Volume 8, No. 2, March 2017 (Special Issue)
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
REVIEW ARTICLE
Available Online at www.ijarcs.info

An Approach of MapReduce Programming Model For Cloud Computing

Bisma Bashir
Department of CSE, SEST
Jamia Hamdard, New Delhi, India
Bismabashir3232@gmail.com

Abstract: Cloud computing is emerging as a new computational area.It delivers computing resources to the users. Cloud computing having many drawbacks such as low scalability,not having support for stream data processing etc, need to be overcome. MapReduce needs to be implemented to overcome the drawbacks of cloud computing. MapReduce having the ability to perform parallel computation and also perform distribution of computation based on several nodes is the good option. This paper provides a comprehensive review and analysis of MapReduce programming model in cloud computing.

Keywords: Cloud Computing; Big Data; MapReduce; Data Distribution; Cloud MapReduce.

I. INTRODUCTION

In today's era large amount of data is being stored on the cloud platform. Cloud computing is an internet-based computing, which provides sheared computer processing resources, data to users, storage devices on demand, etc. being internet-based it enables ubiquitous, on-demand access to a shared pool of configurable computing resources[1]. Cloud computing provides its users to store and process their data on the data centers that may be located far from the users [17]. Companies have to process huge amount of data in a cost-efficient manner. Cloud storage provides online virtualized pool of storage which is generally hosted by third parties. Companies, who require their data to be hosted by the cloud providers, buy or lease storage capacity from the cloud storage providers and use it for their storage needs.

Many organizations face difficulties when dealing with a large amount of data. Data storage capacity and processor computing power constrains are the main challenges behind such application. Large amount of distributed data needs to be processed quickly with good response time and replication at minimum cost. Parallel and distributed computing in a cloud computing environment is the best way for huge data processing. By using MapReduce framework, as a distributed computing paradigm, cloud computing aims at large datasets to be processed on available computer nodes [3].

MapReduce is a programming model, which is composed of a map procedure that performs filtering and sorting and a reduce procedure that performs a summary operation [7]. MapReduce processes and generates large data sets with a parallel, distributed algorithm on a cluster. MapReduce system manages all communication and data transfer between the various parts of the system, and providing for redundancy and fault tolerance. MapReduce optimizes the execution engine and thus the data is processed and accessed easily. When MapReduce is implemented on various nodes, it becomes fast and easy to access data on those nodes. With the help of MapReduce large-scale computations are simplified [2].

The paper is structured as follows: section II gives the literature review of the cloud computing with MapReduce. Section III presents the basic fundamental concept of MapReduce. Section IV gives the concept of MapReduce in cloud computing and section V concludes this paper.

I. LITRATURE REVIEW

Rajkumar et al [2] in there paper proposed Cloud Map Reduce (CMR) to overcome all the drawback of MapReduce framework such as low scalability, does not support edible pricing, stream data processing etc. the implementation results of CMR are shown to be more efficient and development faster than the other implementation of the MR method. **Daneshyar et al [3]**in their paper provides a complete orderly review and analysis of large-scale dataset processing and dataset handling challenges and requirements in a cloud computing environment by using a MapReduce framework and its open-source implementation hadoop. Requirements for MapReduce system to perform large-scale data processing are defined in this paper. There paper also proposed MapReduce framework and one implementation of this framework on Amazon Web services. **Pradeepa et al [4]** in there paper proposed an algorithm corresponding to the MapReduce based on rough theory, which can deal with the massive data. The proposed algorithm can effectively process bigdata. **Gaizhen Yang [5]** proposed in there paper a program based on MapReduce framework that enables distributed programming. Analysis of hadoop architecture and MapReduce working principle are done. It also describes how to perform a MapReduce job in the cloud platform. **Sunny Ranjan [6]** studied the implementation of data mining algorithms based on MapReduce, in order to derive a complete concept about developing such algorithms. **Santhosh voruganti [8]** this paper we implement MapReduce programming model using two components: a Job Tracker (master node) and many Task Trackers (slave nodes). **W.Dai et al [12]** proposed an implementation of decision tree algorithm using MapReduce programming model. Traditional algorithm is transformed into a series of Map and Reduce procedures. To minimize the communication cost some data structures have been designed. Ontology based models are discussed in [16].

