ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

ANALYSIS OF WEB LOGS USING BIG DATA TOOLS

DR.S.SUGUNA[1], M.VITHYA[2]

[1] Assistant Professor, Sri Meenakshi Govt. Arts College for Women (A), Madurai-2,kt.suguna@gmail.com

[2] Lecturer, Sri Meenakshi Govt. Arts College for Women (A), Madurai-2, vithyagopal20@gmail.com

***Abstract*— Web is an important part of organization. Every organization generated huge amount of data from various source. Web mining is the process of discovering the knowledge from the web data. The log files are maintained by the web server. Analyzing web log files has become an important task for E-commerce companies to predict their customer behavior and to improve their business. E-commerce website can generate tens of peta bytes of data in their web log files. So, the large volume of data is called big data. Big data is something so huge and complex that is impossible for handling through traditional system and traditional tools. The analysis of log files is used for learning the user behavior. The analysis of such large web log files are be worked upon by using traditional SQL does not like queries nor can the relational database management system (RDBMS) be used for storage and analysis. So, need parallel processing and reliable data storage system for this huge and complex data. The Hadoop framework provides reliable storage by Hadoop Distributed File System and parallel processing system for large database using Mapreduce programming model. This mechanism helps to process log data in parallel using all the machines in the Hadoop cluster and computes results efficiently.

Index Terms— Big data, Hadoop Framework, Hadoop DFS, Mapreduce Framework, Web Mining, Log files.

I. INTRODUCTION

Web mining is the application of data mining techniques to extract useful knowledge from web data that includes web document, hyperlink between documents, usage logs of web sites etc. Web usage mining is the process of applying data mining techniques to discover usage pattern from the web data, targeted towards various applications. Web usage mining is one of the techniques which play an important role in the personalization of web pages. To analyze the web access information first the web usage data set is collected from the internet and pre-processed the data set like filtering, noise removal etc. The collection of web usage data set gathered from different levels such as server level, client level and proxy level and also from different resources through the web browser and web server interaction using the HTTP protocol[1]. The architecture shown in Fig. 1 of a typical web application is the simplest application has the web and application tiers combined while more complex ones may have multiple web tires to handle static content and security, multiple application tires as well as more data bases.

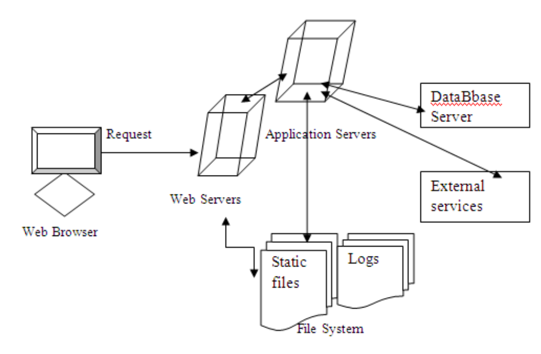


Fig. 1 Architecture for generation of WebLogs

Web log files recorded activity information when a web user submits a request to web server. The main source of raw data is the web access log which is known as Log file [5].A log files contain various parameters which are very useful in recognizing user browsing patterns [6,7].The request information sent by the user via protocol to the web server is recorded in log file. The log file entry contains ip address of the 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 and etc. Log files can be classified into categories depending on the location of their storage that is Web Server Logs and Application Server logs. Web servers maintain at least 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 [9]. Application Server logs can provide a great level of details which used by application developers and analysts to understand how the application is used. Mining the web log file will always be helpful to server and E-commerce companies to increase their online customers by predicting the behavior of their online customer. As the number of customers visiting web sites are increasing the size of the web log file is also increasing [2]. Nowadays the size of web log file is in petabytes. The existing data mining techniques store web log files in traditional DBMS and analyze. RDBMS system can be very expensive and cheaper alternatives like MYSQL cannot scale to the volume of data that is continuously being added. A better technology exists to store terabytes of log data and process it efficiently[8].But in the current scenario the number of online customer's increases day by day and each click from a web page creates on the order hundred bytes data in typical website log file. In large websites handling millions of simultaneous visitors can generate hundred of petabytes of logs per day. So to analyze such big web log file efficiently and effectively we need to develop faster, efficient and effective parallel and scalable data mining algorithm. Also need a cluster of storage devices to store a petabytes of web log data and parallel computing model for analyzing such huge amount of data. Hadoop framework provides reliable clusters of storage facility to keep our large web log file data in a distributed manner and parallel processing features to process a large web log file data efficiently and effectively[3,4].

The paper is organized as follows. Section2 Big data, in section3 Hadoop overview is discussed, section4 shows experimental results analysis and finally section 5 concludes the paper.

