Mining of Datasets using Big Data Technique: Hadoop Platform

Rohit Kumar[1], Daulat Sihag[2], Khushboo Sukhija[3]

[1]M.Tech Scholar, Department of Computer Science Engineering, JCDM College of Engineering, Sirsa
[2]Assistant Professor Department of Computer Science Engineering, JCDM College of Engineering, Sirsa
[3]Assistant Professor, School of Computer Science, Lingayas University, Faridabad

Abstract— BIG DATA IS THE FUTURE OF IT INDUSTRY. *Here see the methodology i.e. ETL process used for analysis of big data by using Hadoop ecosystem. The analysis of big data extracts business values from the raw data and helps in gaining competitive advantage by different organisations. There is a drastic growth of data in the web applications and social networking and such data are said be as Big Data. It requires huge amount of time consumption to retrieve those datasets. It lacks in performance analysis. To overcome this problem the Hive queries with the integration of Hadoop are used to generate the report analysis for thousands of datasets. The objective is to store the data persistently along with the past history of the data set and performing the report analysis of that data set. The main aim of this system is to improve performance through parallelization of various operations such as loading the data, index building and evaluating the queries. Thus the performance analysis is done with parallelization. HDFS file system is used to store the data after performing the MapReduce operations and the execution time is decreased when the number of nodes gets increased. The performance analysis is tuned with the parameters such as the execution time and number of nodes..*

Keywords— Big Data, Hadoop, HDFS.

I. INTRODUCTION

To generate information it requires massive collection of data. The data can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is ‘Data Mining’.

Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important. information in their data warehouses [1][2][3][4].Data mining tools predict future trends and behaviours, helps organizations to make proactive knowledge-driven decisions[2]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [3][5]. Though, data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

1.1 Knowledge Discovery Process

Data mining is one of the tasks in the process of knowledge discovery from the database. The steps in the KDD process contain:

- **Data cleaning:** It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collection.
- **Data integration:** In this stage, multiple data sources, often heterogeneous, are combined in a common source.
- **Data selection:** The data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation:** It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining:** It is the crucial step in which clever techniques are applied to extract potentially useful patterns [1][3].
- **Pattern evaluation:** In this step, interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation:** It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

II. RELATED WORK

D. Abadi analyse the large scale data analysis with the traditional DBMS. The data management is scalable but there is replication of data. Replication of data leads to the fault tolerance [1]. J.ABABI, AVI SILBERCHATZ analyse massive datasets on very large clusters is done within the HadoopDB architecture for the real world application like business data warehousing. It approaches for parallel databases in performance. Still there is no scalability. It consumes huge amount of time for execution [4]. Farah Habib Chan chary analyse large datasets among the clusters of machines are efficiently stored in the cloud storage systems. So that the same information on more than one system could operate the datasets even if any one of the system's power fails [3]. According to IBM the amount of unstructured and multi-structured data within an average organization is about 80% (Savvas, 2011). Taking account the average data growth, annually by 59% (Pettey & Goasduff, 2011), this percentage will likely be much higher in a few years. Not only the volume of data is becoming a problem, also the variety and velocity are issues we need to look at (Russom, 2011). This phenomenon is called "big data" and is identified as one of the biggest IT trends for 2012 (Pettey, 2012) [5][6][7].

Time to market and innovation of new products are nowadays the key factors for enterprises. Data warehouses and BI support the process of making business decisions. These instruments allow the discovery of hidden pattern like asking unknown relations between certain facts or entities. This causes an important change: in the past, the question which has been run against the system was already known at the time of collecting the data, today it is common practice to catch all the data to hold it for questions which will be asked in the future [9]. That is the reason why Big Data is a hot growing topic in information science. To tackle the challenges of Big Data, a new type of technologies has emerged. Most of these technologies are distributed across many machines and have been grouped under the term "NoSQL". NoSQL is actually an acronym that expands to "Not Only SQL". In some ways, these new technologies are more complex than traditional databases, and in other ways they are simpler. There isn't any solution that fits all situations. You must do some compromises. This point will also be analyzed in this thesis. These new systems can scale to vastly larger sets of data, but using these systems require also new sets of technologies. Many of these technologies were first pioneered by two big companies: Google and Amazon. The most popular is probably the MapReduce computation framework introduced by Google in 2004. Amazon created an innovative distributed key-value store called Dynamo. The open source community responded in the year following with Hadoop (free implementation of MapReduce), HBase, MongoDB, Cassandra, RabbitMQ and countless other projects [Mar12]. The heterogeneous mixture learning technology is an advanced technology used in big data analysis. In the above, we introduced difficulties that are inherent in heterogeneous mixture data analysis, the basic concept of heterogeneous mixture learning and the results of a demonstration experiment that dealt with electricity demand predictions. As the big data analysis increases its importance, heterogeneous mixture data mining technology is also expected to play a significant role in the market. The range of application of heterogeneous mixture learning will be expanded broader than ever in the future [8].

III. METHODOLOGY USED

3.1 The Hadoop Platform

Hadoop is an open-source ecosystem used for storing, managing and analysing a large volume of data, designed to allow distributed processing of data sets across thousands of machines.

