Importance of Big Data and Hadoop

J.Suneetha, Student, MCA, KMMIPS, G. Amani, Student, MCA, KMMIPS
Mr. S.Muni Kumar, Assistant Professor, Dept. of MCa, KMMIPS,

ABSTRACT-Users are generating more and more information today, filling an ever growing collection of storage devices. The information may be about customers, web sites, security, or company logistics. As this information grows, so does the need for businesses to sift through the information for insights that will lead to increased sales, better security, lower costs, etc. The Hadoop system was developed to enable the transformation and analysis of vast amounts of structured and unstructured information. It does this by implementing an algorithm called Map Reduce across compute clusters that may consist of hundreds or even thousands of nodes. In this presentation Hadoop will be looked at from a storage perspective. The presentation will describe the key aspects of Hadoop storage, the built-in Hadoop file system (HDFS), and other options for Hadoop storage that exist in the commercial, academic, and open source communities.

Index Terms:-introduction, definition, what comes under big data?, benefits of big data, big data technologies, Operational big data, analytical big data, big data challenges, traditional approach, limitation, Google solution, hadoop, hadoop architecture, map reduce, how does hadoop works, advantages of hadoop, hdfs, features of hdfs, hdfs architecture, goals of hdfs.

Keywords*:* -Hadoop, Big Data, Hbase, Hive, Pig etc

1. INTRODUCTION:-

Importance of Hadoop andBig Data. It is a framework that supports the processing of large data sets in a distributed computing environment. It is designed to expand from single servers to thousands of machines, each providing computation and storage.

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

Big data is a term for data sets that are so large or complex that traditional dataprocessing application softwares are inadequate to deal with them. Challenges include capture, storage, analysis, data duration, search, sharing, transfer, visualization, querying, and updating and information privacy.

2. BIG DATA: DEFINITION

Big Data is a term that represents data sets whose size is beyond the capacity of commonly used software tools to manage and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data in a single data set. It is the term for a collection of data sets, so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.Big data is term which defines three characteristics

Volume:-

Volume refers to amount of data. It refers to the size of data. Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes.

Velocity:-

Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc.

Variety:-

Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value. Today, 70% ofgenerated data are in unstructured format.

What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

- **Black Box Data**:It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.

IJSER

- **Social Media Data:**Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data**:The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data**:The power grid data holds information consumed by a particular node with respect to a base station.
- **Transport Data**:Transport data includes model, capacity, distance and availability of a vehicle.
- **Search Engine Data**: Search engines retrieve lots of data from different databases.

Thus Big Data includes huge volume, high velocity, and extensible variety of data. The data in it will be of three types.

- **Structured data**: Relational data.
- **Semi Structured data**:XML data.
- **Unstructured data**: Word, PDF, Text, Media Logs.

Benefits of Big Data:-

Big data is really critical to our life and its emerging as one of the most important technologies in modern world. Follow are just few benefits which are very much known to all of us:

- Using the information kept in the social network like Face book, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.
- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

Big Data Technologies:-

Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business.

- To harness the power of big data, you would require an infrastructure that can manage and process huge volumes of structured and unstructured data in realtime and can protect data privacy and security.
- There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology:

Operational Big Data:-

- This includes systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.
- NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the pastdecade to allow massive computations to be run inexpensively and efficiently. This makes operational big data workloads much easier to manage, cheaper, and faster to implement.

Analytical Big Data:-

- This includes systems like Massively Parallel Processing (MPP) database systems and Map Reduce that provide analytical capabilities forretrospective and complex analysis that may touch most or all of the data.
- Map Reduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on Map Reduce that can be scaled up from single servers to thousands of high and low end machines.

Big Data Challenges:-

The major challenges associated with big data are as follows:

- Capturing data
- Duration
- Storage
- Searching
- Sharing

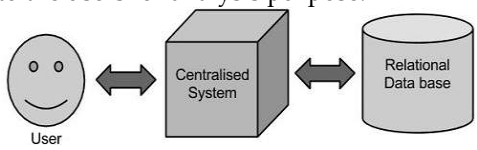
IJSER

- Transfer
- Analysis
- Presentation

To fulfill the above challenges, organizations normally take the help of enterprise servers.

Traditional Approach:-

In this approach, an enterprise will have a computer to store and process big data. Here data will be stored in an RDBMS like Oracle Database, MS SQL Server or DB2 and sophisticated software's can be written to interact with the database, process the required data and present it to the users for analysis purpose.

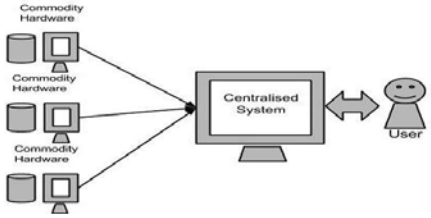


Limitation:-

This approach works well where we have less volume of data that can be accommodated by standard database servers, or up to the limit of the processor which is processing the data. But when it comes to dealing with huge amounts of data, it is really a tedious task to process such data through a traditional database server.

Google's Solution:-

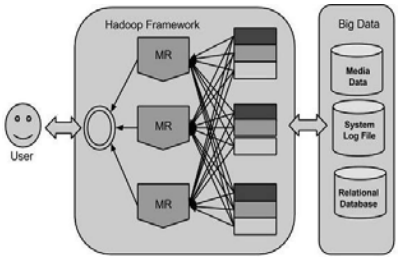
Google solved this problem using an algorithm called MapReduce. This algorithm divides the task into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result dataset.



