iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4178

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

"Analysis of Web Server Log File Using Hadoop"

Sohan Panwar[1], Garima Silakari Tukra[2]
[1]Research Scholar, [2]Assistant Professor, Department of Computer Science & Engg, Truba College Indore, M.P, India.

Abstract: ***Web usage mining is concerned with finding user navigational patterns on the World Wide Web by extracting knowledge from web usage logs. The log files, which in turn give way to, an effective mining and the tools used to process the log files. It also provides the idea of creating an extended log file and learning the user behavior. Analyzing the user activities is particularly useful for studying user behavior when using highly interactive systems. This paper presents the details of the methodology used, in which the focus is on studying the information-seeking process and on finding log errors and exceptions. The next part of the paper describes the working and techniques used by web log analyzer.***
Keywords: Server Web Log files, Map Reduce, Hadoop, Web Mining

I. INTRODUCTION

Web mining is the process in which all data mining techniques are applied on the World Wide Web to extract information. Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data.
Web pages contain enormous amount of information that may not be interested to the user. Weblog data is one of the major sources which contain all the information regarding the users visited links, browsing patterns, time spent on a particular page or link, this information can be used diverse applications like adaptive websites, modified services, customer summary, generate attractive web sites etc. Web Usage Mining (WUM) is the main application of data mining to the web data and estimate the user's visiting behaviors and obtain their interests by inspecting the web log files. The first and crucial step in WUM is preprocessing of log for cleaning data, then be suitable for mining purposes [2].

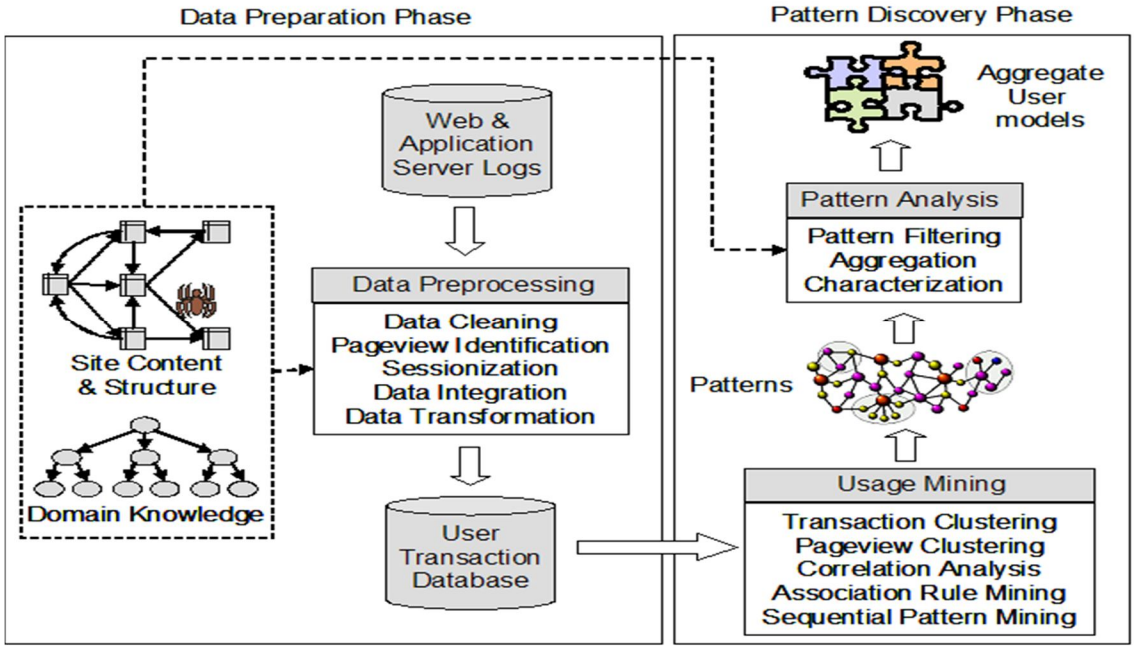


Figure 1 Process of Web Mining

In the web mining, there are three types of mining:

A. Web Structure mining: It discovers useful knowledge from hyperlinks (or links), which represent the structure of the Web.
B. Web Usage mining: It refers to the discovery of user access patterns from Web usage logs, which record every click made by each user.
C. Web Content mining: It extracts or mines useful information or knowledge from Web page contents.