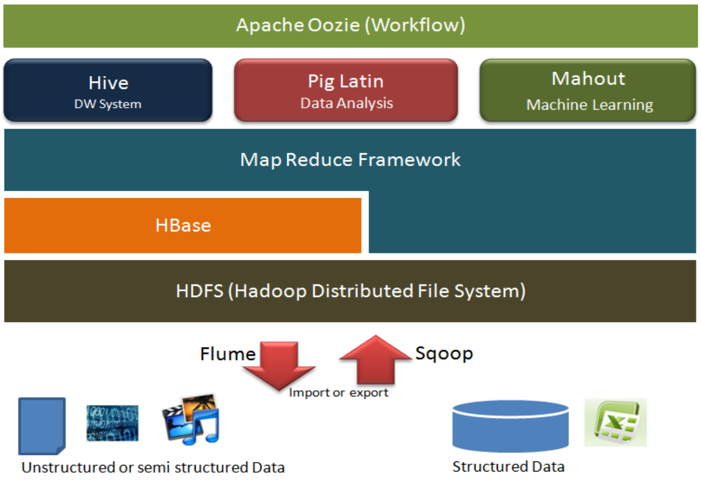


FIG 1. HADOOP ECOSYSTEM

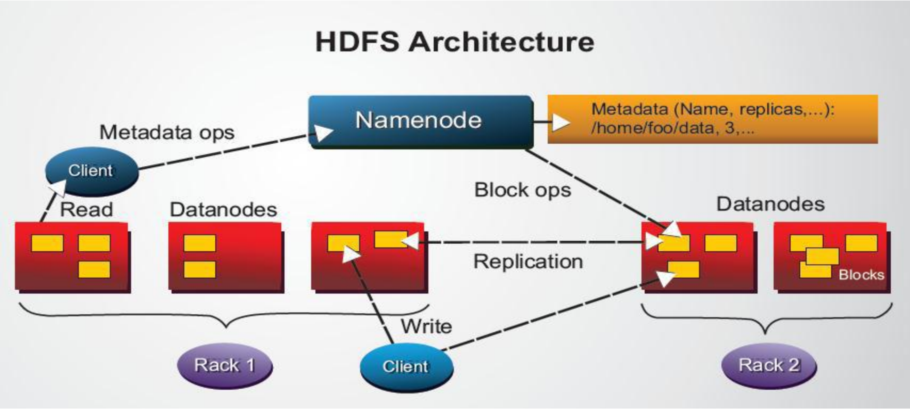


FIG 2. HDFS ARCHITECTURE

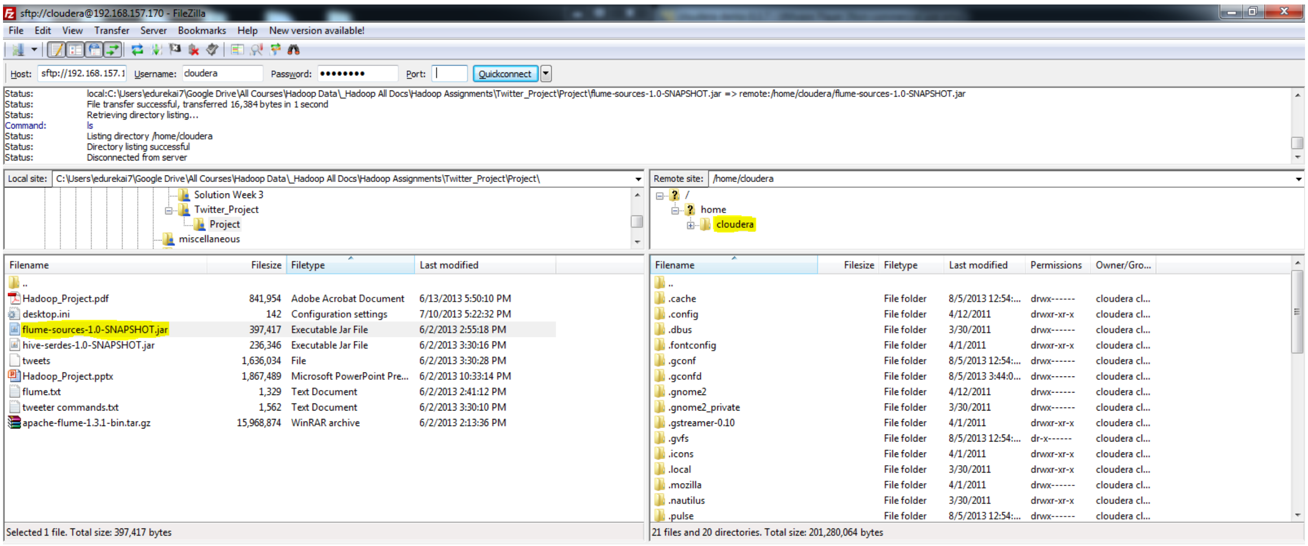
A cluster running Hadoop means running a set of daemons on the different servers of the network. The daemons include:

- **Name Node:** it is the master of HDFS that directs the slave Data Nodes daemons. It is recommended to run the Name Node on its own machine. The negative aspect to the Name Node is that it is a single point of failure.
- **Data Node:** It is responsible of writing and reading HDFS blocks to actual files on the local file system. A Data Node may communicate with other Data Nodes to replicate its data for redundancy.
- **Secondary Name Node:** it is an assistant daemon for monitoring purposes.
- **Job Tracker:** Manages Map Reduce job execution. The Job Tracker daemon determines the execution plan, assigns nodes to different tasks. There is only one Job Tracker per Hadoop cluster.
- **Task Tracker:** Manages the execution of individual tasks on the machine. Communicates to Job Tracker in order to obtain task requests.

3.2 FileZilla

FileZilla is free, cross-platform FTP application software. Consisting of FileZilla Client and FileZilla Server. Binaries are available for *Windows*, *Linux*, and *Mac OS X*. It supports *FTP*, *SFTP*, and *FTPS*.

