Analyzing NASA web server Logs using Hadoop Mapreduce

Pragati Mahale[1], Akshay Bhagwat[1], Priyanka Mahajan[3], Kaveri Jadhav[4], Snehal Bhole[5]
[1, 2, 3, 4, 5] Department of Information Technology
[1, 2, 3, 4, 5] AISSMS IOIT

Abstract*- Traditional database handling tools are not capable to handle rising big data which increasing its volume, variety, velocity day by day. Hadoop is a tool which handles this kind of data effectively. The semi structure log files generated by web server are in large number they are in flat text format. This project deals with performing analysis of NASA web server log using hadoop and visualization of analysis report using R. This enables web admin to monitor traffic on server and region specific request by tracing user session and IP address.*

I. INTRODUCTION

Data is an assemblage of a fact which is present on web server. Now a day's large amount of data is available on internet from which 90% of data is generated in last two-three years either by organizations, groups or by individuals. This data goes on increasing day by days as use of World Wide Web is increases. Traditional databases are unable to manage this large data. Big data is a tool which manage, manipulate and process this data. Big data is collection of high variety, high volume and high velocity information. Big data is alike from databases like RDBMS, which usage Hadoop framework to process terabytes or petabytes of data in les amount of time.

Advanced big data techniques are used to discover the hidden patterns and other information from datasets. Various tools are available to manage, process, manipulate and analyze this data. Some of them are MySQL, Cassandra, MongoDB, and Hadoop framework that includes HDFS, Hbase, Hive and Pig.

Big data includes variety of data in which not only the structure data is presents but semi structure and unstructured data is also presents. Text data, video, images are the unstructured data which is generated in large magnitude.

Multiple researches are done on web Log files some of them are listed below.

Murat Ali [3] proposed smart miner which traced the distinct path accessed by unique user.

Sayalee Narkhede [4] introduced a tool Hadoop Mapreduce which analyze analyzed Log files and generate the statistical report to show page accessed by user. This work is performed on two machine in Hadoop distributed mode and log files are scattered on different nodes.

Milind Bhandare [5] proposed another Log analyzer to analyze different kinds of web server Log files. This is implemented to process multiple queries simultaneously to minimize the response time.

In this project we analyzed the web server Logs of NASA to identify the session using Hadoop Mapreduce in distributed manner. Using Hadoop Mapreduce semi structure log file can be analyze effectively as data is stored on multiple clusters as blocks in Hadoop distribute file system. Session identification gives timestamp which includes IP address of users and totals hits on particular page.

Using identified IP addresses GeoLaction convertor gives exact location of user specified request. The identified result of session identification and GeoLocation convertor are analyzed using R tool.

II. MAPREDUCE PROCESS

Hadoop is a flexible open source framework to store data and to manage large scale computation by running application on cluster of commodity hardware also. It perform multiples of computation on petabytes of data simultaneously. Hadoop works in distributed mode across many clusters and perform parallel computation to speed up the processing. Hadoop splits the large amount of input to the smaller chunks and each chunk can be processed individually on different machines. There are two main components of Hadoop, first is Mapreduce and second is HDFS.

Mapreduce group the logical unit of data together and perform some computation. Mapreduce program can be written in any of the language which is compatible for the developer. Mapreduce breaks the process into two phase i.e. Map phase and Reduce phase. It splits the large file into small chunks which is process by map task. Map function takes the input in the form of key/value pairs, process it and generate zero or more outputs. Map function generate the set of intermediate output which is in key/value pair and given as input to the Reduce task to perform further processing. Map task writes the output in local disk. If the node running the map task fails then Hadoop will automatically reassign that task to another node.

Reduce function takes the input from mapper phase. Usually result of this phase is smaller than input set. Both input and output of mapper and reducer phase is stored in HDFS. Mapreduce suites where data is write once and read many times. The data is stored in HDFS which allows master-slave architecture.

Hadoop cluster consists of one namenode which stores metadata and act as master that manages the blocks present on datanodes. There can be numbers of datanodes which present on

Slave node and frequently report to the namenode about its blocks it stores. It provides actual storage. Jodtracker is responsible for resource management and assign a task to task tracker.

