IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Web Server log Analysis for Unstructured data Using Apache Flume and Pig

A.S. Nagdive[1*], R.M. Tugnayat[2], G.B Regulwar[3], D.Petkar[4]

[1]Department of Information Technology, G.H.R.C.E Nagpur, India
[2]Shri Shankarprasad Agnihotri College of Engineering Wardha, India
[3] Department of Computer Sciences, B.N.C.O.E Pusad, India
[4]AUV Technology Nagpur, India

[*]Corresponding Author: ashlesha.nagdive@gmail.com, Tel.: 9403719799

DOI: https://doi.org/10.26438/ijcse/v7i3.220225 | Available online at: www.ijcseonline.org

Accepted: 23/Mar/2018, Published: 31/Mar/2019

Abstract—Web server normally produces log files. A weblog is a group of connected web pages that consists of a log or daily record of information, particular fields or views which is altered, every now and then, by owner of site, other websites or by website users. This is used to convert the unstructured data of web server log which will be coming from Apache flume into structured format using Pig. An enterprise weblog analysis system based on Hadoop architecture with Hadoop Distributed File System (HDFS), Hadoop MapReduce Software Framework and Pig Latin Language aids the business decision-making process of the system administrators and helps them to collect and identify the potential value which is hidden within such huge data generated by the websites. Such a weblog analysis includes the analysis of an Internet site's entry log as well as provides information about the amount of visitors, days of week and rush hours, views, hits, very often accessed pages, application server traffic trends, performance reports at varying intervals and statistical reports which indicate the performance of program. Web log file is a log file created and stored by a web server automatically. Analyzing such web server access logs files will provide us various insights about website usage. Due to high usage of web, the log files are growing at much faster rate with increase in size. Processing this fast growing log files using relational database technology has been a challenging task these days. Hadoop runs the big data where a massive quantity of information is processed via cluster of commodity hardware. In this paper we present the methodology used in pre-processing of high volume web log files, studying the statics of website and learning the user behaviour using the architecture of Hadoop MapReduce framework, Hadoop Distributed File System, and HiveQL query language Pig.

Keywords—HDFS, Apache Flume, Pig, Hbase, web log server.

I. INTRODUCTION

In today's world of Internet, analysis of log file is becoming essential for studying a client's actions for boosting the promotions as well as purchases. In order to analyze the log data, we obtain required knowledge from this log data with the help of Web mining. Log files are being produced extremely quickly with a speed ranging from about 1 to 10Mb/s for every device. In one day, one data centre is able to produce tens of terabytes of log data. The size of the sets of data is very large. We need a parallel processing system as well as a dependable data storage mechanism to perform the analysis of such huge datasets.

The Hadoop framework gives us dependable data storage with the help of Hadoop Distributed File System as well as Map Reduce calculating standard that acts as a parallel operating setup over huge sets of data. A Hadoop distributed file system splits the initial data as well as provides the initial data fragments across various computers across HDFS cluster. These machines across the hadoop cluster carry blocks of data which enables the processing log data in parallel and evaluates the result efficiently. The superior hadoop methodology is to "Save initially and query afterwards". Initially, Hadoop puts entire data over the Hadoop Distributed File System. After this is completed, Hadoop executes the queries which are in Pig Latin language which enables to lessen the time for reply and the load against the user setup.

Pig is modeled to blend well within explanatory methodology of SQL as well as the procedure-oriented map-reduce methodology associated with either the machine-code or an assembly language which is inflexible, resulting into a large amount of tailor-made consumer computer program that proves difficult for managing as well as reutilizing. After being completely executed, Pig performs the task of compiling Pig Latin in the form of concrete designs. These concrete designs get implemented across Hadoop.[1] Hadoop is a publicly accessible, implementation of map-reduce. An Apache-incubator project such as Pig is open-source and accessible for public usage. Pig drastically lessens the time which is needed for performing generation and implementation related to their data study activities, in contrast with the time taken when Hadoop is utilized alone. A new debugging environment is obtained when Hadoop is integrated with Pig which results in a very high efficiency leading to increased profitability.

