International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume2, Issue 8 (August '2016)

A STUDY ON OVERVIEW OF BIG DATA, HADOOP AND MAP REDUCE TECHNIQUES

V.Shanmugapriya,
Research Scholar,
Periyar University,
Salem, Tamilnadu.

Dr.D.Maruthanayagam,
Assistant Professor,
Sri Vijay Vidyalaya College of Arts & Science,
Dharmapuri,Tamilnadu.

Abstract: We have entered the big data era. Organizations are capturing, storing, and analyzing data that has high volume, velocity, and variety and comes from a variety of new sources, including social media, machines, log files, video, text, image, RFID (Radio Frequency Identification), and GPS (Global Positioning System). These sources have strained the capabilities of traditional relational database management systems and spawned a host of new technologies, approaches, and platforms. Big data is data that exceeds the processing capacity of traditional databases. The data is too big to be processed by a single machine. New and innovative methods are required to process and store such large volumes of data. This paper presents an overview of Big Data, Big Data Analytics and Big Data technologies, Hadoop Distributed File System and MapReduce.

Keywords: Big Data, database, Hadoop, MapReduce and HDFS

I.INTRODUCTION

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. Big data may be important to business – and society – as the Internet has become. Big Data is so large that it's difficult to process using traditional database and software techniques. More data may lead to more accurate analyses. More accurate analyses may lead to more confident decision making. And better decisions can mean greater operational efficiencies, cost reductions and reduced risk. Analyzing big data is one of the challenges for researchers system and academicians that needs special analyzing techniques. Big data analytics is the process of examining big data to uncover hidden patterns, unknown correlations and other useful information that can be used to make better decisions. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data ("big data") to discover patterns and other useful information. Not only will big data analytics help you to understand the information contained within the data, but it will also help identify the data that is most important to the business and future business decisions. Big data analysts basically want the *knowledge* that comes from analyzing the data. **HDFS (*Hadoop Distributed File System)***, is a distributed file system designed to run on commodity hardware. It is inspired by the Google File System. Hadoop [1] [2] is based on a simple data model, any data will fit. HDFS designed to hold very large amounts of data (terabytes or petabytes or even zettabytes), and provide high-throughput access to this information. Hadoop Map Reduce is a technique which analysis big data. MapReduce has recently emerged as a new paradigm for large-scale data analysis due to its high scalability, fine-grained fault tolerance and easy programming model. This term MapReduce actually refers to two separate and distinct tasks map and reduce that Hadoop programs perform.

1.1 What is Big Data?

The term "Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media in 2005 in order to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data. Big Data [3] is the large amounts of data that is collected with time and are difficult to analyze using the traditional database system tools. This data comes from everywhere: posts from social media sites, digital videos and pictures, sensors used to gather climate information, cell phone **GPS** signals, and online purchase transaction records, to name a few. According to MiKE 2.0, the open source standard for Information Management, Big Data is defined by its size, comprising a large, complex and independent collection of data sets, each with the potential to interact. In addition, an important aspect of Big Data is the fact that it cannot be handled with standard data management techniques due to the inconsistency and unpredictability of the possible combinations. [5]

1.2 Big Data Characteristics

Big Data is characterized into four dimensions called 4V's; Volume, Velocity, Variety, Veracity as depicted in Figure 1. Aside, another dimension V (Value/Valor) also used to characterize the quality of the data.

Volume: Volume is concerned about scale of data i.e. the volume of the data at which it is growing. According to IDC [4] report, the volume of data will reach to 40 Zeta bytes by 2020 and increase of 400 times by now. The volume of data is growing rapidly, due to several applications of business, social, web and scientific explorations.

Velocity: The speed at which data is increasing thus demanding analysis of streaming data. The velocity is due to growing speed of business intelligence applications such as trading, transaction of telecom and banking domain, growing number of internet connections with the increased usage of internet, growing number of sensor networks and wearable sensors.

Variety: It depicts different forms of data to use for analysis such as structured like relational databases, semi structured like XML and unstructured like video, text.

Veracity: Veracity is concerned with uncertainty or inaccuracy of the data. In many cases the data will be inaccurate hence filtering and selecting the data which is actually needed is really a 10 cumbersome activity. A lot of statistical and analytical process has to go for data cleansing for choosing intrinsic data for decision making.

