International Journal of Research in Advent Technology, Vol.6, No.5, May 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

Comparative Analysis of Pig and Hive

Shruti Verma[1], Vinod Maan[2]
CSE, College of Engineeringand Technology,Mody University of Science and Technology
Email: verma.shruti500@gmail.com[1] ,vinodmaan@modyuniversity.ac.in[2]

Abstract— Today the size or volume,complexity, variety,rate of growth or veracity of data which organizations handled have reached such unbelievable level that traditional processing and analytical tools failed to process. Big Data is ever growing and can't be determined with respect to its size. To analyze this huge amount of data ,Hadoop can be used. Hadoop is nothing but a framework that is used for processing of large data sets across different clusters.The Tools used to handle this enormous amount of data are Hadoop, Map Reduce, Apache Hive, No SQL etc. Information extraction has recently received significant attention due to the rapid growth of unstructured text data. However, this is computationally intensive and MapReduce and parallel database management systems have been used to analyze large amounts of data. In this paper we familarize big data tools that are used for analysis that is Apache hive and Apache pig .Also, comparison of hive and pig is accomplished on some parameters and during analysis it shows that hive perform better than pig.

Keywords: BigData, Hadoop, HDFS, YARN, Hadoop Ecosystem;

1. INTRODUCTION

The measure of content information develops each day on the Internet, for instance via web-based networking media, news articles or pages. Be that as it may, this information is generally unstructured and its convenience is constrained. In this way, data extraction (IE) is acquainted all together with increment the value of unstructured content. In any case, performing IE assignments is computationally concentrated and MapReduce and parallel database administration frameworks have been utilized to break down a lot of information. A typical method to process extensive arrangements of information is utilizing Apache Hadoop. Hadoop is a Java-executed system that takes into account the appropriated preparing of huge informational indexes crosswise over bunches of PCs. Since composing MapReduce occupations in Java can be difficult, Hive and Pig has been created and functions as stages over Hadoop. Hive and Pig permits clients simple access to information contrasted with executing their own particular MapReduce in Hadoop.The humming word "Bigdata examination" would thus be able to be portrayed as investigation of datasets utilizing diverse investigation strategies.

Data management, processing and storing processes are becoming more difficult with the increased use of digital technology. Because the amount of data increases day by day on the world. This result many company looked for a solution to solve of regarding processes on petabytes of data. The problems are often repeated that the big data problems are that relational databases cannot scale to process the massive volumes of data. The traditional systems are not enough for this solution. In these day Hadoop is often used for data-intensive computing.

1.2. Hadoop

Hadoop is a framework that allows for distributed processing of lasrge data sets across clusters of commodity computers using simple programming models . It is inspired by technical document published by Google.The word hadoop does not have any meaning Doug Cutting discovered Hadoop and named it after his son's yellow-colored toy elephant .

1.3. Hadoop Distributed File System

Traditionally data was stored in a central location and it was sent to the process directly at run time,this method worked well for a limited data however modern systems receive terabytes of data per day and its is difficult for traditional computers for Relational database management system to push high volumes of data to the process.
Hadoop brought a radical approach ,in hadoop the program goes to the data not vice versa ,it initially distributes the data to multiple systems and later runs the computation to wherever the data is located .The HDFS components comprise different servers like NameNode, DataNode, and Secondary NameNode.

- NameNode Server: NameNode and the Secondary NameNode services constitute the master service. The master service is responsible for accepting a job from clients and ensures that the data required for the operation will be loaded and segregated into chunks of data blocks. NameNode knows the data nodes on which all the blocks for a given file exist.

- DataNode Server: Associated with data storage places in the file system and reports to NameNode periodically with lists of blocks they store .It Stores and retrieves blocks when referred by clients or NameNode . Servers read, write requests, performs block creation, deletion, and replication upon instruction from NameNode .
- Secondary NameNode Server: Used for recovery of NameNode in case of NameNode failure.

The key feature of Hadoop HDFS is:

It provides high-throughput access to data blocks .It provides limited interface for managing the file system to allow it to scale and creates multiple replicas of each data block and distributes them on computers throughout the cluster to enable reliable and rapid data access.

2. SYSTEM ARCHITECTURE

The system architecture for performing the comparative study between Pig and Hive is demonstrated in Fig.2.1.

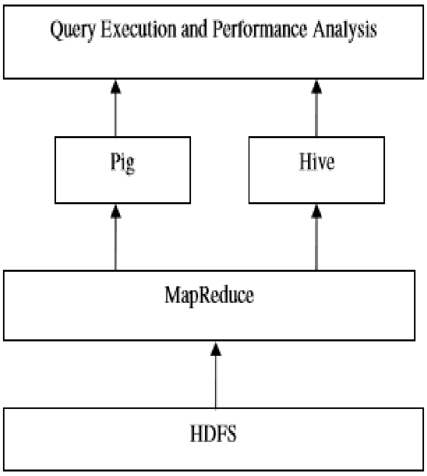


Fig. 2.1 System Architecture

In order to perform the comparative study of both the targeted technologies, a Big Data environment is required to develop first. Keeping in mind the end goal to play out the near investigation of both the focused on innovations, a Big Data condition is required to grow first.The proposed comparative performance study platform is developed using the Hadoop and MapReduce technology. Hadoop is basically a storage technology that scales self for storing huge amount of data as required by the application. Additionally the MapReduce framework provides support to reduce and map the data for the data analytics.

