Volume: 08, Issue: 05 Pages: 61-66 (2017) Special Issue

Comparative Study on Tools and Techniques of Big Data Analysis

B.THILLAIESWARI M.S., M.Phil., B.Ed.,
Assistant Professor, Department of Computer Science,
TBAK College for Women, Kilakarai.
Email:thillaikris@gmail.com

ABSTRACT

Big data is the term for any group of datasets so huge and composite that it becomes difficult to practice using traditional data processing applications. Big data is a set of procedures and technologies that entail new forms of integration to uncover large unknown values from large datasets that are various, complex, and of a immense scale. Analyzing Big Data is a challenging task as it contains huge dispersed file systems which should be fault tolerant, flexible and scalable. There is an immense need of constructions, platforms, tools, techniques and algorithms to handle Big Data. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. In addition, In this work, a tool within the scope of InterIMAGE Cloud Platform (ICP), which is an open-source, distributed context for instinctive image interpretation, is presented. Data Mining Package, is able to execute supervised classification procedures on huge amounts of data, usually referred as big data, on a distributed infrastructure using Hadoop MapReduce. The tool has four classification algorithms implemented, taken from WEKA's machine learning library, namely: Decision Trees, Naïve Bayes, Random Forest and Support Vector Machines (SVM).

Key Words- **Big data computing, Algorithms, Tools, Methods, Issues.**

I.INTRODUCTION

This massive amount of the data is known as **Big Data.** Big data is a axiom, or catch-phrase, operates to designate a huge volume of both structured and unstructured data that is so huge that it's complex to practice using customary database and software techniques. Operations and make quicker, more intelligent decisions .**Big Data**, now a days this term becomes common in IT industries. As there is a huge amount of data lies in the industry but there is nothing before big data comes into picture [3].

An example of big data might be petabytes (1,024terabytes) or Exabyte's (1,024 petabytes) of data comprising of billions to trillions of records of millions of people—all from changed causes (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is naturally lightly structured data that is often inadequate and inaccessible.

1.1Big Data Parameters:

Big Data Analytics that is the handling of the difficult and enormous datasets This data is different from structured data in terms of five parameters –variety, volume, value, veracity and velocity (5V's).

The five V's (volume, variety, velocity, value, veracity) are the challenges of big data management are:

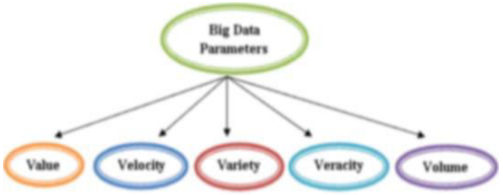

Fig 1: Big Data Parameters

a.Value:

It is a most important v in big data. Value is main tinkle for big data because it is significant for businesses, IT infrastructure system to store large amount of values in database.

b. Velocity:

The data comes at high speed. Sometimes one minute is too late so big data is time sensitive. Some administrations data velocity is central task. The social media posts and credit card trades done in millisecond and files created by this putting in to databases

c. Variety:

Data sources are tremendously heterogeneous. The records comes in many layouts and of any type, it may be structured or unstructured such as text, audio, videos, log files and more. The variations are boundless,

and the data enters the network without having been quantified or qualified in any way.

d. Volume:

Data is ever-increasing day by day of all types ever MB, PB, YB, ZB, KB, TB of information. The data outcomes into large files. Extreme capacity of data is highest dispute of storage. This main issue is determined by decreasing storage cost. Data volumes are predictable to grow 50 times by 2020.

e. Veracity:

The growth in the choice of standards typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dull data. Big data and analytics technologies work with these types of data.

1.2 Architecture:

Big Data are the collection of large amounts of unstructured data. Big Data means enormous amounts of data, such large that it is difficult to collect, store, manage, analyze, predict, visualize, and model the data.

Big Data architecture typically consists of three segments: storage system, handling and analyze.

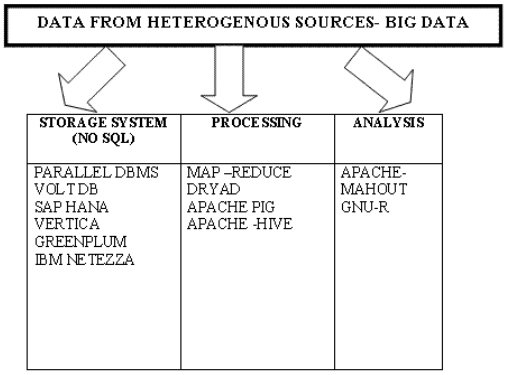

Fig. 2 Architecture of Big Data

Big Data usually vary from data warehouse in architecture; it follows a distributed approach whereas a data warehouse follows a centralized one.