II. FUNDAMENTAL CONCEPT OF MAPREDUCE

Google facilitated MapReduce as a programming framework in order to analyze huge amount of data. With the help of MapReduce developers can perform complex computations in a simple way and can hide the details of data distribution, parallelization and fault tolerance. In MapReduce programming model every map reduce are independent of other ongoing maps and reduces and the operation runs in parallel on different key and values. The most important aspect of MapReduce processing model is that the map tasks are carried on the nodes where the data lives. This ensures that there is very little or no movement of data between nodes [3].

Fundamental phases of MapReduce programming model are [3]:

- Map phase: the input data is divided into M Map functions called as Mapper. Mappers run in parallel and the output of MapReduce is intermediate key and value pairs.
- Shuffle and Sort phase: output from the mappers is partitioned by hashing the output key. Here the number of partitions is equal to the number of reducers. In the shuffle phase, all key and value pairs share the same key that belong to the same division. Each division is stored by a key to merge all values of that key, after partitioning the MapReduce.
- Reduce phase: the output from the second phase are portioned into R Reduce function called as Reducer. Reducers process different in-between keys and also run in parallel.

The above three phases are drawn below [9]:

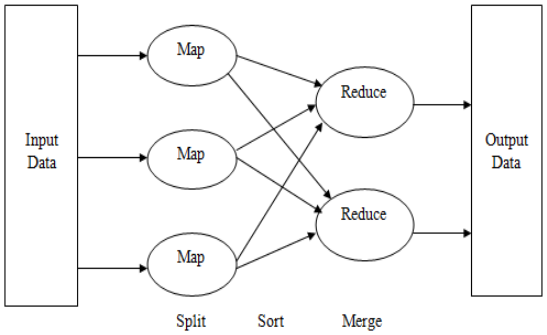

Figure 1: Phases of MapReduce

III. MAPREDUCE IN CLOUD COMPUTING

Cloud computing enable us to store, collect, share and transfer large amounts of data at very high speeds in a flawless and transparent manner that would out of complete necessity order all data to be “totally virtual”. Thus, all data in cloud computing captures the concept of data virtualization through a new programming model which treats all data as a single entity through a process called MapReduce. MapReduce is widely used for big data processing in cloud platforms. Hadoop an open-source implementation of MapReduce hides the complexity of parallel execution across hundreds of servers in a cloud environment. It allows developers to process terabytes of data. How parallel programming work away from the developer is the main reason of using MapReduce with cloud computing [3].

Amazon cloud is using cloud MapReduce as a MapReduce implementation. Cloud MapReduce has three main advantages over other MapReduce implementations built on traditional OS [10].

- Cloud MapReduce implementation is faster than other implementations.
- It has high scalability and failure resistance.
- It has simple line of code.

MapReduce in cloud computing is shown below [11]:

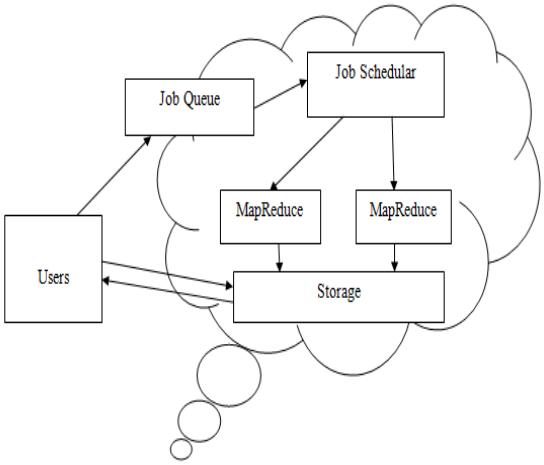

Figure 2: Cloud Computing With MapReduce

IV. CONCLUSION

MapReduce is an easy, effective and flexible tool for parallel data computation. On cloud platform, it has been proven to be a useful tool for distributing the processing over as many processors as possible. After conducting a comprehensive review, this paper concludes that in the number of computing nodes, cloud MapReduce has high scalability and MapReduce simplifies the large-scale data computation. Terabytes of data are processed parallel on various nodes.

