International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Review on Need of MapReduce in Big Data Application

Sushant Shirish Nagavkar[1], Ashishkumar[2]

[1]M.E. Computer Network Student from G.H. Raisoni Collage of Engineering and management, Ahemadnagar, India

[2]Assistant Professor of G.H. Raisoni Collage of Engineering and management, Ahemadnagar, India

Abstract: ***Big-data is broadly distributed data processing applications that operate on large amount of data. The growing power of Big Data assume the significance of analyzing huge amount of data with a frequent and quick rate of growth and change in databases and data warehouses. An inspection of the MapReduce framework is that the framework creates a large amount of transitional data. This survey aims that the modern techniques for parallel query processing using MapReduce. The main fact of data analytics is scalability, due to the huge volume of data that need to be extracted, processed, and analyzed in a sensible fashion. MapReduce is most popular framework for large-scale data analytics mainly due to its salient features like scalability, fault-tolerance, simple programming, and flexibility.***

Keywords: BEA, Big-data, Hadoop, HDFS, MapReduce.

1. Introduction

Big data belongs to datasets whose size is outside the capacity of usual database software tools to capture, store, manage, and analyze [2]. “Big Data", described by the unusual volume of data, data generation velocity, and structural variety of data, support for extensive data analytics form a mainly challenging task [5]. The main aim of big data is to help companies make better business decisions by facilitating data scientists and users to analyze huge volumes of transaction data and other data sources [6]. With the help of predictive analytics and knowledge mining big data can be easily processed but due to unstructured data it may not fit in traditional data warehouse [6]. But traditional data warehouses are unable to handle the processing demands of big data. Apache is founded by Hadoop and it is a software framework for processing large datasets. It uses Hadoop Distributed File System (HDFS) for storage purpose and MapReduce for processing components of Hadoop [2].

For the large-scale processing and analysis of vast data sets MapReduce is the most popular framework. MapReduce programming is useful for processing large datasets. MapReduce uses 2 functions: Map and Reduce function. User can write the Map function which takes input and produces a set of key/value pairs. This all produced values with the same intermediate key I is grouped by the MapReduce library and then passes it to the Reduce function. The Reduce function is also written by the user which accepts and a set of intermediate key I and values for that key. It merges these values to obtain probably smaller set of values [2].

The main function of MapReduce framework executes on a single master machine where input data is preprocessed before map functions are called and/or post process the output of reduce functions. As per need of applications, a pair of map and reduce functions may be executed once or more time. The research area has newly received a lot of attentions for developing MapReduce algorithms for examine big data [7].

2. Big Data

Big data refers to datasets whose size is away from the capacity of typical database software tools to capture, store, manage, and analyze. The possible sources of big data are:

Traditional project data contains customer information from CRM systems, Transactional ERP data, Web store transactions, and common ledger data. Data like Call Detail Records (CDR), weblogs, smart meters, manufacturing sensors, equipment logs, and trading systems data are machine generated data. A customer feedback stream, micro-blogging sites like Twitter and social media platforms like facebook are comes under social data. There are some other sources of data like Health care, Public sector, Retail and Manufacturing [2]. As per report of SINTF, world's 90% data was generated over the past two years. From the last few years due to new and advanced technologies has developed which was responsible for improving the data consumers enthusiasm to create, store and consume data [3].

2.1 Characteristics of Big data

The collection of large and complex data sets that are difficult to process using on-hand database management tools or traditional data processing applications is big data. Big data have some properties which include volume, variety and volume.

- **Volume:** The volume of data at this point is very vast and is produced from a lot of different devices. The data size is generally in terabytes and petabytes. For privacy protection this data needs to be encrypted.
- **Velocity**: This explains the real time characteristic establish in several data sets for example streaming data. The result that fails to spot the precise time is usually of little value.

- **Variety:** Big data consists of a mixture of different types of data i.e. structured, unstructured and semi structured data. The data may be in many different forms like: blogs, videos, pictures, audio files, location information etc[3].

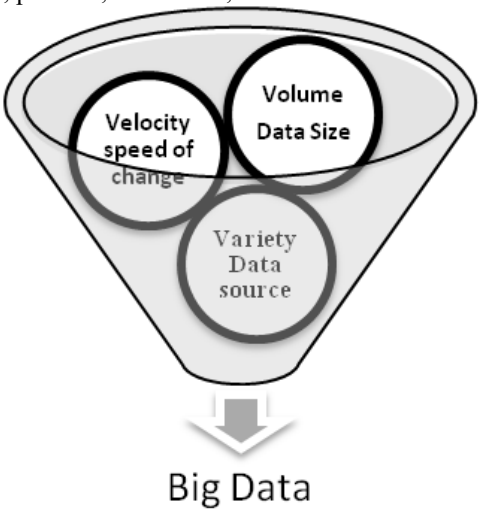

Figure 1: Big data formation characteristics

2.2 Traditional DBMS Vs Big Data

Big Data's scalable technologies need to process well huge amount of data within acceptable times. Big Data scalable technologies contain MPP databases, the Apache Hadoop Framework, the Internet, and archival storage systems [2]. MapReduce is matching to DBMS, not a opposing technology.

