DATA ANALYTICS USING HADOOP

S. SRI RAAGA
B.Tech 3 rd year CSE
Vignan institutes of management and technology for women
raagasakki1@gmail.com

Abstract:
Information technology gives most importance to processing of data. Some petabytes of data is not sufficient for storing large amount of data. Large volume of unstructured and structured data that gets created from various sources such as Emails, web logs, social media like Twitter, Facebook etc. The major obstacles with processing Big Data include capturing, storing, searching, sharing and analysis. By using RDBMS (traditional methods) it is not that easy to process that huge amount data so, for that the tool called hadoop came into picture. It is an open source framework written in Java which supports parallel and distributed data processing and is used for reliable storage of data. It enables the distributing process of large data sets and popular tool for implementing big data analytics. This paper deals with the technology aspects of data analytics for its implementation in organizations and the structure of Hadoop with the details of various components.

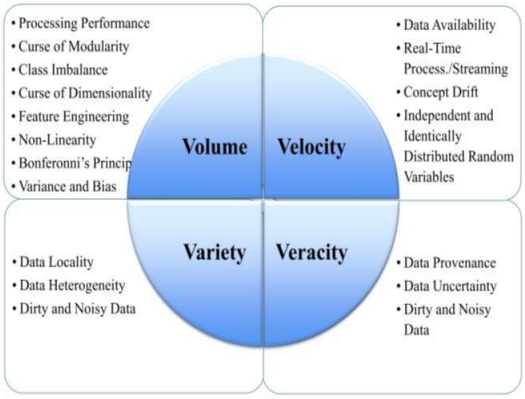
Keywords: *Petabytes , Big Data, Hadoop, web logs , technology, data analytics , data processing .*

i. INTRODUCTION:

Big data analytics is the area where advanced analytic techniques operate on big data sets. It is really about two things, Big data and Analytics and how the two have teamed up to create one of the most profound trends in business intelligence (BI) Map Reduce by itself is capable for analyzing large distributed data sets; but due to the heterogeneity, velocity and volume of Big Data, it is a challenge for traditional data analysis and management tools .database integration and cleaning is much harder than the traditional mining approaches. Parallel processing and distributed computing is becoming a standard procedure which are nearly non-existent in RDBMS. Map Reduce has following characteristics it supports Parallel and distributed processing, it is simple and its architecture is shared-nothing which has commodity diverse hardware (big cluster).Its functions are programmed in a high- level programming language(e.g. Java, Python) and it is flexible. Query processing is done through nosily integrated in HDFS as Hive tool.

ii. CONTEXT:

Big data characteristics:



VOLUME: Determining the value and potential insight and whether it can actually be considered big data or not is done by the size.

VELOCITY: Velocity is defined for the speed at which the data is being generated and processed to meet the demands

VARIETY: Big data draws from the text, audio, images, videos and through data fusion it completes missing pieces

i. IT'S ABOUT VARIETY, NOT VOLUME

The survey indicates companies are focused on the variety of data, not its volume, both today and in three years. The most important goal and potential reward of Big Data analytics initiatives is the ability to analyze diverse data sources.

ii. IMPLEMENTATION EXAMPLES

Big Data for cost reduction: Some organizations that are pursuing Big data believe strongly that for the storage of large data that is structured, Big data technologies like Hadoop clusters are very cost effective solutions that can be efficiently utilized for cost reduction .One company's cost comparison, for example, estimated that the cost of storing one terabyte for a year was $37,000 for a traditional relational database,$5,000 for a database apple appliance, and only $2,000 for a Hadoop cluster.

iii. WHY HADOOP CAME INTO PICTURE:

Traditional Enterprise Approach:

In this approach, an enterprise will have a computer to store and process big data. For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc. In this approach, the user interacts with the application, which in turn handles the part of data storage and analysis.