II. Big Data Characteristics

Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. IDC defines Big Data technologies as a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. Big data characteristics can be described as follows

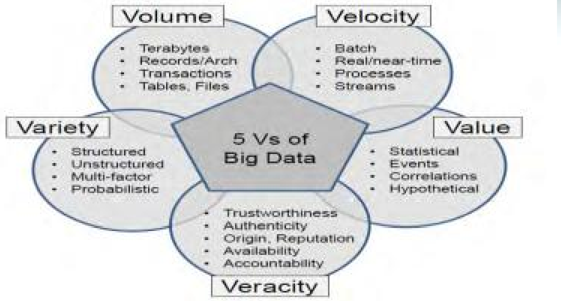


Figure 2. 5V's of Big data

Big data has been defined as early as 2001. Doug Laney, an analyst of META defined challenges and opportunities by increased data with a 5V's model. i.e increase of Volume, Velocity and Variety in his research. In 2011, an IDC report defined Big data as "big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis." With this definition, characteristics of big data can be summarized as five V's i.e., Volume (great volume), Variety (Various Modalities), Velocity (rapid generation) and value (huge value but very low density), Veracity (Quality of the data being captured)[10].

III. Hadoop Overview

Apache Hadoop is an open source project that provides a parallel storage and processing framework that enables customized analytical functions using commodity hardware. It scales out to cluster spanning tens to thousands of server nodes making it possible to process very large amounts of data at a fraction of the cost of data warehouses. The key is the use of commodity servers. Hadoop makes this possible by providing for replication and distribution of the nodes, racks and even data centers.

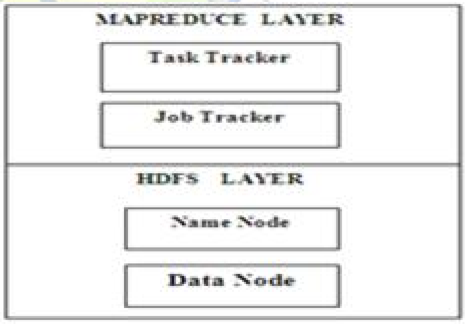


Figure.3. Single Node Hadoop Cluster System Architecture

The Hadoop architecture is divided into two layers: HDFS Layer and MapReduce Layer. Hadoop Distributed File System (HDFS) is a java –based file system that provides scalable and reliable data storage that is designed to span large cluster of commodity servers. MapReduce Layer reads data from, writes data to HDFS storage and processes the data in parallel. Name node keeps track of how weblogs file is broken down into file blocks, which nodes store those blocks. Data node stores the replication of web log file. Job Tracker determines the execution plan by deciding task, and keeps track of all tasks as they are running. Task Tracker is responsible for the execution of individual task on each slave node [11].

A. Hadoop for Log processing

One of the first applications developed using Hadoop was web log processing. Processing the large volume of logs for web companies such as Yahoo and Facebook was a challenge. Fig. 3 shows the architecture of a log processing system using Hadoop. Log files from many different types of servers are fetched via Apache Flume and loaded into a

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

Hadoop cluster. Jobs are scheduled to analyze the logs and generate aggregated summary metrics which are then sent to an external RDBMS and visualization using Business intelligent tool[12].

B. *MapReduce Frame work*

The MapReduce framework consists of two steps namely Map and Reduce step. Master node takes large problem input and slices it into smaller sub problems and distributes these to worker nodes. Worker nodes may do this again and leads to a multi-level tree structure. Worker process smaller problem and hands back to master. In Reduce step Master node takes the answer to the sub problems and combines them in original problem. The MapReduce framework is fault-tolerant. If a node remains silent for longer than the expected interval, a master node makes note and re-assigns the work to other nodes.

C. *Work Flow in MapReduce*

The key to how MapReduce work is to take input as, conceptually, a list of records. The records are split among the different computers in the cluster by Map. The result of the Reducer then takes each set of values that has same key and combines them into a single value. So Map takes a set of data chunks and produces key/value pairs and Reduce merges things, so that instead of a set of key/value pair sets, you get one result. We can't tell whether the job was split into 100 pieces or 2 pieces [13, 14].