One of the architecture arranged describes about adding new 6 rules were in the innovative 12 rules clear in the OLAP system well-defined the procedures [4] of data mining necessary for the analysis of data and defined SDA (standard data analysis) that helped analysis of data that is in gathered form and these were much well timed in evaluation with the decision taken in traditional methods.

On paper an Efficient Technique on Cluster Based Master Slave Architecture Design, the hybrid approach was formed which consists of both top down and bottom up approach. This hybrid approach when compared with the clustering and Apriori algorithm, takes less time in transaction than them.

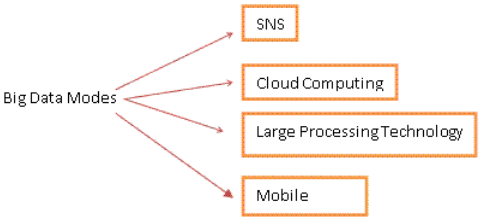

Fig. 3 Big Data modes

The Data Mining termed Knowledge discovery, in work done in Design Principles for Effective Knowledge Discovery from Big Data", its architecture was laid describing extracting knowledge from large data. Data was analyzed using software Hive and Hadoop.

1.3 Algorithm:

Many algorithms were defined earlier in the analysis of large data set. In the opening diverse Decision Tree Learning was used earlier to analyze the big data [6]. The approach is to have a single decision system created from a large and independent n subset of data. Whereas Patil uses a hybrid approach joining both genetic algorithm and decision tree to generate an optimized decision tree thus refining efficacy and performance of computation.

Then clustering techniques approached into existence. Different clustering techniques were being used to examine the data sets [8]. A new algorithm called GLC++ was established for large mixed data set unlike algorithm which contracts with large similar type of dataset. This method could be used with any kind of distance, or symmetric similarity function. Refer the table 1.

2. Technologies and Methods:

There are endless articles, books and periodicals that describe Big Data from a technology perspective so we will instead focus our efforts here on setting out some basic principles and the minimum technology foundation to help relate Big Data to the broader IM domain [12].

2.1. Hadoop:

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes [1]. It distributes the file among the nodes and allows to system continue work in case of a node failure. This approach reduces the risk of catastrophic system failure. Hadoop consists of the Hadoop kernel [10], Hadoop distributed file system (HDFS), maps reduce and related projects are zookeeper, Hbase, Apache Hive. Hadoop Distributed File System consists

of three Components: the Name Node, Secondary Name Node and Data Node.

Table 1: Different Decision Tree Algorithm

AUTHOR'S NAME	TECHNIQUE	CHARACTERISTIC	SEARCH TIME
N. Beckmann, H. -P. Kriegal, R. Schneider, B. Seeger [8]	R-Tree R*-Tree	Have performance bottleneck	O (3D)
S. Arya, D. Mount, N. Netanyahu,	Nearest Neighbor Search	Expensive when searching object is in High Dimensional space	R. Silverman, Grows exponentially with the size of the searching space A. Wu [O log(n) Less Time Consuming
Zhiwei Fu,Fannie Mae	Decision Tree C4-5	Practice local greedy search throughout dataset .	Less Time Consuming
D.V.Patil R.S.Bichkar	GA Tree (Desion Tree + Genetic Algorithm)	Improvement in classification performance and reduction in siza of tree,with no loss in classification accuracy	Improved performance problem like slow memory execution can be reduced
Yen-LingLu, Chin-Shvumg Fawn	Hierarchical Neural Network	High accuracy rate of recognizing data,have high classification accuracy	Less time consuming improved performance

2.1.1 Components of Hadoop:

HBase: It is open source, distributed and Non-relational database system implemented in Java. It runs above the layer of HDFS. It can serve the input and output for the Map Reduce in well-mannered structure.

Oozie: Oozie is a web-application that runs in a java servlet. Oozie use the database to gather the information of Workflow which is a collection of actions. It manages the Hadoop jobs in a mannered way.

Sqoop: Sqoop is a command-line interface application that provides platform which is used for converting data from relational databases and Hadoop or vice versa.

Avro: It is a system that provides functionality of data serialization and service of data exchange. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records.

Chukwa: Chukwa is a framework that is used for data collection and analysis to process and analyze the massive amount of logs. It is built on the upper layer of the HDFS and Map Reduce framework.

Pig: Pig is high-level platform where the MapReduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analyzed that occurs in high level language.

Zookeeper: It is a centralization based service that provides distributed synchronization and provides group services along with maintenance of the configuration information and records.

Hive: It is application developed for data warehouse that provides the SQL interface as well as relational model. Hive infrastructure is built on the top layer of Hadoop that help in providing conclusion, and analysis for respective queries.

