Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Big Data Analysis using Hadoop: A Survey

Rotsnarani Sethy
PhD Scholar, Dept. of Computer Science
Utkal University
Bhubaneswar, India

Mrutyunjaya Panda
Reader, Dept. of Computer Science
Utkal University
Bhubaneswar, India

Abstract— Big data is a collection of large data sets that include different types such as structured, unstructured and semi-structured data. This data can be generated from different sources like social media, audios, images, log files, sensor data, transactional applications, web etc. To process or analyse this huge amount of data or extracting meaningful information is a challenging task now a days. Big data exceeds the processing capability of traditional database to capture, manage, and process the voluminous amount of data. In this paper we first introduce the general background of big data and then focus on hadoop platform using map reduce algorithm which provide the environment to implement application in distributed environment and it can capable of handling node failure.

Keywords— Big Data, Hadoop, HDFS, Map Reduce, Hadoop Components.

I. INTRODUCTION

In general, big data shall mean the datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time. Big Data describes any massive volume of structured, semi structured and unstructured data that are difficult to process using traditional database system such as RDBMS [1]. An example of big data may be Exabyte's (1024 terabytes) of data consisting of trillions of records of millions of people from different sources such as websites, social media, mobile data, web servers, online transactions and so on [2]. In the past, type of information available was limited. There was a well-defined set of technology approaches for managing information. But in today's world, the amount of data has been exploding. It has grown to terabytes and petabytes. Because in every minute, there are 280,000 tweets, more than 100 millions emails are sent. 2 million searching queries in Google, and more than 350 GB of data is processed in Face book in every minute. Some of the applications of big data are in areas such as social media, healthcare, traffic management, banking, retail, education and so on.

A. CHARACTERISTICS OF BIG DATA

As the data is too big and comes from various sources in different form, it is characterized by the following five components:

- **VARIETY**

Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data both from active passive devices. All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems.

- **VOLUME**

The Big word in Big data itself defines the volume. At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future. The social networking sites existing are themselves producing data in order of terabytes everyday and this amount of data is definitely difficult to be handled using the existing traditional systems.

- **VELOCITY**

Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example, the data from the sensor devices would be constantly moving to the database store and this amount won't be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

- **VARIABILITY**

Variability considers the inconsistencies of the data flow. Data loads become challenging to be maintained especially with the increase in usage of the social media which generally causes peak in data loads with certain events occurring.

- **COMPLEXITY**

It is quite an undertaking to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control.

II. TOOLS AND TECHNOLOGIES

For the purpose of processing the large amount of data, the big data requires exceptional technologies. The various techniques and technologies have been introduced for manipulating, analysing and visualizing the big data . There are many solutions to handle the Big Data, but the Hadoop is one of the most widely used technologies [3].

A. HADOOP

Using the solution provided by Google, Doug Cutting and his team developed an Open Source Project called HADOOP. Hadoop runs applications using the Map Reduce algorithm, where the data is processed in parallel with others. Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data. Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework 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.

It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines.

Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop perform. Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).These files are then distributed across various cluster nodes for further processing. HDFS, being on top of the local file system, supervises the processing [8]. Blocks are replicated for handling hardware failure. Checking that the code was executed successfully, performing the sort that takes place between the map and reduce stages, sending the sorted data to a certain computer, writing the debugging logs for each job.

Hadoop has two major layers namely:

- Processing/Computation layer (Map Reduce), and
- Storage layer (Hadoop Distributed File System).

B. MAP REDUCE

Map Reduce is a parallel programming model for writing distributed applications devised at Google for efficient 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 [4]. 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 mapper's 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 [5] 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.

C. 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. HDFS provides file permissions and authentication.Fig.1 briefly describe the HDFS architecture.

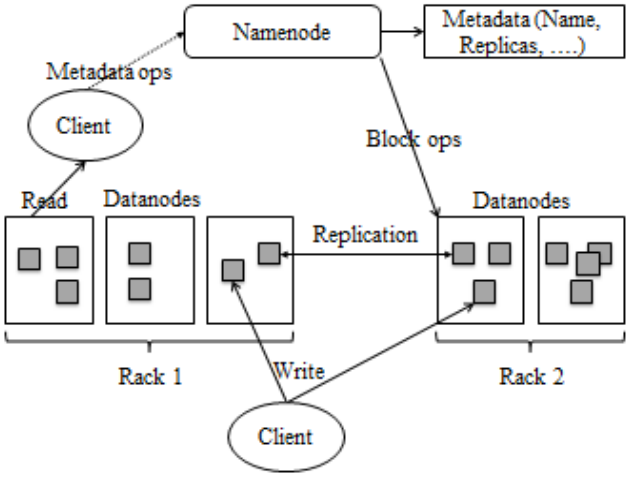

Fig.1 HDFS Architecture .[5]

HDFS 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 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.

