Survey on Big Data and Hadoop

Linta John[1] Smita C Thomas[2] Anju Marium Abraham[3]
[1,2,3]Department of Computer Science & Engineering
[1,2,3]Mount Zion College of Engineering Kadamanitta, Pathanamthitta India

Abstract— To enhance their performances companies across the world uses data since a long time, to make better decisions. In this paper make a survey on Big data Hadoop. For Big Data implementation Hadoop has emerged as a popular tool. For things like startups and online firms Big data may be new. A major challenge has become how to deal with the data and analysis of this data when big growth. The Apache Hadoop Framework has recently attracted a lot of attention for such data-intensive applications. Apache Hadoop Framework Adopted MapReduce, which is a programming model and for the implementation of processing and generating large data sets. For writing Application E.g. Java Map-Reduce, Streaming MapReduce, Crunch, Pig latin, Hive, Oozie etc Hadoop Provides: Distributed File System, Job scheduling, Resource Management Capabilities, and Java API. A map function and a reduce function will specify by user.

Key words: Big Data, Hadoop

I. INTRODUCTION

Big Data the term defines important techniques and technologies that take, store, distribute, manage and analyze larger size of datasets that need high-velocity and different structures. Big data may be structured, unstructured or semi-structured. From different sources data is generated and is arrive in the system at various rates. For processing this large amounts of data with an inexpensive and efficient way, parallelism is used. New architecture, techniques, algorithms are required for big data and also analytics to manage it. Hadoop is the main platform for Big Data structing, and also for solving the problem. Hadoop is an open source software project. Hadoop gives the distributed processing of large data sets across groups of servers. Hadoop is designed to scale up from a single server to thousands of machines. Also provide very high degree of fault tolerance. Big data means the availability of a large amount of data, which is difficult to store, process and uses a traditional database primarily. The data available is large, complex, unstructured and rapidly changing data is very difficult to handle. This may be one of the important reasons for accepting the concept of Big data, and was also first accepted by online companies like Google, eBay, Facebook, LinkedIn etc. Apache Hadoop is a 100% open source. Hadoop is a new way of storing and processing data sets. Instead of depending on to store and process data, Hadoop provides distributed parallel processing of large amounts of data with inexpensive, way and can scale without any limits. By using Hadoop, no data is too big. For the today's connected world, there is more and more data is created every day, Hadoop's main advantage is, that businesses and organizations can find value from data that was recently considered as useless. 'Big Data', is the term refers to data sets there volume, variability, and velocity and this make them difficult to capture, manage, process or analyzed. For analyzing this large amount of data, Hadoop is used. Hadoop is an open source software project that aids the distributed processing. Hadoop is designed to scale from a single server to thousands of machines, with a very high degree of fault tolerance. Hadoop, Map Reduce, Apache Hive, No SQL and HPCC are the technologies used and also these technologies handle massive amount of data in MB, PB, YB, ZB, KB and TB.

II. LITERATURE SURVEY

This paper [1] describes as, Big data is a term that refers to data sets and also combinations of data sets whose size, variability, and rate of growth make them difficult to be captured, managed, processed or analyzed by traditional technologies and tools, like relational databases, within the time limit. Most of the analysts and practitioners currently considered as the data sets which is range from 30-50 terabytes (10 12 or 1000 gigabytes per terabyte) to multiple petabytes (1015 or 1000 terabytes per petabyte) as big data.

The 3 Vs of Big Data

- Volume of the data: Volume is refers as the amount of data. Volume of data stored in an enterprise database have grown from large to large (megabytes and gigabytes to petabytes).
- Variety of the data: There are different types of data and their sources. Data variety varies from structured to unstructured, semi structured, audio, video, XML etc.
- Velocity of the data: Velocity is refers as the speed of data processing. For example time-sensitive processes like catching fraud.

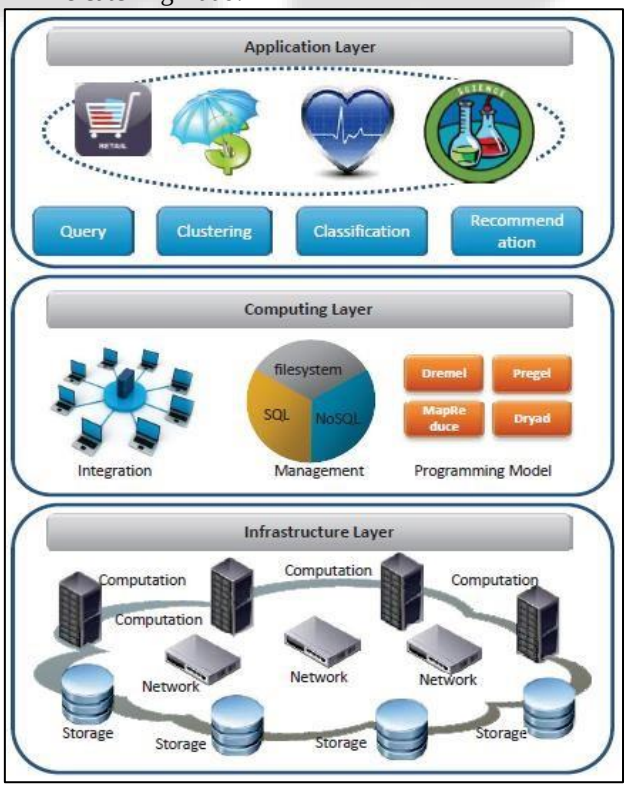

Fig. 1: Layered Architecture of Big Data System

Hadoop is a Programming framework. Hadoop is used to help the processing of large amount of data sets in a distributed environment. Hadoop was developed by Google's MapReduce, which is a software framework. MapReduce will break down the application into various parts. Hadoop Kernel is there in the apache ecosystem. MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper.