Filezilla is used for moving files from local systems to Cloudera directory using FTP client from your Windows.



Host: sftp://192.168.157.1 Username: cloudera Password: •••••••• Port:

```
Status:     local:C:\Users\edurekai7\Google Drive\All Courses\Hadoop Data\_Hadoop All Docs\Hadoop Assignments\Twitter_Project\Project\flume-sources-1.0-SNAPSHOT.jar => remote:/home/cloudera/flume-sources-1.0-SNAPSHOT.jar
Status:     File transfer successful, transferred 16,384 bytes in 1 second
Status:     Retrieving directory listing...
Command:    ls
Status:     Listing directory /home/cloudera
Status:     Directory listing successful
Status:     Disconnected from server
```

Local site: C:\Users\edurekai7\Google Drive\All Courses\Hadoop Data_Hadoop All Docs\Hadoop Assignments\Twitter_Project\Project\

```
Solution Week 3
Twitter_Project
  Project
miscellaneous
```

Remote site: /home/cloudera

```
/
  home
    cloudera
```

Filename	Filesize	Filetype	Last modified
..			
Hadoop_Project.pdf	841,954	Adobe Acrobat Document	6/13/2013 5:50:10 PM
desktop.ini	142	Configuration settings	7/10/2013 5:22:32 PM
flume-sources-1.0-SNAPSHOT.jar	397,417	Executable Jar File	6/2/2013 2:55:18 PM
hive-serdes-1.0-SNAPSHOT.jar	236,346	Executable Jar File	6/2/2013 3:30:16 PM
tweets	1,636,034	File	6/2/2013 3:30:28 PM
Hadoop_Project.pptx	1,867,489	Microsoft PowerPoint Pre...	6/2/2013 10:33:14 PM
flume.txt	1,329	Text Document	6/2/2013 2:41:12 PM
tweeter commands.txt	1,562	Text Document	6/2/2013 3:30:10 PM
apache-flume-1.3.1-bin.tar.gz	15,968,874	WinRAR archive	6/2/2013 2:13:36 PM

Selected 1 file. Total size: 397,417 bytes

Filename	Filesize	Filetype	Last modified	Permissions	Owner/Gro...
..					
.cache		File folder	8/5/2013 12:54:...	drwx------	cloudera cl...
.config		File folder	4/12/2011	drwxr-xr-x	cloudera cl...
.dbus		File folder	3/30/2011	drwx------	cloudera cl...
.fontconfig		File folder	4/1/2011	drwxr-xr-x	cloudera cl...
.gconf		File folder	8/5/2013 12:54:...	drwx------	cloudera cl...
.gconfd		File folder	8/5/2013 3:44:0...	drwx------	cloudera cl...
.gnome2		File folder	4/12/2011	drwx------	cloudera cl...
.gnome2_private		File folder	3/30/2011	drwx------	cloudera cl...
.gstreamer-0.10		File folder	4/1/2011	drwxr-xr-x	cloudera cl...
.gvfs		File folder	8/5/2013 12:54:...	dr-x------	cloudera cl...
.icons		File folder	4/1/2011	drwxr-xr-x	cloudera cl...
.local		File folder	3/30/2011	drwxr-xr-x	cloudera cl...
.mozilla		File folder	4/1/2011	drwx------	cloudera cl...
.nautilus		File folder	3/30/2011	drwxr-xr-x	cloudera cl...
.pulse		File folder	8/5/2013 12:54:...	drwx------	cloudera cl...

21 files and 20 directories. Total size: 201,280,064 bytes

FIG 3. TRANSFERRING FILES FROM WINDOW SERVER TO CLOUDERA

IV. PRESENT WORK

4.1 Hive Component of Hadoop

It Support structured data, e.g., creating tables, as well as extensibility for unstructured data.

