A Work Paper on - Analysing Web Access Logs using Spark with Hadoop

Prachi Gupta[1] Prof. Ritesh Kumar Yadav[2]
[2]Guide & Professor
[1,2]Department of Information Technology
[1,2]SRKU, Bhopal, India

Abstract— Current software application often produce (or can be configured to produce) some auxiliary text files known as log files. Such files are used during various stages of software development, mainly for debugging and profiling purposes. Use of log files helps testing by making debugging easier. It allows to follow the logic of the program, at high level, without having to run it in debug mode. Nowadays, log files are commonly used also at customers installations for the purpose of permanent software monitoring and/or fine-tuning. Log files became a standard part of large application and are essential in operating systems, computer networks and distributed systems.

Keywords: Web Access Logs, Hadoop, Data mining

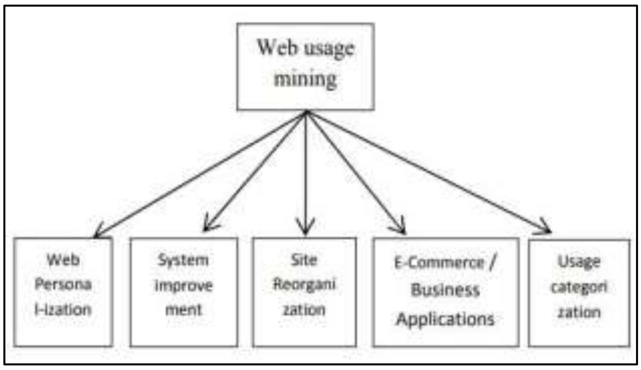
I. INTRODUCTION

Log files area unit usually terribly giant and may have complicated structure. though the method of generating log files is kind of easy and easy, log file analysis may well be an amazing task that needs huge procedure resources, lasting and complicated procedures. This usually results in a standard state of affairs, once log files area unit unceasingly generated and occupy valuable house on storage devices, however no one uses them and utilizes encircled info.

Log files area unit usually the sole manner the way to determine and find a slip-up in software package, as a result of log file analysis isn't tormented by any time-based problems referred to as probe result. this can be Associate in Nursing opposite to Associate in Nursing analysis of a running program, once the analytical method will interfere with time (critical or resource) crucial conditions inside the analyzed program.

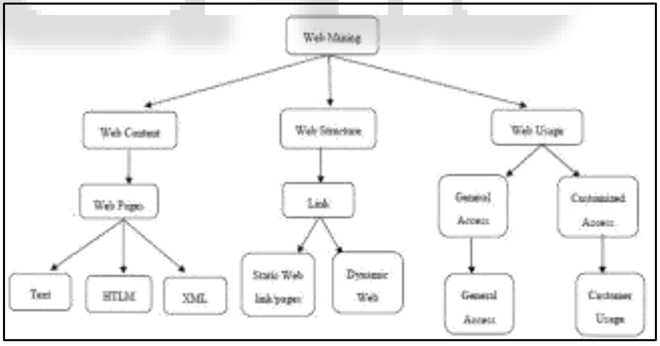
II. WEB PREDICTION

With the quick growth of knowledge on the planet Wide net, finding and retrieving helpful info becomes a really necessary issue. net search engines supply a preferred resolution to the current drawback. Typically, a research engine returns an inventory of websites per their matches to the question. very little info is provided regarding the structure and access frequency of specific computing device containing the online page. an online user might use the hierarchical website list for navigating the online and finding relevant pages. during this dissertation, we tend to propose ANother resolution to the current drawback supported an intelligent agent. rather than providing an inventory of websites, AN agent assists the user in navigating a selected computing device whereas finding out helpful info. The recommendations of the agent area unit supported results of mining journal information and perceptive user behavior. Conceptually, the complete net could also be understood as a graph, during which every website may be a node of the graph and every link is a grip of the graph connecting 2 websites.



III. DATA MINING

Data mining could be a step within the information Discovery in Databases (KDD) method consisting of applying information analysis and discovery algorithms that, at intervals acceptable process potency constraints, manufacture a selected enumeration of patterns over the information. data processing has been with success applied in science, health, marketing, and finance. net mining is that the application of information mining techniques to giant net data repositories. 3 major net mining strategies area unit online page mining, net structure mining and net usage mining. online page mining is that the application of information mining techniques to unstructured data residing in net documents.