IT organizations analyze server logs to answer questions about security and compliance. A server log is a simple text file, which records activity on the server. Computer generated logs that capture data on the operations of a network. Use full for managing network operations, especially for security and regulatory compliance. There are several types of server log website owners are especially interested in access logs which record hits and related information. These logs are in large amount thus resulting collection of large amount of data Big Data. Big data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. It includes huge volume, high velocity, and extensible variety of data. This data can be in structured, semi structured or in unstructured form.

II. RELATED WORK

1) Samneet Singh et. al. this paper presents our cloud provider architecture that explores a search cluster for information indexing and query. In this paper, author proposed a structure that integrates search-headquartered clusters and semantic media wiki by using relaxation APIs to help the exploration of cloud monitoring data. This structure advantages from an internet-based Media-Wiki interface and enables a person to outline the entry to monitoring knowledge and prepare the processing results.

2) JOSEPH et.al. In this paper, the writer offered a distinct performance analysis and analysis for Hadoop Word Count workload utilizing different processors similar to Intel's and AMD's Bobcat E350. Analysis suggests that Hadoop Word Count is compute-sure workload in both map segment and scale down segment. They also conclude that the Intel's ATOM cluster can reap a higher efficiency/watt in comparison with AMD's Bobcat cluster at highest performance. Evaluating Intel's ATOM to Intel's Xeon X5690, the performance/buck for Xeon is better com- pared to the performance/buck for ATOM.

3) Yaxiong Zhao, Jie Wu, and Cong Liu, "Dache: In this paper, author recommends a knowledge-conscious cache framework for big-data functions. A novel cache description scheme and a cache request and reply protocol are designed. We enforce Dache by means of extending Hadoop. Test bed experiment results show that Dache tremendously improves the completion time of MapReduce jobs.

4) Zhuoyao Zhang Ludmila Cherkasova, "In this work, author presents a novel efficiency analysis framework for answering this question. We observe that the execution of every map (lessen) duties consists of distinctive, good-defined knowledge processing phases. Handiest map and scale back services are customized and their executions are consumer-outlined for exclusive MapReduce jobs.

5) Nikzad Babaii Rizvandi et. al. In this paper, we advocate an analytical system to model the dependency between configuration parameters and whole execution time of Map-diminish functions. Our strategy has three key phases: profiling, modeling, and prediction. In profiling, an application is run several occasions with specific units of MapReduce configuration parameters to profile the execution time of the applying on a given platform.

6) Nikzad Babaii et.al. In this paper, author presents an strategy to provision the whole CPU usage in clock cycles of jobs in MapReduce atmosphere. For a MapReduce job, a profile of complete CPU utilization in clock cycles is developed from the job prior executions with distinct values of two configuration parameters e.g., quantity of mappers, and quantity of reducers.

Then, a polynomial regression is used to model the relation between these configuration parameters and whole CPU utilization in clock cycles of the job. We additionally in short learn the have an effect on of input information scaling on measured complete CPU usage in clock cycles. This derived mannequin together with the scaling outcome can then be used to provision the total CPU usage in clock cycles of the same jobs with one of a kind enter information dimension.

III. WEB LOG FILE

The web log is a registry of web pages accessed by different users at different times, which can be maintained at the server-side, client-side or at a proxy server, each having its own benefits and drawbacks on finding the users' relevant patterns and navigational sessions [5].

A. Client Log : it is the client itself which sends to a repository information regarding the user's behavior (can be implemented by using a remote agent (such as Java scripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.

