International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Big Data and Hadoop Framework: A Survey

[1]Prof. Prabhdeep Kaur, [2]Prof. Sapna Shukla
[1,2]University Institute of Computing,
Chandigarh University, Gharuan, Punjab
[1]uic.prabhdeeplobana@gmail.com, [2]uic.sapnashukla@gmail.com

Abstract: *Today's era is the era of big data. This paper documents an attempt that gives a consolidated description of big data while indulging its other unique and defining characteristics by considering definitions from practitioners and academics. In this paper, brief introduction of big data and an overview of Hadoop, which is the core platform of big data and used for processing the data, which uses a map reduce paradigm to process the data, is given. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. Big data environment is used to acquire, organize and analyze the various types of data. There is an observation about Map Reduce framework that framework generates large amount of intermediate data. Therefore, as well as the tasks finishes there is need of throwing that abundant data, because MapReduce is unable to utilize them.*

Keywords: *Big data, Map Reduce, Hadoop, HDFS, Hadoop components, Hive.*

1. Introduction

As Data volumes of social networks, Research, industry is accelerated at rapid speed, need of broader concept is also required and the solution gaining popularity in this area now-a days is "BIG DATA". The term BIGDATA in itself is a wider concept which reflects the image of something in "HUGE". Big data is becoming first priority for new researchers. Big data analytics is treated as a master concept describes the various techniques and technologies to study huge amount of data efficiently. Big data refers to the types of data whose range's mushrooming from petabytes to zettabyttes. Big data is required for the process of analyzing the data in unstructured format. The various big data dimensions are volume, velocity, veracity and variety.

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 [1]. Big data is actually an evolving term that describes any voluminous amount of structured, semi structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, so this term is often used when speaking about petabytes and exabytes of data[2]. Big data is an all-encompassing term for large collection of the data sets so this huge and complex that it becomes difficult to operate them using traditional data processing applications. When dealing with larger datasets, organizations face difficulties in being able to create, manipulate, and manage big data. Big data is particularly a problem in business analytics because standard tools and procedures are not designed to search and analyze massive datasets. An example of big data might be petabytes (1,024 terabytes) or exabytes (1,024 petabytes) of data consisting of billions to trillions of records of millions of people—all from different sources (e.g. Web, sales, customer contact center, social media, mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible[2]. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on"[4].Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks. The world's technological per-capita capacity to store

information has roughly doubled every 40 months since the 1980s;as of 2012, every day 2.5 exabytes (2.5×1018) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization. Big data defined *a*s far back as 2001, industry analyst
Doug Laney (currently with Gartner) articulated the now mainstream definition of big data as the three Vs of big data: volume, velocity and variety. Big data can be characterized by welknown 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily.[2]

2. LITERATURE REVIEW

Xin Luna Dong [12] in 2013 explained challenges of big data integration (schema mapping, record linkage and data fusion). These challenges are explained by using examples and techniques for data integration in addressing the new challenges raised by big data, includes volume and number of sources, velocity, variety and veracity. The advantage of this paper is identifying the data source problems to integrate existing data and systems. The disadvantage of this paper is big data integration such as integrating data from markets, integrating crowd sourcing data, providing an exploration tool for data sources. caching (Dache) framework that made minimum change to the original map reduce programming model to increment processing for big data applications using the map reduce model. It is a protocol, data aware cache description scheme and architecture. The advantage of this paper is, it improves the completion time of map reduce jobs.
Jian Tan [5] 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.
Sagiroglu, S. [6] offered the big data content, its scope, functionality, data samples, advantages and disadvantages along with challenges of big data. The critical issue in relation to the Big data is the privacy and protection. Big data samples describe the review about the environment, science and research in biological area. By this paper, we can conclude that any association in any domain having big data can take the benefit from its careful investigation for the problem solving principle. Using Knowledge Discovery from the Big data convenient to get the information from the complicated data records. The overall appraisal describe that the data is mounting day by day and becoming complex. The challenge is not only to gather and handle the data but also how to extract the useful information from that collected data records. In accordance to the Intel IT Center, there are several challenges related to Big Data which are rapid data growth, data infrastructure, and variety of data, visualization and data velocity.
Garlasu, D. [7] discussed the enhancement about the storage capabilities, the processing power along with handling technique. The Hadoop technology is widely used for the simulation purpose. Grid Computing provides the notion of distributed computing using HDFS. The benefit of Grid computing is the maximum storage capability and the high processing power. Grid Computing makes the big assistance among the scientific research and help the researcher to analyze and store the large and complex data in various formats.
Mukherjee, A. [8] The Big data analysis define the large amount of data to retrieve the useful information and uncover the hidden information. Big data analytics refers to the Map Reduce Framework which is discovered by the Google. Apache Hadoop is the open source platform which is used for the purpose of simulation of Map Reduce Model. In this the performance of SF-CFS is compared with the HDFS with the help of the SWIM by the facebook job traces. SWIM contains the workloads of thousands of jobs with complex and massive data arrival and computation patterns.
Aditya B. [9] 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.