IV. HADOOP

Hadoop is associate open supply, distributed computing framework developed and maintained by the Apache package Foundation written in java.

In hadoop developers will deploy programs written in the other languages or in java for the process of knowledge parallely across multiple artifact machines despite of the very fact that hadoop framework is written in java.

One of the key options of hadoop is that it partitions the computation and knowledge across multiple nodes so makes the applying computation run in parallel on these nodes. necessary options of hadoop square measure redundancy and responsibility which suggests that if any of nodes fails thanks to technical fault or alternative failures, it mechanically creates a backup for that node with none intervention of the operator.

A. MapReduce works

Hadoop divides the job into tasks[6]. There are two types of tasks:

1) Map tasks (Spilts & Mapping)
2) Reduce tasks (Shuffling, Reducing)

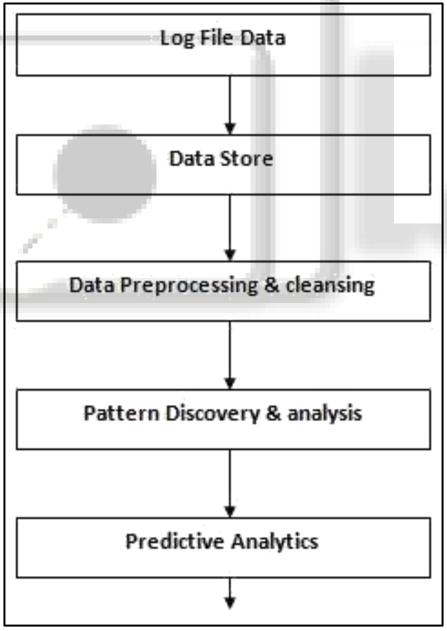
As mentioned above.The complete execution process (execution of Map and Reduce tasks, both) is controlled by two types of entities called a

1) Jobtracker : Acts like a master (responsible for complete execution of submitted job)
2) Multiple Task Trackers: Acts like slaves, each of them performing the job for every job submitted for execution in the system, there is one Jobtracker that resides on Namenode and there are multiple tasktrackers which reside on Datanode.

V. DATA PREPROCESSING

Preprocessing steps are performed prior to classification. They are as follows:

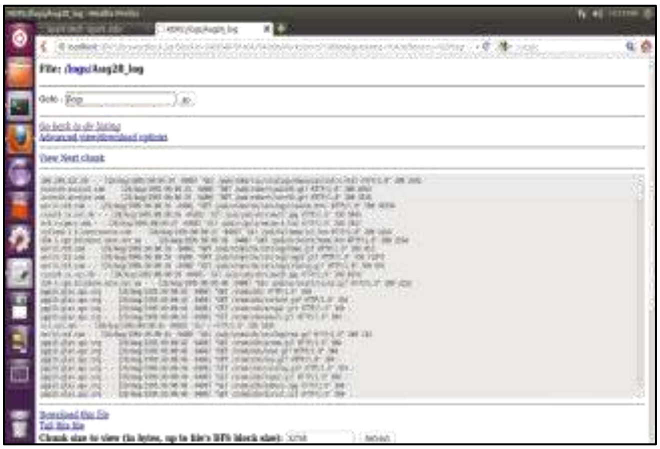
1) Collecting web access logs files from different resources.
2) Convert unstructured data into structured format.
3) Remove unwanted data from the data.
4) Store the preprocessed data for analysis.



Analysis Steps

VI. LOADING DATA INTO HDFS

First we can loading different access lof files in to HDFS, in our dissertation we can analyze web access log which are common access log. Figure 5.1 shows the loading a log file into HDFS. And in this figures we can clearly seen that there is not any structure between the data of these logs file. After loading these different logs file into HDFS we can analyze using spark.



