IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-I, Issue-3 E-ISSN: 2347-2693

Big Data in Cloud Environment

Poornima Sharma[1*], Varun Garg [2], Prof. Randeep Kaur [3], Prof. Satendra Sonare [4]

[1*]M.Tech. Scholar, SS, RGPV University, Bhopal, India, poornimasharma02@gmail.com
[2]M.Tech. Scholar, SS, RGPV University, Bhopal, India, varun.garg04@gmail.com
[3]Asst. Prof., CSE,GGITS Jabalpur, RGPV University, Bhopal, India
[4]Asst. Prof., CSE,GGITS Jabalpur, RGPV University, Bhopal, India

www.ijcaonline.org

Received: 07 Nov 2013 Revised: 14 Nov 13 Accepted: 26 Nov 13 Published: 30 Nov 2013

Abstract- Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. However, big data entails a huge commitment of hardware and processing resources, making adoption costs of big data technology prohibitive to small and medium sized businesses. Cloud computing offers the promise of big data implementation to small and medium sized businesses. Big Data processing is performed through a programming paradigm known as MapReduce. Typically, implementation of the MapReduce paradigm requires networked attached storage and parallel processing. The computing needs of MapReduce programming are often beyond what small and medium sized business are able to commit. Cloud computing is on-demand network access to computing resources, provided by an outside entity. Common deployment models for cloud computing include platform as a service (PaaS), software as a service (SaaS), infrastructure as a service (IaaS) & hardware as a service (HaaS).

Keywords- Big Data, Cloud computing, *Map/Reduce*

I. Introduction

People have talked about Cloud Computing that is nothing else but the services we have used for several years such as hosting service, web email service, document sharing service, and map API service etc. Internet and Web enable to achieve the concept of the cloud computing that is virtual computing resources and repository. Therefore, many web applications have been implemented and provided to the users with web emails, web documents, web services, mashups, web data sources, and data feeds etc. Once the user subscribes to the services, it becomes possible to use the services anytime anywhere with his/her computer.

Cloud computing is more enhanced with its scalability, reliability, and quality of services than the existing services. It is categorized into Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). SaaS is to use a service via Internet without installing or maintaining the software, for example, web email services. PaaS is to have a computing or storage service without purchasing hardware or software, for example, hosting services. IaaS is to have utility computing service that is similar to SaaS but to purchase only the amount of time to use the service like AWS [6, 7, 10]. It is not easy to tell the difference from PaaS and IaaS. But, IaaS users can take its control from the services but PaaS users do not have the right to take the control of the platform. Thus, IaaS users can develop or install programs in more flexible environment. AWS provides S3, EC2, and Elastic MapReduce services for Map/Reduce computation as IaaS and SaaS in cloud computing.

Before Internet and Web did not exist, we did not have enough data so that it was not easy to analyze people, society, and science etc with the limited volumes of data. Contradicting to the past, after Internet and web, it has been more difficult to analyze data because of its huge volumes, that is, tera- or peta-bytes of data, which is called Big Data. Google faced to the issue when collecting Big Data as the existing file systems were not sufficient to store Big Data efficiently. Besides, the legacy computing power and platforms were not enough to compute Big Data. Thus, Google implemented Google File Systems (GFS) and Map/Reduce parallel computing platform, which Apache Hadoop project is motivated from.

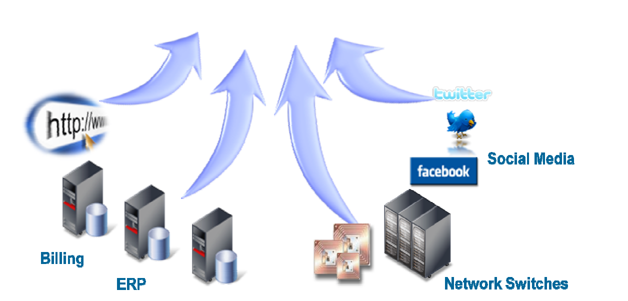
Hadoop is the parallel programming platform built on Hadoop Distributed File Systems (HDFS) for Map/Reduce computation that processes data as (key, value) pairs. Hadoop has been receiving highlights for the enterprise computing because business world always has the big data such as log files for web transactions. Hadoop is useful to process such big data for business intelligence so that it has been used in data mining for past few years. The era of Hadoop means that the legacy algorithms for sequential computing need to be redesigned or converted to

Corresponding Author: *Poornima Sharma*[1*]

Map/Reduce algorithms. Therefore, in this paper, a Market Basket Analysis algorithm in data mining with Map/Reduce is proposed with its experimental result in Elastic Compute Cloud (EC2) ans (Simple Storage Service) S3 of Amazon Web Service (AWS).