HDFS stores the file in rack system. It replicates files on different machines so that client will never goes down. If one of the node gets fail then it will automatically replicates the data blocks on another node. Minimum 3 replicas of file gets store in HDFS that’s why HDFS is highly fault tolerant.

III. LOG ANALISYS USING HADOOP

Hadoop framework is used to process large data in less time. Hadoop performs computation of location of data rather than moving data to the location. It can efficiently manage semi structure data in parallel computational mode

Log files are generated by web server in large quantity which traditional databases cannot handle. Log files are stored in HDFS where processing is made on log to keep track on session accessed by unique user.

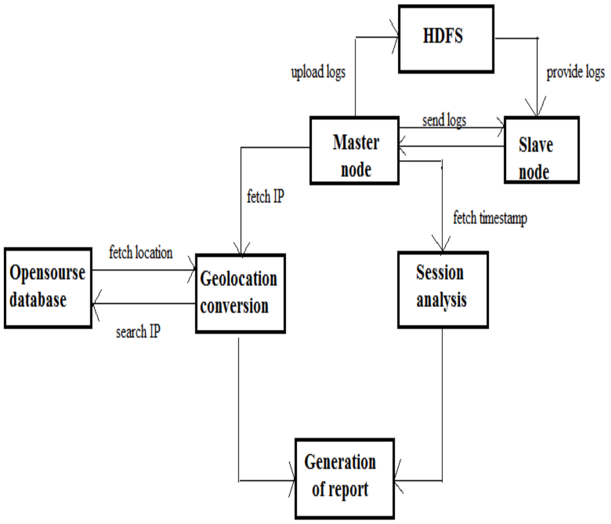
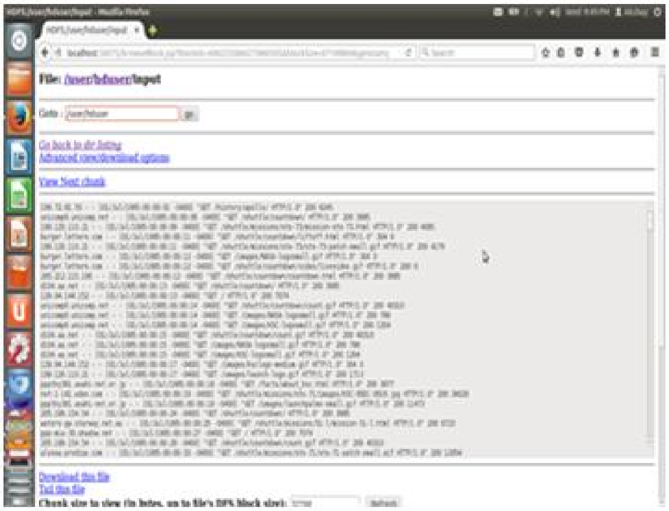


Fig 1. System architecture of system

Data cleaning is the first step implemented in proposed system as a preprocessing step. The NASA log file contain thousands of records which is automatically generated by server. This log may contains the request made by robots, web crawlers or by spider that includes ambiguous, erroneous and incomplete information. To analyze correct log file specific pattern is programmed which will neglect the incorrect data. Incorrect requests are filter out from the NASA log file

The further important task carried out in proposed system is session identification to monitor traffic in network and to identify unique user and unique URLs accessed. Log files that abide in HDFS is given as input to Mapreduce job. Abstract class FileInputFormat is used to give input to the mapreduce. While map task is executing reduce task is set to zero as there is no aggregation function is needed. Input to map task is given through long writable.



```
[illegible]
```

Fig 2. NASA web server logs

Text Input Format is used to break the file into lines as map function takes the input in key/value pair. The key is IP address and the remaining fields are taken as value. In proposed system mapper class is used to separate the IP address from log file. To identify IP address specific pattern is used. If IP address from log file is matches with pattern then IP address is fetched from file. The log file includes the IP address, timestamp, browser version, date and total bytes send for a request.

Session identification is an important model in data mining. Session can be identified from unique IP address and time spent by user on web page from login till log out.

