International Journal of Innovative Research in Information Security (IJIRIS) ISSN: 2349-7017(O)
Volume 1 Issue 3 (September 2014) ISSN: 2349-7009(P)
www.ijiris.com

SURVEY ON BIG DATA PROCESSING USING HADOOP, MAP REDUCE

N.Alamelu Menaka *
Department of Computer Applications

Dr.Jabasheela
Department of Computer Applications

Abstract-We are in the age of big data which involves collection of large datasets.Managing and processing large data sets is difficult with existing traditional database systems.Hadoop and Map Reduce has become one of the most powerful and popular tools for big data processing . Hadoop Map Reduce a powerful programming model is used for analyzing large set of data with parallelization, fault tolerance and load balancing and other features are it is elastic,scalable,efficient.MapReduce with cloud is combined to form a framework for storage, processing and analysis of massive machine maintenance data in a cloud computing environment.

Keywords: Big data, Hadoop, Mapreduce, HDFS,Cloud

I. INTRODUCTION

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, accurate, 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.

Big data is data that is too big, too fast or too hard for existing systems and algorithms to handle.Big Data technology improves performance, facilitates innovation in the products and services of business models, and provides decision making support. Big Data technology aims to minimize hardware and processing costs and to verify the value of Big Data before committing significant company resources.

Properly managed Big Data are accessible, reliable, secure, and manageable. Hence, Big Data applications can be applied in various complex scientific disciplines (either single or interdisciplinary), including atmospheric science, astronomy, medicine, biology, genomics, and biogeochemistry.

II. BIG DATA PROCESSING

Big Data processing techniques analyze big data sets at terabyte or even petabyte scale

A. Hadoop

Hadoop is used for storing and processing big data

Apache Hadoop is a distributed computing framework modeled after Google MapReduce to process large amounts of data in parallel. Once in a while, the first thing that comes to my mind when speaking about distributed computing is EJB. EJB is de facto a component model with remote capability but short of the critical features being a distributed computing framework that include computational parallelization, work distribution, and tolerance to unreliable hardware and software.

Hadoop on the other hand has these merits built-in. ZooKeeper modeled on Google Chubby is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and group services for the Hadoop cluster.

Hadoop Distributed File System (HFDS) modeled on Google GFS is the underlying file system of a Hadoop cluster. HDFS works more efficiently with a few large data files than numerous small files. A real-world Hadoop job typically takes minutes to hours to complete, therefore Hadoop is not for real-time analytics, but rather for offline, batch data processing. Recently, Hadoop has undergone a complete overhaul for improved maintainability and manageability. Something called YARN (Yet another Resource Negotiator) is at the center of this change. One major objective of Hadoop YARN is to decouple Hadoop from MapReduce paradigm to accommodate other parallel computing models, such as MPI (Message Passing Interface) and Spark.

In general, data flows from components to components in an enterprise application. This is the case for application frameworks (EJB and spring framework), integration engines (Camel and Spring Integration), as well as ESB (Enterprise Service Bus) products. Nevertheless, for the data-intensive processes Hadoop deals with, it makes better sense to load a

big data set once and perform various analysis jobs locally to minimize IO and network cost, the so-called "Move-Code-To-Data" philosophy. When you load a big data file to HDFS, the file is split into chunks (or file blocks) through a centralized Name Node (master node) and resides on individual Data Nodes (slave nodes) in the Hadoop cluster for parallel processing.

B. Hadoop Architecture

Hadoop has master slave relationship.

1. Structure of Hadoop Slaves contain two nodes: Task Tracker, Data Node.

a. Task Tracker: Job of task tracker is to process small piece of task that has been given to this particular node.

b. Data Node: Job of this data node is to manage the data that has been given to particular node.

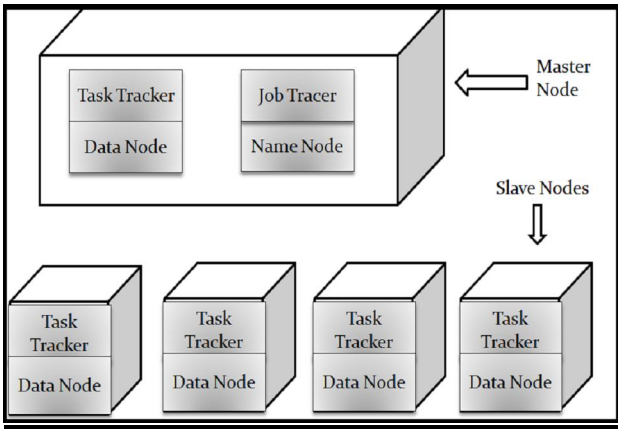

Fig. 1: Stucture of Hadoop

Master has two additional nodes: Name Node, Job tracker.