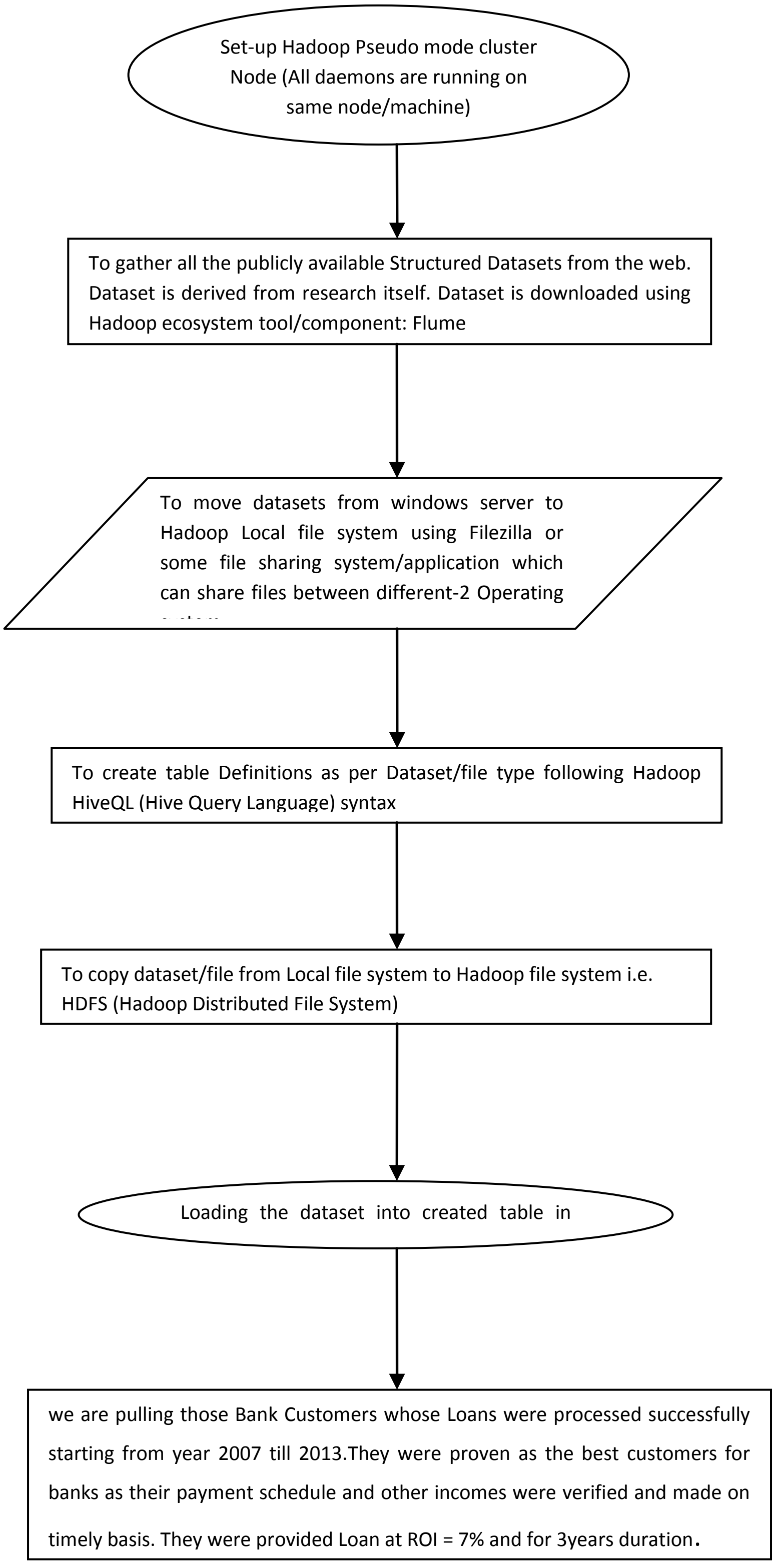
- Hive Query for Creating Tables

Create table user (UserId int, Age int, Gender char) row format delimited fields;

- Hive Query for inserting/loading data into table

Load data Local inpath '/users/local/users.txt' into Table user;

4.2 Flow Chart



Set-up Hadoop Pseudo mode cluster Node (All daemons are running on same node/machine)

To gather all the publicly available Structured Datasets from the web. Dataset is derived from research itself. Dataset is downloaded using Hadoop ecosystem tool/component: Flume

To move datasets from windows server to Hadoop Local file system using Filezilla or some file sharing system/application which can share files between different-2 Operating

To create table Definitions as per Dataset/file type following Hadoop HiveQL (Hive Query Language) syntax

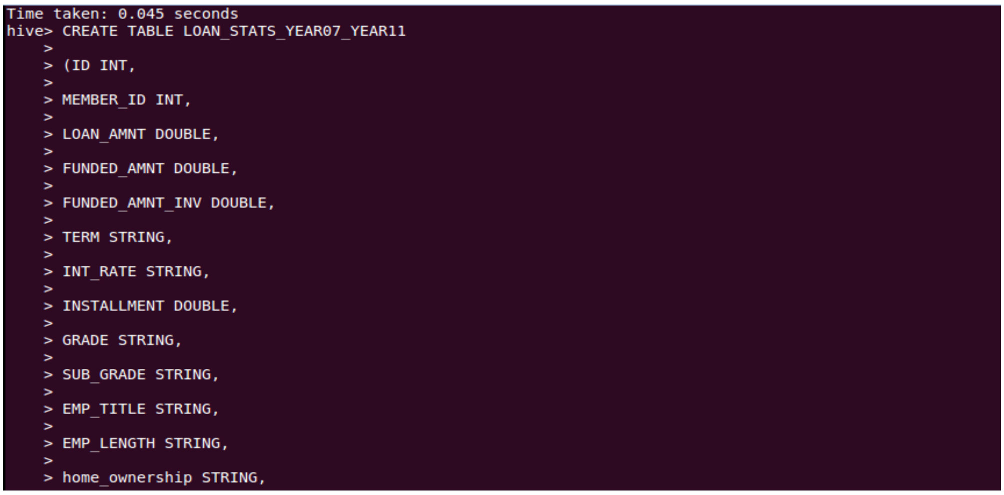
To copy dataset/file from Local file system to Hadoop file system i.e. HDFS (Hadoop Distributed File System)

Loading the dataset into created table in

we are pulling those Bank Customers whose Loans were processed successfully starting from year 2007 till 2013.They were proven as the best customers for banks as their payment schedule and other incomes were verified and made on timely basis. They were provided Loan at ROI = 7% and for 3years duration.