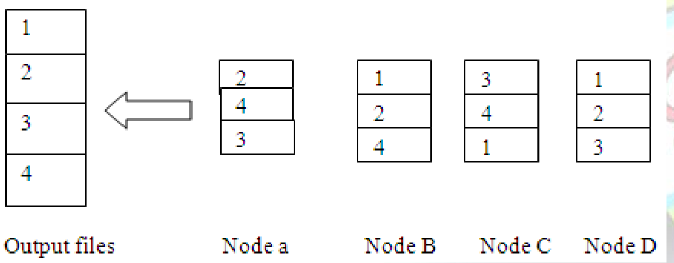


Figure.4. Computation Logic of MapReduce

In this environment, input file is split into four files and each files are stored in different nodes (like Node A,B,C and D). The same file will be stored in different nodes. Here failure of any node never leads to data lose. Data can be shared from any other node.

Map-Reduce Algorithm

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 Map reduce framework sorts the outputs of the maps, which are then input to the reduce task. Typically, both the input and the output of the job are stored in a file system. MapReduce is a 5-step parallel and distributed computation [28].

Step 1: Map () input: he "MapReduce system" designates Map processes, assign the K1 input key value each processor would work on, and provides that processor with all the input data associated with that key value.

Step 2: Map () code: Map () runs exactly once for each K1 key value, generating output organized by key values K2.

Step 3: Shuffle: The MaRreduce system designates Reduce processors, assign the K1 key value each processor would work on, and provides that processor with all the Map-generated data associated with that key value.

Step 4: Reduce () code: Reduce () runs exactly once for each K1 key value produced by the Map.

Step 5: Final output: The MapReduce system collects all the Reduce output, and sorts it by K1 to produce the final outcome.

IV. PROPOSED WORK

The Recipe log files with more than fifteen lakhs log entries are proposed using Hadoop environment. The Fig 5 depicts the MapReduce function of processing web log files and calculating total number of same recipe cook time received by each Recipe Item. The line in The Table 1 shows the sample logs from recipe log file.

```
("id" : { "oid" :  "5160756bcc6202db15" } "name " : "Drop Biscuits and sausage Gravy" , "ingredients" ;
"Biscuits \n 3 cups All-purpose Flour \n 2 Tablespoons  Baking powder \n ½ teaspoon salt \n n1-1/2 stick (3/4
cup) cold Butter, cut Into Pieces\n1-1/4 Cup Butermilk\n Sausage Gravy \n1 pound Breakfast Sausage, Hot or
Mild \n 1/3 cup All-purpose Flour\n 4 cups whole Milk\n ½ teaspoon seasoned salt \n 2 teaspoons Black pepper,
more taste" ,"Url":http://thepioneerwomen.com/cooking/2013/03/drop-biscuts-and-sausage-gravy/", "image" :
"http//static.thepioneer   women.com/cookpng/files/2013/03/bisgrav.jpg",   "time   to   spent",   :   {"$date"
:1365276011104 }, "cooktime" :Pt30M", "source" "thepioneerwomen", "recipeyield" : "12", "datePublished" :
"2013-03-11", "prep  Time" : "PT10M " ,"description" : Late Saturday afternoon.("id" : { "oid" :
"5160756bcc6202db345" } "name " : "Crispy Easter Eggs" , "ingredients" ; "4 Table spoon Butter \n Package
(10 ounces)\n 6 cups rice crispy \nAssorted sprinkles\nSmall chocolate Easter Eggs\n Plastic Easter Eggs"
,"Url":http://thepioneerwomen.com/cooking/2013/03/   Krispy   Easter   Eggs   .com/",   "image"   :
"http//static.thepioneer   women.com/cookpng/files/2013/03/crispy.jpg",   "time   to   spent",   :   {"$date"
:1365276011105 }, "cooktime" :Pt10M", "source" "thepioneerwomen", "recipeyield" : "14", "datePublished" :
"2013-03-12", "prep Time" : "PT11M " ,"description" : Late Sunday afternoon.
```

Table.1. Sample Logs

Figure. 5 shows the actual process sequence to calculate total number of same recipes time by each URL.

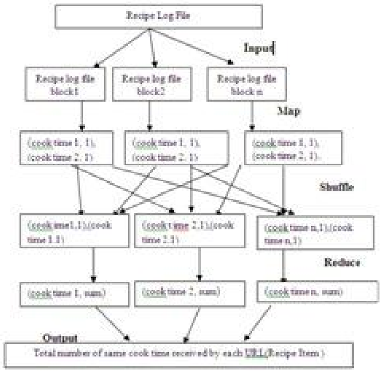
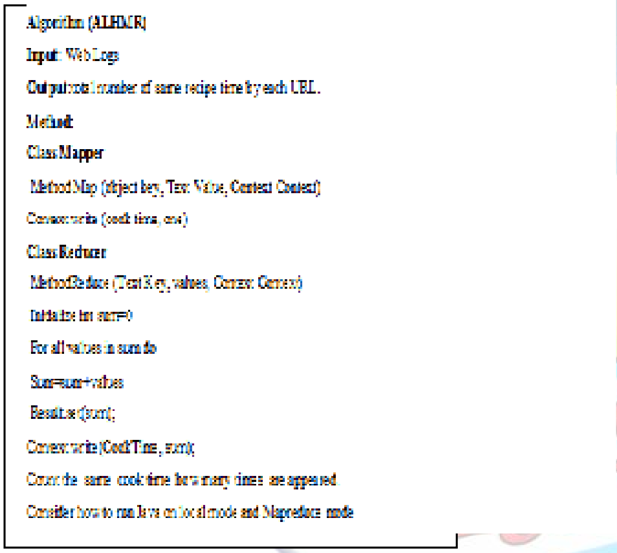