1. **Name Node:** Name node is responsible for keeping the index of which data is residing on which data node. Hence, when application contacts the name node, it tells the application to go to particular node to get required data.

2. **Job Tracker:** Role of job tracker running on master is to break the bigger task into smaller modules and send each module of computation to task tracker. Task trackers will perform small tasks and send result back to job tracker. Job tracker will combine all results and send back to the application

C. Map Reduce

Hadoop MapReduce: software framework for easily writing applications which process vast amounts of data in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner.

In the Map Reduce programming paradigm, the input data is splitted into many pieces and these pieces are given to map processes running on different machines. Then the outputs of these map processes are given to many reduce processes which are the final stages of the execution. This model can be seen

on Figure 1. All input and output data structures are key/value pairs and the framework is able to distribute data between processing nodes based on the key of these pairs. Working with key/value pairs does not imply any restrictions because the key and the value can be any object types.

MapReduce-based distributed systems ensure fast software development, because the developer only needs to care about map and reduce methods. Hadoop is able to run MapReduce algorithms on unlimited number of processing nodes and it optimizes task distribution in a way that data communication overhead is minimal between the machines. In case of hundreds or thousands of processing nodes,it needs to handle faulty machines and network problems, because these events occur quite often in big clusters

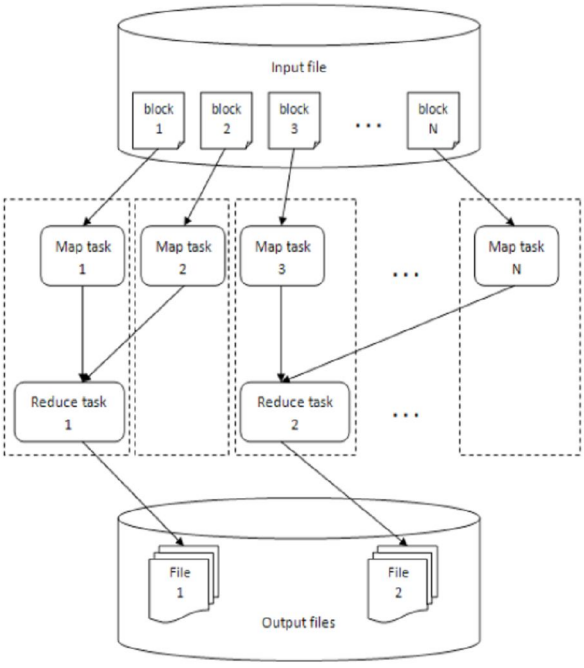

Fig2.Map Reduce Model.

D. MapReduce tasks.

Steps Tasks

(1) Input
  (a.) Data are loaded into HDFS in blocks and distributed to data nodes
  (b) Blocks are replicated in case of failures
  (c) The name node tracks the blocks and data nodes
(2) Job Submits the job and its details to the Job Tracker
(3) Job initialization
  (a)The Job Tracker interacts with the Task Tracker on each data node
  (b) All tasks are scheduled
(4) Mapping
  (a)The Mapper processes the data blocks
  (b) Key value pairs are listed
(5) Sorting the Mapper sorts the list of key value pairs
(6) Shuffling
  (a)The mapped output is transferred to the Reducers
  (b) Values are rearranged in a sorted format
(7) Reduction Reducers merge the list of key value pairs to generate the final result
(8) Result
  (a) Values are stored in HDFS
  (b) Results are replicated according to the configuration
  (c) Clients read the results from the HDFS

III. MANAGEMENT TOOLS

With the evolution of computing technology, immense volumes can be managed without requiring supercomputers and high cost. Many tools and techniques are available for data management, including Google BigTable, Simple DB, Not Only SQL (NoSQL), DataStream Management System (DSMS), MemcacheDB, and Voldemort . However, companies must develop special tools and technologies that can store, access, and analyze large amounts of data in near-real time because Big Data differs from the traditional data and cannot be stored in a single machine. Furthermore, Big Data lacks the structure of traditional data. For Big Data, some of the most commonly used tools and techniques are Hadoop,MapReduce, and Big Table.

These innovations have redefined data management because they effectively process large amounts of data efficiently, cost effectively,and in a timely manner.