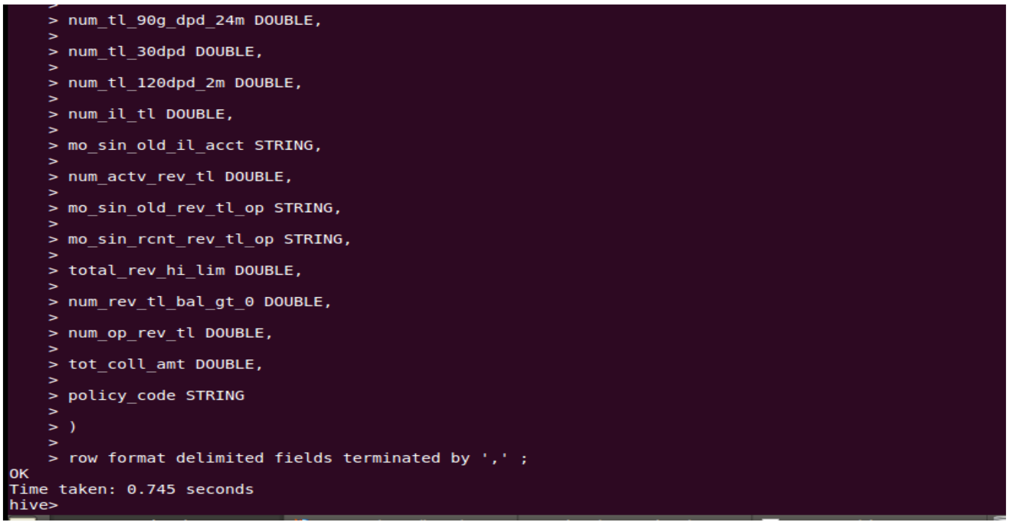
FIG 4. OUTLINE OF PRESENT WORK

V. RESULTS



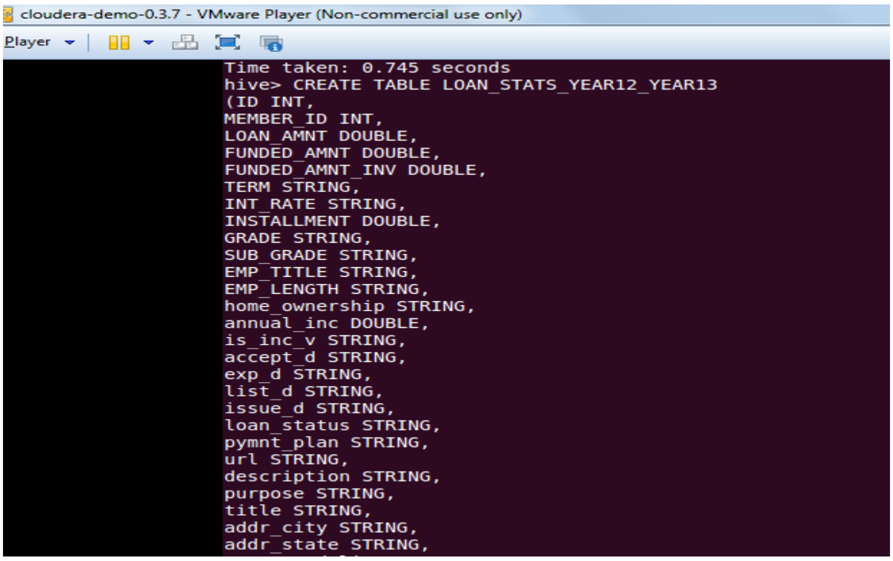
```
Time taken: 0.045 seconds
hive> CREATE TABLE LOAN_STATS_YEAR07_YEAR11
    >
    > (ID INT,
    >
    > MEMBER_ID INT,
    >
    > LOAN_AMNT DOUBLE,
    >
    > FUNDED_AMNT DOUBLE,
    >
    > FUNDED_AMNT_INV DOUBLE,
    >
    > TERM STRING,
    >
    > INT_RATE STRING,
    >
    > INSTALLMENT DOUBLE,
    >
    > GRADE STRING,
    >
    > SUB_GRADE STRING,
    >
    > EMP_TITLE STRING,
    >
    > EMP_LENGTH STRING,
    >
    > home_ownership STRING,
```

FIG 5. CREATE TABLE LOAN_STATS_YEAR07_YEAR11.



```
    > num_tl_90g_dpd_24m DOUBLE,
    >
    > num_tl_30dpd DOUBLE,
    >
    > num_tl_120dpd_2m DOUBLE,
    >
    > num_il_tl DOUBLE,
    >
    > mo_sin_old_il_acct STRING,
    >
    > num_actv_rev_tl DOUBLE,
    >
    > mo_sin_old_rev_tl_op STRING,
    >
    > mo_sin_rcnt_rev_tl_op STRING,
    >
    > total_rev_hi_lim DOUBLE,
    >
    > num_rev_tl_bal_gt_0 DOUBLE,
    >
    > num_op_rev_tl DOUBLE,
    >
    > tot_coll_amt DOUBLE,
    >
    > policy_code STRING
    >
    > )
    >
    > row format delimited fields terminated by ',' ;
OK
Time taken: 0.745 seconds
hive>
```

