IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Big Data: Exploring Hadoop, MapReduce & HDFS in Data Mining

Prof. Datey Zuhaib Khalil[1], Prof. M. Jhansi Lakshmi[2]

M.Tech Computer Science & Engineering (Pursuing), Global Institute of Engineering & Technology Hyderabad,

Lecturer at A. R. Kalsekar Polytechnic Panvel (Navi-Mumbai)[1]

M.Tech Computer Science & Engineering, Associate Prof. Global Institute of Engineering & Technology, Hyderabad[2]

Abstract: The Global digital content created will increase some 30 times over the next ten years – to 35 zetta bytes, this unstoppable increase in data challenges business problems, a big data represents a large and rapidly growing volume of information that is mostly untapped by existing analytical applications and data warehousing systems. To analyze this enormous amount of data Hadoop can be used. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. The technologies used by big data application to handle the massive data are Hadoop, Map Reduce, Apache Hive, No SQL and HPCC. This paper highlights the big data and the new method of hadoop, MapReduce and HDFS to tackle the problem of big data.

Keywords: Data Mining, Big data, Structured data, BI, Big Data analytics, OLAP, EDA, Neural Networks, Hadoop and MapReduce technique, Advantages, Disadvantages.

I. INTRODUCTION

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time, it is important to realize that big data comes in many shapes and sizes. It also has many different uses – real-time fraud detection, web display advertising and competitive analysis, call center optimization, social media and sentiment analysis, intelligent traffic management and smart power grids, all of these analytical solutions involve significant (and growing) volumes of both multi-structured and structured data. Big Data is a term that refers to dataset whose volume (size), complexity and rate of growth (velocity) make them difficult to captured, managed, processed or analyzed by conventional technology and tools such as relational databases. There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big 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.

Experts analyzed two types of big data:

Structured data involved numbers and words that could be easily categorized—generated by network sensors embedded in electronic devices (smartphones and GPS [global positioning system] devices)—and numeric documents such as sales figures, account balances, and transaction data.

Unstructured data included more-complex, narrative information (such as customer reviews and comments from commercial Web sites) as well as photos and multimedia. The connective tissue between those data was natural language and message—requiring keywords that served as searchable terms to uncover patterns of relevance.

II. CHALLENGES OF BIG DATA

A. Heterogeneity and Incompleteness : The computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work. One big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses.

B. Scale : For many decades managing large and rapidly increasing volumes of data has been a challenging issue . In the past, this challenge was alleviated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift happening now: data volume is scaling faster than compute resources, and CPU speeds are static

Now, due to power constraints, clock speeds have largely stalled and processors are being built with increasing numbers of cores. with predictions of "dark silicon"(Dark Silicon refers to the exponentially increasing number of a

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

chip's transistors that must remain passive, or "dark", in order to stay within a chip's power budget), namely that power consideration will likely in the future prohibit us from using all of the hardware in the system continuously. Data processing systems will likely have to actively manage the power consumption of the processor. These unprecedented changes require us to rethink how we design, build and operate data processing components.

C. Timeliness : Longer time needs to analyze the larger data set. However, there are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements in it that meet a specified criterion. It is impractical to Scanning the entire data set to find suitable elements. Rather, index structures are created in advance to permit finding qualifying elements quickly. In doing so, each index structure is designed to support only some classes of criteria. When new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria. New index structures are required to support such queries. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

D. Value: The potential value of Big data is huge. Value is main source for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

E. Veracity: Veracity refers to noise , biases and ad normality When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data.

III. HADOOP: SOLUTION FOR BIG DATA PROCESSING

A. Hadoop is an Apache open source framework written in Java that allows distributed processing of large dataset across cluster of computers using simple programming model Hadoop creates cluster of machines and coordinates work among them . It is designed to scale up from single servers to thousands of machines, each offering local computation and storage Hadoop consists of two component Hadoop Distributed File System(HDFS) and MapReduce Framework.

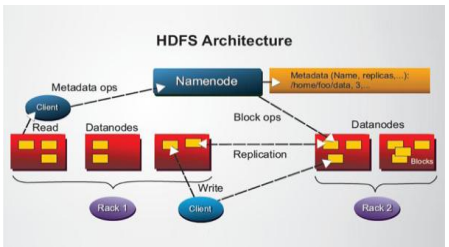

Fig: Hadoop Architecture

B. HDFS (Hadoop Distributed File System) HDFS is a file system designed for storing very large files with streaming data access pattern, running clusters on commodity hardware. HDFS manages storage on the cluster by breaking incoming files into pieces called 'blocks' and stores each blocks redundantly across the pool of the server. HDFS stores three copies of each file by copying each piece to three different servers. Size of each block 64MB. HDFS architecture is broadly divided into following three nodes which are Name Node, Data Node, HDFS Clients/Edge Node.