A. Hadoop components and their functionalities.

Hadoop component Functions
(1) HDFS -Storage and replication
(2) MapReduce- Distributed processing and fault tolerance
(3) HBASE -Fast read/write access
(4) HCatalog -Metadata
(5) Pig -Scripting
(6) Hive- SQL
(7) Oozie- Workflow and scheduling
(8) ZooKeeper- Coordination
(9) Kafka -Messaging and data integration
(10) Mahout- Machine learning

B. Big Data techniques and technologies

To capture the value from Big Data, we need to develop new techniques and technologies for analyzing it. Until now, scientists have developed a wide variety of techniques and technologies to capture, curate, analyze and visualize Big Data. Even so, they are far away from meeting variety of needs. These techniques and technologies cross a number of discipline, including computer science, economics, mathematics, statistics and other expertises. Multidisciplinary methods are needed to discover the valuable information from Big Data. We will discuss current techniques and technologies for exploiting data intensive applications.

We need tools (platforms) to make sense of Big Data. Current tools concentrate on three classes, namely, batch processing tools, stream processing tools, and interactive analysis tools.

Most batch processing tools are based on the Apache Hadoop infrastructure, such as Mahout and Dryad. The latter is more like necessary for real-time analytic for stream data applications.

Storm and S4 are good examples for large scale streaming data analytic platforms. The interactive analysis processes the data in an interactive environment, allowing users to undertake their own analysis of information. The user is directly connected to the computer and hence can interact with it in real time. The data can be reviewed, compared and analyzed in tabular or graphic format or both at the same time. Google's Dremel and Apache Drill are Big Data platforms based on interactive analysis.

Big Data techniques involve a number of disciplines, including statistics, data mining, machine learning, neural networks, social network analysis, signal processing, pattern recognition, optimization methods and visualization approaches.

There are many specific techniques in these disciplines, and they overlap with each other hourly as illustrated in Fig 3.

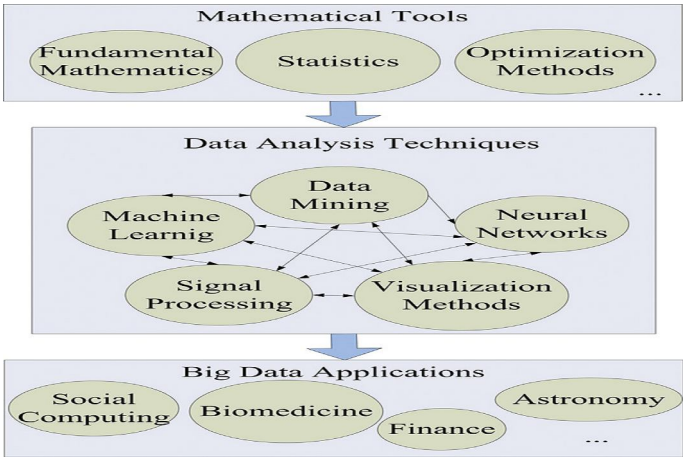

Fig 3. Big Data Techniques

IV. CHALLENGES IN CLOUD

Cloud computing is a highly feasible technology and attract a large number of researchers to develop it and try to apply to

Big Data problems. Usually, we need to combine the distributed MapReduce and cloud computing to get an effective answer for providing petabyte-scale computing CloudView is a framework for storage, processing and analysis of massive machine maintenance data in a cloud computing environment, which is formulated using the Map/Reduce model and reaches real-time response.

Apart from its flexibility, cloud computing addresses one of the challenges relating to transferring and sharing data, because data sets and analysis results held in the cloud can be shared with others.

V. CONCLUSION

This paper explains the big data concept and its importance in business. The technologies used for big data processing mainly Hadoop and Map Reduce. Management tools for big data are explained and also the big data techniques are discussed. Challenges of big data related to cloud is given here.

REFERENCES

[1] Big Data: Survey, Technologies, Opportunities, and Challenges Nawsher Khan,1,2 Ibrar Yaqoob,1 Ibrahi Abaker Targio Hashem,1 Zakira Inayat,1,3 Waleed Kamaleldin Mahmoud Ali,1 Muhammad Alam,4,5Muhammad Shiraz,1 and Abdullah Gani1

[2] The emergence of “big data” technology and analytics Bernice Purcell –Holy Family University.

[3] The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013 by Mike Gualtieri, January 3, 2013

[4] D.Borthakur. The hadoop distributed file system: Architecture and design. Hadoop Project Website, 2007.

[5] C Lam and J. Warren. Hadoop in Action. Manning Publications, 2010.

[6] Arshdeep Bahga, Vijay K. Madisetti, Analyzing massive machine maintenance data in a computing cloud, IEEE Trans Parallel Distrib. Syst. 23 (2012) 1831–1843.

[7] Ching-Yung Lin, Lynn Wu, Zhen Wen, Hanghang Tong, Vicky Griffiths-Fisher, Lei Shi, David Lubensky, Social network analysis in enterprise, Proc. IEEE 100 (9) (2012) 2759–2776.