II. Related Work

A**. Extracting Knowledge from Web Server Logs Using Web Usage Mining:** Use of Internet is increasing day by day. So websites usage growth is also increasing. To maintain the usage data different log files are used with different formats. Web usage mining was used to discover useful knowledge from web server logs. WUM (Web Usage Mining) worked only for single log file format which is W3C format. In this paper, they were use Web Usage Mining technique to extract knowledge from web server logs. The process of WUM was divided in four parts, which were: Data collection, Data Pre-processing, Pattern discovery, Pattern analysis. WUM worked on unstructured data which is very useful. [3]

B. **An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner:** This paper introduces an efficient web mining algorithm for web log analysis. The results obtained from the web log analysis may be applied to a class of problems; from search engines in order to identify the context on the basis of association to web site design of an ecommerce web portal that demands security. The algorithm is compared with its other earlier incarnation called Improved A priori All Algorithm.

E-web log miner provides better performance for time and space complexity compared to earlier techniques which was shown through performance analysis of proposed web mining algorithm. The proposed algorithm, Efficient Web Miner or E-Web Miner can be traced for its valid results and can be verified by computational comparative performance analysis. E-Web Miner reduced the number of database scanning and the candidate sets are found to be much smaller in stage wise comparison with improved. This is successful to be applied in any web log analysis, including information centric network design. As it was effective in analysis of web data, it used only single log file format and also used database transaction.

III. Methodology

Objective of Study:

- Data is growing at an enormous rate which must be used to gain insights. Apache hadoop is open source framework which can process huge amount of data(structured, semi-structured and unstructured).Data analysis (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. To provide Knowledge data from unstructured data set. To convert the unstructured data of web server log which will be coming from Apache flume into structured format using Hive and analyze the data by Pig.
-
- The goal of hadoop is to become the backend of everything. All the data in future will not be kept not on mere sql servers but on a hadoop cluster and retrieval will be done in real time.In order to analyze the log data, we obtain required knowledge from this log data with the help of Web mining. Log files are being produced extremely quickly with a speed ranging from about 1 to 10Mb/s for every device. All the records to be processed are dispersed across the different Application Servers. The Weblogs spread across various Application Servers undergo preliminary collation through Linux FTP protocol which is integrated into a log file. After the files finish cutting, the designated block size is uploaded into the HDFS specified directories through various commands. The data is not changed after the Weblog file is stored in HDFS. The system functions are divided into two categories such as batch analysis and interactive input conditions. System administrator develops a basic program structure which consists of a shell script frame:

(i) To receive different parameters according to the functional needs and
(ii) To call the Pig script which loads log and copies the analysis result that is to be restored across various Linux directories.
System administrator creates various dimensions analysis reports according to the results generated such as sales distribution, Application server flow and data, etc.

IV. STEPS TO PERFORM WEBLOG ANALYSIS USING HADOOP

The various steps to perform a Weblog Analysis using Hadoop are as follows:

A. Data Pre-processing

All the records to be processed are dispersed across the different Application Servers. The Weblogs spread across various Application Servers undergo preliminary collation through Linux FTP protocol which is integrated into a log file.

B. Upload

After the files finish cutting, the designated block size is uploaded into the HDFS specified directories through various commands.

C. Hadoop Processing

The data is not changed after the Weblog file is stored in HDFS. The system functions are divided into two categories such as batch analysis and interactive input conditions. System administrator develops a basic program structure which consists of a shell script frame:

(i) To receive different parameters according to the functional needs and

(ii) To call the Pig script which loads log and copies the analysis result that is to be restored across various Linux directories.

D. Analysis

System administrator creates various dimensions analysis reports according to the results generated such as sales distribution, Application server flow and data, etc.

V. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

HDFS provides a means for greater rate of processing over the data belonging to a program. Hadoop File System which was created by utilizing a DFS design is run over affordable and easy to obtain physical components of a computer. HDFS is different from other distributed systems due to its great fault-resilient nature and its design which utilizes cheap hardware. HDFS carries huge data as well as gives a simpler access mechanism. To save so much large amount of data, the files are saved throughout numerous computers.[2] The acquired files are saved in repeating style to save the system against probable failure of data. HDFS makes programs ready for managing in a parallel manner.