B. Proxy Log: information is stored at the proxy side, thus Web data regards several Websites, but only users whose Web clients pass through the proxy.

C. Server Log: the server stores data regarding requests performed by the client, thus data regard generally just one source. Server Log details are given in Figure 2.

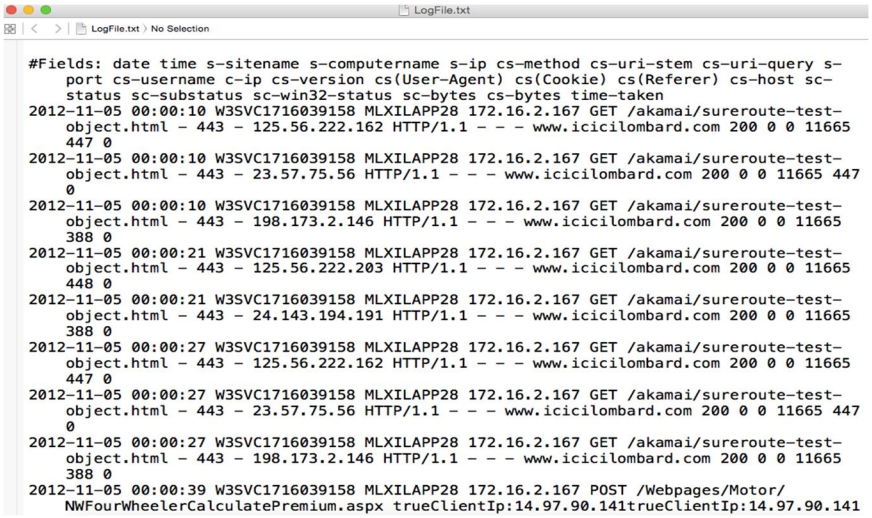


Figure 2 A Sample of Serer Side Web Log

IV. MAP REDUCE

The primary objective of Map/Reduce is to split the input data set into independent chunks that are processed in a completely parallel manner. The Hadoop MapReduce 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.

There is 5-steps parallel and distributed computation:

A. Map input – the "MapReduce system" designates Map processors, assigns the K1 input key value each processor would work on, and provides that processor with all the input data associated with that key value.
B. Map code – Map is run exactly once for each K1 key value, generating output organized by key values K2.
C. Shuffle– the MapReduce system designates Reduce processors, assigns the K2 key value each processor would work on, and provides that processor with all the Map-generated data associated with that key value.
D. Reduce code – Reduce is run exactly once for each K2 key value produced by the Map step.
E. Final output – the MapReduce system collects all the Reduce output, and sorts it by K2 to produce the final output.

V. RESULT

All the experiments in improved performed on a 2.40GHz Intel(R) Core(TM) i7-2430 MB memory, 4GB RAM running on the Unix OS. Programs will be coded on JAVA. For the purpose of this experiment we have utilized the Web server's log file of an e-commerce. We considered a log file corresponding to a two weeks period. This has approximately size of 7 MB.
We calculate computation time for User Agent, IP address Check, Head Request and Proposed Method. And results shown with help of graph. From experiments found that Proposed Method took less computation time compare to Existing methods.