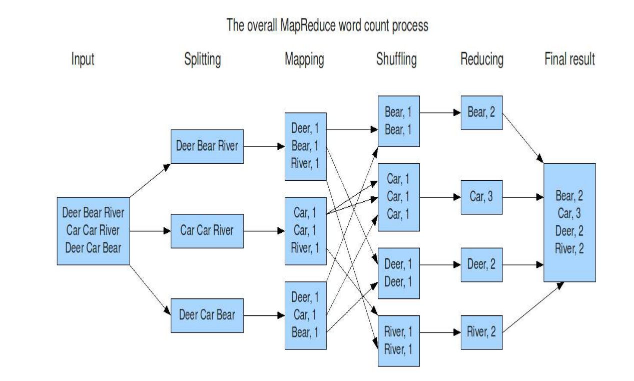
II. Big Data

“Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don’t define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).



III. Map/Reduce in Hadoop

Map/Reduce is an algorithm used in Artificial Intelligence as functional programming. It has been received the highlight since re-introduced by Google to solve the problems to analyze huge volumes of data set in distributed computing environment. It is composed of two functions to specify, “Map” and “Reduce”. They are both defined to process data structured in (key, value) pairs.



Map/Reduce in Parallel Computing [9,11,13]

Map/Reduce programming platform is implemented in the Apache Hadoop project that develops open-source software for reliable, scalable, and distributed computing. Hadoop can compose hundreds of nodes that process and compute peta- or tera-bytes of data working together. Hadoop was inspired by Google's MapReduce and GFS as Google has had needs to process huge data set for information retrieval and analysis [1]. It is used by a global community of contributors such as Yahoo, Facebook, and Twitters. Hadoop’s subprojects include Hadoop Common, HDFS, MapReduce, Avro, Chukwa, HBase, Hive, Mahout, Pig, and ZooKeeper etc. [2].

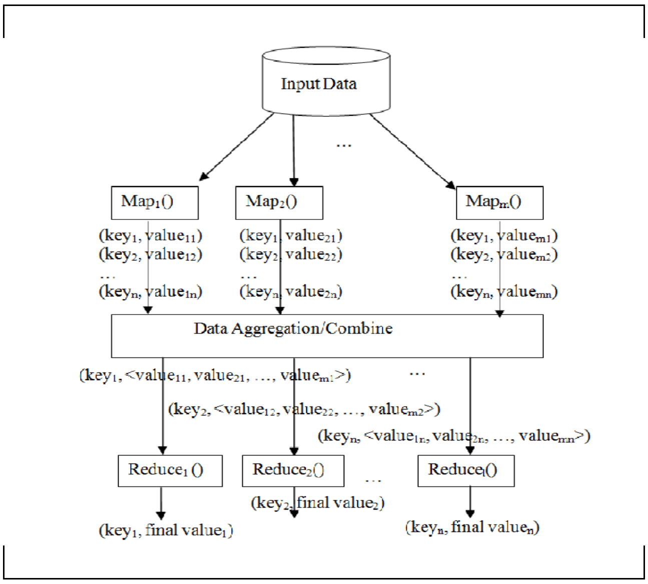


Figure 3.1 Map/Reduce in Parallel Computing

Map/Reduce Flows

The map and reduce functions run on distributed nodes in parallel. Each map operation can be processed independently on each node and all the operations can be performed in parallel. But in practice, it is limited by the data source and/or the number of CPUs near that data. The reduce functions are in the similar situation because they are from all the output of the map operations. However, Map/Reduce can handle significantly huge data sets since data are distributed on HDFS and operations move close to data for better performance [5, 8, 9 ,12].

Hadoop is restricted or partial parallel programming platform because it needs to collect data of (key, value) pairs as input and parallely computes and generates the list of (key, value) as output on map/reduce functions. In map function, the master node parts the input into smaller sub-problems, and distributes those to worker nodes. Those worker nodes process smaller problems, and pass the answers back to their master node. That is, map function takes inputs (k1, v1) and generates <k2, v2> where < >

represents list or set. Between map and reduce, there is a combiner that resides on map node, which takes inputs (k2, <v2>) and generates <k2, v2>. In Figure 3.1, a list of values is collected for a key as (keyn, <value1n, value2n, …, valuemn>) from mappers.

In reduce function, the master node takes the answers to all the sub-problems and combines them in some way to get the output, the answer to the problem [1, 2]. That is, reduce function takes inputs (k2, <v2>) and generates <k3, v3>. Figure 3.1 illustrates

Map/Reduce control flow where each valuemn is simply 1 and gets accumulated for the occurrence of items together in the proposed Market Basket Analysis Algorithm. Thus, for each key, the final value is the total number of values, that is the sum of 1s, as (keyn, final valuen)

The Issues of Map/Reduce

Although there are advantages of Map/Reduce, for some researchers and educators, it is:

1. Need tens-, hundreds-, or thousands-of-nodes to compose Hadoop Map/Reduce platform.
2. If using services of cloud computing, for example, AWS EC2, the overheads mainly come from I/O. That is, it takes long to upload big data to AWS EC2 platform or AWS S3, which is more than computing time.
3. A giant step backward in the programming paradigm for large-scale data intensive applications
4. Not new at all - it represents a specific implementation of well known techniques developed tens of years ago, especially in Artificial Intelligence
5. Data should be converted to the format of (key, value) pair for Map/Reduce, which misses most of the features that are routinely included in current DBMS
6. Incompatible with all of the tools or algorithms that have been built [4].

However, the issues clearly show us not only the problems but also the opportunity where we can implement algorithms with Map/Reduce approach, especially for big data set. It will give us the chance to develop new systems and evolve IT in parallel computing environment. It started a few years ago and many IT departments of companies have been moving to Map/Reduce approach in the states.

IV. Conclusion

Hadoop with Map/Reduce motivates the needs to propose new algorithms for the existing applications that have had algorithms for sequential computation. Besides, it is (key, value) based restricted parallel computing so that the legacy parallel algorithms need to be redesigned with Map/Reduce.

In the paper, the Market Basket Analysis Algorithm on Map/Reduce is presented, which is association based data mining analysis to find the most frequently occurred pair of products in baskets at a store. The data set of the experimental result shows that associated items can be paired in orders 2 and 3 with Map/Reduce approach. Once we have the associated items, it can be used for more studies by statically analyzing them even sequentially, which is beyond this paper.

References

[1]. J. Dean and S. Ghemawa, “MapReduce: Simplified Data Processing on Large Clusters”, Google Labs, OSDI 2004, **(2004)**, pp. 137–150.

[2]. Apache Hadoop Project, http://hadoop.apache.org/.

[3]. B. Stephens, “Building a business on an open source distributed computing”, Oreilly Open Source Convention (OSCON) 2009, **(2009)** July 20-24, San Jose, CA

[4]. W. Kim, “MapReduce Debates and Schema-Free”, Coord, **(2010)** March 3.

[5]. J. Lin and C. Dyer, “Data-Intensive Text Processing with MapReduce”, Tutorial at the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), **(2010)** June, Los Angeles, California

[6]. J. Woo, “Introduction to Cloud Computing”, the 10th KOCSEA 2009 Symposium, UNLV, **(2009)** December 18-19.

[7]. J. Woo, “The Technical Demand of Cloud Computing”, Korean Technical Report of KISTI (Korea Institute of Science and Technical Information), **(2011)** February.

[8]. J. Woo, “Market Basket Analysis Example in Hadoop”, http://dal-cloudcomputing.blogspot.com/2011/03/ market-basket-analysis-example-in.html, **(2011)** March.

[9]. Aster Data, “SQL MapReduce framework”, http://www.asterdata.com/product/advanced-analytics.php.

[10]. Apache HBase, http://hbase.apache.org/.

[11]. J. Lin and C. Dyer, “Data-Intensive Text Processing with MapReduce”, Morgan & Claypool Publishers, **(2010)**.

[12]. GNU Coord, http://www.coordguru.com/.

[13]. J. Woo, D. -Y. Kim, W. Cho and M. Jang, “Integrated Information Systems Architecture in e-Business”, The 2007 international Conference on e-Learning, e-Business, Enterprise Information Systems, e-Government, and Outsourcing, Las Vegas, **(2007)** June 26-29.