Recipe Log File
Input
Recipe log file block1
Recipe log file block2
Recipe log file block n
Map
(cook time 1, 1), (cook time 2, 1)
(cook time 1, 1), (cook time 2, 1)
(cook time 1, 1), (cook time 2, 1).
Shuffle
(cook time1,1),(cook time 1,1)
(cook time 2,1),(cook time 2,1)
(cook time n,1),(cook time n,1)
Reduce
(cook time 1, sum)
(cook time 2, sum)
(cook time n, sum)
Output
Total number of same cook time received by each URL(Recipe Item)

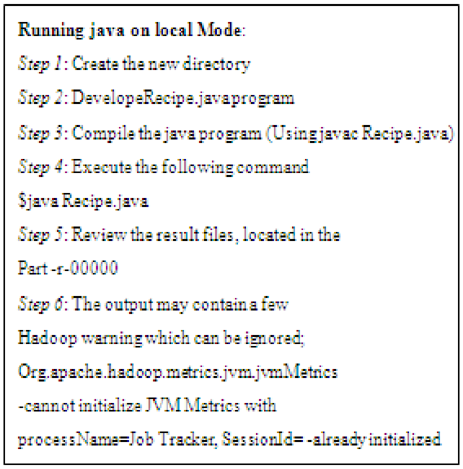
Figure.5. Calculating the total number of same recipe time by each URL (Recipe Item)

The web log file is split into blocks by Hadoop Frame work and stored into Single node cluster. The input to this function is a Recipe log file. For each cook time in the Recipe site, a line will be added into the Recipe Log file. In the Mapper function, each block of the Recipe log file is given as an input to a map function which in turn parse each line using regular expression and emits the Recipe Item as a key along with the value 1 (cook time 1,1),(cook time 2,1)…..(cook time n,1). After mapping the shuffling collects all the (Key, Value) pairs which are having the same cook time from different Mapping function's and forms a group. After this process, Group1 entries will be (cook time 1, 1), (cook time 1, 1)and so an. Group 2 entries will be (cook time 2,1), (cook time 2,1)and so an. Then the reducer function calculates the sum for each cook time group. The result of the reduce function is (cook time 1, sum)…..(Cook time n, sum).



```
Algorithm (ALHMR)
Input: Web Logs
Output: total number of same recipe time by each URL.
Method:
Class Mapper
 Method Map (object key, Text Value, Context Context)
Context.write (cook time, one)
Class Reducer
 MethodReduce (Text Key, values, Context Context)
 Initialize int sum=0
 For all values in sum do
 Sum=sum+values
 Result.set(sum);
Context.write(CookTime, sum);
Count the same cook time how many times are appeared
Consider how to run Java on local mode and Mapreduce mode
```

Consider how to run Java on local mode and MapReduce mode



Running java on local Mode:

Step 1: Create the new directory

Step 2: DevelopeRecipe.javaprogram

Step 3: Compile the java program (Using javac Recipe.java)

Step 4: Execute the following command

$java Recipe.java

Step 5: Review the result files, located in the

Part -r-00000

Step 6: The output may contain a few

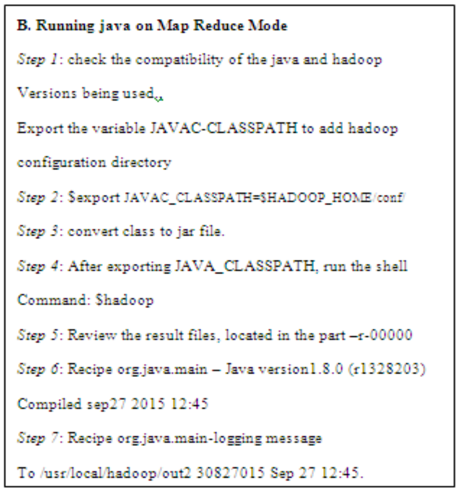
Hadoop warning which can be ignored;

Org.apache.hadoop.metrics.jvm.jvmMetrics

-cannot initialize JVM Metrics with

processName=Job Tracker, SessionId= -already initialized

Java translates the queries into MapReduce jobs and runs the job on the Hadoop cluster. This cluster can be fully distributed cluster.



