International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 4
ISSN: 2454-4248
689 – 693

Generate Analytics from a Product based Company Web Log

B.Tirupathi.Kumar[1], P.Mounitha[2], Sravanthi Nanumasa[3],C.Sudha[4]
[1]Assistant Professor/Dept of IT,Vardhaman college of Engineering,Hyderabad
[2,3,4] Assistant professor/Dept of CSE,Mahatma Gandhi Institute of Technology,Hyderabad
tirupathi.kumar@gmail.com
mounitha.p@gmail.com, sravanthi.nanumasa292@gmail.com, csudhahyd@gmail.com

ABSTRACT: The next generation of industries will be using Big Data to remedy the unsolved data difficulties within the physical global. Big Data analysis may be about constructing systems around the data that is generated. Every department of an organisation consisting of advertising and marketing, finance and HR are actually getting direct get admission to to their own statistics. This is developing a huge activity opportunity and there may be an pressing requirement for the experts to master Big Data Hadoop abilities. Nowadays most of the groups have became to Ecommerce which has grow to be a vital element for business approach and a catalyst for economic improvement. These groups need to predict the evaluation approximately their services and products to tune their commercial enterprise from the customers end. The response from the customers based totally on their sports on the web sites makes a decision the future modifications required to enhance the commercial enterprise values. These companies stores the statistics of all clients in element for destiny analysis which is commonly referred as large statistics, as it's far developing at high costs every day. One of the main programs of large statistics intelligence is Clickstream data which is ideal for e-commerce websites and websites that rely upon clicks. Clickstreams are records of consumer interactions with web sites and other packages. A common technique to load those facts and processing is through the use of traditional databases, however it involves many complexities even as appearing different operations. Here in this paper clickstream records is processed, analysed with the structure of Hadoop the usage of Hortonworks Data Platform (HDP) which offers massive scale processing overall performance and visualized thru strength.

Keywords – *Big data,Hadoop,HDP,Clickstream*

******

I. INTRODUCTION

Web mining is the utility of facts mining strategies to extract useful information from internet information that includes internet file, hyperlink between files, usage logs of internet websites and so on. Web utilization mining is the technique of making use of statistics mining strategies to discover usage sample from the internet statistics. It is one of the techniques to find personalization of web pages. The collection of net usage facts accumulated from specific tiers consisting of server stage, customer degree and proxy degree and additionally from special resources via the net browser and web server interaction the use of the HTTP protocol [1]. But in the contemporary state of affairs the wide variety of on-line client‘s will increase daily and each click on from an internet page creates at the order hundred bytes records in usual internet site log report. When an internet consumer submits request to web server on the equal time person activities are recorded in server aspect. These forms of net get admission to logs are referred to as log report. Request statistics despatched with the aid of the user thru protocol to the internet server that's recorded in log record.The logfiles [2]are contains some entries like ip address of which computer making the request, the visitor data, line of hit, the request method, location and name of the requested file, the HTTP status code, the size of the requested file. Log files can be classified into categories depending on the location of their storage that is web server logs and application server logs. A web server [3] maintains two types of log files: Access log and Error log.The access log records all requests that were made of this server. The error log records all request that failed and the reason for the failure as recorded by the application. A log files have lot of parameters which are very useful to recognizing user browsing patterns [4, 5 6]. Mining the web log file will helpful to server and E-commerce for predicting the behavior of their online customer.

Every day growing online clients as well as increasing the dimensions of net get admission to log [7] .In huge websites coping with tens of millions of simultaneous visitors can generate hundred of peta bytes of logs in line with day. The present information mining techniques shop internet log files in conventional DBMS and examine. RDBMS device can't shop and control the peta bytes of heterogeneous dataset. So, to research such big web log record effectively and efficaciously we need to develop quicker, green and powerful parallel and scalable information mining set of rules. Also want a cluster of storage devices to keep peta bytes of internet log records and parallel computing model for reading huge quantity of statistics. Hadoop framework offers dependable clusters of storage facility to keep our huge web log report statistics in a distributed way and parallel processing capabilities to process a large net log document information efficiently and efficaciously[8,9]. The preprocessed internet logs by using HadoopMapReduce environment is similarly processed for prediction of person subsequent request without

 ISSN: 2454-4248
689 – 693

demanding them to increase the hobby and to reduce the reaction time with ecommerce gadget.