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Hadoop is open- source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers. Hadoop has the following Characteristics,

Reliable: The software is fault tolerant; it expects and handles hardware and software failures

Scalable: Designed for massive scale of processors, memory, and local attached storage

Distributed: Handles replication. Offers massively parallel programming model, Map Reduce.

Hadoop is an Open Source implementation of a large-scale batch processing system.That uses the Map-Reduce framework introduced by Google by leveraging the concept of map and reduces functions that well known used in Functional Programming.

- The following discusses some of these differences Hadoop is particularly useful when: Complex information processing is needed:
- Unstructured data needs to be turned into structured data.
- Queries can't be reasonably expressed using SQL Heavily recursive algorithms.

- Complex but parallelizable algorithms needed, such as geo-spatial analysis or genome sequencing.
- Machine learning:
- Data sets are too large to fit into database RAM, discs, or require too many cores (10's of TB up to PB).

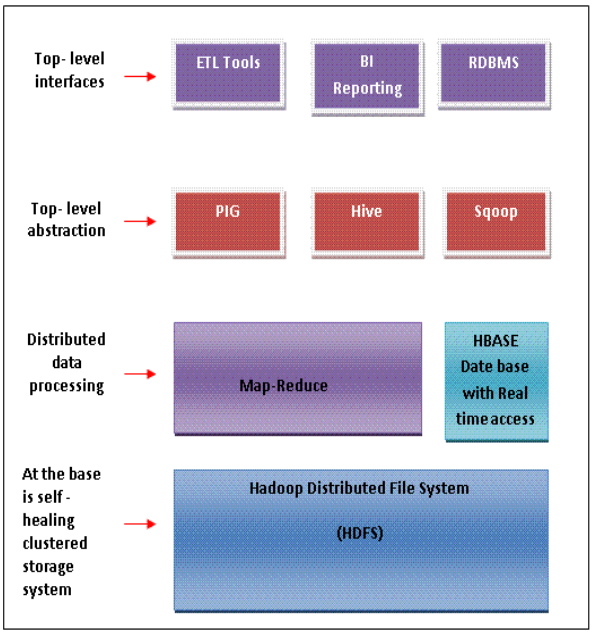

Fig 3:Architechture of Hadoop

Data value does not justify expense of constant real-time availability, such as

- archives or special interest info, which can be moved to Hadoop and remain available at lower cost.
- Results are not needed in real time Fault tolerance is critical.
- Significant custom coding would be required to:
- Handle job scheduling.

- Hadoop was inspired by Google's MapReduce, a software framework in which an application is broken down into numerous small parts. Any of these parts (also called fragments or blocks) can be run on any node in the cluster.

HDFS The Hadoop Distributed File System (HDFS) is the file system component of the Hadoop framework. HDFS is designed and optimized to store data over a large amount of low-cost hardware in a distributed fashion.

Name Node*:* Name node is a type of the master node, which is having the information that means meta data about the all data node there is address(use to talk), free space, data they store, active data node , passive data node, task tracker, job.

Data Node: Data node is a type of slave node in the Hadoop, which is used to save the data and there is task tracker in data node which is used to track on the ongoing job on the data node and the jobs which coming from name node.

2.2 Map Reduce:

Map-Reduce was introduced by Google in order to process and store large datasets on commodity hardware. Map Reduce is a model for processing large-scale data records in clusters. The Map Reduce programming model is based on two functions which are map( ) function and reduce( ) function. Users can simulate their own processing logics having well defined map( ) and reduce( ) functions. Map function performs the task as the master node takes the input, divide into smaller sub modules and distribute into slave nodes. A slave node further divides the sub modules again that lead to the hierarchical tree structure

The Data Node checks for the accurate namespace ID, and if not found then the Data Node automatically shuts down. New Data Nodes can join the cluster by simply registering with the Name Node and receiving the namespace ID. Each Data Node keeps track of a block report for the blocks in its node. Each Data Node sends its block report to the Name Node every hour so that the Name Node always has an up to date view of where block replicas are located in the cluster.

2.2.1 Map Reduce Components:

Name Node: manages HDFS metadata, doesn't deal with files directly.
Data Node: stores blocks of HDFS—defaults replication level for each block.
Job Tracker: schedules, allocates and monitors job execution on slaves—Task Trackers.
Task Tracker: runs Map Reduce operations.

2.2.2 Map Reduce Framework : Map Reduce is a software framework for distributed processing of large data sets on computer clusters. It is first developed by Google .Map Reduce is intended to facilitate and simplify the processing of vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.

