ISSN:2231-0711

Mining Analysis on Customers Data using Big Data tools

P.Sai Prasad
Ph.D. Research Scholar
Dept. of Computer Science
SSSUTMS
Sehore, M.P.

K.Rajesh
Assistant Professor
Dept.of Computer Science
BVSR Engineering College
Chimakurthy, A.P.

Abstract: In this paper the authors presented that how to analyse the Customers data when it is huge and various raw format data. So, for this we have used the Hadoop Big Data tools map-reduce Pig and Hive which makes easier to analyse the various raw formats of data.

Keyword: Hadoop, Big Data, Map- reduce, Hive

INTRODUCTION

Mostly the universe runs on the Business. So they get customers raw data of various formats and they don't know what is the data is there, for that we can use technique of Hadoop and Big data Tools for analysing the data of raw formats. First we should know what is Hadoop , Map reduce and hive.

Hadoop follows Master-Slave architecture. Hadoop uses HDFS to store files efficiently in the cluster. When a file is placed in HDFS it is broken down into blocks, 64 MB block size by default. These blocks are then replicated across the different nodes (*DataNodes*) in the cluster. The default replication value is 3, i.e. there will be 3 copies of the same block in the cluster. We will see later on why we maintain replicas of the blocks in the cluster.

A Hadoop cluster can comprise of a single node (single node cluster) or thousands of nodes.

Hadoop is made up of 2 parts:

1. HDFS – Hadoop Distributed File System
2. MapReduce – The programming model that is used to work on the data present in HDFS.

Hadoop Distributed File System (HDFS) Building blocks of Hadoop:

A. Namenode
B. Datanode
C. Secondary Name node
D. JobTracker
E. TaskTracker

NameNode

The NameNode in Hadoop is the node where Hadoop stores all the location information of the files in HDFS. In other words, it holds the metadata for HDFS.

Secondary NameNode

IMPORTANT – The *Secondary NameNode* is not a failover node for the *NameNode*. The secondary name node is responsible for performing periodic housekeeping functions for the *NameNode*. It only creates checkpoints of the file system present in the *NameNode*.

DataNode

The *DataNode* is responsible for storing the files in HDFS. It manages the file blocks within the node. It sends information to the *NameNode* about the files and blocks stored in that node and responds to the *NameNode* for all filesystem operations.

JobTracker

JobTracker is responsible for taking in requests from a client and assigning *TaskTrackers* with tasks to be performed. The *JobTracker* tries to assign tasks to the *TaskTracker* on the *DataNode* where the data is locally present (Data Locality). If that is not possible it will at least try to assign tasks to *TaskTrackers* within the same rack. If for some reason the node fails the *JobTracker* assigns the task to another *TaskTracker* where the replica of the data exists since the data blocks are replicated across the *DataNodes*. This ensures that the job does not fail even if a node fails within the cluster.

TaskTracker

TaskTracker is a daemon that accepts tasks (Map, Reduce and Shuffle) from the *JobTracker*.

The *TaskTracker* keeps sending a heart beat message to the *JobTracker* to notify that it is alive. Along with the heartbeat it also sends the free slots available within it to process tasks. *TaskTracker* starts and monitors the Map & Reduce Tasks and sends progress/status information back to the *JobTracker*.

MapReduce

MapReduce is the programming model that uses Java as the programming language to retrieve data from files stored in the HDFS. All data in HDFS is stored as files. Even MapReduce was built in-line with another paper by Google.

Google, apart from their papers did not release their implementations of GFS and MapReduce. However, the Open Source Community built Hadoop and MapReduce based on those papers. The initial adoption of Hadoop was at Yahoo Inc., where it gained good momentum and went onto be a part of their production systems. After Yahoo, many organizations like LinkedIn,

Facebook, Snapdeal Netflix and many more have successfully implemented Hadoop within their organizations.

