Research Article

# Analyzing Log Files to Find Hit Count Through the Utilization of Hadoop MapReduce in Cloud Computing Environmen

Anil G,[1*] Aditya K Naik,[1] B C Puneet,[1] Gaurav V,[1] Supreeth S[1]

**Abstract:** Log files which are created by the web serverscontain information about the activities of the visitors like number of visitors and from which domain they are visiting. Terabytes of log files are generated on a daily basis, which can't be efficiently processed by a data storage mechanism. Hadoop cluster deployed on cloud environment provides a reliable solution. The size of the log files keep increasing, the number of data nodes can be expanded in cloud environment.

**Asian Journal of Engineering and Technology Innovation**
**Volume 4, Issue 7**
**Published on: 7/05/2016**

**Cite this article as:** Anil G *et al.*, Analyzing Log Files to Find Hit Count Through the Utilization of Hadoop MapReduce in Cloud Computing Environmen. Asian Journal of Management Sciences, Vol 4(7): 61-65, 2016.

## INTRODUCTION

In today's expanding world, everything is going online. Sectors like public, private and business has seen a significant development in their respective fields. There has been an exponential growth in data over the web. We need to analyze these data to provide better service to various sectors to improve enterprise scenario. Data is heterogeneous in nature and analysis of such data will provide us with important information wherein log files provide an efficient solution. Log files are located in the web server and it contains information about every individual's requests, which is stored in a log entry. The main purpose of using Hadoop MapReduce is to analyze the datasets effectively. In this system, we have implemented Hadoop MapReduce model to analyze log files.

Log files are generated everyday which are in the order of terabytes. Log files contain huge amounts of useful information which can useful to improve business enterprises and future assessment. In order to gain knowledge about the customer's activities, whether he is purchasing the product, if he is finding the application friendly to use or the problems he is facing and how it can be resolved, we need to analyze log files. Thus through log file analysis, we gain insight into all the above mentioned questions and interaction of people with web applications.

## LITERATURE SURVEY

Cloud computing is the recent technological buzzword which can process huge amounts of data within fractions of seconds. We live in a data-centric world, where everything is online. In this data age, many big companies like Yahoo, Google have difficulties in handling large datasets. Google developed [12] MapReduce and Google file system to improve the scalability of data processing. MapReduce was used by various web applications and finally adopted by Hadoop. Log files keep generating a record rate. They are hard to analyze not only because of their huge volume but also due to their contrasting structure. Due to this reason we use Hadoop MapReduce framework which provides reliable data storage for large volumes of log files and parallel distributed processing.

Each and every sector has their own way of putting their business and application online. In the comfort of our home, we can get the weather updates, purchase a product and even perform bank related work. These activities [4] which are performed by the customer are stored in the log files. Later we can analyze these log files to improve the advertising strategies to increase the number of customers, to scale a business and make the application more interesting to use. Analysis of log files helps us to resolve such problems which can be pinpointed and fixed.

Hadoop is a good platform [6] for storing and analyzing tons of data. Two main components of Hadoop [9] are Hadoop distributed file system and MapReduce. HDFS stores petabytes of data by breaking them into small chunks and provides high-performance access to data across Hadoop clusters.Apache Hadoop is an open source project which was developed by Doug Cutting, which allows storage of petabytes of data. Hadoop works well with structured as well as unstructured data, and supports various serialization and data formats.

[1]Reva Institute of Technology and Management, Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Near Border Security Bustop, Bengaluru, Karnataka-560064, India.
E-mail: ashwin@revainstitution.org
*Corresponding author