This paper [2] describes as, Hadoop is a distributed software solution. For data storage and processing it is a scalable fault tolerant distributed system. The two main components in Hadoop are,

A. HDFS

HDFS is a storage system like file system (which is a storage)

B. Map Reduce

This is for retrieval and processing. So HDFS is high bandwidth cluster storage. When a pent byte file is on our Hadoop cluster, HDFS will break down into blocks and then will distributed it to across all of the nodes. This mean if we put our file on hadoop it will make sure that it must has 3 copy of every block and make up that file spread across all the node in our cluster. This is very useful and important, because if we lose a node it has the ability to know what data was there on node and will replicate that blocks that were on that node. It has a name node and a data node, normally one name node per cluster. And also it is essential that the name node is a meta data server and also it must hold in memory If we have multiple rack setup then it will know where block exist and we get data. The data will get from the Map Reduce as name implies. And also there is a Mapper and Reducer programmers. And they will write the mapper function. The functions will go out and inform the cluster what data point is want to retrieve. The data will take and aggregate by the reducer.

- Hadoop Hadoop is a batch processing, here the working on all the data will carried out, then we can say that Map Reduce is working on all of data.
- Pig pig is another one which is built by yahoo, Pig is a high level data flow language, which pull data out of clusters. Now Pig and hive are under the Hadoop Map Reduce job submitted to cluster. This is the beauty of open source framework, people can built, add and community keeps on growing in Hadoop more technologies and projects are added into Hadoop ecosystem.

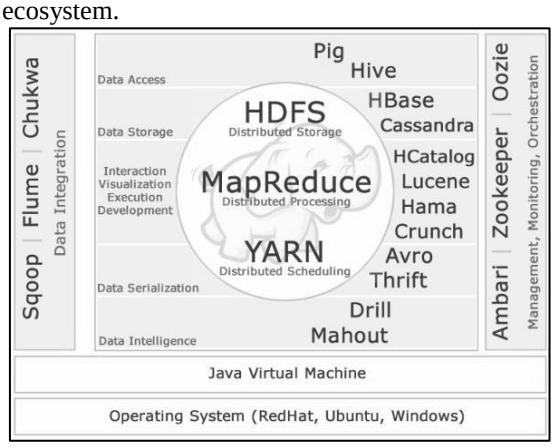

Fig. 2: The image shows the hadoop technology stack.

This paper [3] describes as, Programs which is written in a functional style are automatically get parallelized and executed on a large cluster of commodity machines. The amount of time the system takes are for partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This helps the programmers who are not experience with parallel and distributed systems to easily utilize the resources of a large distributed system. The implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many terabytes of data on thousands of machines. Programmers find the system easy to use:

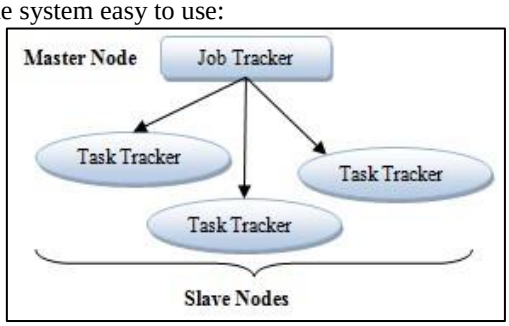

Fig. 3: Hadoop Master Slave Architecture

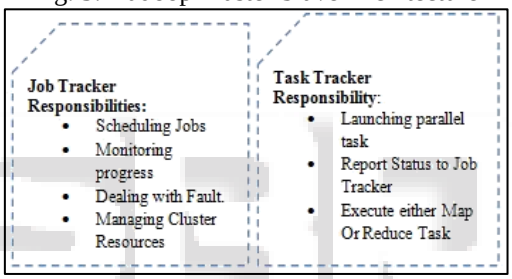

Fig. 4: Hadoop Job and Task Tracker Responsibility

In the initial stage of Hadoop, MapReduce is designed as a master-slave architecture as shown in fig3. The responsibilities of task and job tracker are shown in the fig 4. The JobTracker is the master that manages the cluster resources, scheduling jobs, monitoring progress and dealing with fault-tolerance. In each of the slave nodes, there exists a TaskTracker process. The responsible of the tasktracker is for launching parallel tasks and reporting their status to the JobTracker. The slave nodes are statically divided into computing slots, available to execute either Map or Reduce tasks. The Hadoop community realized the limitations of this static model and recently redesigned the architecture to improve cluster utilization and scalability.

III. CONCLUSION

We have entered an era of Big Data. This paper is an effort to present the basic understanding of BIG DATA and Hadoop. It's usefulness to an organization from the performance perspective. Doug Cutting, Cloudera's chief architect, helped create Apache Hadoop out of necessity as data from the web exploded, and grew far beyond the ability of traditional systems to handle it. Hadoop was initially inspired by papers published by Google outlining its approach to handling an avalanche of data, and has since become the de facto standard for storing, processing and analyzing hundreds of terabytes, and even petabytes of data.

ACKNOWLEDGMENT

We would like to thank, first and foremost, Almighty God, without his support this work would not have been possible. We would also like to thank all the faculty members of Mount Zion college of engineering, for their immense support.

REFERENCES

[1] Harshawardhan S. Bhosale, "A Review paper on Big Data and Hadoop", Volume 4, Issue 10, October 2014 ISSN 2250-3153.
[2] Bijesh Dhyani," Big Data Analytics using Hadoop", Volume 108 – No 12, December 2014.
[3] Poonam S. Patil, "Survey Paper on Big Data Processing and Hadoop Components", Paper ID: OCT14251