File: /logs/Aug28_log

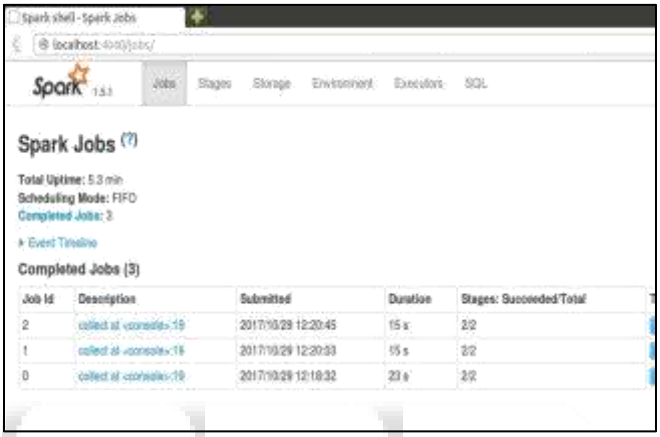
Goto: /logs go

Go back to dir listing
Advanced view/download options

View Next chunk

```
[illegible]
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh



Spark Jobs (?)

Total Uptime: 5.3 min
Scheduling Mode: FIFO
Completed Jobs: 3

▸ Event Timeline

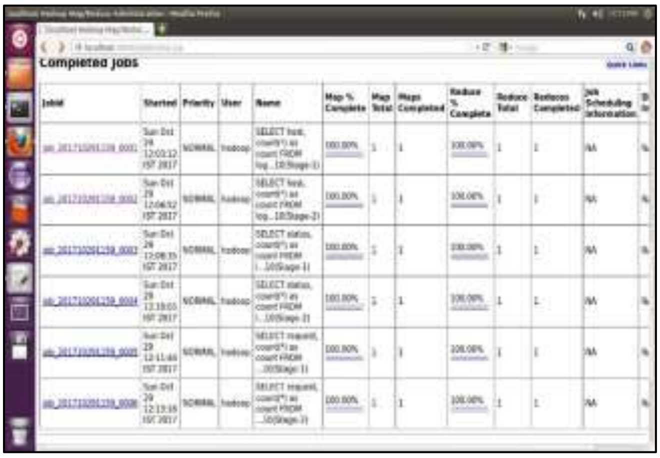
Completed Jobs (3)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total
2	collect at <console>:19	2017/10/29 12:20:45	15 s	2/2
1	collect at <console>:19	2017/10/29 12:20:33	15 s	2/2
0	collect at <console>:19	2017/10/29 12:18:32	23 s	2/2

Time taken by spark framework

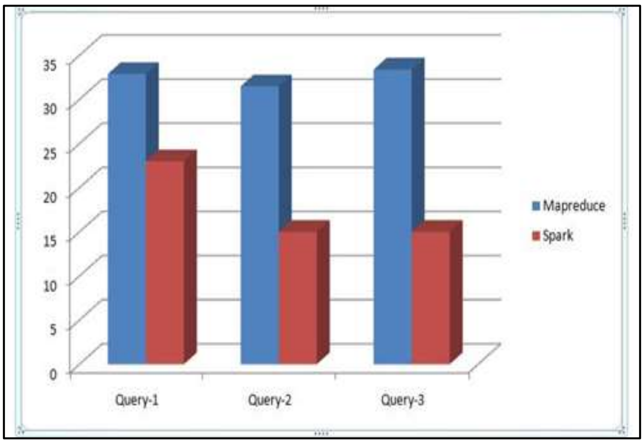
VII. ANALYZING USING MAPREDUCE

We can analyze the complicated log knowledge mistreatment mapreduce framework. And for sql like interface we are able to integrate apache hive on prime of the mapreduce, we are able to write sql question on hive, once storing the log information into HDFS , currently we are able to begin analyzing these complicated log files mistreatment apache hive.



Completed Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information
job_201710291239_0001	Sun Oct 29 12:03:12 IST 2017	NORMAL	hadoop	SELECT host, count(*) as count FROM log...10(Stage-1)	100.00%	1	1	100.00%	1	1	NA
job_201710291239_0002	Sun Oct 29 12:04:32 IST 2017	NORMAL	hadoop	SELECT host, count(*) as count FROM log...10(Stage-2)	100.00%	1	1	100.00%	1	1	NA
job_201710291239_0003	Sun Oct 29 12:08:35 IST 2017	NORMAL	hadoop	SELECT status, count(*) as count FROM l...10(Stage-1)	100.00%	1	1	100.00%	1	1	NA
job_201710291239_0004	Sun Oct 29 12:10:03 IST 2017	NORMAL	hadoop	SELECT status, count(*) as count FROM l...10(Stage-2)	100.00%	1	1	100.00%	1	1	NA
job_201710291239_0005	Sun Oct 29 12:11:46 IST 2017	NORMAL	hadoop	SELECT request, count(*) as count FROM ...10(Stage-1)	100.00%	1	1	100.00%	1	1	NA
job_201710291239_0006	Sun Oct 29 12:13:16 IST 2017	NORMAL	hadoop	SELECT request, count(*) as count FROM ...10(Stage-2)	100.00%	1	1	100.00%	1	1	NA



Execution time taken by spark & mapreduce

VIII. Conclusion

Web usage mining is a process for finding a user navigation patterns in web server access logs. These navigation patterns are further analyze by various data minig techniques. The discovered navigation patterns can be further used for several things like identifying the frequent patterns of the user, predicting the future request of user, etc. and in the recent years there are huge growth in electronic commerce websites like flipkart, amazon, etc. with an huge amount of online shopping websites, it is necessary to notice that how many users are actually reaching to the websites. When user's access any online website, web access logs are generated on the server. Web access logs data helps us to analyze user behavior that contain information like ip address, user name, url, timestamp, bytes transferred. It is very meaningful to analyze the web access logs which helps us in knowing the emergency trends on electronic commerce.