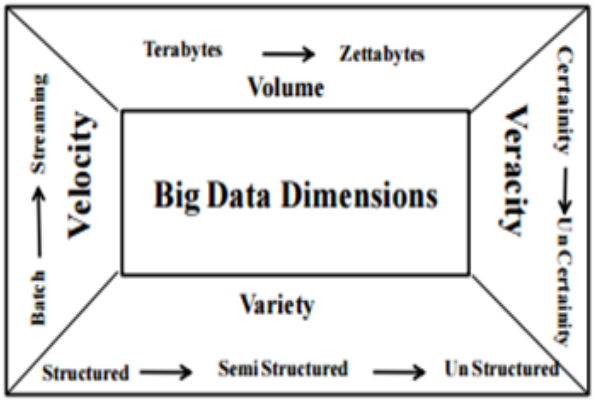


Figure 1: Data Dimensions 4V's

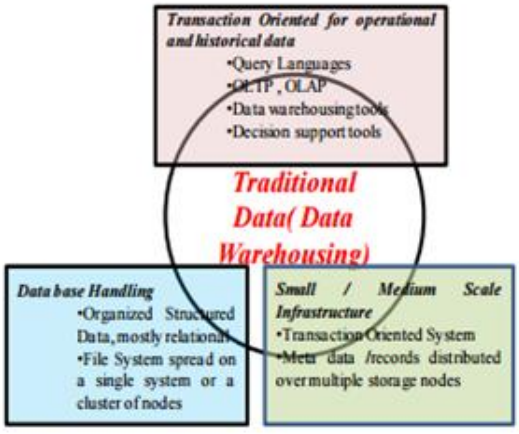
1.3Traditional Database (both Operational OLTP and Warehousing OLAP data)

Bill Inmo [6] described data warehousing as subject oriented, integrated, time-variant, and nonvolatile collection of data, and helping analysts in decision making process. Data warehouse is segregated from the organization's operational database. The operational database undergoes the per day transactions (On Line Transaction Processing – OLTP) which causes the frequent changes to the data on daily basis. Traditional databases typically addresses the applications for business intelligence, however, lack in providing the solutions for unstructured large volumes rapidly changing analytics in business and scientific computing. The several processing techniques under data warehouse are described below.

- Analytical processing involves analyzing the data by means of basic OLAP (Online Analytical Processing) operations, including slice-and-dice, drill down, drill up, and pivoting.
- Knowledge discovery through mining techniques by finding the pattern and associations, constructing analytical models, performing classification and prediction. These mining results can be presented using visualization tools.

Big Database: Big Data addresses the data management and analysis issues in several areas of business intelligence, engineering and scientific explorations. Traditional databases segregate the operational and historical data for operational and analysis reasoning, which are mostly structured. However, Big Data bases address the data analytics over an integrated scale out compute and data platform for unstructured data in near real time. Figure 2 depicts several issues in Traditional data (Data warehousing OLTP/OLAP) and Big Data technologies which are classified into major areas like infrastructure, data handling and decision support software as described below.

- Decision support / intelligent software tools: Big Data technologies address various decision supporting tools for searching the large data volumes and constructs the relations and extract the information based on several analytical methods. These tools would address several machine learning techniques, decision support systems and statistical modeling tools.
- Large scale data handling: rapidly growing data distributed over several storages and compute nodes with multi-dimensional data formats.
- Large scale infrastructure: scale out infrastructure for efficient storage and processing.
- Batch and stream support: capability to handle both batch and stream computation.