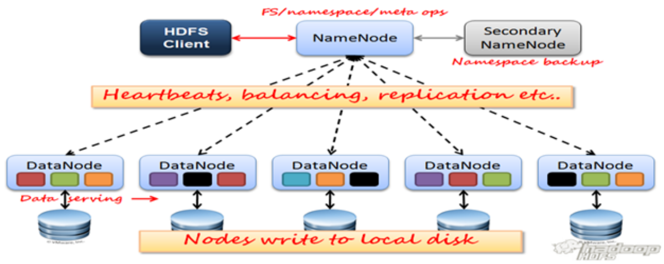


Figure 1.HDFS Architecture

Name Node: A Name Node carries metadata for HDFS. The Name Node saves and maintains information about the position of the files in HDFS. This information maintained by the Name Node is needed for fetching data spread within the hadoop cluster across many computers. The Name Node is software that can be run on computer hardware that is affordable and easy to obtain. The Name node system functions as the main server and performs jobs as mentioned:

a) Management of the file system namespace.

b) Regulation of the client's access to files.

c) Execution of the file or directory setup operations like rename, close, and open.

Data node: A Data node is a component of a computer that is affordable and easy to obtain that executes the duty of saving HDFS files. A Data node is present for each node across a hadoop cluster. The Data nodes handle the process of data-storage across HDFS. They handle the blocks which contain parts of the file at a node. They send the file as well as block information saved within a particular node and provide reply to the Name Node for the entire file system operations. Data nodes execute reading and writing transactions upon file setups, according to user demand. Data nodes carry out the activities like creating a block, deleting a block, as well as replicating a block as per Name Node command[3].

Block: The saving of client data is done inside HDFS files. The file across a file setup is split into a single or multiple fragments and/or saved over distinct Data nodes. Each of the file fragments is a Block. A Block is the minimal data quantity which is conducive for reading and writing by HDFS. By default, 64 MB happens to be the block size.

Apache Flume

Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and transferring large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application.

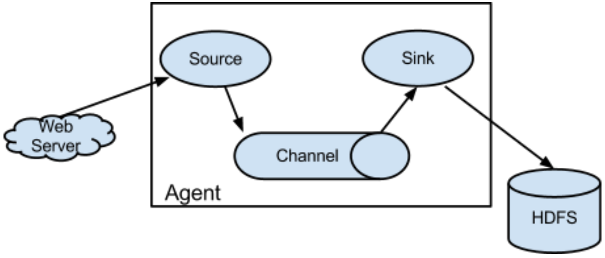


Figure 2. Flume Model

Apache Flume is a tool or service or data ingestion mechanism for collecting, aggregating and transporting large amounts of streaming data such as log files, various sources to a centralized data store. Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data (log data) from various web servers to HDFS.

Using Apache Flume we can store the data in to any of the centralized stores (HBase, HDFS). When the rate of incoming data exceeds the′ rate at which data can be written to the destination, Flume acts as a mediator between data producers and the centralized stores and provides a steady flow of data between them.[2] The transactions in Flume are channel based′ where two transactions (one sender and one receiver) are maintained for each message. It guarantees reliable message delivery. Flume is reliable, fault tolerant, scalable, manageable, and customizable.[3] Flume ingests log data from multiple web′ servers into a centralized store (HDFS, HBase) efficiently. Using Flume, we can get the data from′ multiple servers immediately into Hadoop. Along with the log files, Flume is also used′ to import huge volumes of event data produced by social networking sites like Facebook and Twitter, and e-commerce websites like Amazon and Flipkart. Flume supports a large set of sources and′ destinations types.

VI. IMPLEMENTATION

A.Virtual Machine is created in VMware software.

A virtual machine (VM) is an operating system (OS) or application environment which is installed on software, that imitates dedicated hardware. The user has the same experience on a virtual machine which they would have on dedicated hardware.