Keeping in mind the end goal to play out the near investigation of both the focused on innovations, a Big Data condition is required to grow first
Along these lines, the info information is as a matter of first importance facilitated over the Hadoop vault and after that utilizing the MapReduce structure the information is prepared in Pig and Hive frameworks. The charge line interface is utilized to make inquiries on the information over Pig and Hive with the comparable dataset and the comparative question one by one. In the wake of handling of information and execution of client questions over both the situations the measure of time is evaluated as execution examination of the framework .

2.1. Apache Pig

It is an open-source high level dataflow system developed by yahoo.It is mainly used for analytics.It convert pig scripts to Map-Reduce code thus saving user from writing complex Map-Reduce program.In pig we can write simple queries,these scripts are converted in to mapreduce jobs executed on hadoop by pig, here we don’t have to write map reduce code.

In map reduce, we can write code in java and execute the map reduce code but since everybody is not having the knowledge of java and at the same time we also need some high level language so that we can just get the way of writing the coding and just focus on the logic and write the program, this is why we need Pig. Pig is a high level language. Also,if we are thorough with java then also we have to write lines of code for analyzing data.

2.1.1. Features Of Pig

- Extensible
- Self optimizing
- Easily Programmable

2.1.2. Apache Pig Architecture

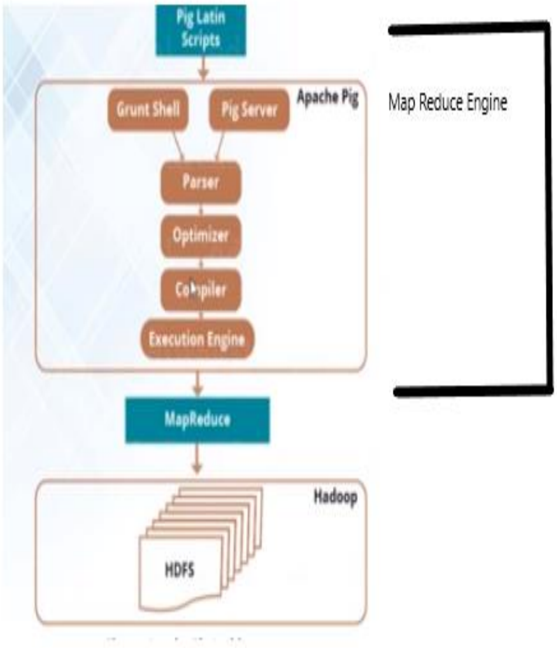


Fig. 2.2 Apache Pig Architecture

We have *pig latin scripts* as shown in Fig.2.2 ,scripts can be executed on *grunt shell* that is the native shell provided by apache pig or we can also submit scripts to the java client on *pig server* ,once these scripts are submitted to the apache pig.It is parsed by the *parser*, optimized by the *optimizer* and *compiler* compiled and converts these scripts to map reduce code and is executed on *Execution Engine* using runtime engine over the hadoop cluster.

2.1.3. Components of Pig

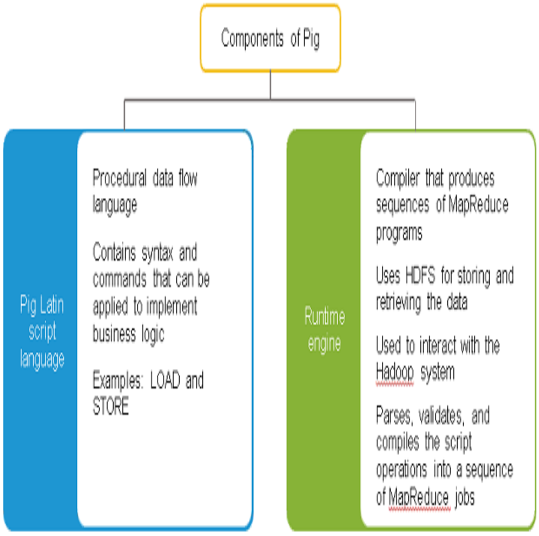

Components of Pig
Pig Latin script language
Procedural data flow language
Contains syntax and commands that can be applied to implement business logic
Examples: LOAD and STORE
Runtime engine
Compiler that produces sequences of MapReduce programs
Uses HDFS for storing and retrieving the data
Used to interact with the Hadoop system
Parses, validates, and compiles the script operations into a sequence of MapReduce jobs

Fig. 2.3 Components of Pig

Pig Latin : Pig latin is language in which we are going to write the code.Earlier, we are writing the code in Java language,once we write the code in pig latin language there is a runtime engine as shown in Fig.2.3, which will convert these scripts to map reduce jobs which are going to run on hadoop system .

Grunt :Its nothing but a shell on which we will going to write our code.

2.1.4. Pig Execution Modes

There are mainly two types of mode in Pig .
Map Reduce Mode: In this mode , Instructions which are executed that will be distributed and perform analysis with large datasets . with large data sets we would prefer Mapreduce mode.The input and output in this mode are present on HDFS.
Command: pig

Local Mode: If we are doing the development then definitely we would like to write the code on sample,we can write our code and scripts in this local mode.the input and output is present on local file system
Command: pig –x local

2.1.5. Pig Increses productivity

10 lines of pig Latin = 200 lines of java
4 hours to write in java = 20 min to write pig Latin

2.1.6. How to Access Pig

There are three ways to access pig tool:

- Grunt:It is a shell we can write our statements in shell.
- Pig Script:We can write pig scripts, it is similar to shell only .we can write multiple statements in a file together,save it with .pig extension and execute that file .
- Via JAVA:there are wasy we can use pig scripts with java

2.1.7. Working of Pig

1. Load data and write pig scripts .
2. Parse the scripts,check the operations and optimize the scripts then execute the statements and submit to hadoop.
3. Results are dumped or stored in HDFS .

2.2. *Hive*

Data warehousing package built on the top of Hadoop.It is used for data analysis[6].Oracle was not able to scale in terms of requirement of the data so

Facebook comes up with a solution and became the early adopter of hadoop platform. Facebook faces many challenges like users are more than 950 million and generate data greater than 500 TB per day.

People used to upload approximately 300 photos per day .So traditional Rdbms is not suitable for such kind of data. In hadoop, there are users who are very good with Structured Query language. So, It was tough to write code in java language and Facebook came up with SQL approach called Hive .Hive provides the sql kind of interface by using that user can write queries and analyze the data . In hive, we can create tables, partitions, schema flexibility .We can write our own custom code in hive.

After congregating the data into HDFS they are analyzed by queries using Hive. Apache Hivedata warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL.

2.2.1. Hive Architecture

In case of hive, data is going to store in hadoop file system .we have the metastore in hive,when we create database,tables and views all those definitions are stored in meta store as shown in Fig 1.3.For hive, there is derby database as metastore is going to store in derby database. Hive supports all the user defined functions.We can store the data in text file,Rc file ,csv file etc.

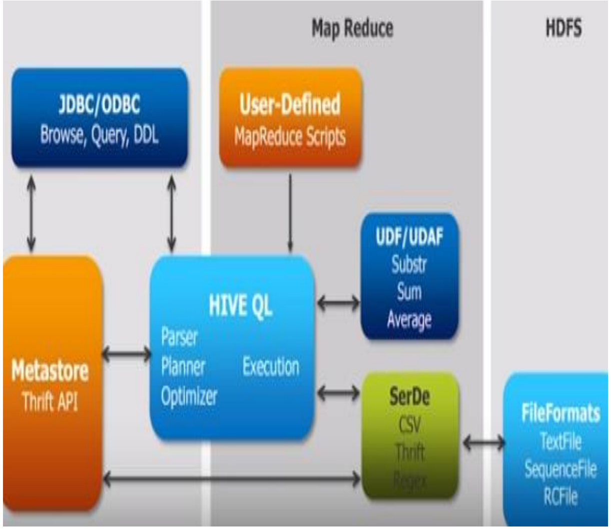


Fig. 2.4 Hive Architecture

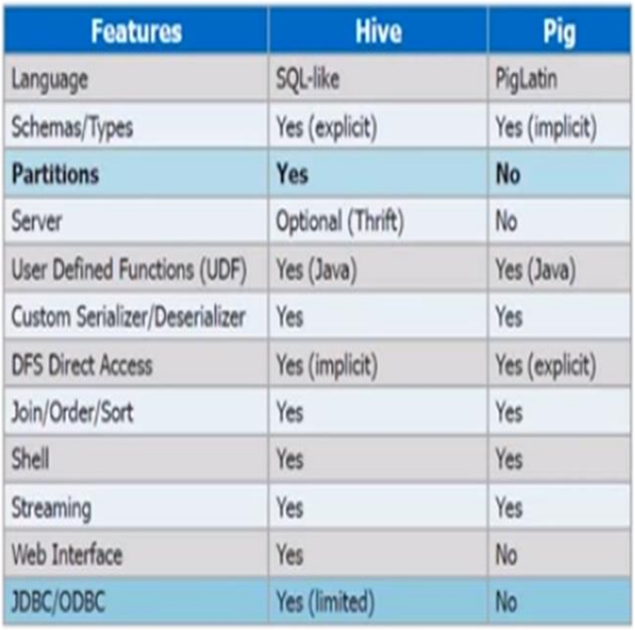
2.2.2. Components of Hive

- *Meta Store*:It stores the meta data about the hive,table definitions, view definitions .
- *Shell*: On shell Hive queries are to be written.
- *Driver*: After submitting the query .Take code and convert it in to a code which hadoop can understand easily.
- *Compiler*:Code is compiled by the compiler .
- *Execution Engine* :It processes the query and generates results as same as MapReduce results.

2.2.3. Working of Hive

- *Execute Query*: The Hive interface sends query to the Driver to execute.
- *Get Plan*: The driver takes the help of query compiler that parses the query to check the syntax and query plan or the requirement of query.
- *Get Metadata*: Metadata came into picture and compiler sends metadata request to Metastore
- *Fetch Result*: After compiling the queries, the execution engine receives the results from Data nodes.
- *Send Results*: The driver sends the results to Hive Interfaces.

3. COMPARISION



Features	Hive	Pig
Language	SQL-like	PigLatin
Schemas/Types	Yes (explicit)	Yes (implicit)
Partitions	Yes	No
Server	Optional (Thrift)	No
User Defined Functions (UDF)	Yes (Java)	Yes (Java)
Custom Serializer/Deserializer	Yes	Yes
DFS Direct Access	Yes (implicit)	Yes (explicit)
Join/Order/Sort	Yes	Yes
Shell	Yes	Yes
Streaming	Yes	Yes
Web Interface	Yes	No
JDBC/ODBC	Yes (limited)	No

Fig. 3.1 Comparision Of Hive and Pig

4. PERFORMANCE ANALYSIS

I have performed the experiment on single system . My machine is having 4GB RAM and has intel core i5 processor.

After setting up the experimental environment, the queries that are listed in Table 3.1 are executed on both PIG and Hive query interfaces and their performance in terms of query execution time is evaluated .

For this we have taken the Consumer Complaints dataset . A shopper dissension can be comprehended as an announcement of disappointment towards an item or an administration. Then again, customer security includes the laws that give purchasers the privilege to enlist their disappointment about any harsh or second rate business hones and that get a sensible determination. Consumer reports are a collection of data about different products and services through reviews and comparisons Purchaser reports are a gathering of information about various items and administrations through surveys and correlations. Such reports help in dissecting the great and the awful side of an item or a service..For analysis consumer complaints datasets we need:-

1. Dataset
   We can collect the consumers complaints dataset, that collectively holds large number of complaints records and opinions.
2. Hadoop
   Hadoop should be configured first as all the mapreduce job will work on hadoop framework, also hadoop comprises of HDFS (hadoop distributed file system) which is used to store such large datasets and mapreduce is used to process these datasets.
3. Bigdata
   Analytical Tools For analyzing these large amount of data we need efficient analytical tools which work on the top of hadoop, apache hive and apache pig through which we can analyze the consumer complaints datasets.

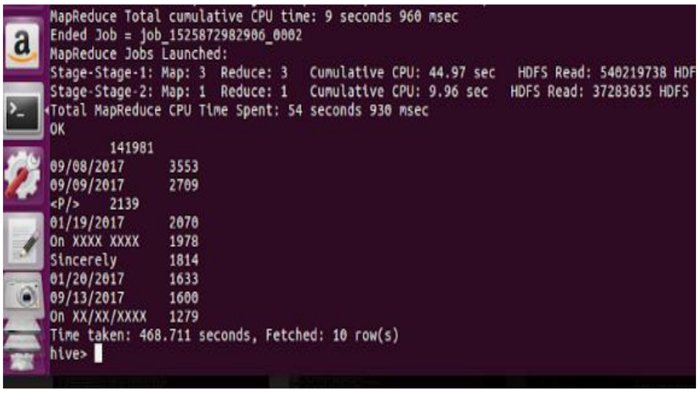
4.1. Experimentation with Hive

The amount of time consumed during input a user query for finding records from the Hive technique is termed here as the query execution time. In order to measure the query execution time, below listed queries are fired on the Hive interface and their performance is observed. After completing the observations first time for all the queries.

Table 4.1 Queries

S.no	Query
1.	Top ten days when Maximum number of complaints registered.
2.	Top twenty issue faced by maximum number of consumers.
3.	Top twenty company on which maximum complaints registered
4.	Top twenty Issue along with company

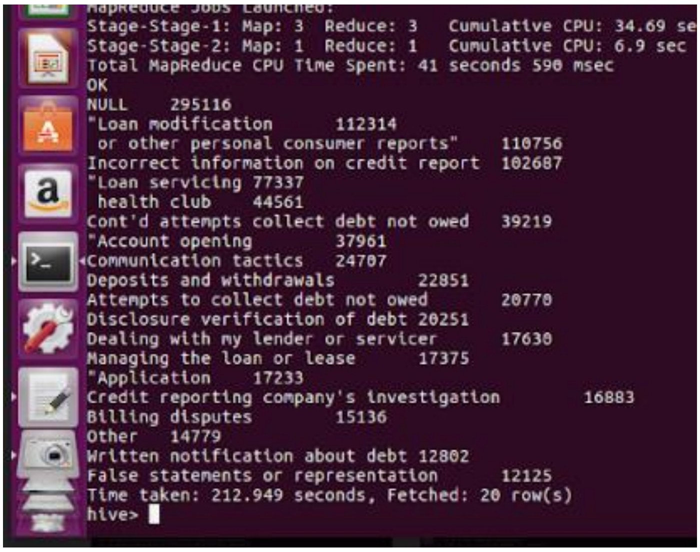
Query 1 :select date_received,count(*) as a from Consumercomplaint group by date_received order by a desc limit 10;



```
MapReduce Total cumulative CPU time: 9 seconds 960 msec
Ended Job = job_1525872982906_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 3   Cumulative CPU: 44.97 sec   HDFS Read: 540219738 HDF
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 9.96 sec   HDFS Read: 37283635 HDFS
Total MapReduce CPU Time Spent: 54 seconds 930 msec
OK
        141981
09/08/2017      3553
09/09/2017      2709
<P/>    2139
01/19/2017      2070
On XXXX XXXX    1978
Sincerely       1814
01/20/2017      1633
09/13/2017      1600
On XX/XX/XXXX   1279
Time taken: 468.711 seconds, Fetched: 10 row(s)
hive>
```

Fig. 4.1 Result

Query 2: select issue,count(*) as a from consumercomplaint group by Issue order by a desc limit 20;

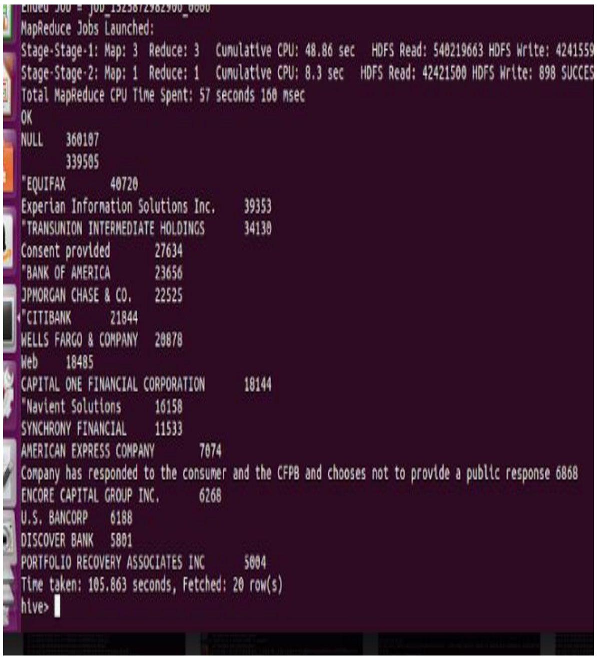


```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 3   Cumulative CPU: 34.69 se
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 6.9 sec
Total MapReduce CPU Time Spent: 41 seconds 590 msec
OK
NULL    295116
"Loan modification      112314
 or other personal consumer reports"    110756
Incorrect information on credit report  102687
"Loan servicing 77337
 health club    44561
Cont'd attempts collect debt not owed   39219
"Account opening        37961
Communication tactics   24707
Deposits and withdrawals        22851
Attempts to collect debt not owed       20770
Disclosure verification of debt 20251
Dealing with my lender or servicer      17630
Managing the loan or lease      17375
"Application    17233
Credit reporting company's investigation        16883
Billing disputes        15136
Other   14779
Written notification about debt 12802
False statements or representation      12125
Time taken: 212.949 seconds, Fetched: 20 row(s)
hive>
```

Fig. 4.2 Result

Query 3

select Company,count(*) as a from Consumercomplaint group by Company order by a desc limit 20;



```
Ended Job = job_1525872982906_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 3  Reduce: 3   Cumulative CPU: 48.86 sec   HDFS Read: 540219663 HDFS Write: 4241559
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 8.3 sec   HDFS Read: 42421500 HDFS Write: 898 SUCCES
Total MapReduce CPU Time Spent: 57 seconds 160 msec
OK
NULL    360107
        339505
"EQUIFAX        40720
Experian Information Solutions Inc.     39353
"TRANSUNION INTERMEDIATE HOLDINGS       34130
Consent provided        27634
"BANK OF AMERICA        23656
JPMORGAN CHASE & CO.    22525
"CITIBANK       21844
WELLS FARGO & COMPANY   20878
Web     18485
CAPITAL ONE FINANCIAL CORPORATION       18144
"Navient Solutions      16158
SYNCHRONY FINANCIAL     11533
AMERICAN EXPRESS COMPANY        7074
Company has responded to the consumer and the CFPB and chooses not to provide a public response 6868
ENCORE CAPITAL GROUP INC.       6268
U.S. BANCORP    6188
DISCOVER BANK   5801
PORTFOLIO RECOVERY ASSOCIATES INC       5004
Time taken: 105.863 seconds, Fetched: 20 row(s)
hive> 
```

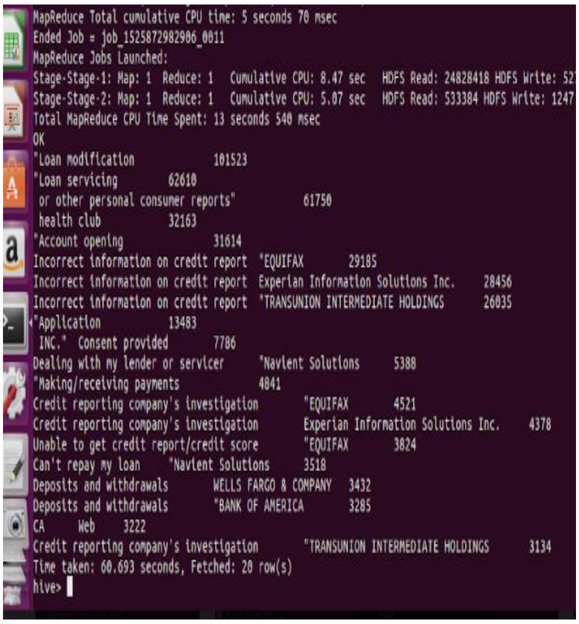
Fig. 4.3 Result

Query 4

create table new as select Company,count(*) as a from Consumercomplaint group by Company order by a desc limit 20;

create table new1 as select Issue,Company from Consumercomplaint where Company in (select Company from new where Company is not null);

select Issue,Company,count(*) as a from new1 group by Issue,Company order by a desc limit 20;



```
MapReduce Total cumulative CPU time: 5 seconds 70 msec
Ended Job = job_1525872982906_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.47 sec   HDFS Read: 24828418 HDFS Write: 52
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.07 sec   HDFS Read: 533384 HDFS Write: 1247
Total MapReduce CPU Time Spent: 13 seconds 540 msec
OK
"Loan modification              101523
"Loan servicing         62610
 or other personal consumer reports"            61750
 health club            32163
"Account opening                31614
Incorrect information on credit report  "EQUIFAX        29185
Incorrect information on credit report  Experian Information Solutions Inc.     28456
Incorrect information on credit report  "TRANSUNION INTERMEDIATE HOLDINGS       26035
"Application            13483
 INC."  Consent provided        7786
Dealing with my lender or servicer      "Navient Solutions      5388
"Making/receiving payments              4841
Credit reporting company's investigation        "EQUIFAX        4521
Credit reporting company's investigation        Experian Information Solutions Inc.     4378
Unable to get credit report/credit score        "EQUIFAX        3824
Can't repay my loan     "Navient Solutions      3518
Deposits and withdrawals        WELLS FARGO & COMPANY   3432
Deposits and withdrawals        "BANK OF AMERICA        3285
CA      Web     3222
Credit reporting company's investigation        "TRANSUNION INTERMEDIATE HOLDINGS       3134
Time taken: 60.693 seconds, Fetched: 20 row(s)
hive> 
```