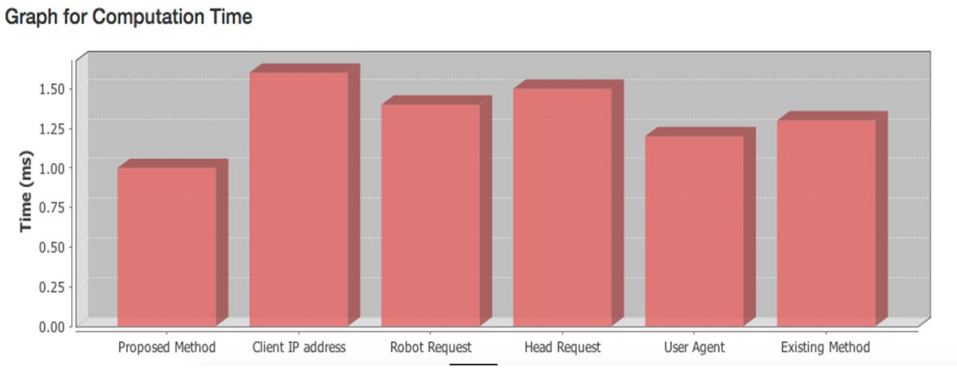


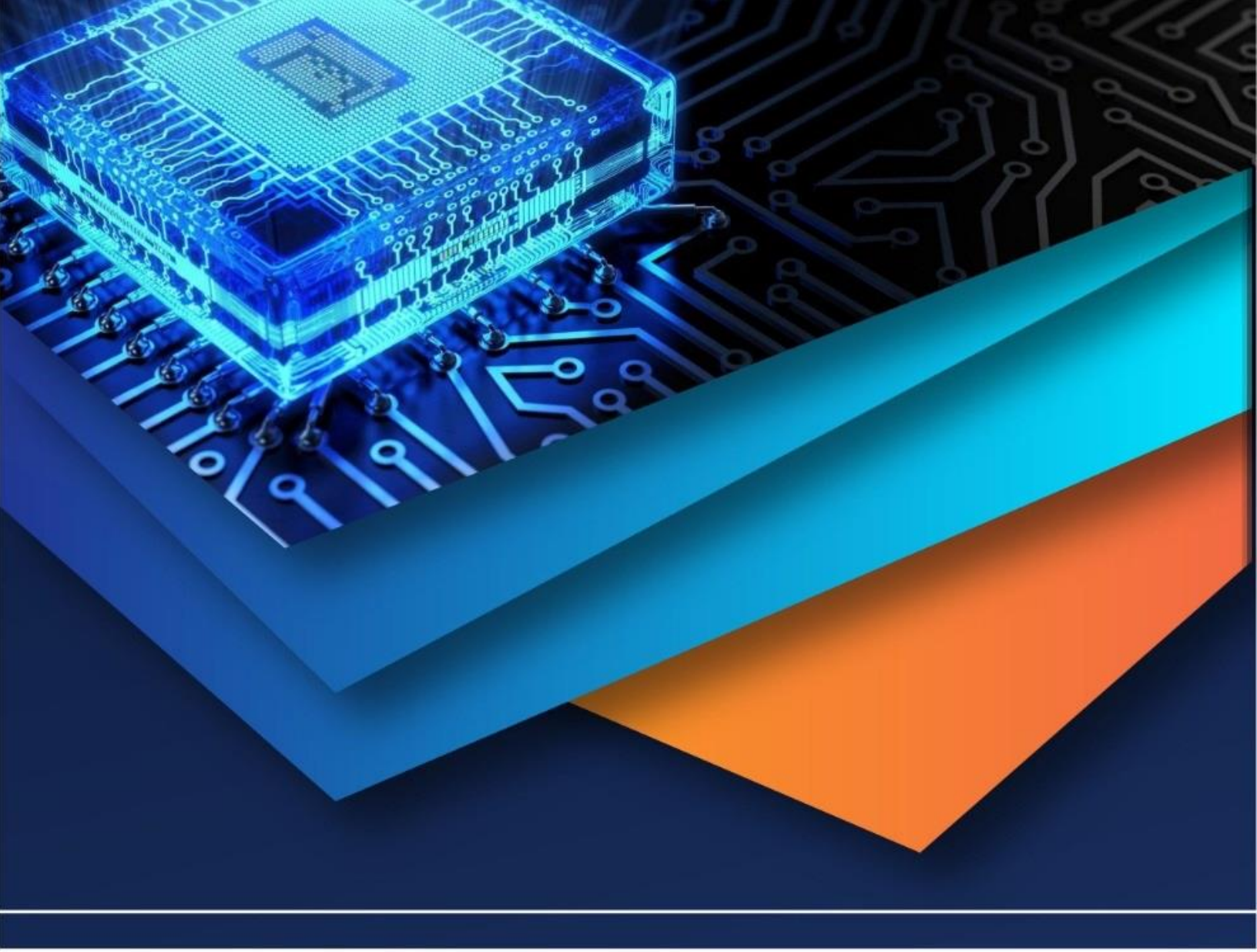
Figure. 3. Computation time Comparison between existing and Proposed Methods

