ISSN:0975-9646

Comparative Analysis Using Hive and Pig on Consumers Data

Pooja Jain[1], Prof. Jay Prakash Maurya[2]
[1]*M-Tech Research Scholar,* [2]*Research Guide, Department of Computer Science & Engineering*
Lakshmi Narain College of Technology Bhopal

***Abstract*— Bigdata refers to the datasets that exceeds the capacity of conventional software systems to share, handle, process the data. Along with the advancement in technologies and applications, an enormous growth in internet users has been realized whereby online user complaints data also raised highly. This consumer complaint data becomes so copious that the service providers can significantly seek such copious data related information and can evidently work on the related issues efficaciously using big data analytics. Rectifying these issues and providing satisfaction to consumers can also help companies to create a good will in the market. These copious datasets can be analysed parallely in multiple ways. In this paper we introduce big data analytical tools Apache hive and Apache pig on the top of hadoop analysing frequent issues, maximum issues registered for particular company etc. Also, comparison of hive and pig is carried out on certain parameters during analysis that exhibits hive performing better than pig.**

Keywords*— Big data, Hadoop, hive, pig, analysis, consumer complaints.

I. INTRODUCTION

Bigdata can be reckoned to be a collection of data so copious that it can't be handled efficaciously using traditional data management techniques. The buzzing word "Bigdata analytics" can thus be described as analysis of datasets using different analysis techniques. Main challenges created by bigdata[1] are shown as 3Vs-

volume– the generation of data has increased from gigabytes to petabytes and terabytes size. Volume denotes this massive data generated from multiple sources that continue to grow.

velocity– velocity denotes the rate at which data is produced which has been changed from gigabytes per day to terabytes per day.

variety– variety refers to the various formats in which data is produced from sources. This data can be in structured or unstructured format.

The best example for a massive parallel processing system is considered as hadoop[2].

Modern day business models face increasing challenges to their marketing strategies regarding consumer satisfaction and protection. The growing influence of ecommerce has made the issue of consumer rights protection even more serious. From a business perspective, it is becoming even more important to build efficient models of trust and reputation. However, the biggest challenge in developing such systems is the fact that trust is mostly considered subjective. Most consumers purchase products or services with certain expectations, and if expectations are not fulfilled, they are likely to choose a different brand or service. Therefore, consumer protection helps drive satisfaction and also enables organizations to reexamine their policies and practices to meet the consumer welfare expectations.

Furthermore, cost variations also act as an important factor in consumer satisfaction. The more a person pays for a product or service, the greater his or her expectations are concerning those purchases. When considering these factors, the rhetoric about consumer protection is usually found to be based on the reality of consumer dissatisfaction (Agasti & Sengupta, 2014) [3].

Usually, an informal complaint is the most common way of registering one's dissent over a product or service in the hope of getting it resolved amicably. But, if such a grievance is not resolved in a satisfactory manner, the consumer may sometimes even register a formal complaint with a third party such as the Better Business Bureau, a country government, the Federal Trade Commission, etc. However, more often than not, this can be a very cumbersome process. The most common form of registering consumer complaints is either through some specific software application or by recording the complaint in a database with spreadsheets or similar tools. Furthermore, with the advent of the Internet and social media, any impediment to an easy way of expressing dissatisfaction with something is greatly reduced, and a larger number of people have access to the complaints. Consumer complaints may assume different meanings in different contexts. What works for a specific domain may not work for others. This creates the need for a system that can tackle different application scenarios dealing with consumer complaints. The main goal of this research is to establish the foundations of such a stable system that can work across an umbrella of businesses and application scenarios with the maximum ease of use.

Hadoop

Hadoop is an open source, distributed computing framework developed and maintained by the Apache Software Foundation written in java. Hadoop [4] components MapReduce processes the data and Hdfs stores large datasets over a cluster. It is used in handling large and complex data which may be structured, unstructured or semi-structured that does not fit into tables. For example, Twitter data falls into the category of "semistructured" data

which can be best stored and analyzed using Hadoop and its underlying file system.

Hadoop Distributed File System

Hadoop Distributed File System (HDFS) is a distributed file system which rests on top of the native file system and is written in java. It is highly fault tolerant and is designed for commodity hardware. HDFS has a high throughput access to application and is suitable for applications with large amount of data. The master-server architecture of HDFS having single name node helps in regulating the file system access. Requests from file system clients are handled by the data nodes. Data is stored as Input splits (blocks) on the underlying file system. The replication factor is set as 3 by default in order to maintain redundancy of data.

Apache Pig

Apache Pig [5] is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

Hive

After congregating the data into HDFS they are analyzed by queries using Hive. Apache Hive[6] data 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.