HDFS (Hadoop Distributed File System)

HDFS is a prime issue of Hadoop and a method to shop the information in allotted way with the intention to compute speedy. HDFS saves records in a block of sixty four MB(default) or 128 MB in size which is logical splitting of data in a Data node (physical storage of statistics) in Hadoop cluster(formation of several Data node that is a group commodity hardware related through single network). All records about records splits in records node called metadata is captured in Name node which is once more a part of HDFS.

Hive

Many programmers and analyst are greater at ease with Structured Query Language than Java or every other programming language for which Hive is created by means of Facebook and later donated to Apache foundation. Hive specially deals with established data which is stored in HDFS with a Query Language much like SQL and known as HQL (Hive Query Language). Hive additionally run Map reduce program in a backend to technique statistics in HDFS but right here programmer has not worry about that backend MapReduce process it will look similar to SQL and result might be displayed on console.

Pig

Similar to HIVE, PIG additionally deals with established facts the use of PIG LATIN language. PIG turned into originally developed at Yahoo to reply similar want to HIVE. It is an alternative supplied to programmer who loves scripting and don't want to apply Java/Python or SQL to procedure information. A Pig Latin software is made from a sequence of operations, or changes, which might be applied to the input statistics which runs MapReduce program in backend to produce output.

Sqoop

SQL to Hadoop and Hadoop to SQL. Sqoop is a tool designed to switch facts between Hadoop and relational database servers. It is used to import information from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop record system to relational databases. It is provided by the Apache Software Foundation.

II. LITERATURE SURVEY

History of Weblog analytics

With the digitization of the world, producing expertise from the raw information aggregated by using internet servers has emerge as more and more essential, specifically in a industrial context. Capabilities from optimizing the consumer experience to information person‘s behavior to providing customized reports are everyday. The first step in gaining insights into consumer‘s conduct on a website is instrumenting the net pages. They can be instrumented the usage of Page Tagging (snippet of JavaScript code that internet packages add to every page) to capture consumer interactions on the web page – and the deeper the instrumentation, the extra the statistics captured, accordingly producing large amounts of statistics. This large amount of weblog statistics provides challenges, consisting of delivery across more than one datacenters and processing at scale to extract insights. The internet analytics system is complex, concerning studying weblogs for information along with URLs accessed, cookies, demographics, places and date/time. This facts is used to analyze internet site site visitors, their utilization, in addition to surfing styles and behavior. To understand the whole image of what is going on at the internet site, the gadget should aid amassing and processing data in actual-time. With a 2nd-by way of-2nd view of vacationer engagement statistics, you'll be able to react right away to visitor developments. However, no longer each insight can be amassed in real-time – like computing bounce remember ratio, unique customers, and so on. So, the system additionally desires to support generating insights in batch. Furthermore, as log facts is amassed and made available for analysis, it's also vital to guide advert-hoc queries through Apache Hive, Impala and so forth, for exploratory analysis to reply questions that arise that one might not recognize beforehand. Following is a high-stage view of 1 manner to architect a device for generating real-time, batch insights and also helping advert-hoc evaluation. The internet site or internet application is instrumented to seize one-of-a-kind consumer interactions on the web page. The instrumentation is then logged to internet server logs. Transport systems like Apache Flume or Apache Kafka are configured to extract logs from log files in actual-time and transport them to a centralized Hadoop cluster. To assist Batch processing to generate insights, the facts is batched and written to HDFS. For real-time processing, the information is immediately fed into a processing device like Apache Storm, Spark, and so on.

