Clickstream Analysis using Hadoop

Harshit Makhecha, Dharmendra Singh, Bhagirath Prajapati, Priyanka Puvar

Computer Engineering Department
A D Patel Institute of Technology,
Gujarat, India

Abstract

E-Commerce websites generates huge churns of data due to large amount of transactions taking place every second and so their inventory should be updated as per transactions very quickly to remain stable in these competitive market. Analyzing web log files has become one of the important task for E-Commerce companies to predict their customer behavior. Clickstream data is very important part of big data marketing as it will tell what customers click on and purchase or (do not purchase). The primary focus of the paper is to prepare web log analysis system which will depict trends based on the users browsing mode using Hadoop MapReduce and handling heterogeneous query execution on log file.

Introduction

As we know all over the world, E-Commerce industry is growing rapidly. Data is increasing exponentially and tremendously all over the world. The Data which is collected, stored and analyzed from multiple data sources is the biggest challenge for most E-Commerce industries and also lot of data is waiting to be analyzed. For example, web-log file is that one type of data which has to be analyzed to predict the customer patterns of buying or visiting that product page. The web log file contains different useful information like IP address of the computer making the request (i.e. the visitor), the name of the requested file, its location, the HTTP status code, browser of that system etc.

Now to increase the profits of E-commerce industries, mining that web log file will always be helpful, because by mining the web log file, E-commerce companies predict the behavior of their online customers. Personalized experience, including content and promotions are offered by E-commerce companies by predicting their online customer behavior. Recommendations can also be provided on the basis of their browsing behavior. By mining the web log file, E-commerce companies can do a lot more. The size of the web log file is also increasing day by day as the number of customer visiting e-commerce websites are also increasing. Pattern discovery data mining techniques are already available to analyze the web log files. But in the current trend, day by day online customers are increasing and each click from a web page creates the order of 150 bytes in a simple web log file. At the same time, many large websites are handling simultaneous customers which generates hundreds of petabytes of data stored in web log file.

It is said that 90% of the world's data is generated in last two years alone only. Every day, 2.8 quintillion bytes of data gets generated. Twitter generates around 13 TB of data every day. Few years ago, companies were generating data and all others were consuming data, but now model is changed as now all of us are generating data and all of us are consuming data[1].

The various sources of data generation are Social Media, Scientific instruments, Mobile devices, Sensor Technology etc. This all data will constitute into a Big Data. In information technology, Big Data is a collection of data sets so large and complex that is becomes difficult to process and store using on-hand existing tools and technologies.

The Data which is generated is divided into 3 types[3].

1. Structured Data: - The data which is in tabular form.
2. Unstructured Data: - The data which is not in organized form. Metadata, Twitter tweets, and other social media posts are good examples of unstructured data.
3. Semi-Structured Data: -It is a form of structured data but do not forms a formal structure of data model. Example: - xml files[6].

Clickstream Data and its uses: -

Clickstream Data: - A user leaves information behind while visiting a website forms a clickstream data. It is captured in website log files in semi-structured format. It contains information such as visitor's ip address, date and time stamp, user-id that uniquely identifies the visitor and destination URL's of the visited pages.

Uses of clickstream Data: -

The one of the original use of Hadoop was to store and process the massive volume of clickstream data. Now all types of enterprises uses Hadoop to refine and analyze data and they can then answer business questions such as:

1. What is the most efficient path for a visitor to research a product, and then buy it?
2. What products do visitors tend to buy and what are they most likely to buy in the future?
3. Where should I spend resources on fixing and enhancing the user experience on my website?

Related Work

When we compare SQL DBMS and Hadoop MapReduce, it is suggested that Hadoop MapReduce performs better than SQL DBMS. The traditional data base management system cannot handle a large dataset. So for this purpose only, we need to have Big Data technologies like Hadoop Framework. For Big Data analysis, Hadoop MapReduce is used in many areas. To analyze web log file, Hadoop is a good platform as the size of the web log is increasing day by day. Apache Hadoop is an open source project created by Doug cutting and developed by the Apache Software Foundation. When we want to store large scale data, Hadoop platforms allows us on thousands of nodes and analyze it. Normally Hadoop cluster consists of thousands of nodes which store multiple blocks of log files. Log files are fragmented into blocks and these blocks are evenly distributed over a hundreds of cluster by Hadoop. After that blocks are also replicated over the multiple nodes to achieve reliability and fault tolerance[2].