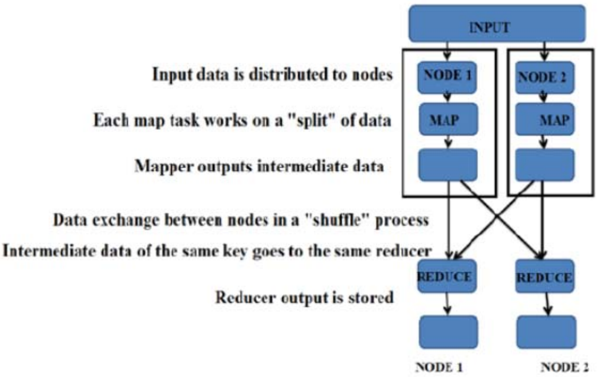

INPUT
Input data is distributed to nodes
Each map task works on a "split" of data
Mapper outputs intermediate data
Data exchange between nodes in a "shuffle" process
Intermediate data of the same key goes to the same reducer
Reducer output is stored
NODE 1
NODE 2
MAP
MAP
REDUCE
REDUCE
NODE 1
NODE 2

Fig.1.MapReduce Job submission flow mechanism

In this paper author's analyzed online data with the help of Big Data technology and visualized

Analysis of Online Shopping Data

Big Data technology analyzes the any type of data whether it is structured or unstructured. In Big Data technology everyone can analysis the data in a few second that takes more time in traditional one.

In the past for analysis no one used the Big Data technology to analyze and visualize the large scale online data like gigabytes or terabytes. Therefore we need parallel techniques. With the help of Big Data technology, we can easily tackle the huge amount of online data in parallel.

Hadoop distribution platform for implementing Big Data technologies i.e. Hadoop, Pig and Hive etc. It provides virtual Linux environment for map reducing. Apache Hadoop is a framework running on the top of the cloudera .Apache Hadoop is used for processing the huge data across the cluster of commodity computers using simple programming model. Map Reduce is generally used for splitting the task across processor. Map Reduce is written in Java programming language. For running ClouderaCDH5, VM Player has been used to execute the ClouderaCDH5 machine as a virtual framework. Local machine must have required 4GB RAM for successfully configure the ClouderaCDH5 machine on VM Player.

Algorithm for analysis of Online Data

An algorithm used for analysis of online data has two phases i.e. mapper phase and reducer phase.

Mapper

The Mapper code reads the input files as <Key,Value> pairs and emits key value pairs.

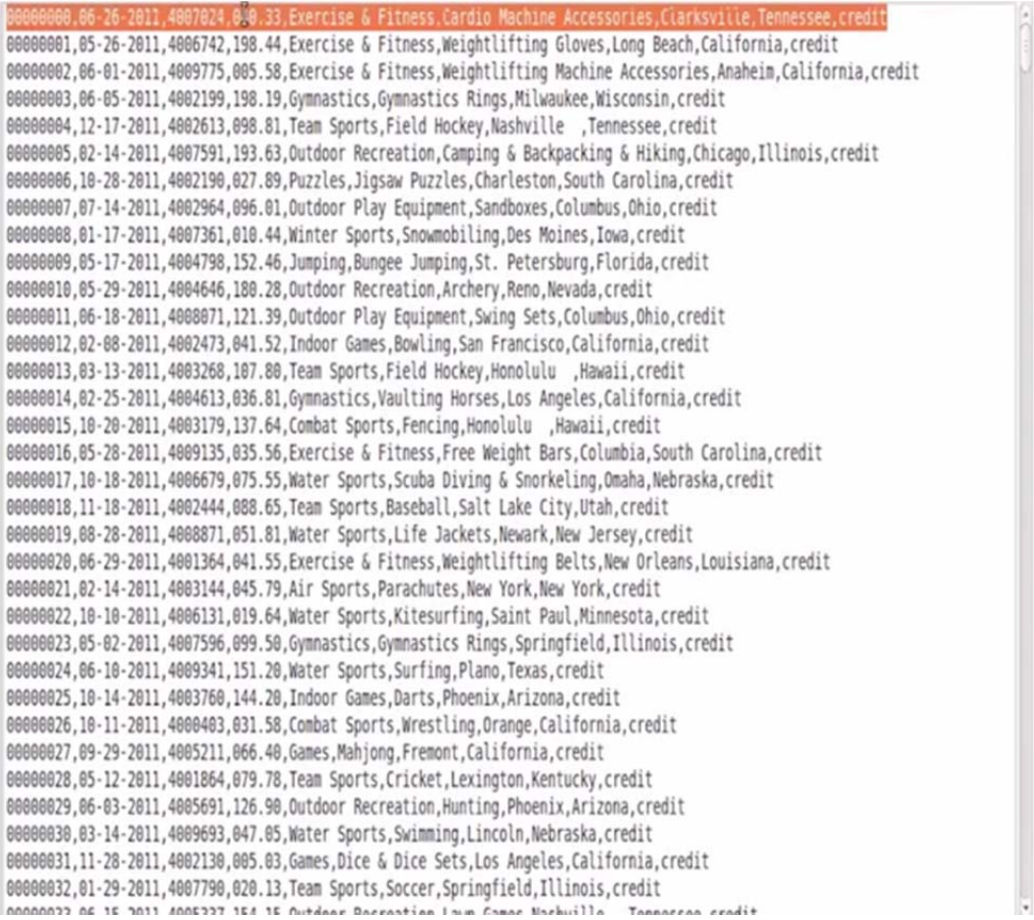
Sort and shuffling is done Reducer

The Reducer code reads the outputs generated by the different mappers as <Key,Value> pairs and emits key value pairs.

SIMULATION WORK

Simulation work has been initialized by installation of clouderaCDH3 on windows.

Take some file named as retail.txt



```
00000000,06-26-2011,4007024,040.33,Exercise & Fitness,Cardio Machine Accessories,Clarksville,Tennessee,credit
00000001,05-26-2011,4006742,198.44,Exercise & Fitness,Weightlifting Gloves,Long Beach,California,credit
00000002,06-01-2011,4009775,005.58,Exercise & Fitness,Weightlifting Machine Accessories,Anaheim,California,credit
00000003,06-05-2011,4002199,198.19,Gymnastics,Gymnastics Rings,Milwaukee,Wisconsin,credit
00000004,12-17-2011,4002613,098.81,Team Sports,Field Hockey,Nashville  ,Tennessee,credit
00000005,02-14-2011,4007591,193.63,Outdoor Recreation,Camping & Backpacking & Hiking,Chicago,Illinois,credit
00000006,10-28-2011,4002190,027.89,Puzzles,Jigsaw Puzzles,Charleston,South Carolina,credit
00000007,07-14-2011,4002964,096.01,Outdoor Play Equipment,Sandboxes,Columbus,Ohio,credit
00000008,01-17-2011,4007361,010.44,Winter Sports,Snowmobiling,Des Moines,Iowa,credit
00000009,05-17-2011,4004798,152.46,Jumping,Bungee Jumping,St. Petersburg,Florida,credit
00000010,05-29-2011,4004646,180.28,Outdoor Recreation,Archery,Reno,Nevada,credit
00000011,06-18-2011,4008071,121.39,Outdoor Play Equipment,Swing Sets,Columbus,Ohio,credit
00000012,02-08-2011,4002473,041.52,Indoor Games,Bowling,San Francisco,California,credit
00000013,03-13-2011,4003268,107.80,Team Sports,Field Hockey,Honolulu  ,Hawaii,credit
00000014,02-25-2011,4004613,036.81,Gymnastics,Vaulting Horses,Los Angeles,California,credit
00000015,10-20-2011,4003179,137.64,Combat Sports,Fencing,Honolulu  ,Hawaii,credit
00000016,05-28-2011,4009135,035.56,Exercise & Fitness,Free Weight Bars,Columbia,South Carolina,credit
00000017,10-18-2011,4006679,075.55,Water Sports,Scuba Diving & Snorkeling,Omaha,Nebraska,credit
00000018,11-18-2011,4002444,088.65,Team Sports,Baseball,Salt Lake City,Utah,credit
00000019,08-28-2011,4008871,051.81,Water Sports,Life Jackets,Newark,New Jersey,credit
00000020,06-29-2011,4001364,041.55,Exercise & Fitness,Weightlifting Belts,New Orleans,Louisiana,credit
00000021,02-14-2011,4003144,045.79,Air Sports,Parachutes,New York,New York,credit
00000022,10-10-2011,4006131,019.64,Water Sports,Kitesurfing,Saint Paul,Minnesota,credit
00000023,05-02-2011,4007596,099.50,Gymnastics,Gymnastics Rings,Springfield,Illinois,credit
00000024,06-10-2011,4009341,151.20,Water Sports,Surfing,Plano,Texas,credit
00000025,10-14-2011,4003760,144.20,Indoor Games,Darts,Phoenix,Arizona,credit
00000026,10-11-2011,4000403,031.58,Combat Sports,Wrestling,Orange,California,credit
00000027,09-29-2011,4005211,066.40,Games,Mahjong,Fremont,California,credit
00000028,05-12-2011,4001864,079.78,Team Sports,Cricket,Lexington,Kentucky,credit
00000029,06-03-2011,4005691,126.90,Outdoor Recreation,Hunting,Phoenix,Arizona,credit
00000030,03-14-2011,4009693,047.05,Water Sports,Swimming,Lincoln,Nebraska,credit
00000031,11-28-2011,4002130,005.03,Games,Dice & Dice Sets,Los Angeles,California,credit
00000032,01-29-2011,4007790,020.13,Team Sports,Soccer,Springfield,Illinois,credit
00000033,06-15-2011,4005337,154.15,Outdoor Recreation,Lawn Games,Nashville  ,Tennessee,credit
```

Fig 2.Raw Data

First create DataBase

Hive>create database onlinetransaction;

Create table in Database

Hive(onlinetransaction)>create table transaction(txnno INT,txndate STRING,custno INT,amount DOUBLE,category STRING,product STRING,city STRING,state STRING,spendby STRING) row format delimited fields terminated by ‘,’ stored as textfile;

Loading data from local

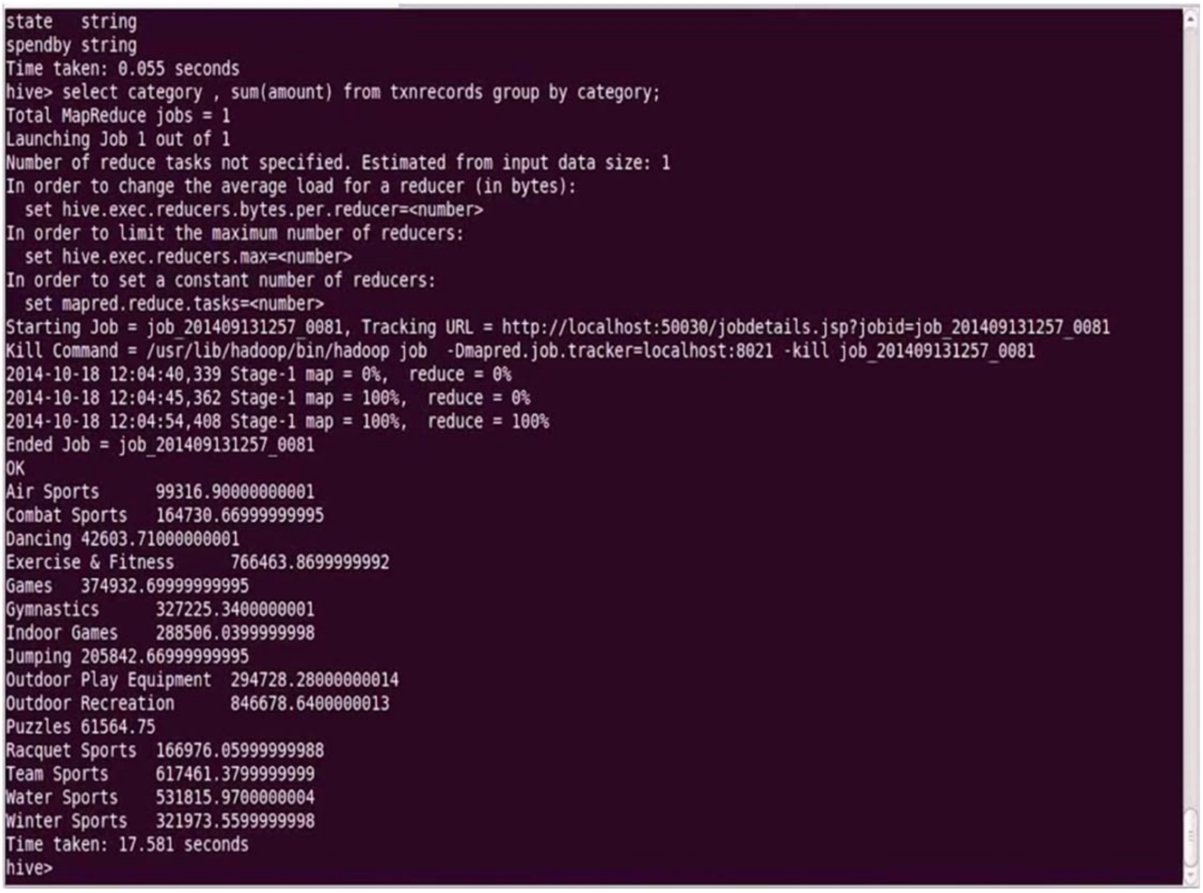
Hive>load data local in path ‘home/cloudera/desktop/retail.txt' overwrite into table transaction;

Operations on Online Raw Data

If we want category wise total amount

Hive>Select category, Sum(amount) from transaction group by category;

After that, these analyzed data presented on Hadoop and connected directly to Big Data Visualization tool i.e. Tableau tool. All the data have collected in the Big Data tool.



```
state   string
spendby string
Time taken: 0.055 seconds
hive> select category , sum(amount) from txnrecords group by category;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201409131257_0081, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201409131257_0081
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201409131257_0081
2014-10-18 12:04:40,339 Stage-1 map = 0%,  reduce = 0%
2014-10-18 12:04:45,362 Stage-1 map = 100%,  reduce = 0%
2014-10-18 12:04:54,408 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201409131257_0081
OK
Air Sports      99316.90000000001
Combat Sports   164730.66999999995
Dancing 42603.71000000001
Exercise & Fitness      766463.8699999992
Games   374932.69999999995
Gymnastics      327225.3400000001
Indoor Games    288506.0399999998
Jumping 205842.66999999995
Outdoor Play Equipment  294728.28000000014
Outdoor Recreation      846678.6400000013
Puzzles 61564.75
Racquet Sports  166976.05999999988
Team Sports     617461.3799999999
Water Sports    531815.9700000004
Winter Sports   321973.5599999998
Time taken: 17.581 seconds
hive>
```

Fig 3: showing the amount by category wise

CONCLUSION

In this study, we have analyzed and visualized the online data of India using Big Data technology for a given data in fraction of second which was not possible by traditional techniques. In this paper authors are thus using Big Data technology to better understand the data and minimize the potential damage that an online data can cause.

REFERENCES

[1] ApacheHive .https://hive.apache.org/index.html.

[2] Prajapati V (2014) Big Data Analytics with R and Hadoop. Shroff Publishers & Distributors Pvt Ltd.

[3] Cloudera.Http://www.cloudera.com/content/cloudera/en/products-and-Services/cdh.html.

[4] Hadoop: The Definitive Guide by Tom White, 3rd Edition, O’reilly

[5] Hadoop in Action by Chuck Lam, MANNING Publ.

[6] Hadoop for Dummies by Dirk deRoos, Paul C.Zikopoulos, Roman B.Melnyk,Bruce Brown, Rafael Coss

[7] Hadoop in Practice by Alex Holmes, MANNING Publ.

[8] Hadoop MapReduce Cookbook,Srinath Perera, Thilina Gunarathne