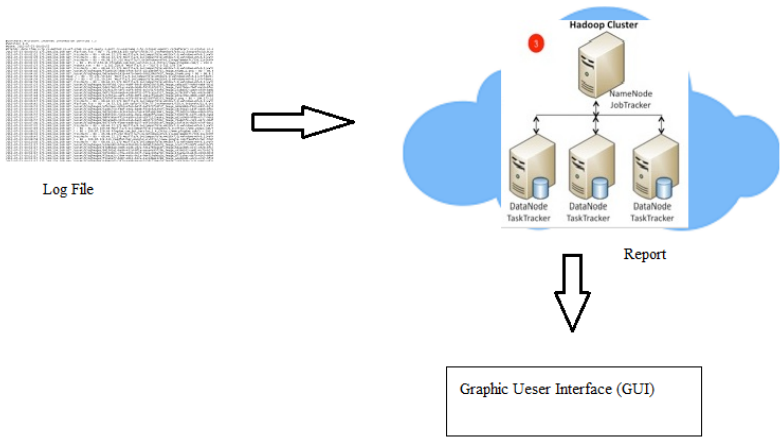


**Figure 1:** System Architecture

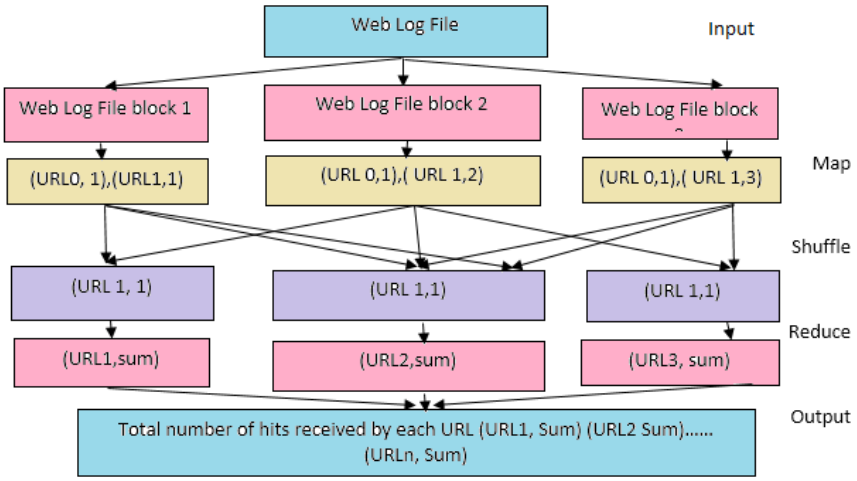


**Figure 2:** Calculating log files hit based on URL

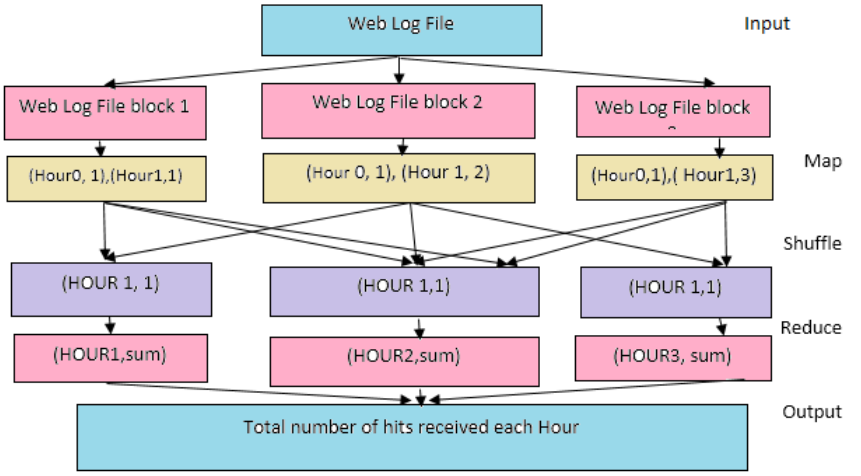


**Figure 3:** Calculating log files hit on each hour

Hadoop is distributed from the ground up and we can add more nodes and increase the capacity.

A HDFS cluster [5] primarily consists of a NameNode that manages the file system metadata and DataNodes that store the actual data. One of the major aspects MapReduce programming model is that it divides tasks in a manner that allows their execution in parallel. Parallel processing allows multiple nodes to take on these divided tasks, so that they run entire programs in less time. A MapReduce [1, 6] job generally splits the input data-set into various chunks which are later processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then provided as inputs to the reduce tasks. Typically both the input and the output of the job are stored in a file-system.