- Parallel DBMS are capable for querying of large data sets.
- MR-style systems are for composite analytics and ETL tasks.
- Parallel DBMS need data to fit into the relational model of rows and columns.
- In distinguish, the MR model doesn't need that data files remain to a schema defined using the relational data model. That is, the MR programmer is free to charge structure of their data in every way or even to have no structure at all [2].

3. What is Hadoop?

Hadoop has been efficiently used by many corporations including AOL, Amazon, Facebook, Yahoo and New York Times to run their applications on clusters. For example, AOL uses it for analyzes the behavioral samples of their users to offer requisite services. MapReduce parallel processing framework is from Apache Hadoop which is an open source implementation of the Google's. Hadoop isolates the fine points of parallel processing, in which it includes data distribution to processing nodes, restarting failed subtasks and consolidation of results after computation. Using this framework developer can write parallel processing programs that spotlight on their division problem, rather than parallelization issues. Hadoop consist of 1) Hadoop Distributed File System (HDFS) and 2) Hadoop MapReduce: a software framework [4].

3.1 Characteristics of Hadoop

- **Scalable–** As per demand of additional nodes those can be added without making any change in data formats, the way of data loading and in the way the jobs or application are written.
- **Cost effective–** Hadoop introduce the especially parallel computing to commodity servers. The output is a considerable reduce in cost which in turn makes it inexpensive to model all the data.
- **Flexible–** Hadoop is schema less so it can take any type of data like structured or not and it can accept data from number of different sources. This data can be joined and aggregated in random ways allowing detailed analysis than any other system can provide.
- **Fault tolerant–** If any node is lost, the system is able to redirects work to another node of the data and continues processing without losing a beat.

3.2 Hadoop Architecture

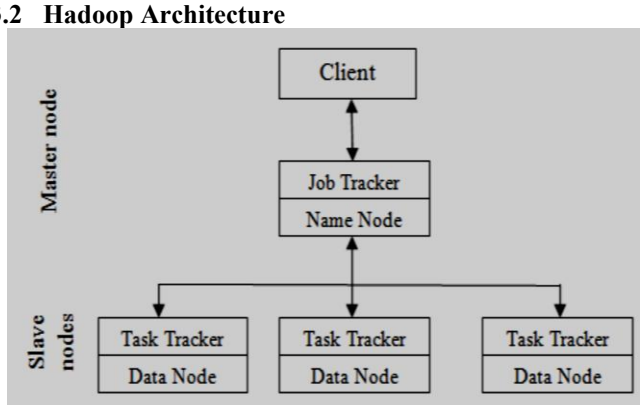

Figure 2: Hadoop Architecture

One of the primary components of Hadoop is HDFS. This clusters and designed like Master-slave architecture. The Master (NameNode) is responsible for managing the file system namespace and files operations like opening, closing, renaming and directories and also verifies the mapping of blocks to DataNodes along with flexible access to files by clients. Slaves (DataNodes) are responsible for performing read and write operation from the clients beside with block creation, deletion, and replication upon instruction from the Master (NameNode) [2].

Key benefits of Hadoop
1) Capable to large files.
2) Capable to be parallelism.
3) Flexible implementation platform.
4) HDFS also run already existing file system.

4. Mapreduce

For processing and generating large dataset MapReduce programming model is used. Users can write a map function that processes a key/value pair to produce a set of intermediate key/value pairs, and a reduce function is capable of merging this all intermediate values associated with the same intermediate key.
Programs written in this functional manner are automatically parallelized and executed on a large group of service

machines. At the execution-time system partition the input data, schedules the program's execution across a set of machines, handles machine failures and manage the necessary inter-machine communication. This permit programmers not including any knowledge with parallel and distributed systems to easily develop the resources of a large distributed system. Simplicity is the feature of MapReduce Model. Due to Map () and Reduce () can be written by users. Input data are collection of group in a distributed file system. Programs are added into a distributed-processing framework [2].

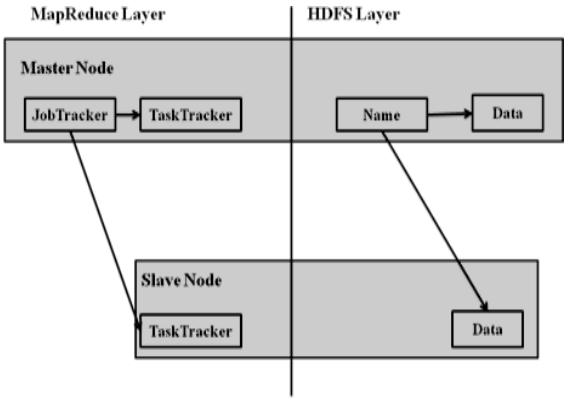

Figure 3: Hadoop simple cluster graphic

MapReduce is popular due to its simple programming interface and excellent performance while applying a large range of applications. Such applications receive a huge amount of input data and also called as "Big-data applications". As shown in Fig. 4.2, input data is first divide and then supply to workers in the map phase. Individual data substances are called records. Each worker got this splinted input from MapReduce system to produces records. Intermediate results generated in the map phase are shuffled and sorted by the MapReduce system and are then give to the workers in the reduce phase. Final results are produced by number of reducers and written to the disk.