Workflows the use of MapReduce/Apache Pig/Apache Spark are created to cleanse log records and generate insights periodically. The output facts produced is then written again to HDFS. These scheduled scripts certainly analyze the logs on numerous dimensions and extract the outcomes. The results are by using default stored onto HDFS, but we can also use garage implementation for other repositories additionally inclusive of Apache HBase, MongoDB, etc. Real-time aggregates are saved in HBase. The insights in HBase also are available for creating realtime dashboards. Hive is set up to expose the uncooked internet log and output of facts analysis to be accessed using SQL. Schemas for web log and insights need to be modeled and maintained. Reporting tools can then access the outcomes or do exploratory evaluation on net log records the usage of widely to be had tools.

Web Log Cleaning for Mining of Web Usage Patterns

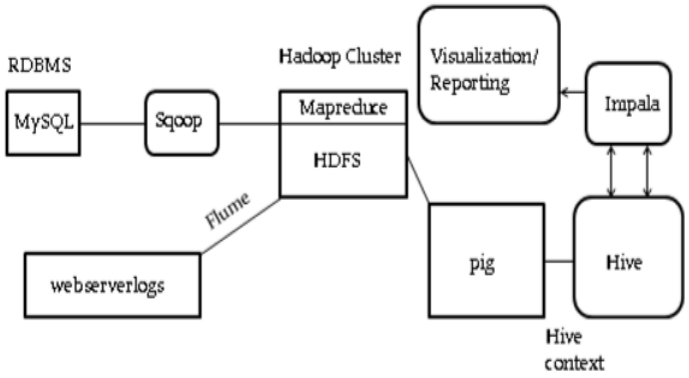
In [10] proposed a new technique for preprocessing of net log statistics and the affiliation regulations are being hired to extract the useful styles. Log files are the high-quality supply to predict the consumer conduct, to analyze usage pattern via those phases together with pattern discovery and

 ISSN: 2454-4248
689 – 693

analysis section. In [11] proposed data mining techniques like the first phase of preprocessing and to find out consumer get admission to patterns from web log. They discussed area extraction and statistics cleaning algorithms and proved web log mining can be used for numerous programs consisting of net personalization, web page advice, web page development and so on. In [12] analyzed a few important factors like statistics exploration, pastime and possibilities of customers.
In[13]discussed to discover the frequent usage through the client and their experimental have a look at unearths a few interesting patterns thru affiliation rule mining set of rules and FP growth algorithm. They proved affiliation rule mining have some difficulty and suitable for least amount of data set however FP boom have minimum problem and suitable for huge information set with out any user interaction. In [14] mentioned the significance of records preprocessing strategies and user consultation identity techniques for any transaction documents. In [15] proposed some filtering methods primarily based on statistical attributes to discover rules or patterns. They proved tools do no longer imply which internet usage mining algorithms are used but provide powerful graphical visualizations of the effects. In paper [16] implemented k-way clustering algorithm first then implemented association rule mining method on clustered information for pattern discovery. They find out drawbacks, generation of beside the point policies. In [17]proposed method to discover same user consultation. They were grouped the same information primarily based on two similarities including user similarity and session similarity and can beneficial for group the identical internet customers. In [18] net log analyzer tool is used for studying usage sample from net logs. They proved the ones gear are useful to web administrator a good way to improve the net website online overall performance thru the upgrades of contents, shape, presentation and transport. The paper shape is as follows. In section3 proposed architecture is discussed. Section4 suggests experimental outcomes evaluation and segment four concludes the paper.

III. ARCHITECTURE

The data is collected from the web server using flume and data from RDBMS is collected by using Sqoop. The entire collected data is stored in the Hadoop Storage called HDFS i.e., Hadoop Distributed File System. This stored data is processed using map reduce followed by PIG. This is again processed through HIVE and the output of the HIVE is stored in SQL format. This SQL format data is converted into reporting or visualization using Impala.