Maximum time limit for session is 30 minutes. Is response doesn't come from user side or session time is exceeds then new session number is generated for same IP address. Date function is used to calculate the difference between similar IP address. ParseLog class takes the timestamp from mapper class as parameter and compare the parameter with object of SimpleDareFormat. GetHours (), getDay (), getDate () methods of date class is used to extract the hour, day and date from timestamp.

Log file consists of various fields and using IP address unique user is identified. Once all the key are found, combiner will combine all the value of specific key. After this reduce task aggregate the result and total count of IP address stored in HDFS.

GeoLocation convertor is an open source database containing the list of coordinates for IP addresses. It takes the identified session as input and mapped the address to the longitude, latitude, country and city etc. to convert the IP address inti coordinate IP address needs to search in open source database. Binary search algorithm is implemented in proposed system to get the exact location of IP address from the open source

IV. RESULT AND INTERPRETATION

Hadoop framework provides five services namely Namenode, Datanode, Tasktracker, Jobtracker and secondary Namenode. In pseudo distribute system all nodes run on same local machine. In this mode Jobtracker and Tasktracker are set to idle mode. Namenode keep all information about the data. Preprocessed log file is stored in HDFS. Mapper and reducer class is used to fetch the IP addresses from log file. Map task is responsible to bring one set of data to another set. Without HDFS it is difficult to manage and analyze large volume of data. In HDFS data is stored in block system where default size of each block is 64 MB.

4.1 Pseudo distributed mode

Hadoop framework provides five services namely Namenode, Datanode, Tasktracker, Jobtracker and secondary Namenode. In pseudo distribute system all nodes run on same local machine. In this mode Jobtracker and Tasktracker are set to idle mode. Namenode keep all information about the data. Preprocessed log file is stored in HDFS. Mapper and reducer class is used to fetch the IP addresses from log file. Map task is responsible to bring one set of data to another set. Without HDFS it is difficult to manage and analyze large volume of data. In HDFS data is stored in block system where default size of each block is 64 MB.

4.2 Fully distributed mode

All the services of Hadoop framework is used in fully distributed mode. In proposed system cluster is created with three to four machines of similar configuration. Hadoop assigns a job to all daemons. In this mode cluster of different machine is created from which one act as master and others act as slave. Actual data is stored in datanode at slave machine

Master node runs the services such as Namenode, Datanode, Tasktracker and Jobtracker. Slave node are responsible to run Tasktracker and Datanode. In distributed cluster if one node fails another node provides the result and operate without losing data. Master node stores the IP addresses of slave machine. Namenode stores the metadata which mean it stores location of actual data in database. Datanode sends the continuous heartbeat signal to namenode that indicate the datanode still working. If datanode fails to send the heartbeats signal in particular limit of time then namenode consider datanode is fail and assign its responsibilities to another node

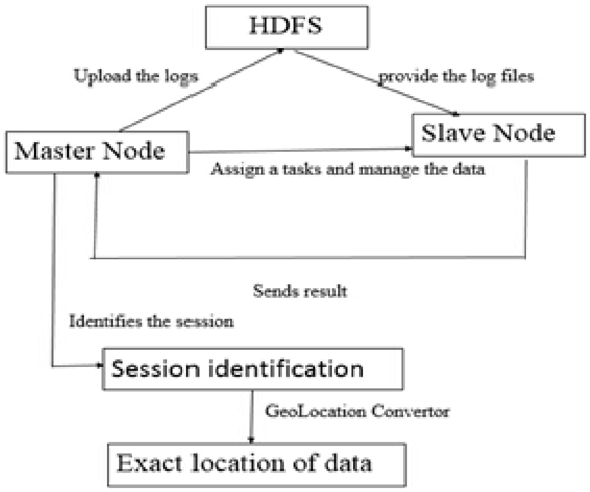


Fig 3. Data processing in Hadoop file system

Jobtracker assigns a job to tasktracker. Tasktracker perform actual tasks on location of data. Tasktracker also sends the heartbeat signal to jobtracker. If tasktracker is fails at any single point then jobtracker reassign its task to another node.