VI. CONCLUSION

In this study, we discuss the Web Server Log Processing that uses Hadoop for improving the performance of a database management system (DBMS)-based analysis service system that processes big data. Traditional log processing systems are not suitable for processing big data and providing service because of their disadvantage in consuming more time for processing and analyzing. We introduced a distributed parallel platform, Hadoop ecosystem, for improving the performance of the system by minimizing the processing time in analyzing big data. This study explained the method of changing an existing log analysis service system to a distributed parallel-based environment system to address the problems encountered during the processing of big data. We optimized the system by using Hadoop ecosystem to improve the performance while processing big data.

REFERENCE

[1] Samneet Singh and Yan Liu,"A Cloud Service Architecture for Analyzing Big Monitoring Data",ISSNll1007-0214ll05/10llpp55-70 Volume 21, Number 1, February 2016
[2] JOSEPH A. ISSA, "Performance Evaluation and Estimation Model Using Regression Method for Hadoop WordCount", Received November 19, 2015, accepted December 12, 2015, date of publication December 18, 2015, date of current version December 29, 2015.
[3] Yaxiong Zhao, Jie Wu, and Cong Liu, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework",ISSNll10070214ll05/10llpp39-50 Volume 19, Number 1, February 2014
[4] Zhuoyao Zhang Ludmila Cherkasova, "Benchmarking Approach for Designing a MapReduce Performance Model", ICPE'13, April 21-24, 2013
[5] Nikzad Babaii Rizvandi, Albert Y. Zomaya , Ali Javadzadeh Boloori, Javid Taheri1, "On Modeling Dependency between MapReduce Configuration Parameters and Total Execution Time", 2012
[6] Nikzad Babaii Rizvandi, Javid Taheri1, Reza Moraveji, Albert Y. Zomaya, "On Modelling and Prediction of Total CPU Usage for Applications in MapReduce Enviornments", 2011
[7] Extracting WebLog of Siam University for Learning User Behavior onMapReduce -2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012) .
[8] Mining of Web Server Logs in a Distributed Cluster Using Big Data Technologies -(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No. 1, 2014.
[9] Tom White: Hadoop, "The Definitive Guide (1st edn.)", O'Reilly Media, Inc., United States of America, 2009.
[10] Hadoop MapReduce Change Log. Release0.22.1 – Unreleased. http://hadoop.apache.org/mapreduce/docs/r0.22.0/changes.h tml >, Accepted 02012012.
[11] Web Log Analysis for Security Compliance Using Big Data- Volume 5, Issue 3, March 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering
[12] M. D. Kunder. World wide web size - daily estimated size of the world wide web. http://www.worldwidewebsize.com/, 2011. Last Visit: 2011 November.
[13] H. Liu and V. Ke_selj. Combined mining of web serverlogs and web contents for classifying user navigation patterns and predicting users' future requests. Data Knowl. Eng., 61:304{330, May 2007.
[14] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. E_ective personalization based on association rule discovery from web usage data. In Proceedings of the 3rd international workshop on Web information and data management, WIDM '01, pages 9{15, New York, NY, USA, 2001.
[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.
[16] Miki Nakagawa and Bamshad Mobasher, (2003)"Impact of site characteristics on Recommendation Models Based on Association Rules and Sequential Patterns", Proceedings of the IJCAI'03 Workshop on Intelligent Techniques for Web Personalization, Acapulco, Mexico, August 2003.
[17] F. Khalil, J. Li, and H. Wang. A framework for combining markov model with association rules for predicting web page accesses. In Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006), pages 177–184,



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY