A Survey 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— 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.
Keywords: HDFS, Hadoop, Web Access Logs

I. INTRODUCTION

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.

II. LITERATURE REVIEW

The projected system as delineate in figure1 consists of three stagesare involving manus one is log preprocessing, second is analysis and prophetic perfecting is last stage of the proposed design. take into account the sample logs from direction log come in table. Each entry contains object id, name, ingredients, URL for this page, image, time spent, date, source, direction yield, date revealed, cooktime, prep time, description, standing code and scientific discipline address for each every} entry.

III. PREPROCESSING PHASE

This part is a very important to get rid of unwanted log entries form input log files. victimization internet logs we will predict user‘s next request while not distributing them. however not all details in web logs area unit acceptable for the aim of mining navigation patterns. therefore log desires improvement before it may be used for prediction. take into account the improvement algorithmic program (CLE_ ALG). The main purpose of log preprocessing is to cut back amount of data set from original amount and reduce the prediction process time.

In this part cleanedlogs area unit processed to come up with logs with counts supported direction id and direction preparation time victimization Hadoop and Mapreduce. Log files area unit collected from many different varieties of server area unit fetched via Apache flume and loaded into a Hadoop cluster. Jobs area unit regular to analyze the logs and generate collective outline metrics and mental image victimization business intelligent tools.

MapReduce processes these blocks in a very parallel manner. this surroundings, computer file is split into four file and every file is keep in numerous nodes (like node1, node 2, node 3 and node 4). constant file are keep in numerous nodes. Here failure of any node ne'er ends up in information lose. information may be shared from the other node. The practicality of plotter is input downside go smaller sub issues and distributes these to employee nodes. Reducer step master node takes the answer to sub issues combines them into original downside.

MapReduce algorithmic step for numeration the cook time frequency from direction log files is shown below (AHMR). The input to the current operate may be a direction log file. for every cook time within the direction web site, a line are additional into the direction Log file. within the plotter operate, every block of the direction log file is given as associate degree input to a map operate that successively dissect each line victimization regular expression and emits the direction Item as a key at the side of the worth one.

IV. SIMULATION ENVIRONMENT

For analyzing a log data we can configure apache hive for analyze the log data comes from web.
Mining and Analysis Requirement
We have following requirements for mining and analysis are:
1) Apache Hadoop
2) Apache Spark
3) Apache Hive

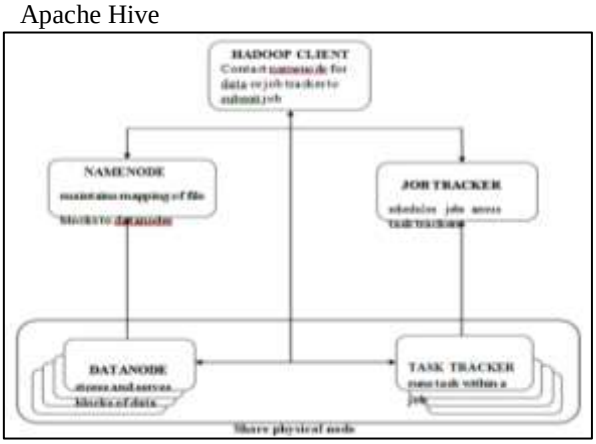

Fig. 1:

V. COMMODITY HARDWARE

Hadoop doesn't need big-ticket, extremely reliable hardware to run on. It's designed to run on clusters of trade goods hardware (commonly out there hardware out there from multiple vendors) that the possibility of node failure across the cluster is high, a minimum of for giant clusters. HDFS is meant to hold on operating while not a plain interruption to the user within the face of such failureMap tasks (Spilts & Mapping)

A. 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.

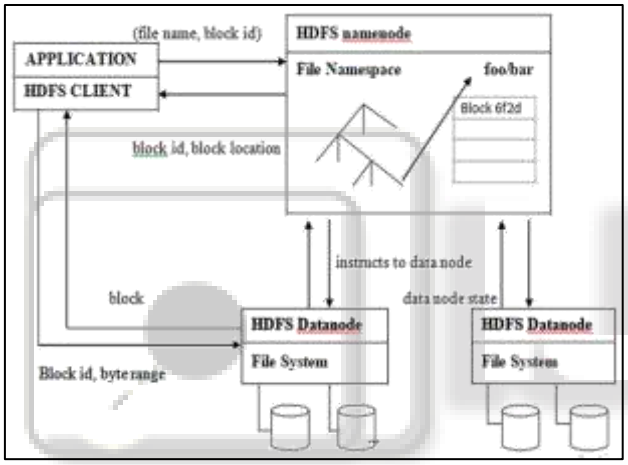

Fig. 2: Architecture of HDFS

VI. ARCHITECTURE OF HIVE

The following component diagram depicts the architecture of Hive:

