Enhancing the Efficiency of Web Data Mining using Cloud Computing

Tripti Lamba*
Leena Chopra**

Abstract

Data Mining is the process of discovering actionable information from raw data, which helps to enhance the capability of existing business process. Due to the unrestricted use of Internet by individuals ubiquitously, limitless data has to be stored and maintained on servers. World Wide Web is a group of massive amount of information resources, interconnected files on Internet. Mining the valuable information from this huge source is the main area of concern. In cloud computing web mining techniques and applications are major areas to focus on. Another name for cloud Computing is a distributed computing over the Network. Cloud computing doesn't require to deploy the application on local computer as it directly delivered the hosted services over the internet. The objective of the paper is to study the Map-Reduce programming model and the Hadoop development platform of cloud computing and to ensure efficiency of Web mining using these parallel mining algorithms.

Keywords: Data Mining, Web mining, Cloud Computing, map-reduce

I. Introduction

A) Web Mining

Extensive version of data mining can be termed as web mining. On web data is stored in a heterogeneous manner in a semi-structured or unstructured form due to which mining on web is difficult as compared to traditional data mining. Web data mining is used to extract useful information or facts from Web Usage logs[2], Web Hyperlinks, Web Page contents. Different types of web Mining are:

- Web structure Mining
- Web Content Mining
- Web Usage Mining [4]

The process of extracting the information on Web is called Web content mining. In Web Mining, data collection is a substantial task especially for Web Structure and Web content mining, and involves crawling a large number of Web pages[3]. The Internet has today changed computing to distributed computing or cloud computing. All the major Social Media sites: Twitter, Facebook, Linked In, and Google+ contains abundance of information are today on cloud platform. For instance Tweets happen every millisecond on Twitter, they happen at the "speed of thought". This data is available for consumption all the time. The data on Twitter ranges from small tweets to long conversational dialogues to interest graphs etc. Now which data mining technique to apply, how to find association or correlation or how to cluster the data based on their similarity, so as to gain efficiency in the platform of cloud computing is the research area.

Problems associated with Web Mining

1. **Scalability:** The database is huge and it contains large dataset so mining interesting rules adds on to uninterested rules that are huge. There is no efficient algorithm for extracting useful pattern from the huge database.
2. **Type of Data**: The data on Web is heterogeneous[5]. Web cleaning is the most important process and is very difficult for semi structured data and unstructured data. According to researchers 70% of the time is spent on data pre-processing.

Tripti Lamba*
Research Scholar
Jagan Nath University, Jaipur, India
Leena Chopra**
Research Scholar
Amity Univesity, Noida, India

3. **Efficiency:** Mining rules from semi structure and unstructured as in the semantic web is a great challenge. Lot of time and memory consumption leads to decreased efficiency.
4. **Security:** The data on web is accessed publicly. There is no data that is hidden, so this is another challenge in Web Mining.

B) Cloud Computing

The computer resources these days are consumed as utility by various companies the same manner one consumes electricity or a rented house. There is no need to fabricate and retain computing infrastructures in-house. There are three types of cloud private, public and hybrid. Cloud services are mainly categorized into three types: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service(IaaS)[8]. There are various benefits of Cloud, some of which are mentioned below:

- **Self-service provisioning:** It all depends on the end users, which type of services they yearn for. Users can revolve around multiple computing assets for almost any type of workload on-demand.
- **Elasticity:** Companies can scale up as computing needs increase and then scale down again as demands decrease.
- **Pay per use:** There is a flexibility of using the services and computing resources as per the need of demand of the user. This facility permits users to pay only for the resources and workloads they utilize.

Cloud computing is most impressive technology because it is cost efficient and flexible. Cloud Mining's Software as Service (SaaS) is used for implementing Web Mining, as it reduces the cost and increases the security. Compared to all the other web mining techniques, Web usage mining is immeasurably used and have known productive outcomes[7].