B. Firstly, web server is created which is use to create server logs.

C. Hadoop is installed which is used to store server logs.

D.Installation of FLUME tool is required to store unstructured server logs into HDFS.

A company has tons of services running on multiple servers. Also lots of data (logs) produce by them, that need to analyze together. In order process logs, we need a reliable, scalable, extensible and manageable distributed data collection service which can perform flow of unstructured data (logs) from one location to another where they will process (in HDFS). Apache flume is an open source data collection service for transferring the data from source to destination.

Apache Flume is the most reliable, distributed, and available service for systematically collecting, aggregating, and moving large amounts of streaming data (logs) into the Hadoop Distributed File System (HDFS). Based on streaming data flows, it has a simple and flexible architecture. It is highly fault-tolerant and robust and with best reliability mechanisms for fail-over and recovery. Flume allows data collection in batch as well as streaming mode.

E .Installation of Pig tool.

F. Pig tool is required to convert unstructured logs into structured one. Pig is better than Hive to process unstructured data. Therefore the data is first cleansed with Pig and then processed with Hive. This makes Pig a good fit for this use case as well, since it supports data with partial or unknown schemas, and semi-structured or unstructured data.

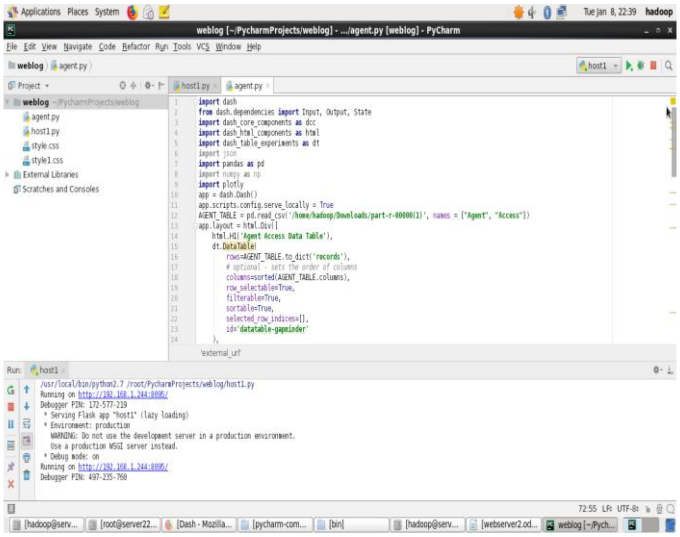
G .HBASE is installed to store structured data

H. Data is ready for the query.

I. The data is now ready to manipulate using Pig

J. Data is analyze using different commands in Pig shell.

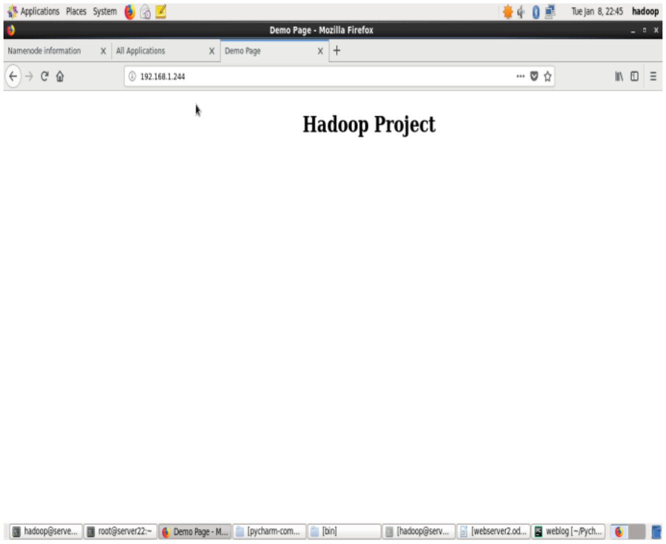
VII. RESULT AND CONCLUSION



```
import dash
from dash.dependencies import Input, Output, State
import dash_core_components as dcc
import dash_html_components as html
import dash_table_experiments as dt
import json
import pandas as pd
import numpy as np
import plotly
app = dash.Dash()
app.scripts.config.serve_locally = True
AGENT_TABLE = pd.read_csv('/home/hadoop/Downloads/part-r-00000(1)', names = ["Agent", "Access"])
app.layout = html.Div([
    html.H1('Agent Access Data Table'),
    dt.DataTable(
        rows=AGENT_TABLE.to_dict('records'),
        # optional - sets the order of columns
        columns=sorted(AGENT_TABLE.columns),
        row_selectable=True,
        filterable=True,
        sortable=True,
        selected_row_indices=[],
        id='datatable-gapminder'
    ),
```