Architectural Diagram of Hadoop

IV. SYSTEM IMPLEMENTATION

Installing VMware Workstation:

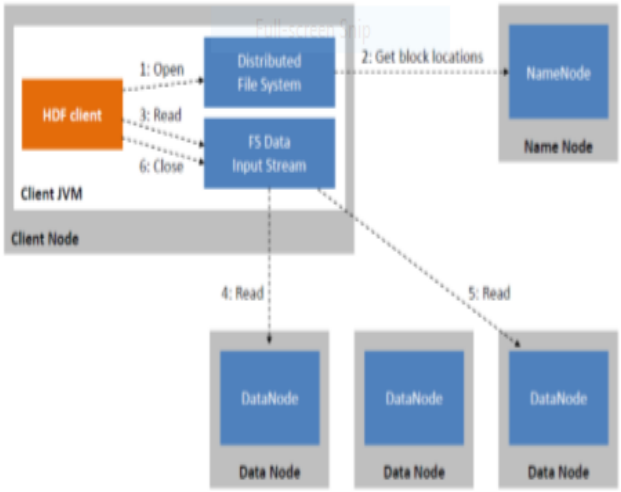
VMware Workstation is a program that allows you to run a virtual computer within your physical computer. The virtual computer runs as if it was its own machine[20]. A virtual machine is great for trying out new operating systems such as Linux, visiting websites you don't trust, creating a computing environment specifically for children, testing the effects of computer viruses, and much more. You can even print and plug in USB drives. Read this guide to get the most out of VMware Workstation.

HDFS

The Hadoop Distributed File System (HDFS) is designed to store very large datasets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks[19].

HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

 ISSN: 2454-4248
689 – 693

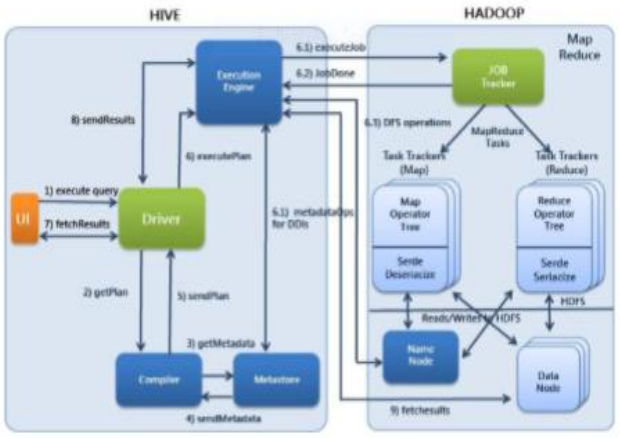


HDFS Workflow

HDF client is opened from Distributed File System and gets the block locations from NameNode. Now the data is read from FS data Input Stream which is read from different DataNode and the client node is closed.

HIVE

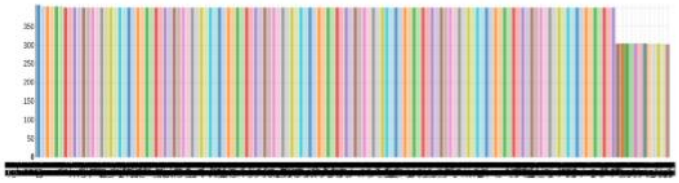
Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.



Query flow in hive

The process in which the flow of the queries is controlled in the hive.

V. EXPERIMENTAL RESULTS



Bar chart from the weblogs when status is not equal to 200

In the above figure bar chart from the weblogs when it is unsuccessful.

Bar charts from weblogs for the query

In the above figure it shows bar charts from weblogs for the query

VI. CONCLUSION AND FUTURE WORK

Analyzing weblogs could be very essential to many agencies, but it is a difficult venture that requires special equipment due to large volumes of statistics and the complexity of analysis. To advantage maximum from the answer, you need with a purpose to manner information in actual-time, batch and execute advert-hoc queries and greater. The traditional weblogs analysis solution on Hadoop can be quite complicated and require making distinctive technologies to paintings together, while an integrated platform along with CDAP can help with most tough elements. In this paper have proposed blog analysis prediction the usage of massive data surroundings. Map lessen is framework for extraordinarily parallel and distributed systems across huge dataset .By the use of map reduce with Hadoop allows in disposing of scalability bottleneck. This type of era used to analyze large records units has ability to wonderful enhancement to blog evaluation. Hence we can estimate the rating a of the web sites, its request and status actual date of request and views based totally at the facts obtained from the website. This helps for the humans to enhance the sites based at the suggestions and critiques with the aid of the users. Analyzing massive statistics has been at the tip of many a technologist's tongue for the past couple of years. This evaluation is described as the future for businesses trying to gain insights into commercial enterprise operations and discover crime styles.