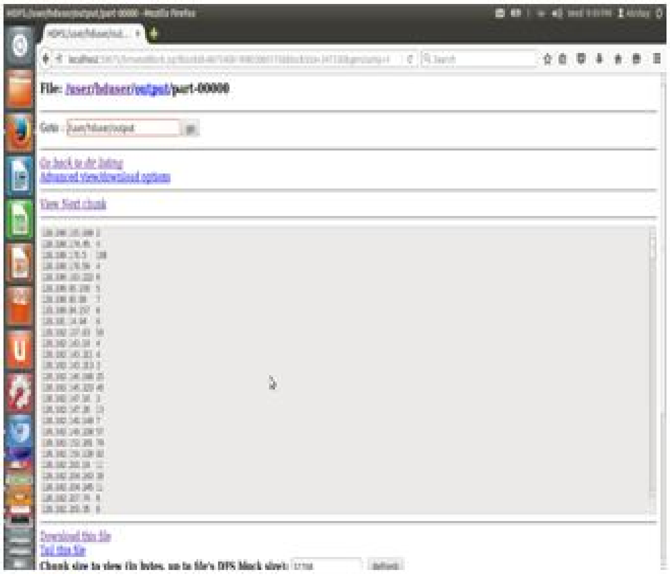
HDFS replicates the data three times as minimum replication factor is three. This replication is stored on slave node. Master node identify slave node by using its IP addresses. Secondary namenode which does not act as a namenode but in certain point of condition if namenode is fails then secondary namenode is responsible to assign all tasks again to namenode.

Log data is divided into blacks and processed on different machines. Master node assign tasks to slave node. At slave node mapper and reducer phase completes the processing and send result back to the master node. To identify the different session one particular pattern is used to fetch the IP addresses from log file

ParseLog class takes the timestamp as parameter and compare the parameter with object of SimpleDateFormat. GetHours (), getDay (), getDate () methods of date class is used to extract the hour, day and date from timestamp.

Timestamp is used to calculate the incoming request on web sites. Using this timestamp exact count of request is get to know.

Using this we plot a graph in R tool which shows the request coming on sites in particular hour of time.



File: /user/hduser/output/part-00000

Go back to dir listing
Advanced view/download options

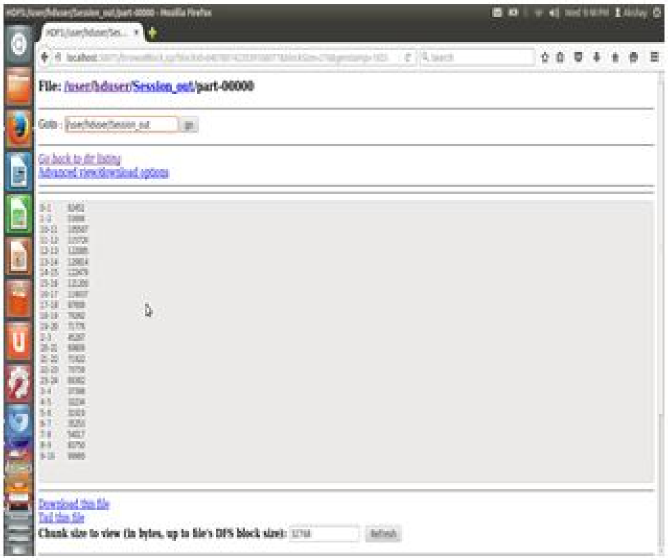
View Next chunk

```
[illegible]
```

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

Fig 4. Unique IP addresses with count

GeoLocation (), map (), toString () class of GeoLocation convertor, converts the result of session identification to longitude and latitude. IP addresses are search in open source database using binary search algorithm.