II. LITERATURE REVIEW

According to [7], Big data analytical capabilities using cloud delivery models could ease adoption for many industry, and most importantly could be cost saving, it could simplify useful insights that provide them with different kinds of competitive advantage. Many companies to provide online Big Data analytical tools some of the top most companies like Amazon Big data Analytics Platform, HIVE web based Interface, SAP Big data Analytics, IBM InfoSphere BigInsights, TERADATA Big Data Analytics, 1010data Big Data Platform, Cloudera Big Data Solution etc. Those companies analyze huge amount of data with help of different type of tools and also provide easy or simple user interface for analyzing data.

This research paper [8] demonstrates use of Apache pig on real world problem. This research paper also provides analysis on data collect from news web links for creating new ways to showcase the news updates. This paper provides information on representing news updates in various categories. Authors have showcased analysis on utilization of Apache Pig tool and RSS as input tool for appropriate results.

This research study [9] provides information on analysis of large scale data by adopting three tools HIVE, PIG and Map Reduce with Hadoop, it take web logs, which contain customer habit like shopping social media network on the basis of that data result Authors have indicated that HIVE and Pig are more quick to develop whereas Map Reduce takes longer time to run.

A. Consumer complaints and complaints handling

According to the report of Chinese e-commerce research, selling fake products, information disclosure, delivery delays, network bulk, returns and refund difficulty, canceling orders difficultly, price fraud, service attitude became the most hottest complaints about online shopping in China in the first half of 2015[10]. Complaint behaviors can provide various positive implications for vendors, including guidance to develop innovative products or services and opportunities to redress consumers' problems [11]. A complainant, thus, chooses to communicate with the firm regarding a problematic consumption experience, rather than simply withdrawing from being a customer, giving the firm an opportunity to provide
some form of remedy and/or to take some corrective action regarding its processes, through creating satisfied customers, that is, the customer response path, and through improving business processes and practices based on insights derived from customer complaints, that is, the organizational learning path[11][12]. Customer complaints are a daily reality of business in any virtually industry and are typically stored in data warehouses, making them easily accessible to managers and an extraordinary recovery may turn complaining customers into loyal ones and generate more goodwill than if the failure had not occurred in the first place[13]. Evidence is offered to show that only a small percentage of dissatisfied customers ever communicate with the store[14], thus, service providers should encourage their customers to complain if they experience service failures and effectively take measures to hand complaints, because complaints may encourage suppliers to improve goods and services and, thus, produce some lasting benefit[15] and complaint handling as a form of remedy for the failure service, related to whether the post-service is successful or not.

B. Theory of Justice and Satisfaction with Complaint Handling

As a basis for understanding the process of complaining and its outcomes, contemporary studies on complaint management offer substantial evidence supporting the importance of three specific dimensions of perceived fairness in complaint handling: distributive, procedural, and interactional fairness. The first dimension distributive fairness refers to the allocation of benefits and costs among the parties involved in a transaction. The second dimension, procedural fairness, concerns the complaint handling policies and procedures used by a firm. Interactional fairness, the third dimension, involves how company representatives treat and communicate with the customer during a complaint[16][17].

C. Consumer loyalty

In traditional offline transaction, consumer loyalty is viewed as the strength of the relationship between an individual's relative attitude and repeat patronage[18], this

concept of consumer loyalty also is suitable for the ecommerce. Loyalty is usually measured by two outcome dimensions: consumer repurchase intention and positive WOM. Credible WOM is highly important for future transactions in the on-line environment[19]. Faced with more and more fiercely competing, effective, friendly and stable relationship with consumer is very significant for online vendors successfully competing among lots of competitors.

D. Antecedents of Loyalty

Service failures can cause dissatisfaction among customers, which may lead to customer complaints and a loss of loyalty in future purchases. To avoid such problems, firms act to rectify service failures through the process of service recovery. Depending on how firms handle service recovery, results may vary massively: from losing an angry customer to retaining a satisfied, relieved one, who may still be willing to purchase again in the future[20].

In general, many studies suggest that compensating customers after a service failure leads to more favorable consumer responses, either by dissipating their anger and dissatisfaction or by enhancing their overall experience. Satisfaction with complaint handling was key to consumer recommendation of the service to others [21]. Effective measures of service recovery can strengthen the customer's trust in the quality of products or services, and develop customer loyalty[20]. Other research suggests that service recovery efforts to remedy service failures are crucial to maintaining relations with existing customers.

A satisfactory public company response is therefore not only crucial in terms of customer retention, but also in the form of increasing corporate reputation and brand equity generated by third-party online consumers who read about positive complaint resolution and hence may perceive less risk and cognitive dissonance.

