IPASJ International Journal of Computer Science (IIJCS)
Web Site: http://www.ipasj.org/IIJCS/IIJCS.htm
A Publisher for Research Motivation ..... Email:editoriijcs@ipasj.org
Volume 5, Issue 11, November 2017 ISSN 2321-5992

Role of Map reduce Algorithm to Improve the Web data Access

T. Mylsami [1], Dr. B.L. Shivakumar [2]

[1] Department of Computer Science and Information Technology,
Dr. G.R. Damodaran College of Science, Coimbatore, Tamil Nadu, India

[2] Principal, Sri Ramakrishna Polytechnic College, Coimbatore, Tamil Nadu, India

Abstract

In recent days the data generation in web based is enormous and even it keep on upgrading every minute. Among the various sources of data generation through web source play an important role.Due to increase of web users, the data generation, data transfer and storage are increasing in unpredictable in nature. The overcome the issue in web data generation, the user can effectively accesses those data in easy way through Map reduce algorithm. The Map Reduce algorithm will sort out the different problems in web and it produce the better results in web access.

1. Web mining

Web mining is the techniques used to discover patterns from the World Wide Web. It deals either organized and unstructured data or information from browser activities, server logs, and webpage accessed patterns. Web mining divided into three different categories such as Web usage mining, Web content mining and Web structure mining.

Web mining uses the techniquesand algorithms used to extract information directly from the Web.The extracted web data from the sources like Web documents, services, Web content, hyperlinks and server logs. The key role of the Web mining is check for patterns in Web data by collecting and analyzing information in order to reach the required webpage.

Web mining is a sub division of data mining; it focused on the World Wide Web as the key data source, it includes data components from Web content, server logs to everything. The contents of data mined from the Web may be a collection of facts. Normally Web pages are consisting of text, structured data such as text, lists, tables, images, video and audio.

Types of Web mining:

- Web content mining — The process of mining useful information from the contents of Web pages, which are mostly text, images and audio/video files. In among various techniques maximum of cases the user prefers to use Natural Language Processing (NLP) and information retrieval.
- Web structure mining — The goal of Web structure mining is to generate structural summary about the Web site and Web page. The process of analyzing the nodes and connection structure of a website through the use of graph theory. There are two things that can be obtained from the structure of a website in terms of how it is connected to other sites and also how each web page is connected.
- Web usage mining — Web usage mining relies on data captured behind the scene in server logs and databases.It belongs to the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of online user.This is the process of extracting patterns and information from server logs to gain insight on user activity.It includes various activities like user log, how many click happened, which link accessed and other related details.

2. Functions of Map reduce Framework

MapReduce framework is a brainwashing model related to data and specifically it used forprocessing and generating big data sets with a parallel, distributed algorithm.Map Reduceconceptmainly splits the input datainto independent elements and later processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the

maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system.

MapReduce is a software framework designed specifically for distributed processing of large data sets. Map/Reduce clusters of commodity hardware. It is a sub-project of the Apache Hadoop project. The framework takes care of scheduling tasks, monitoring them and re-executing any failed tasks.

- The map task is done by means of Mapper Class
- The reduce task is done by means of Reducer Class.

Mapper class takes the input, tokenizes it, maps and sorts it. The output of Mapper module is used as input by Reducer module, which in turn searches matching pairs and reduces them for final output.

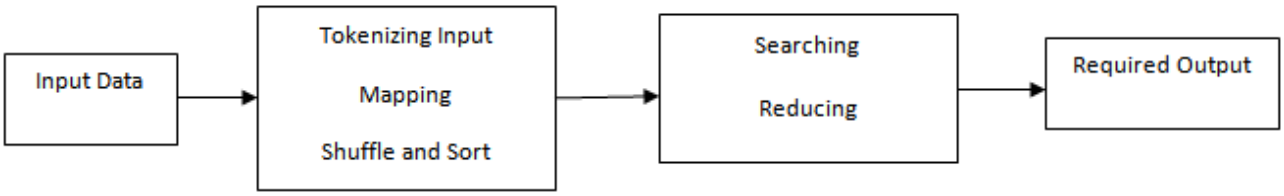

Figure 1: Basic flow of Map Reduce

The MapReduce algorithm contains two important tasks, namely Map and Reduce.