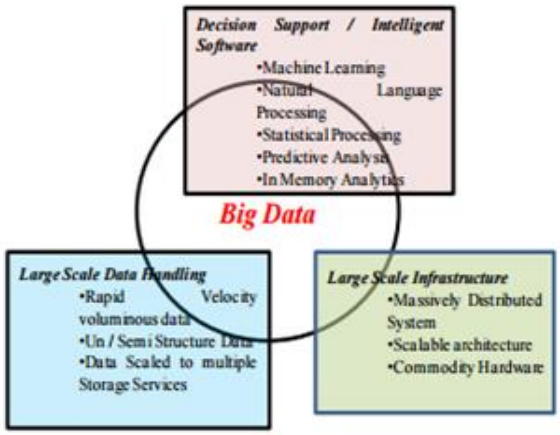


Figure 2: Big Data Vs Traditional Data (Data Warehousing) models

1.4. Managing and Analyzing Big Data

The most important question that arises at this point of time is how do we store and process such huge amount of data; most of which is raw, semi structured, and may be unstructured. Big data platforms are categorized depending on how to store and process them in a scalable, fault tolerant and efficient manner [7]. Two important information management styles for handling big data are relational DBMS products enhanced for systematic workloads (often known as analytic RDBMSs, or ADBMSs) and non-relational techniques (sometimes known as NOSQL systems) for handling raw, semi structured and unstructured data. Non-relational techniques can be used to produce statistics from big data, or to preprocess big information before it is combined into a data warehouse.

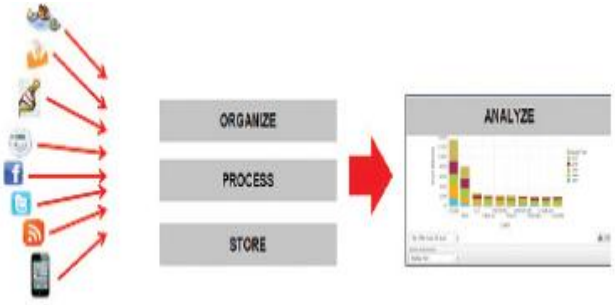


Figure 3: Big Data Management

Analytics – Big Data techniques

Analytics is the process of analyzing the data using statistical models, data mining techniques and computing technologies. It combines the traditional analysis techniques and mathematical models to derive information. Analytics and analysis performs the same function, however, analytics is the application of science to analysis. Big Data Analytics refers to a set of procedures and statistical models to extract the information from a large variety of data sets. A few major Big Data Analytics application areas are discussed below.

Text analytics: The process [9] of deriving information from text sources. The text sources forms of semi-structured data that include web materials, blogs and social media postings (such as tweets). The technology within text analytics comes from fundamental fields including linguistics, statistics and machine learning. In general, modern text analytics uses statistical models, coupled with linguistics theories, to capture patterns in human languages such that machines can understand the meaning of texts and perform various text analytics tasks. Text mining in the area of sentiment analysis helps organizations uncover sentiments to improve their customer relationship management.

In **Memory analytics:** In memory analytics [11] is the process which ingests the large amounts of data from a variety of sources directly into the system memory for efficient query and calculation performance. In-memory analytics is an approach for querying data when it resides in a computer's random access memory (RAM), as opposed of querying data stored on physical disks. This results in vastly shortened query response times, allowing business intelligence (BI) applications to support faster business decisions.

Predictive analysis: Predictive analysis [10] is the process of predicting future or unknown events with the help of statistics, modeling, machine learning and data mining by analyzing current and historical facts.

Graph analytics: Graph analytics [8] studies the behavior analysis of various connected components, especially useful in social networking web sites to find the weak or strong group

1.5 Technologies for Big Data

Big data has great potential to produce useful information for companies which can benefit the way they manage their business solutions. MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster [12]. MapReduce is divided into two broad steps:

- Map : Mapper performs the task of filtering and sorting
- Reduce: Reducer performs the task of summarizing the result. There can be multiple reducers to parallelize the aggregations.

Some other technologies for handling big data

1. Parallel Computing – It involves the processing data on several machines simultaneously, each running its own OS, memory, computation speed and works on different parts of data. The output is communicated via message passing. Thus parallel computing helps in reducing the time for analysis of big data greatly.

2. Distributed File System – One or more central servers store files that can be accessed, with proper authorization rights, by any number of remote clients in the network. The

distributed system uses a uniform naming convention and a mapping scheme to keep track of where files are located. When the client device retrieves a file from the server, the file appears as a normal file on the client machine, and the user is able to work with the file in the same ways as if it were stored locally on the workstation. When the user finishes working with the file, it is returned over the network to the server, which stores the now-altered file for retrieval at a later time.

3. Apache Hadoop - It is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It can be scaled up from a single server to thousands of machines and with a very high degree of fault tolerance. Instead of relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.

4. Data Intensive Computing – It is a class of parallel computing application which uses a data parallel approach to process big data. This works based on the principle of collocation of data and programs or algorithms used to perform computation. Parallel and distributed system of interconnected stand alone computers that work together as a single integrated computing resource is used to process / analyze big data.