System Architecture

Figure[1] shows the cluster configuration of Hadoop System. There are two nodes in the cluster. One node is called slave node and another one is called master node. The architecture is divided in two layers. MapReduce layer and Hadoop Distributed File System(HDFS) layer. HDFS is a java-based file distribution system which provides reliable and scalable data storage that is designed to span large clusters of servers. Name Node will keep track of how web log data is fragmented into file blocks, which nodes store those blocks. Replication of web log file will be stored by Data node. The execution plan for which files to be processed, assigns nodes to different tasks, and tracks all the running task. On each slave node, Task Tracker is responsible for the execution of individual tasks[4].

Implementation

First of all, we will have to create a database using Hive shell and after creating database we will have to add Apache SERDE(Serializer/Deserializer) jar file shown in figure[2]. SerDe is short for Serializer/Deserializer. Hive uses the SerDe interface for IO. The interface handles both serialization and deserialization and also interpreting the results of serialization as individual fields for processing.

A SerDe allows Hive to read in data from a table, and write it back out to HDFS in any custom format[5]. Anyone can write their own SerDe for their own data formats.

After that create one external table, in which that sample web log dataset is to be loaded into HDFS and Hadoop Hive. After ingesting data set into HDFS, fire a query which extracts top 5 products which has maximum visits by customer. When we will run this query, the map reduce job will be submitted which we can see in browser also. So we can get top 5 trends of product name followed by their number of visits shown in figure [3].

Conclusion

This paper thus provides insights to process and analyze web log data and how can we mine the data we like. We can also provide recommendations on the basis of clickstream analysis. It can also respond to the competitive market rapidly by changing the prices of its products every 2 minutes (if required) whilst other retailers change the prices of the products manually and It will provide better predictive analysis and lot more.

References

[1] What is big data: - IBM?

[2] “Why Big Data is a must in E-Commerce”, Guest post by Jerry Jao, CEO of Retention Science. http://www.bigdatalandscape.com/news/why-big-data-is-a-must-in-ecommerce

[3] Tom White, (2009) “Hadoop: The Definitive Guide. O’Reilly”, Scbastopol, California.

[4] Apache-Hadoop, http://Hadoop.apache.org

[5] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai, “ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING”, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011

[6] https://en.wikipedia.org/wiki/Semi-structured_data

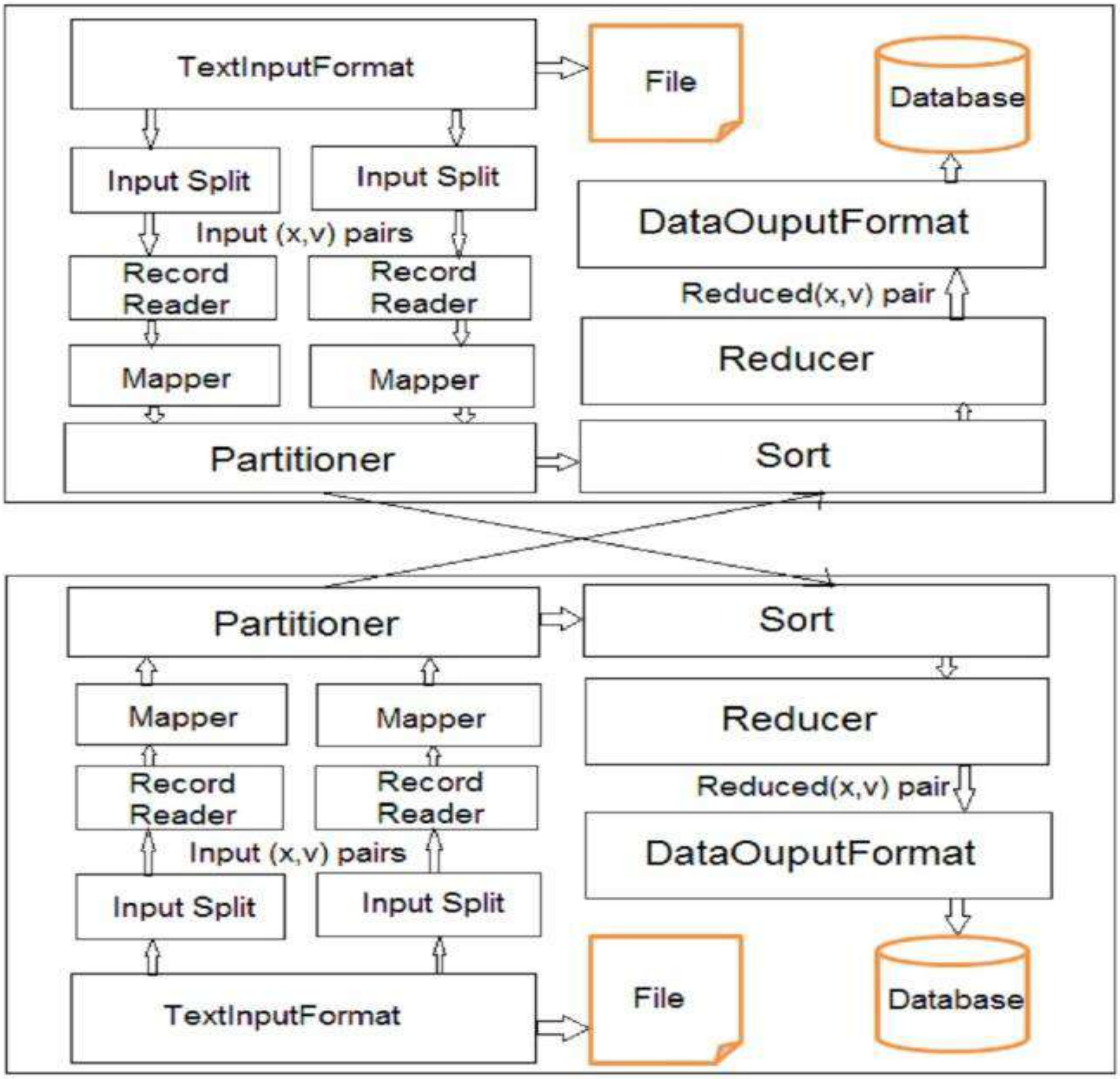
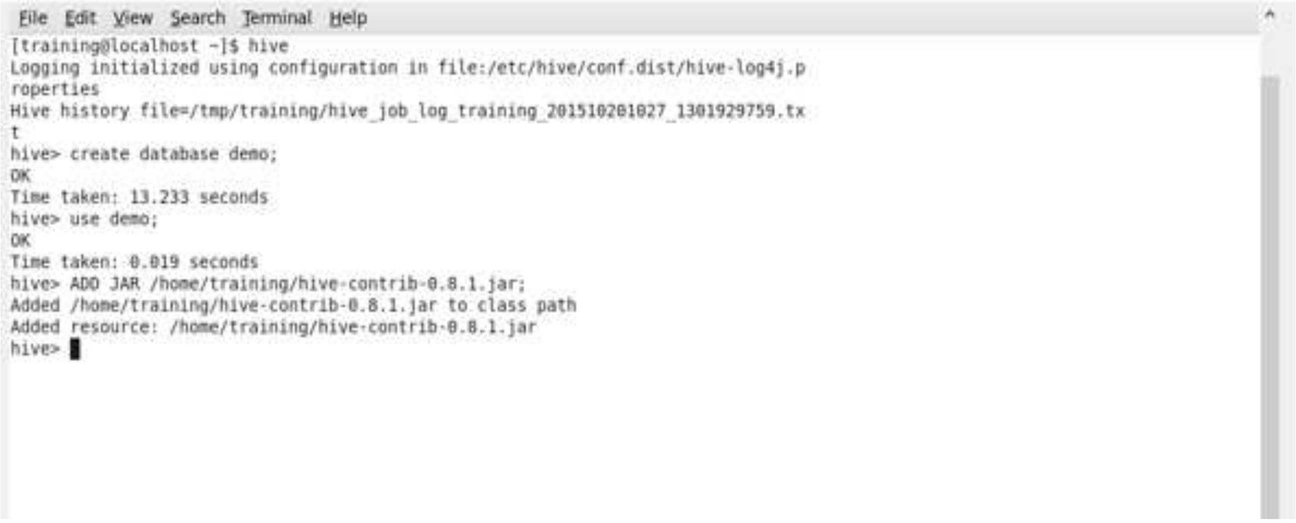
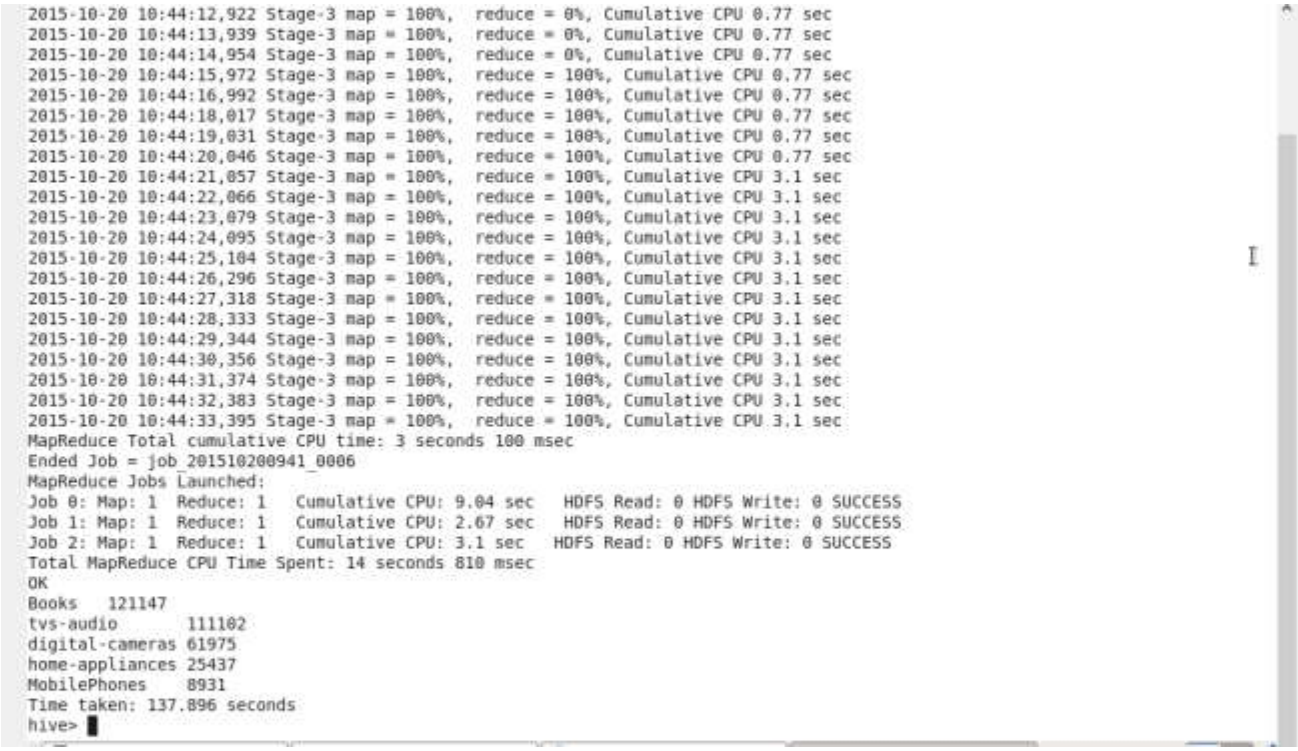