Fig 4.4 Result

The amount of time consumed during input a user query for finding records:

Table 4.2 Execution Time taken by PIG

S.No.	Time Taken(min)
Query 1	7.811
Query 2	3.549
Query 3	1.764
Query 4	6.001

4.2. Experimentation with Pig

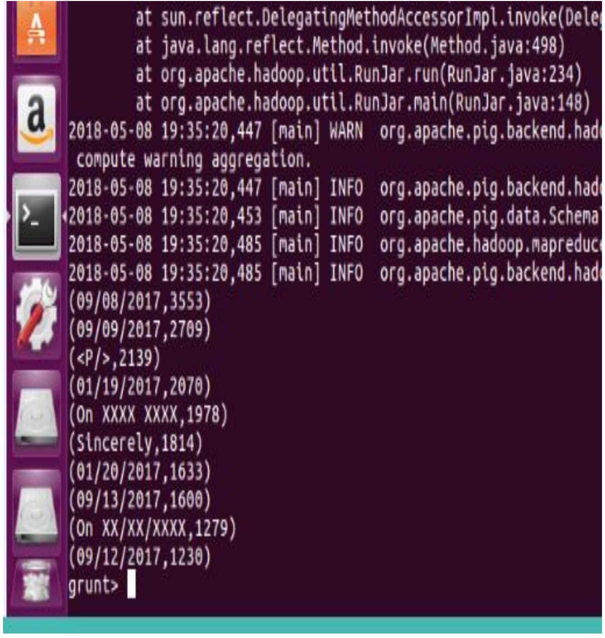
The time required to execute the user request by the user input query is termed as query execution time of Pig.

Table 4.3 Queries

S.no	Query
1.	Top ten days when Maximum number of complaints registered.
2.	Top twenty issue faced by maximum number of consumers.
3.	Top twenty company on which maximum complaints registered
4.	Top twenty Issue along with company

Query 1:

```
 A = Load the data using Pig Storage as (Schema);
 B = foreach A generate date_received as date;
 C = filter B by date is not null;
 D = group C by date; E = foreach D generate group,
COUNT(C.date);
 F = order E by $1 DESC;
 Result = LIMIT F 10;
 Dump Result;
```

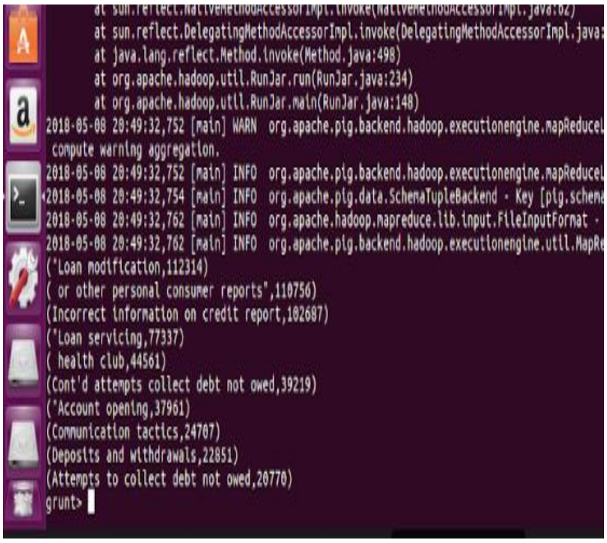


```
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(Dele
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:234)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
2018-05-08 19:35:20,447 [main] WARN  org.apache.pig.backend.had
 compute warning aggregation.
2018-05-08 19:35:20,447 [main] INFO  org.apache.pig.backend.had
2018-05-08 19:35:20,453 [main] INFO  org.apache.pig.data.Schema
2018-05-08 19:35:20,485 [main] INFO  org.apache.hadoop.mapreduc
2018-05-08 19:35:20,485 [main] INFO  org.apache.pig.backend.had
(09/08/2017,3553)
(09/09/2017,2709)
(<P/>,2139)
(01/19/2017,2070)
(On XXXX XXXX,1978)
(Sincerely,1814)
(01/20/2017,1633)
(09/13/2017,1600)
(On XX/XX/XXXX,1279)
(09/12/2017,1230)
grunt> 
```

Fig. 4.5 Result

Query 2:

```
G = foreach A generate Issue;
H= filter G by issue is not null;
I = group H by Issue;
 J = foreach  I generate group, COUNT(C.Issue);
K = order J by $1 DESC;
Result = LIMIT F 10;
Dump Result;
```