Conceptually the map and reduce functions gave by the user have connected Types:

Map (k1, v1) → list (K2, v2)

Reduce (K2, list (v2)) → list (v2)

I.e., the input keys and values are drawn from a different domain than the output keys and values. Furthermore, the intermediate keys and values are from the same domain as the output keys and values [9].

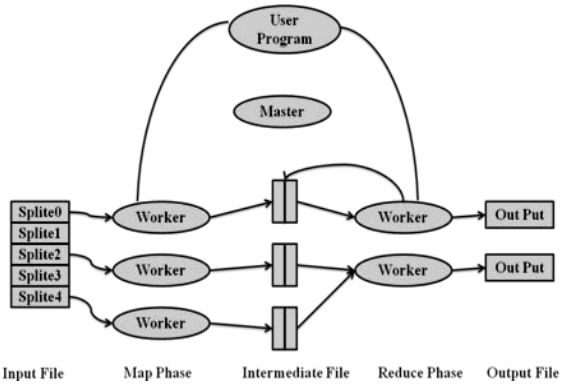

Figure 4: MapReduce programming model [1].

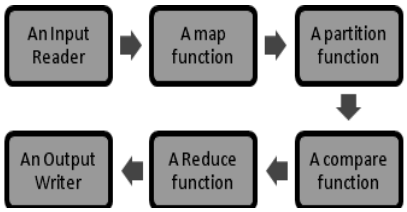

Figure 5: MapReduce data flow

5. Conclusion

This paper analyzes the concept of big data and how it differs from traditional database. It also clearly identifies the Hadoop environment, its architecture and how it can be implemented using MapReduce functions. So, it is sure that this paper helps the researches to understand the basic concepts of big data, Hadoop and MapReduce to move further. But this system has some errors like an observation of the MapReduce framework is that the framework generates a large amount of intermediate data. Such abundant information is thrown away after the tasks finish, because MapReduce is unable to utilize them. Once the results are produced then it will do same tasks for similar query for iterative execution of same query this produces the time consuming task and processing work more.

In our future work, researcher must able to utilize this produced results data for further processing of similar query and must be reduce the processing time and processing cost of query to get better performance form system and system also able to use the produced results.

References

[1] Yaxiong Zhao*, Jie Wu, and Cong Liu, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework", February 2014

[2] D.Usha [Á*] and Aslin Jenil A.P.S[Á],"A Survey of Big Data Processing in Perspective of Hadoop and Mapreduce", [Á]Hindustan University, Accepted 05 March 2014.

[3] Kudakwashe Zvarevashe1, Mainford Mutandavari2, Trust Gotora3 ,"A Survey of the Security Use Cases in Big Data", International Journal of Innovative Research in Computer and Communication Engineering(An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 5, May 2014

[4] B.Thirumala Rao Associate Professor Dept. of CSE Lakireddy Bali Reddy College of Engineering Dr. L.S.S.Reddy Professor & Director Dept. of CSE Lakireddy Bali Reddy College of Engineering, "Survey on Improved Scheduling in Hadoop MapReduce in Cloud Environments", International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011
[5] Christos Doulkeridis. Kjetil Norvag, "A Survey of Large-Scale Analytical Query Processing in MapReduce."
[6] N. Monica1, K. Ramesh Kumar2. "Survey on Big Data by Coordinating Mapreduce to Integrate Variety of Data" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Volume 2 Issue 11, November 2013 www.ijsr.net
[7] Kyuseok Shim Seoul National University shim@ee.snu.ac.kr "MapReduce Algorithms for Big Data Analysis."
[8] Jun Wang, Qiangju Xiao, Jiangling Yin, and Pengju Shang Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32826 USA "DRAW: A New Data-gRouping-AWare Data Placement Scheme for Data Intensive Applications With Interest Locality."
[9] Jeffrey Dean and Sanjay Ghemawat jeff@google.com, sanjay@google.com Google, Inc. "MapReduce: Simplified Data Processing on Large Clusters."
[10] Tyson Condie, Neil Conway, Peter Alvaro, Joseph M. Hellerstein UC Berkeley Khaled Elmeleegy, Russell Sears Yahoo! Research "MapReduce Online."

Author Profile

Sushant Nagavkar received the B.E. degree in Information Technology with first class from DKTE college of Engineering and textile Ichalkaranji under Shivaji University Kolhapur in 2013. Now I am with GHRCEM college of Engineering and Management, Ahemadnagar, Maharashtra appearing M.E. degree in Computer Networks.