Figure [1]



```
[training@localhost ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
Hive history file=/tmp/training/hive_job_log_training_201510201027_1301929759.tx
t
hive> create database demo;
OK
Time taken: 13.233 seconds
hive> use demo;
OK
Time taken: 0.019 seconds
hive> ADD JAR /home/training/hive-contrib-0.8.1.jar;
Added /home/training/hive-contrib-0.8.1.jar to class path
Added resource: /home/training/hive-contrib-0.8.1.jar
hive> 
```

Figure [2]



```
2015-10-20 10:44:12,922 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.77 sec
2015-10-20 10:44:13,939 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.77 sec
2015-10-20 10:44:14,954 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 0.77 sec
2015-10-20 10:44:15,972 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 0.77 sec
2015-10-20 10:44:16,992 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 0.77 sec
2015-10-20 10:44:18,017 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 0.77 sec
2015-10-20 10:44:19,031 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 0.77 sec
2015-10-20 10:44:20,046 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 0.77 sec
2015-10-20 10:44:21,057 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:22,066 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:23,079 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:24,095 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:25,104 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:26,296 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:27,318 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:28,333 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:29,344 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:30,356 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:31,374 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:32,383 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
2015-10-20 10:44:33,395 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 3.1 sec
MapReduce Total cumulative CPU time: 3 seconds 100 msec
Ended Job = job_201510200941_0006
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 9.04 sec   HDFS Read: 0 HDFS Write: 0 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 2.67 sec   HDFS Read: 0 HDFS Write: 0 SUCCESS
Job 2: Map: 1  Reduce: 1   Cumulative CPU: 3.1 sec   HDFS Read: 0 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 810 msec
OK
Books   121147
tvs-audio       111102
digital-cameras 61975
home-appliances 25437
MobilePhones    8931
Time taken: 137.896 seconds
hive> 
```

Figure[3]