B. Running java on Map Reduce Mode

Step 1: check the compatibility of the java and hadoop

Versions being used.

Export the variable JAVAC-CLASSPATH to add hadoop

configuration directory

Step 2: $export JAVAC_CLASSPATH=$HADOOP_HOME/conf

Step 3: convert class to jar file.

Step 4: After exporting JAVA_CLASSPATH, run the shell

Command: $hadoop

Step 5: Review the result files, located in the part –r-00000

Step 6: Recipe org.java.main – Java version1.8.0 (r1328203)

Compiled sep27 2015 12:45

Step 7: Recipe org.java.main-logging message

To /usr/local/hadoop/out2 30827015 Sep 27 12:45.

V. Experimental Results

This section discusses the results obtained from the experiment

A. *Experimental setup*

To calculate the total number of Recipe based on cook time received by each Recipe Item, a single node Hadoop cluster is set up with the configurations with Ubuntu 14.04 operating system, Hadoop version 2.6.0, and Single node cluster 192.168.2.1 and dataset Amazon Recipe Logs of 1 Terabyte. Before executing the MapReduce code in the single node cluster environment, the Recipe log file is loaded into the HDFS of Hadoop framework. Fig 6 shows the contents of the output directory named number of cook time by recipe items in HDFS. The output is stored in afile called part r-000000.

Figure.6. Number of Cook time by Recipe Item Logs output directory in HDFS

MapReduce function is used to calculate the total number of Recipe based on cook time. HMR algorithm is executed in 52423 milliseconds using Map Reduce Environment. The number of Mapper task launched is 5 and Reduce task launched is 2. Time taken by map task is 22 seconds and reduce task is 32 seconds.

VI. Conclusion

This paper describes a detailed view of processing big data such as Recipe log file with more than fifteen lacks of logs using Hadoop frame work. It gives a description of how log file is processed using mapreduce and how Hadoop framework is used for parallel computation of log files. Data from different source and places are loaded into HDFS for further processing. We proved that processing big data with the help of Hadoop environment leads to minimum computation time and store large amount of data compare than traditional data base system.. .

References

[1] M.Santhanakumar and C.Christopher Columbus, "Web Usage Analysis of Web pages UsingRapidminer", WSEAS Transactions on computers, E-ISSN: 2224-2872,

[2] S.Saravanan and B.UmaMaheswari, "Analyzing Large Web Log Files in A Hadoop DistributedCluster Environment", International Journal of Computer Technology & Applications, vol.5, pg:1677-1681.

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

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

[5] 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.

[6] 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.

[7] 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

[8] A white paper by OrzotaInc, "Beyond Web Application Log Analysis using Apache Hadoop".

[9] G. Arumugam, S. Suguna, "Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs",IEEE Explorer,Pages: 1-6, ISBN:978-2-9532-4431-1, June 2009.

[10] Xindong Wu AT, "Data Mining with Big data", IEEE Transactions on knowledge and data Engineering,vol 26,no.1,January 2014

[11] ShreyasKudale, AdvaitKulkarni and A.LeenaA.Deshpande, "Predictive Analysis Using Hadoop: A Survey",International Journal of Innovative Research in Computer and Communication Engineering, vol.1,Issue.8, October 2013,

[12] Laurel Thejas Souza and Grish U R, " Error Log Analytics using Big data And Mappreduce", International journal of computer science and Information Technologies, vol. 6 (3),pg: 2364-2367,2007.

[13] PrajaktaDange and Dr. Deven Shah, " Web Log Analysis for security Compliance Using Big Data",International Journal of Advanced Research in Computer Science and Software Engineering , vol,5, Issue.3,March 2015.

[14] S.SiddharthAdhikari ,DeveshSaraf, Mahesh Revanwar and Nikhil Ankam, "Analysis of Log Data and Statistics Report Generation Using Hadoop",International Journal of Innovative Researchin Computer and Communication Engineering, vol.2, Issue. 4, April2014.

[15] VidyullathaPellakuri and r.D.RajeswaraRao,"HadoopMapreduce Framework in Big Data Analytics" ,International Journal of Computer Trends and Technology vol,8 ,Issue.3, Feb2014

[16] Ted Garcia,"Analysis of Big data Technologies and Method",IEEE Computersociety,2011.