1. Name Node It is centrally placed node, which contains information about Hadoop file system . The main task of name node is that it records all the metadata & attributes and specific locations of files & data blocks in the data nodes. Name node acts as the master node as it stores all the information about the system .and provides information which is newly added, modified and removed from data nodes.

2. Data Node It works as slave node. Hadoop environment may contain more than one data nodes based on capacity and performance. A data node performs two main tasks storing a block in HDFS and acts as the platform for running jobs.

3. HDFS Clients/Edge node HDFS Clients sometimes also know as Edge node . It acts as linker between name node and data nodes. Hadoop cluster there is only one client but there are also many depending upon performance needs .

C. Hadoop Cluster:

In a Hadoop cluster, data is distributed to all the nodes of the cluster present on which data can be loaded as shown in fig. 4. The Hadoop Distributed File System (HDFS) will do this distribution of large data files into chunks which are managed by different nodes in the cluster [9]. An active monitoring system then re-replicates the data in response to system failures (if occurs) which can provide partial storage. Even though the file chunks are replicated and distributed across number of machines, they form a single namespace, so their contents are universally accessible.

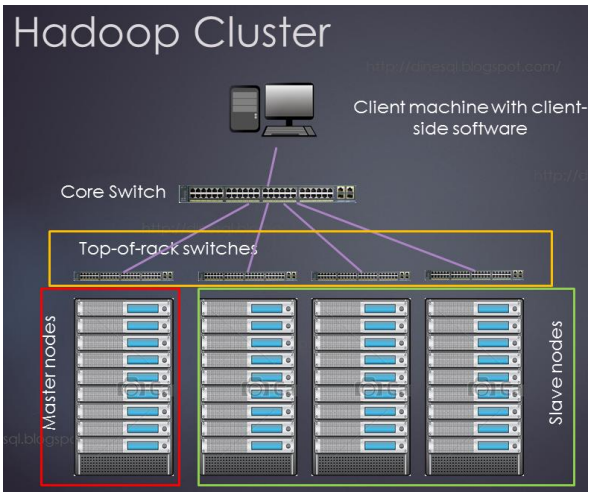

Fig: Hadoop Cluster

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Data is conceptually record-oriented in the Hadoop programming framework. Since files are spread across the distributed file system, each compute process which is running on a node operates on a subset of the data. This strategy of moving computation to the data allows Hadoop to achieve high data locality which in turn results in high performance.

D. MapReduce:

It is a distributed processing framework where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. MapReduce is defined as a programming model for processing and generating large sets of data. There are two phases in MapReduce, the "Map" phase and the "Reduce" phase. The system splits the input data into multiple chunks, each of which is assigned a map task that can process the data in parallel.

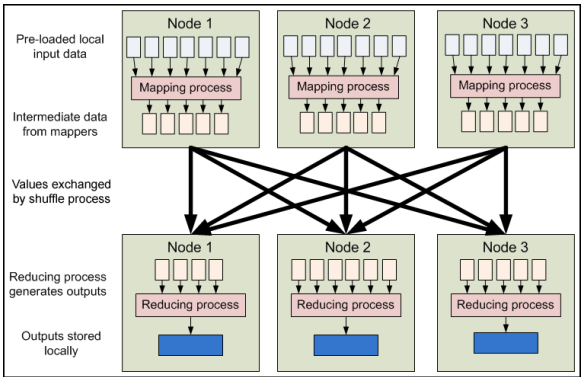

Fig: How Map Reduce Works

Each map task reads the input as a set of (key, value) pairs and produces a transformed set of (key, value) pairs as the output. The framework shuffles and sorts outputs of the map tasks, sending the intermediate (key, value) pairs to reduce task, which groups them into final results.

IV. CONCLUSION

In this paper I have explained the concept of Big data, operational and Analytical processing systems. I have also discussed about the problems and challenges of handling big data, the paper highlights how Hadoop and its techniques like Map Reduce and HDFS can simplify and solve the problems of big data. In this paper I have tried to cover all the details of Hadoop and its core components like Map Reduce and HDFS.

V. REFERENCES

1) Apache Hadoop: http://Hadoop.apache.org
2) Hadoop Distributed File System, http://hadoop.apache.org/hdfs
3) Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data.(Bhawna Gupta
4) Big Data by DT Editorial Services.
5) Dean, J. and Ghemawat, S., "MapReduce: a flexible data processing tool", ACM 2010.
6) Big Data Challenges and Opportunities by Roberto V. Zicari.
7) Survey Paper on Big Data and Hadoop by Varsha B.Bobade.
8) Introduction to Hadoop and MapReduce by Ramesh Sekharan and Peeyush Bishnoi from Yahoo.