III. PROBLEM DEFINITION

A consumer complaint can be understood as a statement of dissatisfaction towards a product or a service. On the other hand, consumer protection comprises of the laws that give consumers the right to register their dissatisfaction about any abusive or inferior business practices and that get a reasonable resolution. Consumer reports are a collection of data about different products and services through reviews and comparisons. Such reports help in analyzing the good and the bad side of a product or a service. Consumer reports also include research about a product or service to highlight its advantages and disadvantages. However, such reports are usually limited in scope and applicability. Most of these reports are based on results from in-house laboratory tests and experiments. This limits consumers from submitting their reviews or complaints about their purchases. Even the reporting agencies that allow a consumer to submit his or her concerns usually do not collate the user feedback in a meaningful way.

Although the concept of consumer complaints analysis has huge applicability and impact, there are no existing software patterns for developing a software system that explicitly deals with every aspect concerning it.

IV. PROPOSED WORK

For analysing these large and complex data a power tool is required, we are using hadoop [22] which is a open source implementation of mapreduce, a processing framework designed for deep analysis and transformation of very large data.

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[6] which work on the top of hadoop, apache hive and apache pig through which we can analyze the consumer complaints datasets.

V. PROPOSED METHODOLOGY:

Steps or Algorithm Steps will follow are:

Step 1: first we collect consumer complaints datasets from web resources.

Step 2: After collection we can load that consumers complaint datasets using hadoop command line.

Step 3: The datasets are stored into HDFS which is very reliable for storing huge or complex data size.

Step 4: The consumers complaint datasets are processed by mapreduce which is a processing engine in the hadoop framework .

Step 5: we can analyse these consumers complaints with the help of bigdata analytical tools which can work on top of the hadoop and in the backend the hadoop will process the datasets.

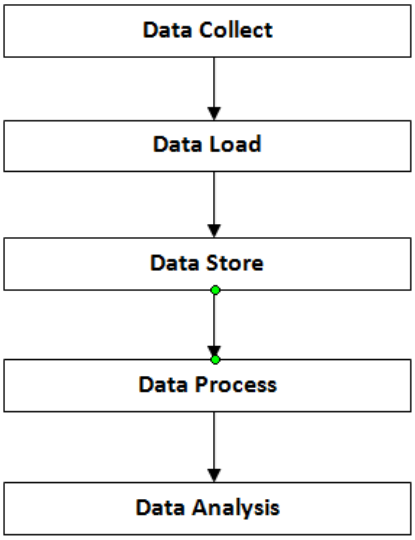

Fig 1. ANALYSIS STEPS

VI. RESEARCH QUESTIONS

Some of the problem statements along which the analysis has been done in this paper:-

Query-1. Top ten days when Maximum number of complaints registered:- Here we can analyze the average number of complaints registered per day and the maximum numbers of complaints registered on which date.

Query-2. Top twenty issue faced by maximum number of consumers:- Here we can analyze the issue which is faced by maximum consumer and we can find top twenty most popular issue.

Query-3. Top twenty company on which maximum complaints registered:- Here we can find those companies on which maximum complaints registered by the consumers.

Query-4. Top twenty issue along with company:- Here we can find the issue which is faced by most of the consumer for a particular company.

Hadoop Cluster Mode: Pseudo-Distributed hadoop cluster mode is used where hadoop daemons run on the local machine.

Data Set: The data was in .csv format, that is each line represents data record and each record has one or more field separated by comma.

Data Set Description: The data set includes the following fields

```
hive> describe complaints;
OK
date_received           string
product                 string
sub_product             string
issue                   string
sub_issue               string
consumer_complaint_narrative    string
company_public_response string
company                 string
state                   string
zipcode                 bigint
tags                    string
consumer_consent_provided       string
submitted_via           string
date_sent_to_company    string
company_response_to_consumer    string
timely_response         string
consumer_disputed       string
complaint_id            bigint
Time taken: 0.213 seconds, Fetched: 18 row(s)
hive>
```

Fig 2. Dataset Description

Tools and Technologies Used:

1. Hadoop
2. Pig
3.Hive

VII. EXPERIMENTAL FINDINGS

Query-1. Top ten days when Maximum number of complaints registered.