File: /user/hduser/Session_out/part-00000

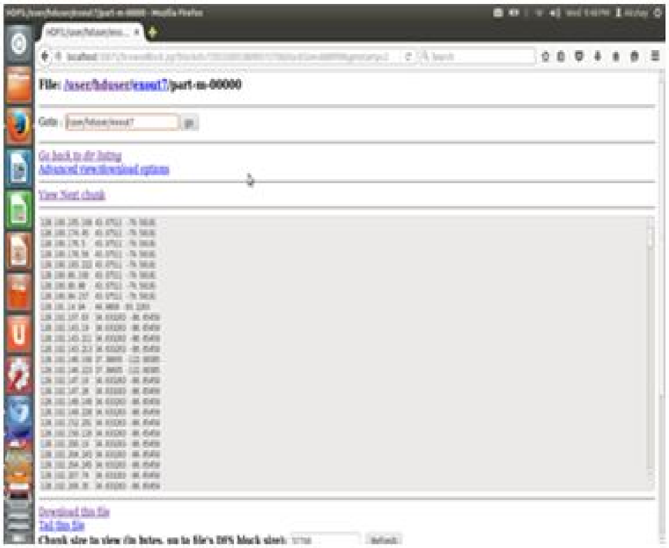
Go back to dir listing
Advanced view/download options

```
0-1    [illegible]
1-2    [illegible]
10-11  [illegible]
11-12  [illegible]
12-13  [illegible]
13-14  [illegible]
14-15  [illegible]
15-16  [illegible]
16-17  [illegible]
17-18  [illegible]
18-19  [illegible]
19-20  [illegible]
2-3    [illegible]
20-21  [illegible]
21-22  [illegible]
22-23  [illegible]
23-24  [illegible]
3-4    [illegible]
4-5    [illegible]
5-6    [illegible]
6-7    [illegible]
7-8    [illegible]
8-9    [illegible]
9-10   [illegible]
```

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

Fig 5. Session identification

Mapreduce task is divided into record reducer, mapper and practitioner and reducer. Record reader reads a data from log file with IP Fig 6. IP addresses and its coordinates address as key and remaining data as value. It pass the data to the mapper phase. Map tasks group all the similar type of data and reduce task is set to zero while map task is executing. Reducer has three primary phases such as shuffle, sort and reducer. Practitioner controls the partitioning of keys of intermediate map output.Total number of practitioner is



File: /user/hduser/exout7/part-m-00000

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):

Fig 6. IP addresses and its coordinates

Same as total number of reduce task. The key is used to drives the partition.

Log file is analyzed using R. R is a free software environment to produce statistical report. The statistical report of session identification and user specific request location is set on world map using R tool.

V. CONCLUSION

Big data technology is completely integrated with real domain problem. This paper shows the implementation of big data technologies where framework handle large amount of data in distributed cluster. Log data is stored in HDFS and the data is manipulated using session identification algorithm to explore the traffic on internet and to know the location specific request. Unique ID key is used to track the user behavior. GeoLocation convertor gives exact location of incoming request. Both algorithm process on NASA web server log. From result it is concluded that Hadoop framework has high performance

REFERENCES

[1] Ruchi Verma, Sathyan R Mani, "Use of Big Data Tehnologies in Capital Markets," 2012 Infosys Limited, Bangalore, India.

[2] James Manyika, Brad Brown et.al, "Big Data: The next frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, June 2011.

[3] Murat Ali, Ismail Hakki Toroslu, "Smart Miner: A New Framework for mining Large Scale Web Usage Data," WWW 2009, April 20-24. 2009 Madrid, Spain. ACM 978-1-60558-487-4/09/04.

[4] Sayalee Narkhede and Tripti Baraskar, "HMR Log Analyzer: Analyze Web Application Logs over Hadoop MapReduce," International Journal of UbiComp (IJU) vol.4, No.3, July 2013.

[5] Milind Bhandare, Vikas Nagare et al., "Generic Log Analyzer Using Hadoop Mapreduce Framework," International Journal of Emerging Technology and Advanced Engineering (IJETAE), vol.3, issue 9, September 2013.

[6] http://www.web-datamining.net/

[7] Kanchan Sharadchandra Rahate et al., "A Novel Technique for Parallelization of Genetic Algorithm using Hadoop," International Journal of Engineering Trends and Technology (IJETT), vol.4, issue 8, August 2013.

[8] T. K. Das et al., "BIG Data Analytics: A Framework for Unstructured Data Analysis," International Journal of Engineering and Technology (IJET) vol 5, No 1, Feb-Mar 2013.

[9] Joseph McKendrick, "Big Data, Big Challeneges, Big Opportunities: 2012 IOUG Big Data strategies survey," September 2012)