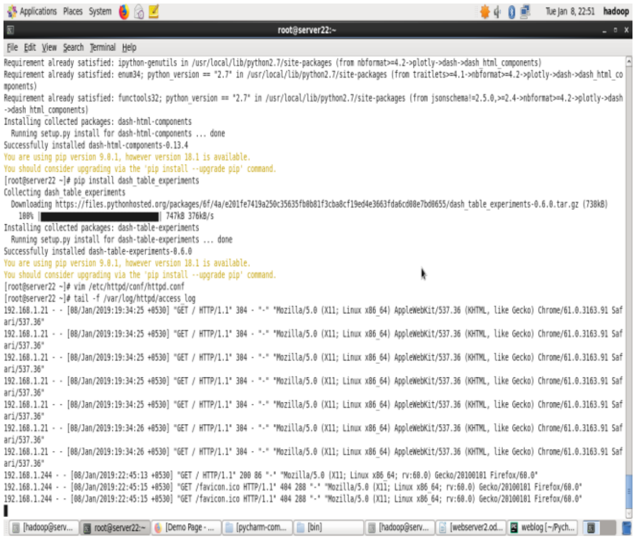
```
/usr/local/bin/python2.7 /root/PycharmProjects/weblog/host1.py
Running on http://192.168.1.244:8095/
Debugger PIN: 172-577-219
 * Serving Flask app "host1" (lazy loading)
 * Environment: production
   WARNING: Do not use the development server in a production environment.
   Use a production WSGI server instead.
 * Debug mode: on
Running on http://192.168.1.244:8095/
Debugger PIN: 497-235-768
```