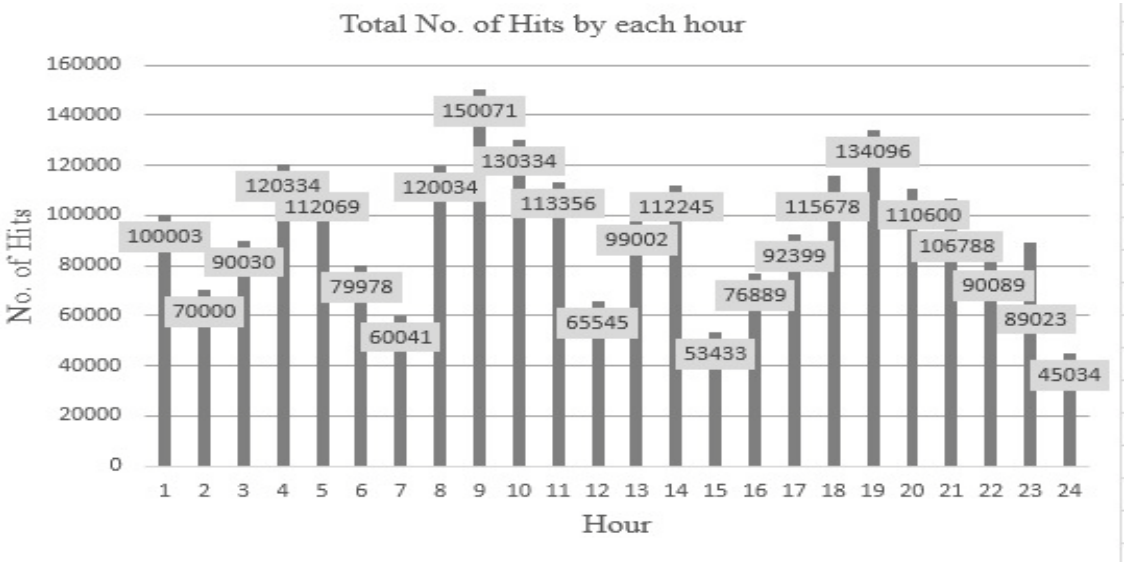


**Figure 4:** Hits received by a web site in each hour (Bar chart)

| | HOST | HITS |
|---|---|---|
| 1 | 134.76.89.104 | 124 |
| 2 | 67.109.56.78 | 324 |
| 3 | 75.220.13.112 | 221 |
| 4 | 88.105.10.122 | 178 |
| 5 | 78.56.137.145 | 67 |
| 6 | 80.54.67.125 | 268 |
| 7 | 56.78.78.129 | 23 |
| 8 | 90.74.138.93 | 45 |
| 9 | 81.45.76.231 | 98 |
| 10 | 61.80.94.251 | 142 |
| 11 | 85.167.107.212 | 63 |
| 12 | 116.78.82.108 | 11 |
| 13 | 204.90.56.87 | 54 |
| 14 | 84.57.93.106 | 8 |

**Figure 5:** Table data on number of hits for each hour

## PROPOSED SYSTEM

In Figure 1 Hadoop Cluster is created on cloud environment to carry out the experiment. Amazon EC2 is used for cloud environment and four instances are created with Ubuntu14.04 operating system. These four instances are used as nodes of Hadoop Cluster. One node is used as master (name) node and other three are used as data (slave) nodes.

The Log file is a raw text file with fields containing IP address, Time Stamp, HTTP Request, etc.Python cherrypy server is run on the master node. The log file is uploaded to

the server which will put the log file in HDFS. Hadoop splits the log file in small chunks of file of equal size and these chunks are distributed over multiple nodes in the cluster. These blocks are processed in parallel manner using MapReduce. Pig script processes the log files distributed across Hadoop Cluster. It will produce the result in csv format. JavaScript is used to process the csv files to produce in report with pie chart, bar chart, table format, etc. The server will return these report to the user.

This Figure 2 describes the MapReduce function of processing log file and calculating the total number of hits received by each URL. Log file is given as input to the function. A line is added to the log file for each hit in the web site. The line in the log file contains the following fields: client IP address, User name, Server Name, date, time, request method, requested resource, HTTP version, HTTP Status and Bytes sent. Hadoop Framework splits the log files into blocks and stored into data nodes. In the mapper function. To the map function we give each block of log file as input, which parses each line using regular expression and emits the URL as a key along with the value 1 (URL1,1), (URL2,1), (URL3,1),….,(URLn,1). After mapping is done, the shuffling collects all the (Key, Value) pairs which are having the same URL from different mapping function's and forms a group. After this process, Group1 entries will be (URL1,1), (URL1,1), (URL1,1) and so on. Group2 entries will be (URL2,1), (URL2,1) and so on. Later the reduce function determines the sum for each URL group. The result of the reduce function is (URL1,sum), (URL2,sum),…(URLn,sum).