- The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).
- The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.

MapReduce Phases: -The following is the architecture diagram for Map Reduce algorithm and mentioning the different phases.

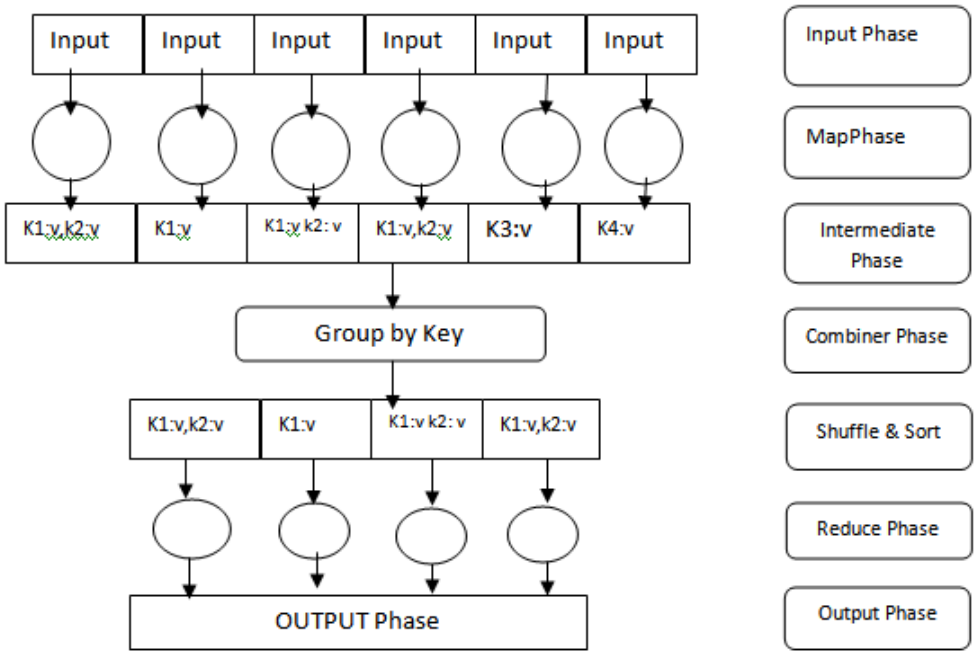

Figure 2: The basic structure of Map reduce framework

- **Input Phase**–In this phase the each record has been given as input file and sends the record to Map phase in the form of key-value pairs.
- **Map**–It is a user-defined function and it takes a series of key-value pairs.It processes each one of pairs and then to generate zero or more key-value pairs.
- **Intermediate Keys** − The key-value pairs generated by the map phase are called intermediate keys.
- **Combiner**–Combiner is a type of local Reducer that groups similar data from the map phase into identifiable sets. It takes the intermediate keys from the map as input and applies a user-defined code to aggregate the values in a small scope of one map phase and act as an optional one in Algorithm.
- **Shuffle and Sort** − Reducer task starts with the Shuffle and Sort process. It downloads and processes the grouped key-value pairs onto the local machine. The individual key-value pairs are sorted by key into a larger data list. The data list groups the equivalent keys together so that their values can be iterated easily in the Reducer task.
- **Reducer** − The Reducer takes the grouped key-value paired data as input and runs a Reducer function on each one of them. The data can be aggregated, filtered, and combined in a number of ways, and it requires a wide range of processing. Once the execution is over, it gives zero or more key-value pairs to the output step.
- **Output Phase** − In output phase, we have an output formatter that translates the final key-value pairs from the Reducer function.

3. Basic Operation of Map Reduce

In recent decades the growth of Internet data is huge and the data from various sources also high in numbers. Due the different data file format, it a difficult for every managing person in the fields.The key problem in the following areas for huge data in internet.