OTHER DIFFERENT COMPONENTS [7] OF HADOOP ARE:

Apache Pig : software for analysing 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.

HBase 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. HBase is an example of a NoSQL 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.

Oozie: Oozie is a java based web-application that runs in a java servlet. Oozie 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.

Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption [9]. Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

III. DATA MINING FOR BIG DATA

Data mining is the process of extracting information from a large data sets and transform it into an understandable form for further use. Data mining can be used in such a case where database is large and the classification of such a data is difficult. There are many techniques used in data mining to process and mine the uncertain data [10]. Clustering is the important technique in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups called clusters. The research in Big Data analysis using data mining especially with clustering technique still considered to be young, which attracts many researchers to conduct further research in this potential area.

IV. LITERATURE REVIEW

The authors [1] pointed out that handling of huge data using earlier RDBMS tools is little bit complex, hence feels the necessity of alternate tools that can handle such a huge data which is usually referred to as 'big data**".** In this, the authors argued that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc.The authors [2] discussed the analysis of big data and they stated that Data is generated through many sources like business processes, transactions, social networking sites, web servers, etc. and remains in structured as well as unstructured form . Processing or analysing the huge amount of data or extracting meaningful information is a challenging task. The term "Big data" is used for large data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many peta bytes of data in a single data set. Difficulties include capture, storage, search, sharing, analytics and visualizing. The Authors [3] **have** done a lot of experiment on the big data problem. At last he found that the hadoop cluster, Hadoop Distributed File System (HDFS) for storage and map reduce method for parallel processing on a large volume of data. The Authors [4] emphasizes on a prominent data processing tool Map Reduce survey which will help in understanding various technical aspects of the Map Reduce framework. In this survey, the author expresses different views on Map Reduce framework and introduces its optimization strategies. Author also hands a challenge on parallel data analysis with Map Reduce framework. The Authors [5] defines big data Problem using Hadoop and Map Reduce" reports the experimental research on the Big data problems in various domains. It describe the optimal and efficient solutions using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage data and Map Reduce framework for parallel processing to process massive data sets and records. The Authors [6] discussed an overview of big data's concept, tools, techniques, applications, advantages and challenges. They used Hadoop technology for the implementation purpose. The authors have briefly discussed about HDFS and Map Reduce technology to process massive data sets and records. The Authors [7] Stated that streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, allowing organizations to react quickly when problem appear or detect to improve performance. Huge amount of data is created everyday termed as "big data". The tools used for mining big data are apache hadoop, apache pig, cascading, scribe, storm, apache Hbase, apache mahout, MOA, etc. Thus, he instructed that our ability to handle many Exabyte's of data mainly dependent on existence of rich variety dataset, technique, software framework. The Authors [8] discussed that the big data refers to a collection of vast amount of structured, unstructured and semi structured data that are very difficult to manage, process and to store using common database management system. To store and manage this huge amount of different data the data storage technique are used such as clustered network attached storage (NAS) and object based storage. The hadoop architecture is best in this case to manage and different data structure using map reduce method. The Authors [9] talks about the theoretical assumptions, that improves the performance of Hadoop/map reduce and purposed the optimal reduce task assignment schemes that minimize the fetching cost per job and performs the both simulation and real system deployment with experimental evolution. The advantage of this paper is improves the performance of large scale Hadoop clusters. The disadvantage of this paper is environmental factors such as network topologies effect on a reduce task in map reduce clusters. The Authors [10] Discussed about the Data Mining and some Clustering Techniques for the criteria's of big data. They also stated that with the beginning in the era of big data, the data is increasing at rapid speed not only in size but also in variety. There come challenges and difficulties to handle such large amount of data with the growing data. Big data exhibits different characteristics like volume, variety, variability, value, velocity and complexity due to which it is very difficult to analyse data and obtain information with traditional data mining techniques.

V. CONCLUSION AND FUTURE WORK

Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Today, Data is generated from various different sources and can arrive in the system at various rates. To process these large amounts of data is a big issue today. In this paper we discussed Hadoop tool for Big data in detail. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. We also discussed some hadoop components which are used to support the processing of large data sets in distributed computing environments. In future we can use some clustering techniques and check the performance by implementing it in hadoop.

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 Hadoop and Map Reduce"** Volume 4, Issue 5, May 2014" 27

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

[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.