2.3. Hive:

Hive is a distributed agent platform, a decentralized system for building applications by

networking local system resources. Apache Hive data warehousing component, an element of cloud-based Hadoop ecosystem which offers a query language called Hive QL that translates SQL-like queries into Map Reduce jobs automatically. Applications of apache hive are SQL, oracle, IBM DB2.

2.4. No-SQL:

No-SQL database is an approach to data management and data design that's useful for very large sets of distributed data. These databases are in general part of the real-time events that are detected in process deployed to inbound channels but can also be seen as an enabling technology following analytical capabilities such as relative search applications. developer in advance.

It is useful when enterprise need to access huge amount of unstructured data. There are more than one hundred No SQL approaches that specialize in management of different multimodal data types (from structured to non-structured) and with the aim to solve very specific challenges [11]. Data Scientist, Researchers and Business Analysts in specific pay more attention to agile approach that leads to prior insights into the data sets that may be concealed or constrained with a more formal development process.

.

2.5. HPCC:

HPCC is an open source platform used for computing and that provides the service for handling of massive big data work flow. HPCC system is a single platform having a single architecture and a single programming language used for the data simulation. HPCC system was designed to analyze the enormous amount of data for the purpose of solving complex problem of big data.

HPCC system is based on enterprise control language which has the declarative and on-procedural nature programming language the main components of HPCC are:

HPCC Data Refinery: Use parallel ETL engine mostly.

HPCC Data Delivery*:* It is massively based on structured query engine used. Enterprise Control Language distributes the workload between the nodes in appropriate even load.

3. Big Data Issues:

Security has always been an issue when data privacy is considered. Data integrity is one of the primary components when preservation of data is considered [11]. Access and sharing of Data which is not meant for public, has to be protected. For this type of security many researchers have been done. Security has always been an issue when data are considered.

They provide the architecture to store the data. Uses cloud computing to make the data unavailable to the intruder [12] .Data integrity is one of the primary components when preservation/security is considered. Hash functions were primarily used for preserving the integrity of the data. The drawback of using hash function is that a single hash can only identify the integrity of the single data string. And because of this drawback, it becomes impossible to locate the exact position within the string where the change has been occurring. The solution to overcome the above problem is to split the data string into the block and then protect each block by the hash function. Privacy Issues in Public Social Media, the very idea of privacy to the people who are using social media was explained.

4. Conclusion:

Due to Increase in the amount of data in the field of genomics, meteorology, biology, environmental research, it becomes difficult to handle the data, to find Associations, patterns and to analyze the large data sets. In this paper we have also discussed the challenges of Big data (volume, variety, velocity, value, veracity) and various advantages and a disadvantage of these Technologies .This paper discussed an architecture using Hadoop HDFS distributed data storage, real-time No SQL databases, and MapReduce distributed data processing over a cluster of commodity servers. The commercial impacts of the Big data have the potential to generate significant productivity growth for a number of vertical sectors. The main goal of our paper was to make a survey of various big data handling techniques those handle a massive amount of data from different sources and improves overall performance of systems. Growing talent and building teams to make analytic-based decisions is the key to realize the value of Big Data.

Reference:

[1] Ivanka Valova, Monique Noirhomme, "Processing Of Large Data Sets: Evolution, Opportunities And Challenges", Proceedings of PCaPAC08 .

[2] Joseph McKendrick, "Big Data, Big Challenges, Big Opportunities: 2012 IOUG Big Data Strategies Survey", IOUG, Sept 2012.

[3] Nigel Wallis, "Big Data in Canada: Challenging Complacency for Competitive Advantage", IDC, Dec 2012.

[4] "Big Data for Development: Challenges and Opportunities", Global Pulse, May 2012
Yuri Demchenko ―The Big Data Architecture Framework (BDAF) Outcome of the Brainstorming Session at the University of Amsterdam 17 July 2013.

[5] Amogh Pramod Kulkarni, Mahesh Khandewal, ―Survey on Hadoop and Introduction to YARN.

International 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).

[6] Sagiroglu, S.Sinanc, D., Big Data: A Review,2013, 20-24.
Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, ―Survey Paper On Big DataInternational Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.

[7] Margaret Rouse, April 2010―unstructured data.

[8] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013.

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

[10] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), ―Big Data Framework2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499

[11] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N ―Analysis of Big Data using Apache Hadoop and Map Reduce Volume 4, Issue 5, May 2014.

[12] Suman Arora, Dr.Madhu Goel, ―Survey Paper on Scheduling in Hadoop. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.

[13] Aditya B. Patel, Manashvi Birla and Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce," in Proc. 2012 Nirma University International Conference On Engineering.

[14] Jimmy Lin ―Map Reduce Is Good Enough? The control project, IEEE Computer 32 (2013).