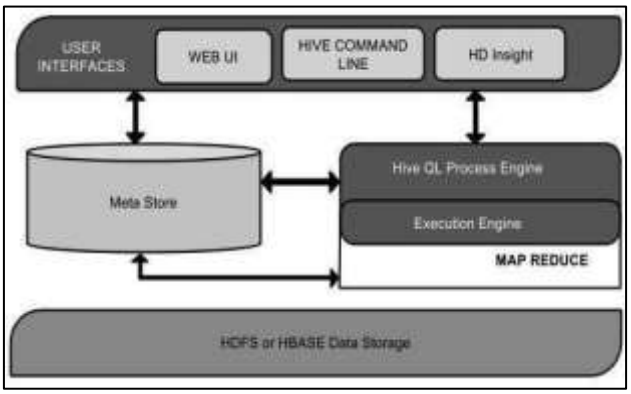

Fig. 3:

VII. PROBLEM DESCRIPTION

One of the challenges faces most frequently by those people within the field of usability is finding sensible information regarding user behavior quickly, accurately, and, in most cases, cheaply. In Associate in Nursing surroundings wherever several stakeholders question the come on investment in usability, some within the business have developed fascinating ideas geared toward gathering user information. One such plan is that the analysis of server log files to collect data regarding user behavior. On the surface, it's simple to know the gravitation towards server logs: They're purportedly an information supply that portrays what folks do on a web site. Server logs purportedly show what folks click on, that pages they read, and the way they get from page to page.

VIII. PROPOSED WORK

For storing these large and complex data we need a powerful tool [10], we introduces apache hadoop which is a open source framework for storing large datasets. And for analyzing these large datasets we uses apache spark framework and also analyze the same with mapreduce framework.

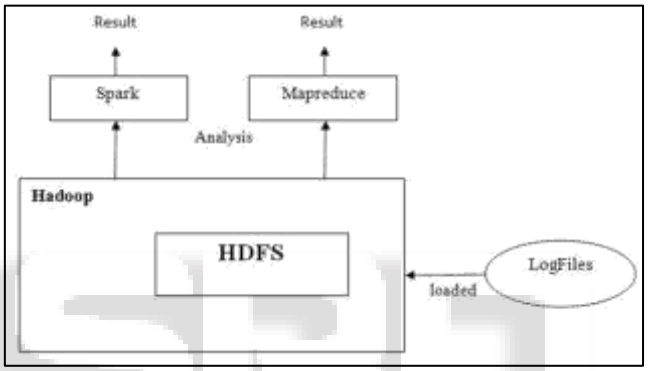

Fig. 4: Proposed system workflow

IX. CONCLUSION

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. These ecommerce websites generates petabytes of log data every day which is not possible by traditional tools and techniques to store and analyze such log data. In these dissertation we proposed an hadoop framework which is very reliable for storing such huge amount of data in to HDFS and then we can analyze the unstructured logs data using apache spark framework to find user behaviour. And in these paper we can also analyze the log data using mapreduce framework and finally we can compare the performance on spark and mapreduce framework on analyzing the log data.

REFERENCES:

[1] Dr.S.Suguna, M.Vithya, J.I.Christy Eunaicy, "Big Data Analysis in E-commerce System Using HadoopMapReduce" in 2016 IEEE.
[2] Mrunal Sogodekar, Shikha Pandey, Isha Tupkari, Amit Manekar, "BIG DATA ANALYTICS: HADOOP AND TOOLS" in 2016 IEEE Bombay Section Symposium (IBSS).

[3] Mohammed Hamed Ahmed Elhiber and Ajith Abraham, "Access Patterns in Web Log Data: A Review" in Journal of Network and Innovative Computing ISSN 2160-2174, Volume 1 (2013) pp. 348-355.
[4] Gerald Stermsek, Mark Strembeck, Gustaf Neumann, "A User Profile Derivation Approach based on Log-File Analysis" in Institute of Information Systems, New Media Lab Vienna University of Economics and BA, Austria.
[5] Bijesh Dhyani, Anurag Barthwal, "Big Data Analytics using Hadoop" in International Journal of Computer Applications (0975 – 8887) Volume 108 – No 12, December 2014.
[6] M. Dhavapriya, N. Yasodha , "Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table" in International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 1, Jan - Feb 2016.
[7] Sandeep Kumar Dewangan*, Shikha Pandey† and Toran Verma, "A Distributed Framework for Event Log Analysis using MapReduce" in 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)
[8] McKinsey, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey & Company, 2011, http: //www.mckinsey.com/.