1. Data storage 2. Data Processing 3. Data Management.

The same problem is occurred in web related application like, Data generation in webpage, data processing, data retrieval and data storage. To solve the various issues in web page concepts, the Map Reduce algorithm will perform in maximum possible ways and giving better results in time. Specifically it executes in real time applications and managing the various applications for regular work flow through Map reduce algorithm.

The following is the sample diagram for function operational of Map reduce component.

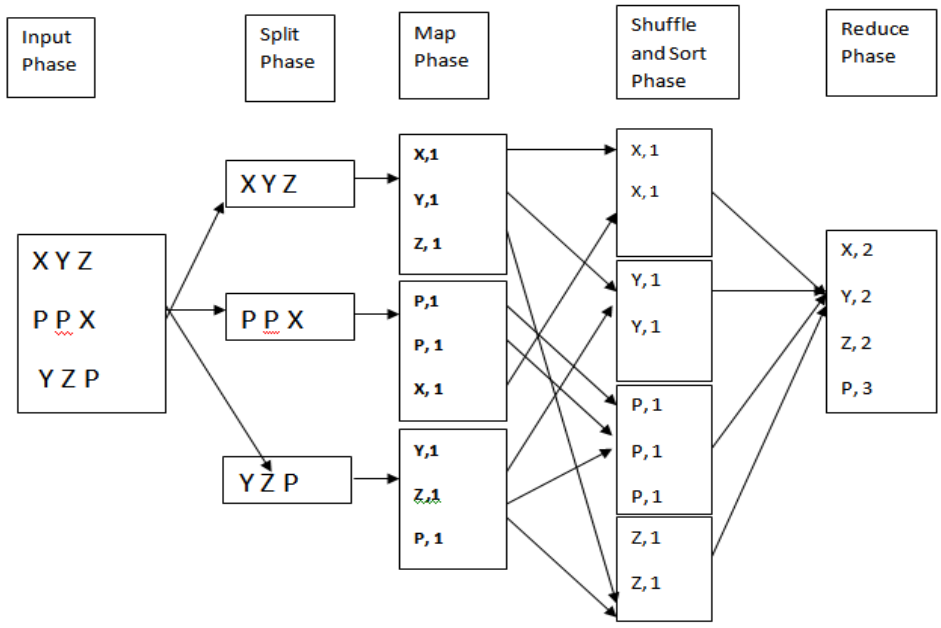

Figure3: Map Reduce framework

From the above figure 3 explains the overall functions of Map reduce functions applied in various applications. Consider that above input phase of X, Y, Z, P is web pages. Each web page as give as input for Map reduce framework and to reach the output as expected one in minimum steps.

The following steps to be consider for effective search of output.

- The Input of all possible web pages to be directly given to Process.
- Second it split the web pages in to similar category in groups and names it.
- Each category of web page is assigned with key value in map phase.
- The entire key valued website has been shuffle and sort done in next phase.
- Finally the reduce web page has been grouped and forward for the next processing stage.

Advantages of Map Reduce programming

- **Scalability–**It is an ability to store as well as distribute large data sets across sufficiently of servers. Those servers will be aninexpensive one and it can be operated in parallel and with each addition of newer servers it adds more processing power.

- **Cost-effective solution–It is**highly scalable structure and implies that it comes across as a very cost-effective solution for businesses that need to store ever growing data.

- **Flexibility– Inall the application areas,** MapReduce concept access to different new sources of data, it operates on different types of data and data maybelong to structured or unstructured. Even though it allows the user to generate value from all of the data and can be accessed by the user in easier manner.

- **Fast -** The Map Reduce system uses a storage method as distributed file system, and it implements a mapping system to locate data in a clustering. The tools used for data processing and generating in MapReduce programming, are available same servers. Due this reason it allows the user for faster processing of data.

- **Parallel processing- The key factorexecute on** MapReduce programmingit divides the tasks in a manner that allows their execution in parallel. It mainly consumed the timing for all the cases.

- **Simple model of programming - T**he most importantactivityis thatbased on a simple programming model. It allows programmers to develop MapReduce programs that can handle tasks with more ease and efficiency.