Figure 3 : Flume Agent performance



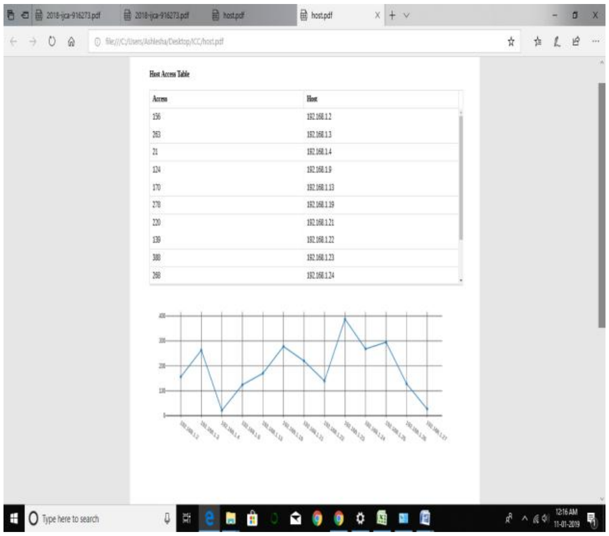
Hadoop Project

Figure 4. Web Page Model


Applications Places System
Tue Jan 8, 22:51 hadoop
root@server22:~
File Edit View Search Terminal Help
[hadoop@serv... root@server22:~ [Demo Page - ... [pycharm-com... [bin] [hadoop@serv... [webserver2.od... weblog [~/Pych...

```
Requirement already satisfied: ipython-genutils in /usr/local/lib/python2.7/site-packages (from nbformat>=4.2->plotly->dash->dash_html_components)
Requirement already satisfied: enum34; python_version == "2.7" in /usr/local/lib/python2.7/site-packages (from traitlets>=4.1->nbformat>=4.2->plotly->dash->dash_html_co
mponents)
Requirement already satisfied: functools32; python_version == "2.7" in /usr/local/lib/python2.7/site-packages (from jsonschema!=2.5.0,>=2.4->nbformat>=4.2->plotly->dash
->dash_html_components)
Installing collected packages: dash-html-components
  Running setup.py install for dash-html-components ... done
Successfully installed dash-html-components-0.13.4
You are using pip version 9.0.1, however version 18.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
[root@server22 ~]# pip install dash_table_experiments
Collecting dash_table_experiments
  Downloading https://files.pythonhosted.org/packages/6f/4a/e201fe7419a250c35635fb0b81f3cba8cf19ed4e3663fda6cd08e7bd0655/dash_table_experiments-0.6.0.tar.gz (738kB)
    100% |████████████████████████████████| 747kB 376kB/s
Installing collected packages: dash-table-experiments
  Running setup.py install for dash-table-experiments ... done
Successfully installed dash-table-experiments-0.6.0
You are using pip version 9.0.1, however version 18.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
[root@server22 ~]# vim /etc/httpd/conf/httpd.conf
[root@server22 ~]# tail -f /var/log/httpd/access_log
192.168.1.21 - - [08/Jan/2019:19:34:25 +0530] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.91 Saf
ari/537.36"
192.168.1.21 - - [08/Jan/2019:19:34:25 +0530] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.91 Saf
ari/537.36"
192.168.1.21 - - [08/Jan/2019:19:34:25 +0530] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.91 Saf
ari/537.36"
192.168.1.21 - - [08/Jan/2019:19:34:25 +0530] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.91 Saf
ari/537.36"
192.168.1.21 - - [08/Jan/2019:19:34:25 +0530] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.91 Saf
ari/537.36"
192.168.1.21 - - [08/Jan/2019:19:34:26 +0530] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.91 Saf
ari/537.36"
192.168.1.21 - - [08/Jan/2019:19:34:26 +0530] "GET / HTTP/1.1" 304 - "-" "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/61.0.3163.91 Saf
ari/537.36"
192.168.1.244 - - [08/Jan/2019:22:45:13 +0530] "GET / HTTP/1.1" 200 86 "-" "Mozilla/5.0 (X11; Linux x86_64; rv:60.0) Gecko/20100101 Firefox/60.0"
192.168.1.244 - - [08/Jan/2019:22:45:15 +0530] "GET /favicon.ico HTTP/1.1" 404 288 "-" "Mozilla/5.0 (X11; Linux x86_64; rv:60.0) Gecko/20100101 Firefox/60.0"
192.168.1.244 - - [08/Jan/2019:22:45:15 +0530] "GET /favicon.ico HTTP/1.1" 404 288 "-" "Mozilla/5.0 (X11; Linux x86_64; rv:60.0) Gecko/20100101 Firefox/60.0"
```

Figure 5.Raw Data



Host Access Table

Access	Host
156	192.168.1.2
263	192.168.1.3
21	192.168.1.4
124	192.168.1.9
170	192.168.1.13
278	192.168.1.19
220	192.168.1.21
139	192.168.1.22
388	192.168.1.23
268	192.168.1.24

Figure 6:Analysis Number of access and Host

Web sites are one of the important means for organizations for making advertisements. In order to get outlined results for a specific web site, we need to do log examination that helps enhance the business methodologies and also produce measurable reports. In this project with the help of Hadoop framework web server log files are analyzed. Data gets stored on multiple nodes in a cluster so the access time required is reduced. MapReduce works for large datasets giving efficient results. Using visualization tool for log analysis will give us graphical reports indicating hits for web pages, client‗s movement, in which part of the web site clients are interested. From these reports business groups can assess what parts of the site need to be enhanced, who are the potential clients, what are the regions from which the site is getting more hits, and so on. This will help organizations plan for future marketing activities. Log analysis can be done using many different techniques however what is important is response time. Hadoop MapReduce model provides parallel distributed processing and reliable data storage for huge volumes of web log files. Hadoop‗s ability of moving processing to data rather than moving data to processing helps enhance response.