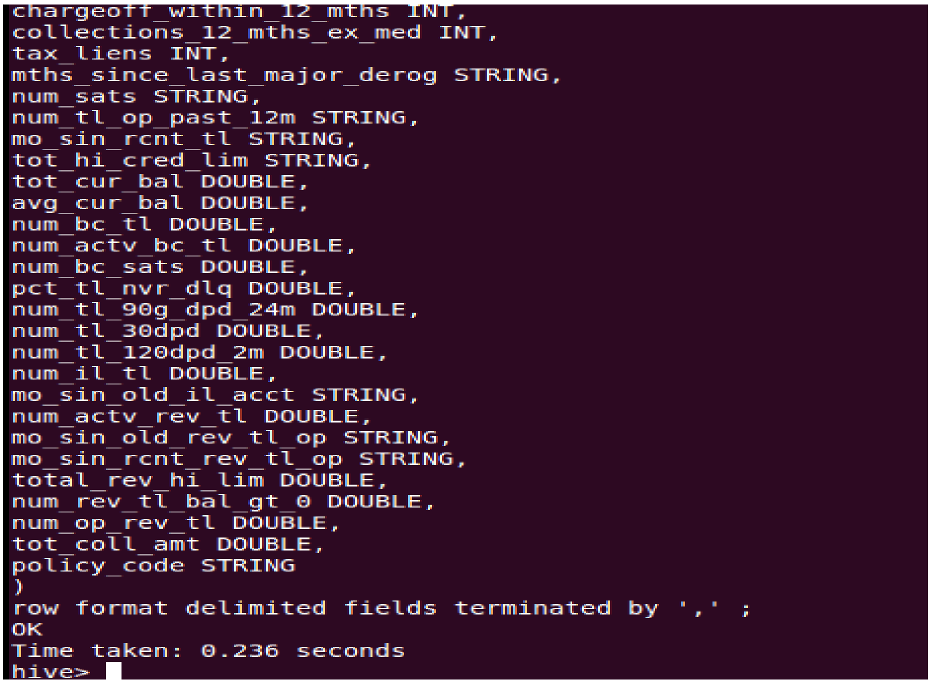
FIG 6. CREATE TABLE LOAN_STATS_YEAR07_YEAR11



cloudera-demo-0.3.7 - VMware Player (Non-commercial use only)
Player

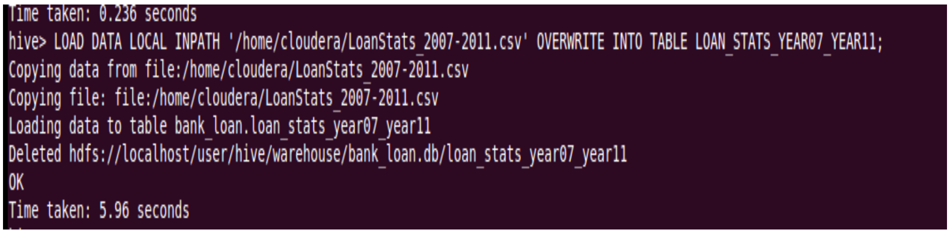
```
Time taken: 0.745 seconds
hive> CREATE TABLE LOAN_STATS_YEAR12_YEAR13
(ID INT,
MEMBER_ID INT,
LOAN_AMNT DOUBLE,
FUNDED_AMNT DOUBLE,
FUNDED_AMNT_INV DOUBLE,
TERM STRING,
INT_RATE STRING,
INSTALLMENT DOUBLE,
GRADE STRING,
SUB_GRADE STRING,
EMP_TITLE STRING,
EMP_LENGTH STRING,
home_ownership STRING,
annual_inc DOUBLE,
is_inc_v STRING,
accept_d STRING,
exp_d STRING,
list_d STRING,
issue_d STRING,
loan_status STRING,
pymnt_plan STRING,
url STRING,
description STRING,
purpose STRING,
title STRING,
addr_city STRING,
addr_state STRING,
```

FIG 7. CREATE TABLE LOAN_STATS_YEAR112_YEAR13



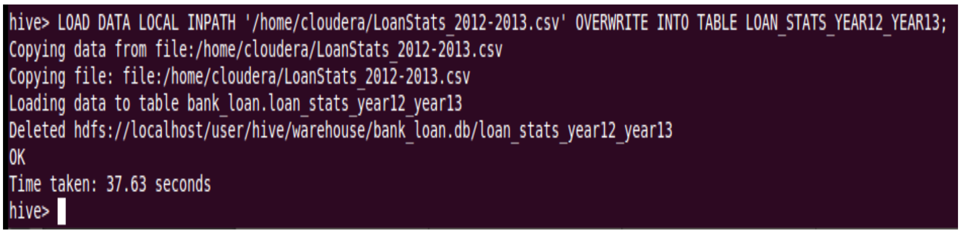
```
chargeoff_within_12_mths INT,
collections_12_mths_ex_med INT,
tax_liens INT,
mths_since_last_major_derog STRING,
num_sats STRING,
num_tl_op_past_12m STRING,
mo_sin_rcnt_tl STRING,
tot_hi_cred_lim STRING,
tot_cur_bal DOUBLE,
avg_cur_bal DOUBLE,
num_bc_tl DOUBLE,
num_actv_bc_tl DOUBLE,
num_bc_sats DOUBLE,
pct_tl_nvr_dlq DOUBLE,
num_tl_90g_dpd_24m DOUBLE,
num_tl_30dpd DOUBLE,
num_tl_120dpd_2m DOUBLE,
num_il_tl DOUBLE,
mo_sin_old_il_acct STRING,
num_actv_rev_tl DOUBLE,
mo_sin_old_rev_tl_op STRING,
mo_sin_rcnt_rev_tl_op STRING,
total_rev_hi_lim DOUBLE,
num_rev_tl_bal_gt_0 DOUBLE,
num_op_rev_tl DOUBLE,
tot_coll_amt DOUBLE,
policy_code STRING
)
row format delimited fields terminated by ',' ;
OK
Time taken: 0.236 seconds
hive>
```

FIG 8. CREATE TABLE LOAN_STATS_YEAR12_YEAR13



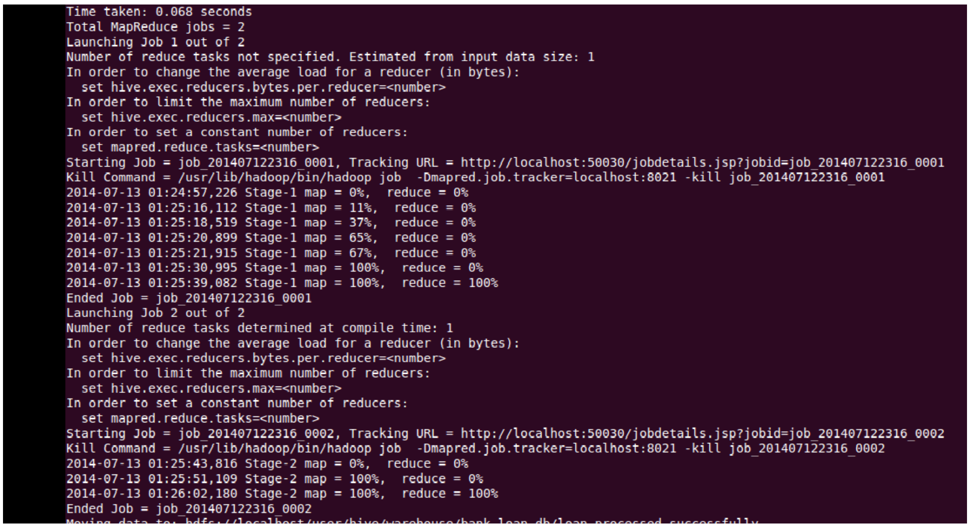
```
Time taken: 0.236 seconds
hive> LOAD DATA LOCAL INPATH '/home/cloudera/LoanStats_2007-2011.csv' OVERWRITE INTO TABLE LOAN_STATS_YEAR07_YEAR11;
Copying data from file:/home/cloudera/LoanStats_2007-2011.csv
Copying file: file:/home/cloudera/LoanStats_2007-2011.csv
Loading data to table bank_loan.loan_stats_year07_year11
Deleted hdfs://localhost/user/hive/warehouse/bank_loan.db/loan_stats_year07_year11
OK
Time taken: 5.96 seconds
```

FIG 9. LOADING DATA INTO TABLE "LOAN_STATS_YEAR07_YEAR11



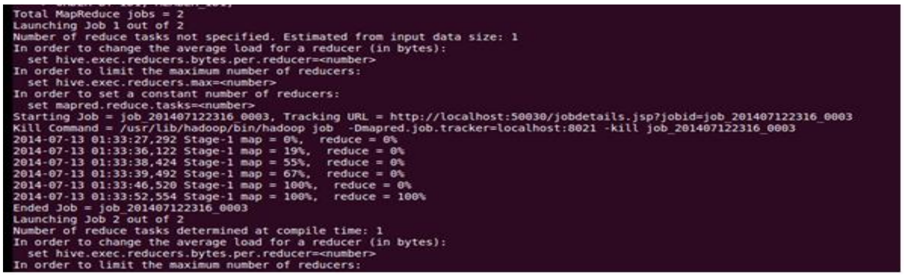
```
hive> LOAD DATA LOCAL INPATH '/home/cloudera/LoanStats_2012-2013.csv' OVERWRITE INTO TABLE LOAN_STATS_YEAR12_YEAR13;
Copying data from file:/home/cloudera/LoanStats_2012-2013.csv
Copying file: file:/home/cloudera/LoanStats_2012-2013.csv
Loading data to table bank_loan.loan_stats_year12_year13
Deleted hdfs://localhost/user/hive/warehouse/bank_loan.db/loan_stats_year12_year13
OK
Time taken: 37.63 seconds
hive>
```

FIG 10. LOADING DATA INTO TABLE "LOAN_STATS_YEAR12_YEAR13"



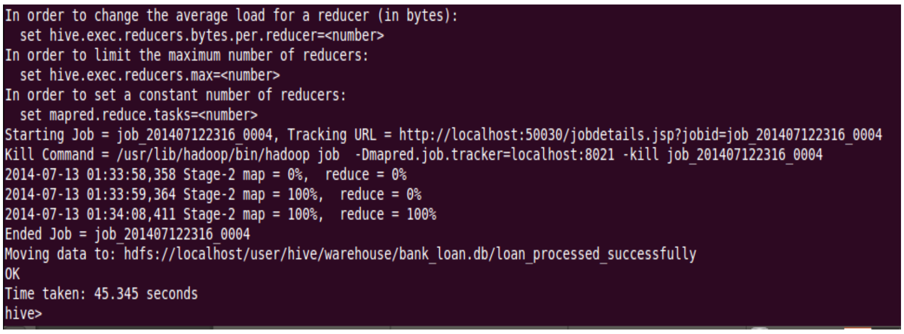
```
Time taken: 0.068 seconds
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number 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_201407122316_0001, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201407122316_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201407122316_0001
2014-07-13 01:24:57,226 Stage-1 map = 0%,  reduce = 0%
2014-07-13 01:25:16,112 Stage-1 map = 11%,  reduce = 0%
2014-07-13 01:25:18,519 Stage-1 map = 37%,  reduce = 0%
2014-07-13 01:25:20,899 Stage-1 map = 65%,  reduce = 0%
2014-07-13 01:25:21,915 Stage-1 map = 67%,  reduce = 0%
2014-07-13 01:25:30,995 Stage-1 map = 100%,  reduce = 0%
2014-07-13 01:25:39,082 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201407122316_0001
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 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_201407122316_0002, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201407122316_0002
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201407122316_0002
2014-07-13 01:25:43,816 Stage-2 map = 0%,  reduce = 0%
2014-07-13 01:25:51,109 Stage-2 map = 100%,  reduce = 0%
2014-07-13 01:26:02,180 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_201407122316_0002
Moving data to: hdfs://localhost/user/hive/warehouse/bank_loan.db/loan_processed_successfully
```