3. BIG DATA AND HADOOP

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes. 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. In which application is broken into smaller parts (fragments or blocks).Apache Hadoop consists of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive. Hadoop Distributed File System (HDFS)

consists of three Components: the Name Node, Secondary Name Node and Data Node [10].

A. Components of Hadoop [3]:

 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.

 Pig: Pig is high-level platform where the Map Reduce 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.

B. 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 largescale 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 slave node processes the base problem and passes the result back to the master Node. The Map Reduce system arrange together all intermediate pairs based on the intermediate keys and refer them to reduce() function for producing the final output. Reduce function works as the master node collects the results from all the sub problems and combines them together to form the output.

Map(in_key,in_value)---
>list(out_key,intermediate_value)

Reduce(out_key,list(intermediate_value))---
>list(out_value)

The parameters of map () and reduce () function is as follows:

map (k1,v1) ! list (k2,v2) and reduce (k2,list(v2)) ! list (v2)

A Map Reduce framework is based on a master slave architecture where one master node handles a number of slave nodes. Map Reduce works by first dividing the input data set into even-sized data blocks for equal load distribution. Each data block is then assigned to one slave node and is processed by a map task and result is generated. The slave node interrupts the master node when it is idle. The scheduler then assigns new tasks to the slave node. The scheduler takes data locality and resources into consideration when it disseminates data blocks. Map Reduce always manages to allocate a local data block to a slave node. If the effort fails, the scheduler will assign a rack-local or random data block to the slave node instead of local data block. When map() function complete its task, the runtime system gather all intermediate pairs and launches a set of condense tasks to produce the final output. Large scale data processing is a difficult task, managing hundreds or thousands of processors and managing parallelization and distributed environments makes is more difficult. Map Reduce provides solution to the mentioned issues, as is supports distributed and parallel I/O scheduling, it is fault tolerant and supports scalability and it has inbuilt processes for status and monitoring of heterogeneous and large datasets as in Big Data. It is way of approaching and solving a given problem. Using Map Reduce framework the efficiency and the time to retrieve the data is quite manageable. To address the volume aspect, new techniques have been proposed to enable parallel processing using Map Reduce framework [10]. Data aware caching (Dache) framework that made slight change to the original map reduce programming model and framework to enhance processing for big data applications using the map reduce model [11]. The advantage of map reduce is a large variety of problems are easily expressible as Map reduce computations and cluster of machines handle thousands of nodes and fault-tolerance. The disadvantage of map reduce is Real-time processing, not always very easy to implement, shuffling of data, batch processing.

Map Reduce Components:

1. **Name Node**: manages HDFS metadata, doesn't deal with files directly.

2. **Data Node**: stores blocks of HDFS—default replication level for each block: 3.

3. Job Tracker: schedules, allocates and monitors job execution on slaves—Task Trackers.
4. Task Tracker: run Map Reduce operations.

4. CONCLUSION AND FUTURE WORK

The objective of this paper is a survey and overview provided on big data, its characteristics, challenges and issues as well as the opportunities in big data. This paper initiates a review to examining the traditional view of data analytics and big data analytics. The authors identified various issues related to big data which includes its storage, management and processing. The authors also reviewed the hadoop frameworks and its components, HDFS and map Reduce. The future research work will focus on identifying the various effective data layout s to increase the Hadoop Map Reduce jobs and to acquire the more understanding of the issues and challenges which are related with big data.

REFERENCES

[1] 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.
[2] Apache HBase. Available at http://hbase.apache.org
[3] Apache Pig. Available at http://pig.apache.org
[4] Parmeshwari P. Sabnis, Chaitali A.Laulkar , "*SURVEY OF MAPREDUCE OPTIMIZATION METHODS*", ISSN (Print): 2319- 2526, Volume -3, Issue -1, 2014
[5] Jian Tan; Shicong Meng; Xiaoqiao Meng; Li ZhangINFOCOM, "Improving ReduceTask data locality for sequential MapReduce" 2013 Proceedings IEEE ,1627 - 1635
[6] Sagiroglu, S.; Sinanc, D., "Big Data: A Review", 2013,20-24
[7] Garlasu, D.; Sandulescu, V.; Halcu, I.; Neculoiu, G,"A Big Data implementation based on Grid Computing", Grid Computing, 2013, 17-19
[8] Mukherjee, A.; Datta, J.; Jorapur, R.; Singhvi, R.; Haloi, S.; Akram, "Shared disk big data analytics with Apache Hadoop", 2012,18-22.
[9]Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", 2012, 6-8
[10] Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1,2010, 72-77.
[11] Wang, J.; Xiao, Q.; Yin, J.; Shang, P. Magnetics, "DRAW: A New Data-gRouping-AWare Data Placement Scheme for Data Intensive Applications With Interest Locality"IEEE Transactions (Vol: 49), 2013, 2514 – 2520.
[12] Dong, X.L.; Srivastava, D. Data Engineering (ICDE)," Big data integration" IEEE International Conference on , 29(2013) 1245–1248.