References

[1] Daniel E. Olivares, Claudio A. Cañizares et al. ''A Centralized Optimal Energy Management System for Microgrids ''IEEE PES General Meeting ,July 2011.

[2] He Cai, Guoqiang Hu et al., ''The adaptive distributed observer approach to the cooperative output regulation of linear multi-agent systems'',ELSEVIER Automatica, Vol. 75,Pp. 299–305, August 2017.

[3] S. Anand, B. G. Fernandes, and J. M. Guerrero, "Distributed control to ensure proportional load sharing and improve voltage regulation in lowvoltage DC microgrids," IEEE Trans. Power Electron., vol. 28, no. 4, pp. 1900–1913, Apr. 2013.

[4] Mehrdad Yazdanian, and Ali Mehrizi-Sani ''Distributed Control Techniques in Microgrids'' IEEE Transactions On Smart Grid,vol.5,no.6,pp.2901-2909,nov 2014.

[5] Rodrigo A F. Ferreira1,2, Henrique AC. ''Analysis of Voltage Droop Control Method for dc Microgrids'' IEEE International Conference on Industry Applications ,pp.1-6,nov2012.

[6] S. Grillo, M. Marinelli, S. Massucco, And F. Silvestro, "Optimal Control Strategy Of A Battery-based Storage System To Improve Renewable Energy Integration In Distribution Networks," IEEE Trans. Smart Grid, vol. 3, No. 2, Pp. 950–958, Jun. 2012.

[7] Xiong Liu, Peng Wang ''A Hybrid AC/DC Microgrid and Its Coordination Control'' IEEE Transactions On Smart Grid, Vol.2,No.2,pp.278-286June 2011.

[8] J. Vasquez, J. M. Guerrero, M. Savaghebi, J. Eloy-Garcia, and R. Teodorescu, "Modeling, analysis, and design of stationary reference frame droop controlled parallel three-phase voltage source inverters," IEEE Trans. Ind. Electron., vol. 60, no. 4, pp. 1271–1280, Apr. 2013.

[9] K. Jaehong, J. M. Guerrero, P. Rodriguez, R. Teodorescu, and N. Kwanghee, "Mode adaptive droop control with virtual output impedances for an inverter-based flexible AC microgrid," IEEE Trans. Power Electron, vol. 26, no. 3, pp. 689–701, Mar. 2011.

[10] Changhee Cho, Member, Jin-Hong Jeon, ''Active Synchronizing Control of a Microgrid'' IEEE Transactions On Power Electronics, vol.26, no.12, pp.3707-3719, dec 2011.

[11] Zaheeruddin , Munish Manas et al. ''Renewable energy management through microgrid central controller design: An approach to integrate solar, wind and biomass with battery'' ELSEVIER Energy Reports Vol. 1, Pp 156-163 November 2015.