FIG 11. EXECUTION



```
Total MapReduce jobs = 2
Launching Job 1 out of 2
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_201407122316_0003, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201407122316_0003
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201407122316_0003
2014-07-13 01:33:27,292 Stage-1 map = 0%,  reduce = 0%
2014-07-13 01:33:36,122 Stage-1 map = 19%,  reduce = 0%
2014-07-13 01:33:38,424 Stage-1 map = 55%,  reduce = 0%
2014-07-13 01:33:39,492 Stage-1 map = 67%,  reduce = 0%
2014-07-13 01:33:46,520 Stage-1 map = 100%,  reduce = 0%
2014-07-13 01:33:52,554 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201407122316_0003
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 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:
```

FIG 12. EXECUTION



```
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_201407122316_0004, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201407122316_0004
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201407122316_0004
2014-07-13 01:33:58,358 Stage-2 map = 0%,  reduce = 0%
2014-07-13 01:33:59,364 Stage-2 map = 100%,  reduce = 0%
2014-07-13 01:34:08,411 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_201407122316_0004
Moving data to: hdfs://localhost/user/hive/warehouse/bank_loan.db/loan_processed_successfully
OK
Time taken: 45.345 seconds
hive>
```

FIG 13. EXECUTION

VI. CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

In this present work, Data mining is performed by using the Hadoop Ecosystem Approach. The presented work is about to performed data mining on large loan data sets using Hive component of Hadoop Ecosystem. Here, the hive queries are performed for mining the useful data like "we are pulling those Bank Customers whose Loans were processed successfully starting from year 2007 till 2013.They were proven as the best customers for banks as their payment schedule and other incomes were verified and made on timely basis. They were provided Loan at ROI = 7% and for 3years duration." This information can be mined in less time because of parallelization feature of Hadoop ecosystem.

So from our data analytics we conclude that those customers are best market for Banks in future and can be given priority over other customers. Companies can create separate operational data store (ODS) to make inventory of those customers for faster search & processing of loan.

6.2 Future Work

In this present work, we have performed data mining on large structured data (big data) by executing hive queries on hive component of Hadoop Ecosystem.

Our future work will focus on Analysis or mining of unstructured data like images, audios, videos, graphs using map reduce component of Hadoop Ecosystem.

Obtained results from the system shows that the effective Data mining is been performed..

REFERENCES

[1] D.Abadi,"Data management in the cloud: Limitations and opportunities,"*IEEE Transactions on Data Engineering, Vol.32, No.1, March 2009.*

[2] Y. Xu, P. Kostamaa, and L. Gao. "Integrating Hadoop and Parallel DBMS", *Proceedings of ACM SIGMOD, International conference on Data Management, New york, NY, USA 2010.*

[3] Farah Habib Chan chary,"Data Migration: Connecting databases in the cloud" *IEEE JOURNAL ON COMPUTER SCIENCE, vol no-40, page no-450-455, MARCH 2012.*

[4] Kamil Bajda at el."Efficient processing of data warehousing queries in a split execution environment", *JOURNAL ON DATA WAREHOUSING, vol no-35, ACM, JUNE 2011.*

[5] Savvas, A. (2011, October 25). IBM: Businesses unable to analyze 90 percent of their data. ComputerworldUK. Retrieved August 9, 2012.

[6] Pettey, C., & Goasduff, L. (2011, June 27). Gartner Says Solving “Big Data” Challenge Involves More Than Just Managing Volumes of Data. Stamford: Gartner.

[7] Russom, P. (2011). Big Data Analytics. TDWI Research.

[8] FUJIMAKI Ryohei, MORINAGA Satoshi**,”** The Most Advanced Data Mining of the Big Data Era”, *NEC TECHNICAL JOURNAL Vol.7 No.2/2012FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] Pavlo, A., Paulson, E., Rasin, A., Abadi, D.J., DeWitt, D.J., Madden, S., Stonebraker, “A Comparison of Approaches to Large-Scale Data Analysis”, In: SIGMOD '09.

[10] Das et al., 2010, Das S., Sismanis Y., Beyer K.S., Gemulla R., Haas P.J., McPherson J., Ricardo: Integrating R and Hadoop, *In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data (SIGMOD '10), 2010, pp. 987-998*

[11] IBM 2012, what is big data: Bring big data to the enterprise, http://www-01.ibm.com/software/data/bigdata/, IBM.

[12] Labrinidis and Jagadish 2012, A. Labrinidis and H. Jagadish, Challenges and Opportunities with Big Data, In *Proc. of the VLDB Endowment, 5(12):2032-2033, 2012*

[13] Wu X. and Zhu X. 2008, Mining with Noise Knowledge: Error-Aware Data Mining, *IEEE Transactions on Systems, Man and Cybernetics, Part A, vol.38, no.4, pp.917-932.*

[14] Badrul Chowdhury, Tilmann Rabl, Pooya Saadatpanah,” A BigBench Implementation in the Hadoop Ecosystem” available at http://msrg.org

[15] Wei Fan, Albert Bifet,” Mining Big Data: Current Status, and Forecast to the Future” *SIGKDD Explorations Volume 14, Issue 2*

[16] Kalyani M Raval**, “** Data Mining Techniques”, *International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 10, October 2012*

[17] Jim (Zhanwen) Li1, Siyuan He2, Liming Zhu1,3,Xiwei Xu,” Challenges to Error Diagnosis in Hadoop Ecosystems”**,**School of Computer Science and Engineering, University of New South Wales, Sydney, Australia