framework in Hadoop"

[4] International Journal of Applied Information Systems

[5] (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA 2nd National Conference on Innovative Paradigms in Engineering & Technology (NCIPET 2013) – www.ijais.org

[6] Suman Arora, Dr.Madhu Goel, "Survey Paper on Scheduling in Hadoop" International Journal of Advanced

[7] Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014

[8] Wang, F. et al. Hadoop High Availability through Metadata Replication. ACM (2009).

[9] B.Thirumala Rao, Dr. L.S.S.Reddy, "Survey on

[10] Improved Scheduling in Hadoop MapReduce in Cloud Environments", International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011

[11] Amogh Pramod Kulkarni, Mahesh Khandewal, "Survey on Hadoop and Introduction to YARN", International

[12] Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

[13] Vishal S Patil, Pravin D. Soni, "HADOOP SKELETON & FAULT

TOLERANCE IN HADOOP CLUSTERS", International Journal of Application or Innovation in Engineering & Management (IJAIEM)Volume 2, Issue 2, February 2013 ISSN 2319 - 4847

[14] Sanjay Rathe, "Big Data and Hadoop with components like Flume, Pig, Hive and Jaql" International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV

[15] Yaxiong Zhao, Jie Wu and Cong Liu, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework",TSINGHUA SCIENCE AND TECHNOLOGY ISSNl1007-0214l 05/10l lpp39-50 Volume 19, Number 1, February 2014

[16] Parmeshwari P. Sabnis, Chaitali A.Laulkar , "SURVEY OF MAPREDUCE OPTIMIZATION METHODS", ISSN (Print): 2319- 2526, Volume -3, Issue -1, 2014

[17] Puneet Singh Duggal ,Sanchita Paul ," Big Data Analysis: Challenges and Solutions", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV

[18] Chen He,Ying Lu,David Swanson, "Matchmaking: A New MapReduce Scheduling Technique", EECS Department, University of California, Berkeley, Tech. Rep.,April 2009.