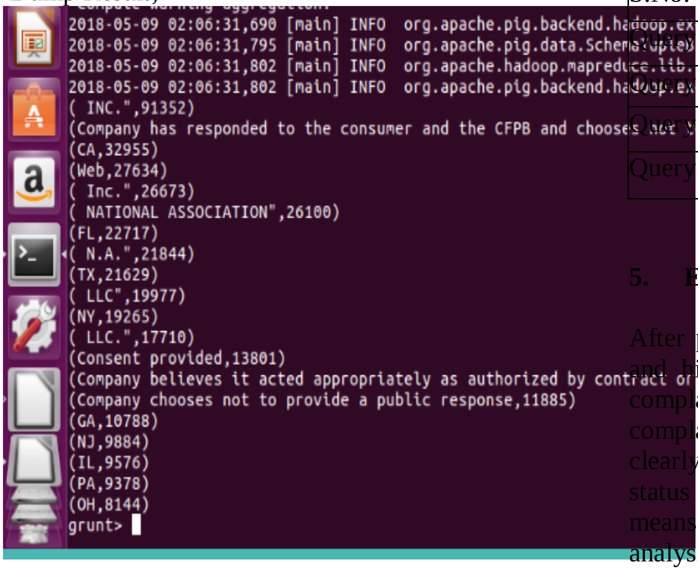


```
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:234)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
2018-05-08 20:49:32,752 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceL
 compute warning aggregation.
2018-05-08 20:49:32,752 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceL
2018-05-08 20:49:32,754 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schema
2018-05-08 20:49:32,762 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat -
2018-05-08 20:49:32,762 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRe
("Loan modification,112314)
( or other personal consumer reports",110756)
(Incorrect information on credit report,102687)
("Loan servicing,77337)
( health club,44561)
(Cont'd attempts collect debt not owed,39219)
("Account opening,37961)
(Communication tactics,24707)
(Deposits and withdrawals,22851)
(Attempts to collect debt not owed,20770)
grunt>
```

Fig. 4.6 Result

Query 3:

```
B = foreach  A generate Company;
C = filter B by Company is not null;
D = group C by Company;
E = foreach D generate group, COUNT(C.Company);
F = order E by $1 DESC;
Result = LIMIT  F  20;
 Dump Result;
```

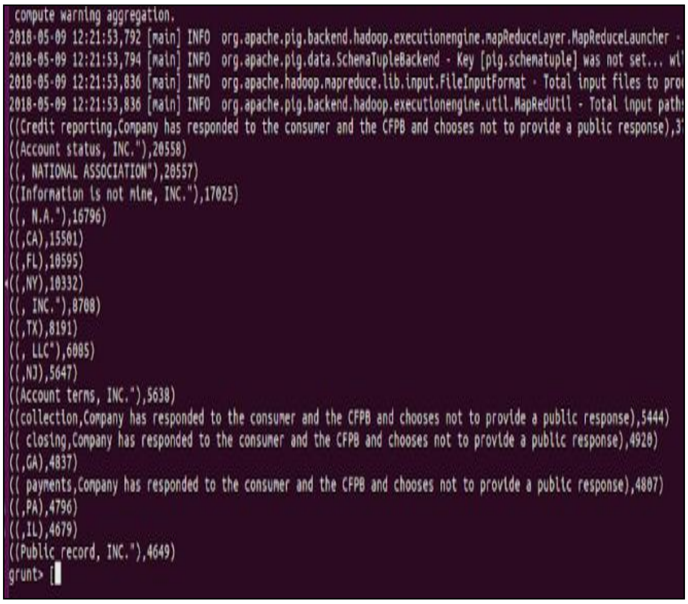


```
2018-05-09 02:06:31,690 [main] INFO  org.apache.pig.backend.hadoop.ex
2018-05-09 02:06:31,795 [main] INFO  org.apache.pig.data.Schem
2018-05-09 02:06:31,802 [main] INFO  org.apache.hadoop.mapreduce.lib.
2018-05-09 02:06:31,802 [main] INFO  org.apache.pig.backend.ha
( INC.",91352)
(Company has responded to the consumer and the CFPB and choose
(CA,32955)
(Web,27634)
( Inc.",26673)
( NATIONAL ASSOCIATION",26100)
(FL,22717)
( N.A.",21844)
(TX,21629)
( LLC",19977)
(NY,19265)
( LLC.",17710)
(Consent provided,13801)
(Company believes it acted appropriately as authorized by contract or
(Company chooses not to provide a public response,11885)
(GA,10788)
(NJ,9884)
(IL,9576)
(PA,9378)
(OH,8144)
grunt>
```

Fig. 4.7 Result

Query 4:

```
W = foreach A generate Issue,Company;
 X = group W by (Issue,Company);
Y = foreach X generate group, COUNT(W.Company);
Z = order Y by $1 DESC;
Final_result = LIMIT Z 20;
Dump Final_Result
```



```
 compute warning aggregation.
2018-05-09 12:21:53,792 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher -
2018-05-09 12:21:53,794 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... wi
2018-05-09 12:21:53,836 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to proc
2018-05-09 12:21:53,836 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input path
((Credit reporting,Company has responded to the consumer and the CFPB and chooses not to provide a public response),3
((Account status, INC."),20558)
((, NATIONAL ASSOCIATION"),20557)
((Information is not mine, INC."),17025)
((, N.A."),16796)
((,CA),15501)
((,FL),10595)
((,NY),10332)
((, INC."),8708)
((,TX),8191)
((, LLC"),6085)
((,NJ),5647)
((Account terms, INC."),5638)
((collection,Company has responded to the consumer and the CFPB and chooses not to provide a public response),5444)
(( closing,Company has responded to the consumer and the CFPB and chooses not to provide a public response),4920)
((,GA),4837)
(( payments,Company has responded to the consumer and the CFPB and chooses not to provide a public response),4807)
((,PA),4796)
((,IL),4679)
((Public record, INC."),4649)
grunt> [
```

Fig. 4.8 Result

The amount of time consumed during input a user query for finding records:

Table 4.3 Execution Time taken by PIG

S.No.	Time Taken(min)
Query 1	14
Query 2	12
Query 3	18
Query 4	21

5. EXPERIMENTAL RESULT ANALYSIS

After performing operations on the dataset using pig and hive, we can find the frequent issues, average complaints and the company on which maximum complaints registered, from the analysis result we can clearly examine the consumers need and the company status if the company having maximum complaints means company is not good at its services, so the analysis result can help industries, corporation and individual for taking any decision regarding company, issues and many things.