 ISSN: 2454-4248
689 – 693

VII.REFERRENCES

[1]. M.Santhanakumar and C.Christopher Columbus, ―Web Usage Analysis of Web pages UsingRapidminer‖, WSEAS Transactions on computers, EISSN: 2224-2872, vol.3, May 2015.

[2]. ShailyG.Langhnoja ,MehulP.Barot and DarshakB.Mehta, ―Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery ―,International Journal of Data Mining Techniques and Applications, vol.2 ,Issue.1, June 2013.

[3]. Web server logs ://http. Sever side log.org.

[4]. Nanhay Singh, Achin Jain, Ram and Shringar Raw, ―Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques‖, International Journal of Data Mining & Knowledge Process(IJDKP) vol.3, Issue.4, July 2013.

[5]. J.Srivastava et al, ―Web usage Mining: Discoveryand Applications of usage patterns from Web Data―, ACM SIGKDD Explorations, vol.1, Issue. 2, pp.12-23, 2000.

[6]. S.Saravanan and B.UmaMaheswari, ―Analyzing Large Web Log Files in A HadoopDistributedCluster Environment‖, International Journal of Computer Technology & Applications, vol.5, pp. 1677-1681.

[7]. K.V.Shvachko, ― TheHadoop Distributed File System Requirements‖, MSST ‘10 Proceeding of the 2010 IEEE 26th Symposium on Mass Storage System and Technologies(MSST).

[8]. Apache Hadoop ://http://hadoop.apache.org.

[9]. A white paper by OrzotaInc, ―Beyond Web Application Log Analysis using Apache Hadoop‖.

[10]. Resul Das, Ibrahim Turkoglu, ―Extraction of Interesting Patterns through Association Rule Mining for Improvement of website Usability‖ International of Electrical & Electronics Engineering, vol 9, issue 2, 2010.

[11]. Amit Pratap Singh, Jain Dr.R.C., ―A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation‖ International Journal of Emerging Trends & Technology in Computer Science, vol 3, issue 3 ,May 20143.

[12]. The int Aye, ―Web Log Cleaning for Mining of Web Usage Patterns‖ , IEEE 2011.

[13]. Ms Shashi Sahu, Leena Sahu ― A Survey on Frequent Web Mining with Improving Data Quality of Log Cleaner‖ International Journal of Advanced Research in Computer Engineering & Technology, vol 4, issue 3, March 2015.

[14]. Rahul Mishra, Abha Choubey, ―Comparative Analysis of Apriori Algorithm and 61 Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data‖ International Journal of Computer science and Information Technologies, vol 3, 2012.

[15]. Pani .S.K, Panigraphy.L, Sankar.V.H, Bikram Keshari Ratha, Padhi .A.K, ―Web Usage Mining: A Survey on Pattern Extraction from Web Log‖ International Journal of Instrumentation Control & Automation, vol 1, issue 1, 2011.

[16]. Suresh.R.M, Padmajavalli.R, ― An Overview of data preprocessing in Data and Web Usage mining‖, IEEE 2006.

[17]. Maryam Jafari, ShahramJamali., ― Discovering Users Access Patterns for web Usage Mining from Web Log Files‖ Journal of Advanced in Computer Research vol 4, issue 3, August 2013.

[18]. Preeti Sharma & Sanjay Kumar, ― An Approach for Customer Behavior Analysis using Web Mining‖ International Journal of Internet Computing, vol 1, issue 2 2011 ISSN No. 2231 – 6965.

[19]. https://hortonworks.com/apache/hdfs.

[20]. http://downloads.vmware.com/d/info/desktop_downloads/vmware_workstation/7.0