Above diagram shows various commodity hardwares which could be single CPU machines or servers with higher capacity.

Hadoop:-

Doug Cutting, Mike Cafarella and team took the solution provided by Google and started an Open Source Project called HADOOP in 2005 and Doug named it after his son's toy elephant. Now Apache Hadoop is a registered trademark of the Apache Software Foundation.Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different CPU nodes. In short, Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data.



Hadoop is an Apache open source framework written in java that allows distributed processing of large datasetsacross clusters of computers using simple programming models. A Hadoop frame-worked application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

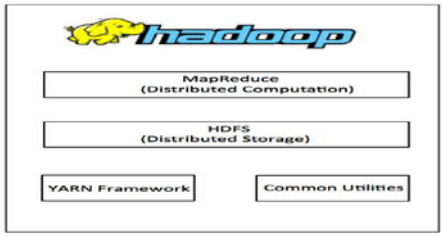
Hadoop Architecture:-

Hadoop framework includes following four modules:

- **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries providefile system and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce:** This is YARN-based system for parallel processing of large data sets.

We can use following diagram to depict these four components available in Hadoop framework.

IJSER



Since 2012, the term "Hadoop" often refers not just to the base modules mentioned above but also to the collection of additional software packages that can be installed on top of or alongside Hadoop, such as Apache Pig, Apache Hive, Apache HBase, Apache Spark etc.

MapReduce:-

Hadoop **MapReduce** is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.The term MapReduce actually refers to the following two different tasks that Hadoop programs perform:

The Map Task: -This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples (key/value pairs).

The Reduce Task:-This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.Typically both the input and the output are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.The MapReduce framework consists of a single master **JobTracker** and one slave **TaskTracker** per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves TaskTracker execute the tasks as directed by the master and provide task-status information to the master periodically.The JobTracker is a single point of failure for the Hadoop MapReduce service which means if JobTracker goes down, all running jobs are halted.

Advantages of Hadoop:-

- Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
- Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
- Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
- Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware.

Hadoop Distributed File System:-

Hadoop can work directly with any mountable distributed file system such as Local FS, HFTP FS, S3 FS, and others, but the most common file system used by Hadoop is the Hadoop Distributed File System (HDFS).The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters (thousands of computers) of small computer machines in a reliable, fault-tolerant manner.HDFS uses a master/slave architecture where master consists of a single **NameNode** that manages the file system metadata and one or more slave **DataNodes** that store the actual data.

A file in an HDFS namespace is split into several blocks and those blocks are stored in a set of DataNodes. The NameNode determines the mapping of blocks to the DataNodes. The DataNodes takes care of read and write operation with the file system. They also take care of block creation, deletion and replication based on instruction given by NameNode.

Features of HDFS:-

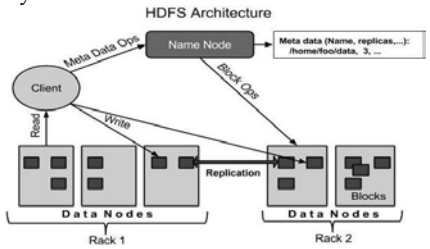
- It is suitable for the distributed storage and processing.
- Hadoop provides a command interface to interact with HDFS.
- The built-in servers of namenode and datanode help users to easily check the status of cluster.
- Streaming access to file system data.

IJSER

- HDFS provides file permissions and authentication.

HDFS Architecture:-

Given below is the architecture of a Hadoop File System.



HDFS follows the master-slave architecture and it has the following elements.

Namenode:-

The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software. It is a software that can be run on commodity hardware. The system having the namenode acts as the master server and it does the following tasks:

- Manages the file system namespace.
- Regulates client's access to files.
- It also executes file system operations such as renaming, closing, and opening files and directories.

Datanode:-

The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. For every node (Commodity hardware/System) in a cluster, there will be a datanode. These nodes manage the data storage of their system.

- Datanodes perform read-write operations on the file systems, as per client request.
- They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.

Block:-

Generally the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks. In other words, the minimum amount of data that HDFS can read or write is called a Block. The default block size is 64MB, but it can be increased as per the need to change in HDFS configuration.

Goals of HDFS:-

- **Fault detection and recovery**: Since HDFS includes a large number of commodity hardware, failure of components is frequent. Therefore HDFS should have mechanisms for quick and automatic fault detection and recovery.
- **Huge datasets**: HDFS should have hundreds of nodes per cluster to manage the applications having huge datasets.
- **Hardware at data**: A requested task can be done efficiently, when the computation takes place near the data. Especially where huge datasets are involved, it reduces the network traffic and increases the throughput.

8. CONCLUSION:-

Currently we are passing through Big Data phase. This paper focused on concept of Big Data alongwith 3 Vs. it also present problems and challenges while processing Big Data. In order to gain from Big Datathese challenges must be addressed. The paper describes various pros and cons of Hadoop as a Big Datamanagement tool. Although, Hadoop with its ecosystem is a powerful solution for handing Big Data. But still, Hadoopdoesn't soundsgood for frequently changing data.

REFERENCES:-

http://www.lavastorm.com
https://gigaom.com
http://wikibon.org
http://link.springer.com

IJSER