In our Experiment we likewise presented hive which is more helpful when contrasted with pig on examination of .csv datasets. We can state that hive perform speedier when contrasted with pig based on different parameters, additionally the above question comes about exhibit that the execution time taken by hive is less when contrasted with pig. What's more, the mapreduce occupations created by hive are less when compared with pig whereby the execution time is less in hive. Another advantage of utilizing hive is number of lines of code, which are more in pig however in hive just a single line of inquiry is adequate.

Table 5.1 Execution time taken by hive and pig

Execution Time Taken(in min)	**PIG**	**HIVE**
Query 1	14	7.811
Query 2	12	3.549
Query 3	18	1.764
Query 4	21	6.001

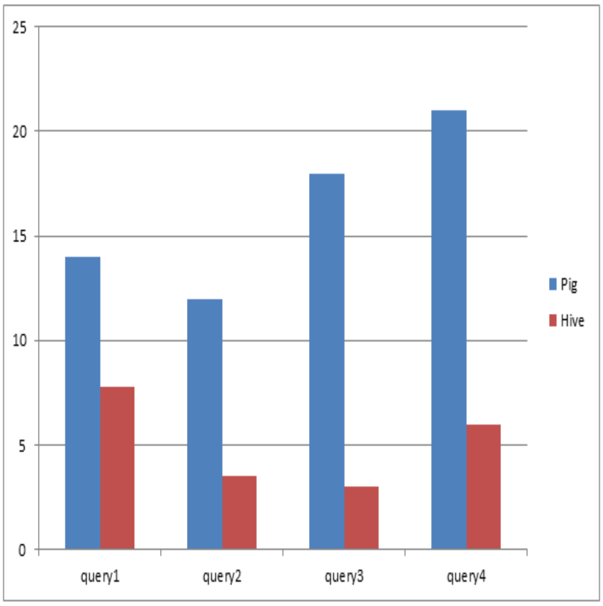


Fig.4.9 Result

6. CONCLUSION

Hadoop Mapreduce is currently a well known decision for performing expansive scale information examination. Bigdata examination utilizing pig and hive reveals insight into critical issues looked by purchasers and enables the establishments or partnerships to redress these issues, to give appropriate fulfillment to the buyers, change in administrations, to keep beware of issues and to develop cooperative attitude in the market. Then again, it gives buyers to recognize legitimately among the organizations and make the specialist co-op determination overwhelmingly.

In light of the parameters like execution time, number of mapreduce occupations, lines of code it has been inspected that hive holds preferable and effective over pig. Based on the parameters like execution time, number of mapreduce jobs, lines of code it has been examined that hive holds better and efficient than pig.

REFERENCES

[1]Apache Hadoop: http://Hadoop.apache.org

[2]Dean, J. and Ghemawat, S., "MapReduce: a flexible data processing tool", ACM 2010.

[3]DeWitt & Stonebraker, "MapReduce: A major step backwards", 2008.

[4].Hadoop Distributed File System http://hadoop.apache.org/hdfs

[5]HadoopTutorial:http://developer.yahoo.com/hadoop/tutorial/mo dule1.html

[6]J. Dean and S. Ghemawat, "Data Processing on Large Cluster", OSDI '04, pages 137–150, 2004

[7]J. Dean and S. Ghemawat,"MapReduce: Simplified Data Processing on Large Clusters", p.10, (2004).

[8]Jean-Pierre Dijcks, "Oracle: Big Data for the Enterprise", 2013

[9]J. Christy Jackson, V. Vijaya kumar, Md. Abdul Quadir, and C. Bharathi, "Survey on Programming Models and Environments for Cluster, Cloud, and Grid Computing that defends Big Data," 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), ELSEVIER, 2015.

[9]Dataset that is used in this project, Available: https://github.com/jasondbaker/seis734.

[10]F. Provost, T. Fawcett, "Data Science and its relationship to Big Data and data-driven decision making," University of Massachusetts Amherst, DOI: 10.1089/big.2013.1508, March 2013.

[11]Munesh Kataria, Ms.Pooja Mittal, "Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql," IJCSMC, Vol. 3, July 2014, pp. 759 – 765.

[12]Hadoop Computing Solutions, www-01.ibm.com/software/data/infosphere/hadoop

[13]Hadoop Core, www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html

[14]Figure2.1,2.2,2.3 pig components, pig and hive architectures, system architecture reprinted from White T., Hadoop: The Definitive Guide,Third Edition, 2012.

[15]Dirk deRoos, Chris Eaton, George Lapis, PauZikopoulos and Tom Deutsch, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data.McGraw Hill Osborne Media;1 edition(October19,2011).

[16]Dataset that is used in this project, Available: https://github.com/jasondbaker/seis734.

[17]James M. Harris, and Dr. Cynthia, and Z.F. Clark, "Strengthening Methodological Architecture with Multiple Frames and Data Sources," Proceedings 59th ISI World Statistics Congress, Hong Kong, August 2013.

[18]Figure 4.1,4.2,4.3,4.4,4.5,4.6,4.7,4.8 result of execution of Query Screenshots taken from system .