In future, various optimization algorithms such as genetic algorithm, ant colony optimization algorithm, etc can be used to optimize the MapReduce procedure. By optimizing the MapReduce procedure, both data access and implementation becomes easy.

V. ACKNOWLEDGMENT

I would like to express sincere thanks to my supervisor Dr Farheen Siddiqui for her valuable guidance in this paper.

VI. REFERENCES

[1] "Cloud computing", En.wikipedia.org, 2017. [Online] Available: https://en.wikipedia.org/wiki/cloud_computing. [Accessed: 25- Feb- 2017].

[2] M.NRajkumar, S. Balachandar, V.Venkatesakumar, T.Mahadevan,“Survey On MapReduce In Cloud Computing”, *International Journal of Computer Science and Mobile Applications*, Vol.2, 12 December- 2014.

[3] S. Daneshyar, M.Razmjoo, "Large-scale Data Processing Using MapReduce In Cloud Computing Environment", *International Journal on Web Service Computing (IJWSC)*, Vol.3, No.4, December 2012.

[4] A. Pradeepa, Dr.A.S.Thanamani,"Hadoop File System and Fundamental Concept Of MapReduce Interior and Closure Rough Set Approximations", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, No. 10, October 2013.

[5] G. Yang, "The Application of MapReduce in the Cloud Computing", *International Symposium on Intelligence Information Processing and Trusted Computing*, 2011.

[6] S. Ranjan, "Large-scale Data Mining Analytics Based on MapReduce" *Institute of Parallel and Distributed Systems*, 2014.

[7] "MapReduce", En.wikipedia.org, 2017. [Online]. Available: http://en.wikipedia.org/wiki/MapReduce.

[8] S.voruganti,"Map Reduce a Programming Model for Cloud Computing Based On Hadoop Ecosystem", *(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 5 , 2014.

[9] 2017. [Online]. Available: https://www.google.co.in/serch?q=fundamental+phases+of+MapReduce.

[10] "Google Code Archive - Long-Term Storage For Google Code Project Hosting.". *Code.google.com*. N.p., 2017.

[11] "Cloud Computing Patterns - Dzone Cloud". *dzone.com*. N.p., 2017.

[12] Wei Dai and Wei Ji. "A MapReduce Implementation of c4.5 Decision Tree Algorithm". *International journal of database theory and application*, Vol. 7, 2014, 49-60.

[13] T.Bazaz, A.Khalique, "A Review On Single Sign On Enabling Technologies and Protocols". *International journal Of Computer Applications*, Vol. 151, No. 11, 2016.

[14] H.Fayaz, A.Khalique, "A Review On Sociological Impacts Of Social Networking". International journal Of Engenering Applied Sciences and Technology, Vol. 1, pp. 6-12, 2016.

[15] B. Bashir, A.Khalique, "A Review on Security versus Ethics". *International journal Of Computer Applications*, Vol. 151, No. 11, October 2016.

[16] Farheen Siddiqui "State Of Art Ontological Infrastructure For Cloud Computing" International Journal of Computer & Organization Trends (IJCOT) – Volume 36 Number1 – October 2016 ISSN: 2249-2593 Pg22-26

[17] Farheen Siddiqui "Applications for Big Data in of Intelligent Distributed Processing "IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. I (Nov. - Dec. 2016), PP 61-64 DOI: 10.9790/0661-1806016164