- **Security and Authentication -** Security is anessential aspect of any application to protect the data from unauthorized user. Due the security the vast amount of data to safe and can be maintained for many a years for future access.

Disadvantages

- Programming model is very restrictiveand lack of central data can be preventive.

- Joins of multiple datasets are tricky and slow

- In case of cluster operations like debugging, distributing software, collection logs are too hard to mange.

4. Conclusion

The Map Reduce concept has been playing an important in data handling applications. The MapReduce is simple but provides good scalability and fault-tolerance for massive data processing.It substitute DBMS application and even for datawarehousing. So the Map Reduce framework has applied in various applications to fulfill the objectives in minimum steps. Like the same in Web page access the Map Reduce concepts has solved the web access in short time and reduce the unwanted linked pages for user. Finally the Map Reduce algorithm play an efficient role in all web related applications and data transaction related fields.

References

[1] Dr. Siddaraju, Sowmya CL, Rashmi K, Rahul M, " Efficient Analysis of Big Data using Map Reduce Framework", International Journal of Recent Development in Engineering and Technology, Vol.2, Issue-6, June2014, pp 64-68.
[2] Kyog Ha Lee, Yoon-Joon Lee, "Parallel Data Processing with MapReduce: A Survey", SIGMOD Record, Dec 2011, Vol. 40, No.4, pp 11 – 20.
[3] ShafaliAgarwal, ZebaKhanam, " Map Reduce: A Survey paper on Recent Expansion", International Journal of Advanced Computer Science and Applications, Vol. 6, No.8, 2015. pp – 209-215.
[4] P.Sudha, Dr.R.Gunavathi, " ASurevy Paper on Map Reduce in Big data", International Journal of Science and Research, Vol.5, Issue.9, Sep 2016, pp 1103-1107.
[5] ShitalSuryawanshi, V.S.Wadne, "Big data Mining using Map Reduce: A Survey paper", IOSR Journal of Computer Engineering, Vol.16, Issue.6, Dec-2014, pp 37-40.
[6] Tripti Mehta, NehaMangla, " A Survey paper on Big Data Analytics using Map Reduce and Hive on HadoopFramewrok", International Journal of recent Advances in Engineering and Technology, Vol.4, Issue.2, Feb 2016, pp 112-118.
[7] M.Srilekha, A.Santhoshi, " Page Rank Algorithm in Map Reducing for Big Data", International Journal of Cocneptions on Computing and Information Technology, Vol.3, Issue1, Apr 2015.
[8] MadhaviVaidya, "Parallel Processing of Cluster by Map Reduce", International Journal of Distributed and Parallel Systems, Vol.3, No.1, Jan 2012. Pp 167- 179.
[9] Jeffrey Dean and Sanjay Ghemawat, "Map Reduce : Simpliied Data Processing on Large Clusters", OSDI 2004.

AUTHOR

Mr. T.Mylsami, Assistant Professor, Department of Computer Science and Information Technology, Dr.G.R.Damodaran College of Science, Coimbatore. He obtained his Master Degree from Indira Gandhi National Open University, New Delhi, India in 2007 and he obtained Master of Philosophy (Data Mining) from Bharathiar University, Coimbatore 2011. He has more than 10 years of academic experience in college. His research interests include Data Mining, Web mining and Software testing.

Dr.B.L.Shivakumar holds a Ph.D. in Computer Science from Bharathiar University, Coimbatore. He has more than 20 years of academic experience in various positions in reputed colleges. At present he is working as Principal at Sri Ramakrishna Polytechnic College, Coimbatore. He has 12 years of research experience and has 56 research publications to his credit in reputed journals and conferences. He has successfully guided six Ph.D. Research Scholars from leading Universities and presently 6 students are pursuing Ph.D., under his guidance. He has written seven books and 38 articles related to Computer in Tamil daily "Dina Thanthi" in Computer Jallam. He is recipient of Bharat Jyoti award conferred by The India International Friendship Society, New Delhi and NSS Best Programme Officer award by Bharathiar University. His interest includes Computer Forensic Science, Network Security and Cloud computing. He is a member in a number of Academic Bodies and Professional Societies.