Limitation

This approach works fine with those applications that process less voluminous data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data. But when it comes to dealing with huge amounts of scalable data, it is a big hectic task

iv. SOLUTIONS:

*map reduce programming model extended to support iterative applications

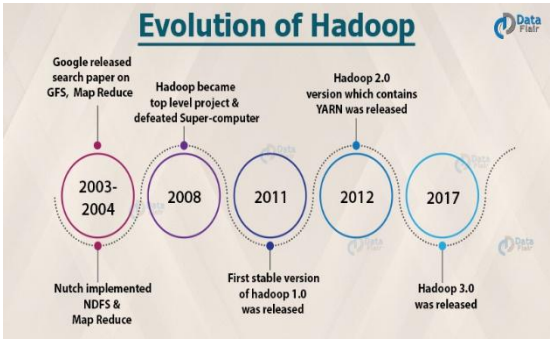
*supports parallel process, map reduce and iterative applications-a large and a useful subset of large-scale data intensive computations

*simple and easy to use

*suitable for efficient execution in cloud environments

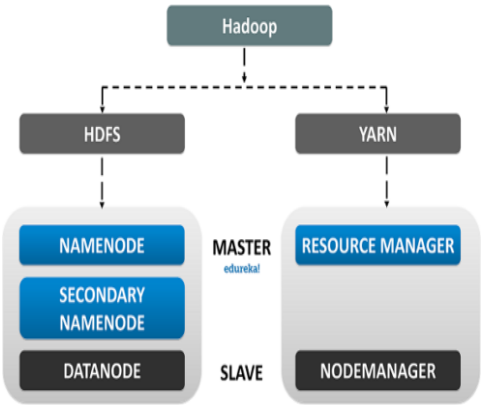
*map collectives-improve the usability of the iterative map reduce model

v. HISTORY OF HADOOP:



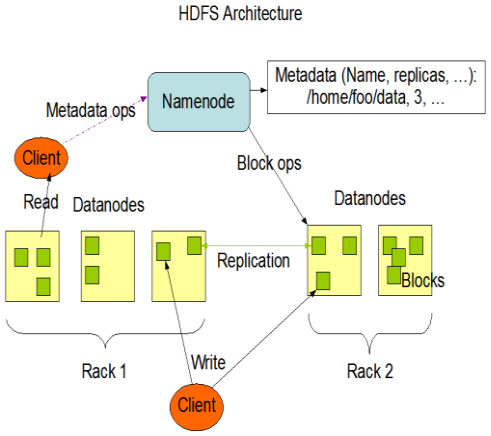
Hadoop was started with Doug Cutting and Mike Camarilla in the year 2002 when they both started to work on Apache Notch project. Apache Notch project was the process of building a search engine system that can index 1 billion pages. After a lot of research on Notch, they concluded that such a system will cost around half a million dollars in hardware, and along with a monthly running cost of $30, 000 approximately, which is very expensive. So, they realized that their project architecture will not be capable enough to the workaround with billions of pages on the web. So they were looking for a feasible solution which can reduce the implementation cost as well as the problem of storing and processing of large datasets.

vi. COMPONENTS OF HADOOP:



Apache Pig: software for analyzing large data sets that consists of a high-level language similar to SQL for expressing data analysis programs, coupled with infrastructure for evaluating these programs. It contains a compiler that produces sequences of Map- Reduce programs. Base non-relational columnar distributed database designed to run on top of Hadoop Distributed File system (HDFS). It is written in Java and modelled after Google's Big Table. Base is an example of a nosily data store. Hive: it is Data warehousing application that provides the SQL interface and relational model. Hive infrastructure is built on the top of Hadoop that help in providing summarization, query and analysis. Cascading: software abstraction layer for Hadoop, intended to hide the underlying complexity of Map Reduce jobs. Cascading allows users to create and execute data processing workflows on Hadoop clusters using any JVM-based language. Avro: it is a data serialization system and data exchange service. It is basically used in Apache Hadoop. These services can be used together as well as independently. Big Top: It is used for packaging and testing the Hadoop ecosystem. Oozy: Oozy is a java based web-application that runs in a java servlet. Oozy uses the database to store definition of Workflow that is a collection of actions. It manages the Hadoop jobs. So there are many advantages of hadoop that are: 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

i. HADOOP ARCHITECTURE:



At its core, Hadoop has two major layers namely:

(a) Processing/Computation layer (Map Reduce), and

(b) Storage layer (Hadoop Distributed File System).