II. MAPREDUCE

MapReduce is a framework for efficiently processing the analysis of big data on a large number of servers. It was developed for the back end of Google's search engine to enable a large number of commodity servers to efficiently process the analysis of huge numbers of webpages collected from all over the world. Apache [13] [14] developed a project to implement MapReduce, which was published as open source software (OSS), this enabled many organizations, such as businesses and universities, to tackle big data analysis. It was originally developed by Google and built on well-known principles in parallel and distributed processing. Since then Map Reduce was extensively adopted for analyzing large data sets in its open source flavor Hadoop.

2.1. What is MapReduce?

MapReduce is generally defined as a programming model for processing and generating large sets of data **[**15**]**. The user specifies a map function that takes a so called key/value pair as an input to generate a set of intermediate key/value-pairs (see Figure 4). Then, the reducer function merges all of the intermediate values associated with the same intermediate key **[**15**]**.

In detail, input data can be given by a list of records or words that are represented by (key, value) pairs, as shown in Figure 4. There are two phases in MapReduce, the "Map" phase and the "Reduce" phase. First, the data is distributed and processed independently from other data items between different nodes called, "mappers." These mappers are represented by the rectangular figures under the "Map" column. Next, the mapper outputs intermediate (key, value) pairs which then enter the "Reduce" phase represent by the "Reduce" column. The data having the same key (word) are bundled together and processed into the same node which is done by the "reducer." The reducer then combines different values with the same key into nodes and sums up the intermediate results from the mapper. The results are then put into the distributed file system.

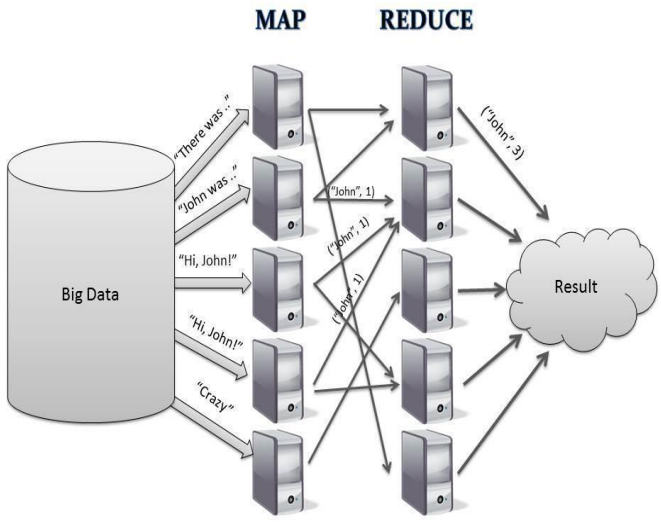
The programs written using this model are automatically parallelized, since the mapper- and reducer-functions can be executed in parallel among a cluster of machines **[**2**]**. It functions primarily on the principle that large problems can be divided into smaller ones.

Figure 4, MapReduce Overview Diagram [16]

III. HADOOP

Apache Hadoop is a popular open-source framework used to process large sets of data that are sent across a cluster of computers using the MapReduce programming model [17**]**. The framework of Hadoop is designed to work on thousands of machines and provides high availability. Depending on the number of machines, the more machines there are in a cluster, it is more likely a machine node in the cluster would fail. Thus, the library itself is designed to monitor and handle failures at the application layer. Delivering a highly available service on top of a cluster of computers may be more prone to failures **[**17**]**. The architecture of Apache Hadoop consists of two core components, which are needed to implement a MapReduce use case.

3.1 Hadoop Architecture



At its core, Hadoop has two major layers namely:

(a) Processing/Computation layer (MapReduce), and

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

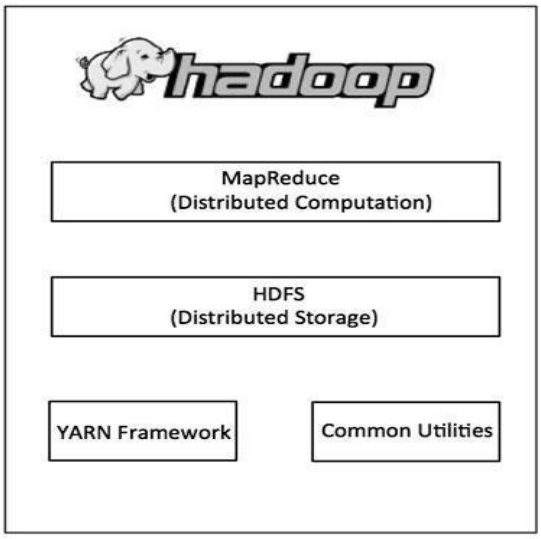


Figure 5: Apache Hadoop Architecture

(a)MapReduce

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

(b) Hadoop Distributed File System

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.