References

[1] Babak Yadranjiaghdam, Nathan Pool, Nasseh Tabrizi, *"A Survey on Real-time Big Data Analytics: Applications and Tool"*, 2016 International Conference on Computational Science and Computational Intelligence

[2] P. Muthulakshmi1, S. Udhayapriya , *"A Survey on Big Data Issues and Challenges"*,International Journal of Computer Sciences and Engineering, Vol.-6, Issue-6, Jun 2018 E-ISSN: 2347-2693

[3] SayaleeNarkhede and TriptiBaraskar, *"HMR Log Analyzer: Analyze Web Application Logs over HadoopMapReduce*," International Journal of UbiComp (IJU) vol.4, No.3, July 2013.

[4] Mirghani. A. Eltahir ; Anour F. A. Dafa-Alla*," Extracting knowledge from web server logs using web usage minning",* Published in: 2013 International Conference On Computing, Electrical And Electronic Engineering (Icceee)

[5] https://en.wikipedia.org/wiki/Apache_Hadoop

[6] Dr.S.Suguna, M.Vithya,J.I.ChristyEunaicy, *"Big Data Analysis in E-commerce System Using HadoopMapReduce"*in 2016 IEEE.

[7] G.S.Katkar, A.D.Kasliwal, *"Use of Log Data for Predictive Analytics through Data Mining"*, Current Trends in Technology and Science, ISSN: 2279-0535. Volume: 3, Issue: 3(Apr-May 2014).

[8] Savitha K, Vijaya M S, *"Mining of web server logs in a distributed cluster using big data technologies*", International Journal of Advanced Computer Science and Applications, Vol.5, NO.1, 2014

[9] Mahendra Pratap Yadav ; Pankaj Kumar Keserwani ; Shefalika Ghosh Samaddar, "*An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner"* 2012 1st International Conference on Recent Advances in Information Technology (RAIT)

[10] Xianjun Ni, *"Design and Implementation of Web log Minning"* International conference of computer engineering and technology 2009

[11] Apache-Hadoop,http://www.hadoop.apache.org

Authors Profile

1.Prof.Ashlesha S. Nagdive pursed Bachelor of Engineering from Amravati university, in 2008 and Master of Engineering from GHRCE Nagpur University in year 2011.She is currently pursuing Ph.D. from Amravati university and currently working as Assistant Professor in Department of IT GHRCE since 2010. She is a member of IEEE. He has published many research papers in reputed international journals. Her main research work focuses on, Big Data, Hadoop, Data Analytics, Data Visualization, She has 8 years of teaching experience and 4 years of Research Experience.

Dr. R.M Tugnayat, Principal of Shri Shankarprasad Agnihotri college of Engineering Wardha. He has completed his PhD from Nagpur university. He has more than 20 years of teaching experience and Research Experience. He is a member of IEEE.He has publications in various International conferences and Journals. Subject of Expertice Software Engineering, BigData, Computer Networks.

Prof.Ganesh Regulwar is currently pursuing Ph.D. from Amravati university and currently working as Assistant Department of Computer Sciences, B.N.C.O.E Pusad, India. He is a member of IEEE. He has published many research papers in reputed international journals. Her main research work focuses on, Software Engineering, Software Testing. He has more than 15 years of teaching experience and 4 years of Research Experience.

Mr. Dany Petkar Currently Working as Technical Manager at AUV Technology from May 2016.1years 9 months experience in IBM India Pvt. Ltd. PUNE as Server Engineer for Linux 1 year and 1 months Experience in M Intergraph Systems Pvt.Ltd1years’ experience in Jetking Computer Hardware & Networking Institute as trainer for Networking, Linux, Windows Server 2008 &CCNA.6 months experience as System Engineer at Alankit Assignments Limited.1 year 2 months experience as Technical Trainer at Chester Information Achiever, Nagpur. Subject of Expertise is Big Data Hadoop.