C) Web Mining and Cloud Computing

One of the mostly used technologies in Web Mining is Web Usage Mining[1]. Web Usage mining using Cloud Computing is majorly adopted these days due to its reduced cost efficiency and flexibility[6]. However, in spite of improved movement and attention, there are considerable, continual concerns about cloud computing that ultimately compromise the vision of cloud computing as a new IT procurement model. Fundamentally Cloud Mining is novel approach to faced search interface for your data. The major challenge which is a security of web mining is been offered by SaaS (Software-as-a Service) and used for dropping the cost which is termed as cloud mining technique. It's been targeted to change the existing framework of web mining to generate an influential framework by Hadoop and map Reduce communities for projecting analytics. [9]

In the next section we have discussed how to use Map/ Reduce Model in Cloud Computing and what are the various benefits of using this model.

II. Cloud Computing and Map/ Reduce Model

The term cloud is a representation designed for the Internet, an intellection of the Internet's fundamental infrastructure that helps to spot the point at which accountability moves from the user to an external provider. Cloud Computing is one of the most captivating areas where lots of services are being utilized. The main objective of Cloud computing is to fully utilize the resources dispersed at various places[10]. Map/ Reduce model which is a programming model, proposed by Google is used for processing voluminous data sets. Map/Reduce Model processes around 20 petabytes of data in a single day. This model is gaining more popularity in cloud computing these days[11][12]. Map/ Reduce model is used for parallel and disseminated processing of huge data sets on clusters[13]. Some of the applications of Map/Reduce are:

At Google:

- Index building for Google Search
- Article clustering for Google News
- Statistical machine translation

At Yahoo!:

- Index building for Yahoo! Search
- Spam detection for Yahoo! Mail

At Facebook:

- Ad optimization
- Spam detection

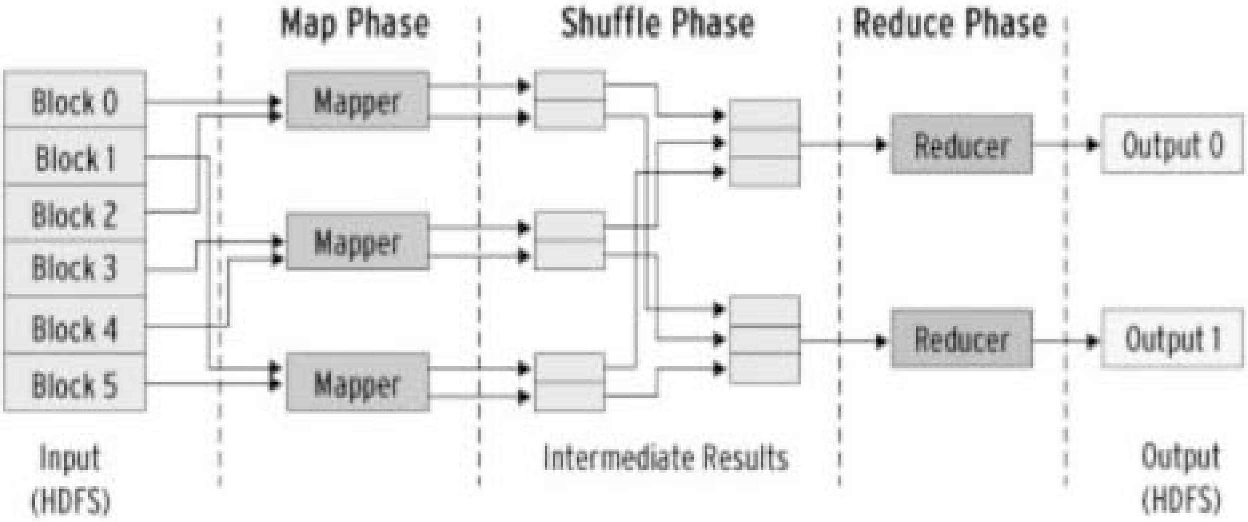

Fig. 1 Map/ Reduce System Framework[14]

A) Advantages of Map/Reduce Framework:

The main advantage of the MapReduce framework is its fault tolerance, where periodic reports from each node in the cluster are expected when work is completed. A task is transferred from one node to another. If the master node notices that a node has been silent for a longer interval than expected, the main node performs the reassignment process to the frozen/ delayed task. Some of the advantages [15] of Map/ Reduce Framework are mentioned below:

Scalability and Distributed Processing: Hadoop platform that utilizes Map/Reduce framework is extremely scalable. It has the capability to accumulate and distribute large data sets across ample of servers which operates in parallel which leads to reduced cost.

Flexibility: It operates on Structured and Unstructured data from variety of sources like email, e-commerce, social media, etc.

Fast: This framework works on Distributed architecture so huge amount of data ranging from Terabytes to petabytes. It takes minutes to process terabytes of data, and hours for petabytes of data.

Security and Authentication: Security is the major area of concern in almost every field. MapReduce works with HDFS and HBase security which allows only access to only authenticated users.

B) Map/ Reduce System Framework

The basic architecture of Map/Reduce is mentioned in Fig. 1[14] Map/ Reduce involve two basic steps:

- Map: performs filtering and sorting and
- Reduce :performs a summary operation

The input and output are in the form of key-value pairs. After the input data is partitioned into splits of appropriate size, the map procedure takes a series of key-value pairs and generates processed key-value pairs, which are passed to a particular reducer by a certain partition function; later after the data sorting and shuffling, the reduce procedure integrates the results. The scalability achieved using MapReduce to implement data processing across a large volume of CPUs with low implementation costs, whether on a single server or multiple machines, is a smart proposition.

III. Conclusion

Cloud Computing is definitely one of the widely used technologies as it is cost efficient and flexible. Web Usage Mining uses Cloud Computing Service SaaS (Software as a Service) to increase the security and reduce the cost. In this paper we have discussed the basic Map/Reduce model and its advantages. The future work will focus on new ways to improve the current model so as to aim at more accurate and faster approach for Web Usage mining, based on Cloud Computing.

References

1. M. U. Ahmed and A. Mahmood, "Web usage mining:," International Journal of Technology Diffusion, vol. 3, no. 3, pp. 1–12, Jul. 2012.

2. S. K. Pani, et.al L "Web Usage Mining: A Survey On Pattern Extraction From Web Logs", International Journal Of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.

3. Singh, Brijendra, and Hemant Kumar Singh. "Web data mining research: a survey." In Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on, pp. 1-10. IEEE, 2010.

4. J Vellingiri, S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology .Volume 11 Issue 4 Version 1.0 March 2011.

5. Li, J., Xu, C., Tan, S.-B, "A Web data mining system design and research". Computer Technology and Development 19: pp. 55-58, 2009

6. Robert Grossman , Yunhong Gu, "Data mining using high performance data clouds: experimental studies using sector and sphere", Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, August 24-27, 2008

7. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining," ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, p. 12, Jan. 2000.

8. Khanna, Leena, and Anant Jaiswal. "Cloud Computing: Security Issues And Description Of Encryption Based Algorithms To Overcome Them." International Journal of Advanced Research in Computer Science and Software Engineering 3 (2013): 279-283.

9. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using modelbased clustering. In In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,pages 280{284, Boston, Massachusetts, 2000.

10. Zhu, W., & Lee, C. (2014). A new approach to web data mining based on cloud computing. Journal of Computing Science and Engineering, 8(4), 181–186. doi:10.5626/jcse.2014.8.4.181

11. "MapReduce." Wikipedia. N.p.: Wikimedia Foundation, 11 Jan. 2017. Web. 2 Jan. 2017.

12. Divestopedia, and Securities Institute. What is MapReduce? - definition from Techopedia. Techopedia.com, 2017. Web. 2 Jan. 2017.

13. Posted, and Margaret Rouse. What is MapReduce? - definition from WhatIs.com. SearchCloud Computing, 25 June 2014. Web. 2 Jan. 2017.

14. Hornung, T., Przyjaciel-Zablocki, M., & Schätzle, A. (2017). Giant data: MapReduce and Hadoop » ADMIN magazine. Retrieved January 10, 2017, from http://www.admin-magazine.com/HPC/Articles/MapReduce-and-Hadoop

15. Lee, K.-H., Lee, Y.-J., Choi, H., Chung, Y. D., & Moon, B. (2012). Parallel data processing with MapReduce. ACM SIGMOD Record, 40(4), 11. doi:10.1145/2094114.2094118