Apart from the above mentioned two core components, Hadoop framework also includes the following two modules:

- **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules.
- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.

3.2 .How Does Hadoop Work?

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.

- 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.
- 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.

IV.CONCLUSION

In this paper an overview of big data has been given. It also states some of the advantages of big data in our society, some technologies being used. As we have entered an era of Big Data, processing large volumes of data has never been greater. Through better Big Data analysis tools like Map Reduce over Hadoop and HDFS, guarantees faster advances in many scientific disciplines and improving the profitability and success of many enterprises.

REFERENCES

[1]. Hadoop,"PoweredbyHadoop,"http://wiki.apache.org/hadoop/PoweredBy.

[2]. Hadoop Tutorial,YahooInc., https://developer.yahoo.com/hadoop/tutorial/index.html

[3]. Marcus R. Wigan, Roger Clarke, "Big Data's BigUnintendedConsequences" Published by the IEEE Computer Society, pp 46-53

[4]. P. Zikipoulos, T. Deutsch, D. Deroos, Harness the Power of Big Data, 2012, http://www.ibmbigdatahub.com/blog/harness-power-big-databook-excerpt

[5]. MIKE 2.0, Big Data Definition, http://mike2.openmethodology.org/wiki/Big_Data_Definition

[6]. W. H. Inmon, Building the Data Warehouse, John Wiley & Sons, 2005.

[7]. http://hive.apache.org/

[8]. Neo4j Graph Database, http://www.neo4j.org.

[9]. C. C. Aggarwal, C. Zhai, Probabilistic Models for Text Mining: in Mining Text Data, Kluwer Academic Publishers, Netherlands, 2012, pp.257-294.

[10]. C. Nyce, Predictive Analytics white paper, American Institute for CPCU/Insurance Institute of America,

http://www.theinstitutes.org/doc/predictivemodeling whitepaper.pdf, 2007.
[11]. In-Memory Analytics, Leveraging Emerging Technologies for Business Intelligence, Gartner Report, 2009.
[12]. http://en.wikipedia.org/wiki/MapReduce
[13]. Apache Hive, http://hive.apache.org/

[14]. Apache Giraph Project, http://giraph.apache.org/
[15]. Dean, Jeffery, and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." Google Research Publication (2004): 1-13. Google. Web. 18 May 2014. http://research.google.com/archive/mapreduce.html.
[16]. Fries, Sergie. "MapReduce: Fast Access to Complex Data." Data Management and Data Exploration Group. 1 Jan. 2014. Web. 15 July 2014. <http://dme.rwth-aachen.de/en/research/projects/mapreduce>.
[17]. "Apache Hadoop." Hadoop. 30 June 2014. Web. 8 June 2014. <http://hadoop.apache.org/>.

AUTHORS PROFILE

V.Shanmugapriya received her M.Phil Degree from Periyar University, Salem in the year 2007. She has received her M.C.A Degree from Madurai Kamaraj University, Madurai in the year 2002. She is working as Assistant Professor, Department of Computer Science, PGP College of Arts & Science, Namakkal, Tamilnadu, India. She is pursuing her Ph.D Degree at Periyar University. Salem, Tamilnadu, India. Her areas of interest include Big Data and Data Mining.

Dr.D.Maruthanayagam received his Ph.D Degree from Manonmanium Sundaranar University, Tirunelveli in the year 2014. He has received his M.Phil, Degree from Bharathidasan University, Trichy in the year 2005. He has received his M.C.A Degree from Madras University, Chennai in the year 2000. He is working as Assistant Professor, Department of Computer Science, Sri Vijay Vidyalaya College of Arts & Science, Dharmapuri, Tamilnadu, India. He has above 14 years of experience in academic field. He has published 1 book, 15 International Journal papers and 21 papers in National and International Conferences. His areas of interest include Grid Computing, Cloud Computing and Mobile Computing.