MAP REDUCE:

Map Reduce is a parallel programming model for writing distributed applications devised at Google for processing of large amounts of data (multiterabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The Map Reduce program runs on Hadoop which is an Apache open-source framework. It is a processing technique and a program model

for distributed computing based on java. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name Map Reduce implies, the reduce task is always performed after the map job. The major advantage of Map Reduce is that it is easy to scale data processing over multiple computing nodes. Under the Map Reduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the Map Reduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the Map Reduce model. The stages of Map Reduce Program Generally Map Reduce paradigm is based on sending the computer to where the data resides! Map Reduce program executes in two stages, namely map stage and reduce stage.

· Map stage: The map or mappers job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

· Reduce stage: This stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS during a Map Reduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster. The

Framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes. Most of the computing takes place on nodes with data on local disks that reduces the network traffic. After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

. 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 commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. It is suitable for the distributed storage and processing. Hadoop provides a command interface to interact with HDFS. The built-in servers of name node and data node help users to easily check the status of cluster.

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 HDFS provides file permissions and authentication. DFS follows the master-slave architecture and it has the following elements.

· **NAME NODE**

The name node is the commodity hardware that contains the GNU/Linux operating system and the name node software. It is software that can be run on commodity hardware. The system having the name node acts as the master server and it does the following tasks: Manages the file system namespace. Regulates client's access to files and It also executes file system operations such as renaming, closing, and opening files and directories.

· **DATA NODE:**

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

· **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 write is called a Blow

ii. APPLICATIONS OF BIG DATA ANALYTICS:

Usage of big data analytics in the manufacturing field:

It provides an infrastructure for transparency and performs inconsistency and availability. Acoustics, vibration, pressure, current, voltage and controller data are the different types of sensory data to predict the manufacture, vast amount of sensory data is needed.

Usage of big data in healthcare

Big data analytics helps the healthcare to improve personalized medicine and prescriptive analytics, clinical risk intervention, waste and care variability reduction, automated external and internal reporting of patient data. Healthcare systems are not trivial and generated in level of data.

iii. 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

iv. WHO USES HADOOP?

1. IBM
2. Yahoo
3. Facebook
4. Hp
5. Intel
6. Ebay

7. Netflix

8. Twitter

v. CONCLUSION:

The need to process enormous quantities of data has never been greater. Not only are terabyte- and petabyte-scale datasets rapidly becoming commonplace, but there is consensus that great value lies buried in them, waiting to be unlocked by the right computational tools. Big Data analysis tools like Map Reduce over Hadoop and HDFS, promises to help organizations better understand their customers and the marketplace, hopefully leading to better business decisions and competitive advantages. For engineers building information processing tools and applications, large and heterogeneous datasets which are generating continuous flow of data, lead to more effective algorithms for a wide range of tasks, from machine translation to spam detection.

REFERENCES :

[1] 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{23 2013).

[2] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache

[3] Aditya B. Patel, Manashvi Birla, Ushma Nair, (6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop

[4] Kyong-Ha Lee Hyunsik Choi "Parallel Data Processing with Map Reduce: A Survey" SIGMOD Record, December 2011 (Vol. 40, No. 4)

[5] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, "Shared disk big data analytics with Apache Hadoop", 2012, 18-22

[6] D.Rajasekar, C.Dhanamani, S.K. Sandhya "A Survey on Big Data Concepts and Tools" Volume 5 Issue 2 (February.2015) IJETAE-2250-2459.

[7] Albert Bifet "Mining Big Data In Real Time"Informatica 37 (2013) 15–20 DEC 2012.

[8] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013. 1994 2/13/04

[9] Dong, X.L.; Srivastava, D. Data Engineering (ICDE)," Big data integration"IEEE International Conference on , 29(2013) 1245–1248

[10] Kosha Kothari, Ompriya Kale "Survey of various Clustering Techniques for Big Data in Data Mining" Volume 1, Issue 7, 2014 IJIRT-2349-6002.