Figure 3 describes the MapReduce function of processing log file and calculating the total number of hits received in every hour. Log file is given as input to the function. The web log file is split into blocks. To the map function we give each block of log file as input, which parses each line using regular expression and emits hour as a key along with the value 1 (Hour0,1), (Hour1,1), (Hour3,1),….,(Hour23,1). After mapping is done, the shuffling collects all the (Key, Value) pairs which are having the same hour from different mapping function's and forms a group. After this, Group1 will be (Hour0,1), (Hour0,1), (Hour0,1) and so on. Group2 will be (Hour1,1), (Hour1,1) and so on. . Later the reduce function determines the sum for each hour group. The result of the reduce function is (Hour0, sum), (Hour1, sum), (Hour23, sum).

## EXPERIMENTAL SETUP

To analyze the log file of web sites and to produce the report with total number of hits received by each URL and by a web site in each hour, and etc.

In AWS cloud a Hadoop Cluster is created with the following configuration

| | |
|---|---|
| Cloud Service | AWS (Amazon Web Service) |
| Operating System | Ubuntu 14.04 |
| Nodes | Name Node: 54.67.83.75<br>Data Node1: 54.67.44.230<br>Data Node2: 52.53.203.173<br>Data Node3: 54.193.38.46 |
| Data | Web log file |

## RESULTS OF EXPERIMENT

Figure 4 shows the number of hits received by a web site in each Hour which is represented pictorially. From the graph, we can see that during 9th hour the website has been accessed the most.

Figure 5 shows the number of hits received by a web site which is represented pictorially. From the graph, we can see that the ip address 67.109.56.78 has visited the website the most.

## CONCLUSION

We need to analyze the log files in order to understand the customer activities and improve the business strategies. We have proposed best fit MapReduce programming model on Hadoop multimode cluster which is deployed on top of cloud environment to analyze log files. The bar chart gives the statistical record of analysis for various parameters in the log file. In future, we can increase the number of nodes in the cluster, which increases the performance of the system.

## REFERENCES AND NOTES


1. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. Sixth Symp. Operating System Design and Implementation (OSDI '04), pp. 137-150, Dec. 2004.
2. D. Jiang, B.C. Ooi, L. Shi, and S. Wu, "The Performance of MapReduce: An In-Depth Study," Proc. VLDB Endowment, vol. 3, no. 1, pp. 472-483, 2010.
3. Apache Hadoop Project, http://Hadoop.apache.org/, 2013.
4. An Efficient Market Basket Analysis Technique with Improved MapReduce Framework on Hadoop: An E-commerce Perspective (International Conference on Database Systems for Advanced Applications, 2012 7238, 258-271, LNCS, Springer)
5. Debajyoti Mukhopadhyay, Chetan Agrawal, Devesh Maru, Pooja Yedale, Pranav , "Addressing Name Node Scalability Issue in Hadoop Distributed File System Using Cache Approach," ICIT '14 Proceedings of the 2014 International Conference on Information Technology Pages 321-326.
6. Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N Prasad.M.R , "Analysis of Bidgata using Apache Hadoop and Map Reduce," May 2014 International Journal of Advanced Research in Computer Science and Software Engineering.

7. IBM 2012, What is big data: Bring big data to the enterprise,http://www-01.ibm.com/software/data/ bigdata/,IBM.
8. Amazon EC2, http://aws.amazon.com/ec2/
9. Jens Dittrich and Jorge-Arnulfo Quiane-Ruiz, "Efficient Big Data Processing in Hadoop MapReduce," Proceedings of the VLDB Endowment VLDB Endowment Volume 5 Issue 12, August 2012 Pages 2014-2015.
10. Data Mining with BigData by Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding , 1041-4347/13/$31.00 © 2013 IEEE
11. S. Ghemawat, H. Gobioff, and S.-T. Leung. The Google file system. In SOSP, pages 29–43, 2003
12. Gillick D., Faria A., DeNero J., MapReduce:Distributed Computing for Machine Learning, Berkley, December 18, 2006.