Using pig:-

A = Load the data set using Pig Storage;
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

```
2017-03-29 13:12:23,805 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2017-03-29 13:12:23,806 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(",76433)
(01/19/2017,1673)
(On XXXX XXXX,1377)
(01/20/2017,1269)
(08/27/2015,963)
(06/26/2014,916)
(09/20/2016,915)
(07/06/2016,914)
(08/26/2015,912)
(07/26/2016,899)
grunt>
```

Fig 3. Query-1 result using pig

Using Hive:-

select date_received,count(*) as a from complaints group by date_received order by a desc limit 10;

```
Stage-Stage-1: Map: 2  Reduce: 2   Cumulative CPU: 23.51 sec   HDFS Read: 282145529 HDFS Write: 1734188 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.03 sec   HDFS Read: 17239091 HDFS Write: 140 SUCCESS
Total MapReduce CPU Time Spent: 30 seconds 540 msec
OK
"       76433
        56428
01/19/2017      1673
On XXXX XXXX    1377
01/20/2017      1269
08/27/2015      963
06/26/2014      916
09/20/2016      915
07/06/2016      914
08/26/2015      912
Time taken: 217.571 seconds, Fetched: 10 row(s)
hive>
```

Fig 4. Query 1 result using hive

Query-2. Top twenty issue faced by maximum number of consumers.

Using pig:-

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

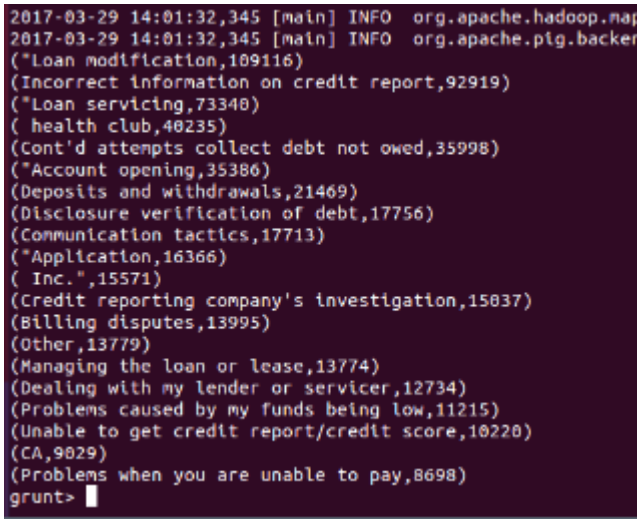

Fig 5. Query-2 result using pig

Using Hive:-

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

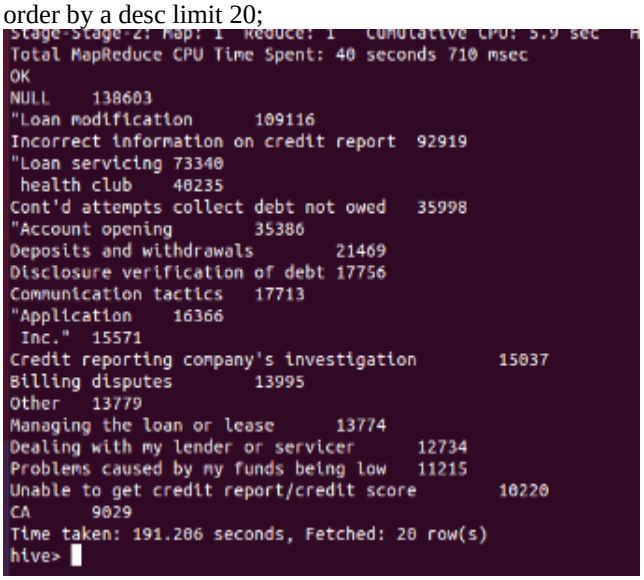

Fig 6. Query-2 result using hive

Query 3:- Top twenty company on which maximum complaints registered

Using Pig:-

A = Load the data set using Pig Storage;
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

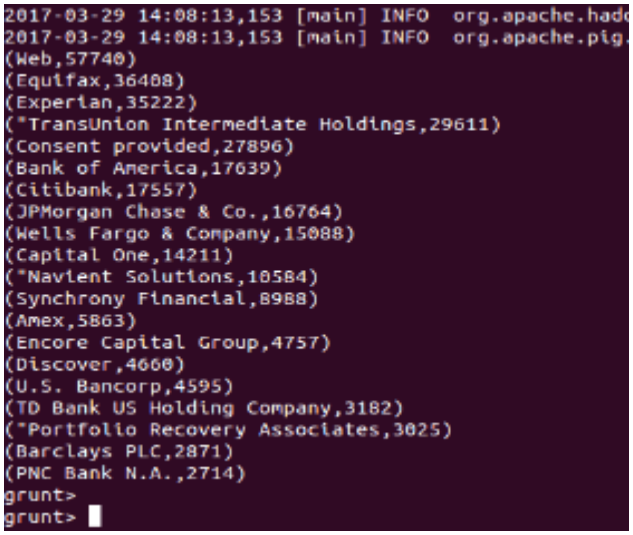

Fig 7. Query-3 result using pig

Using Hive:-

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

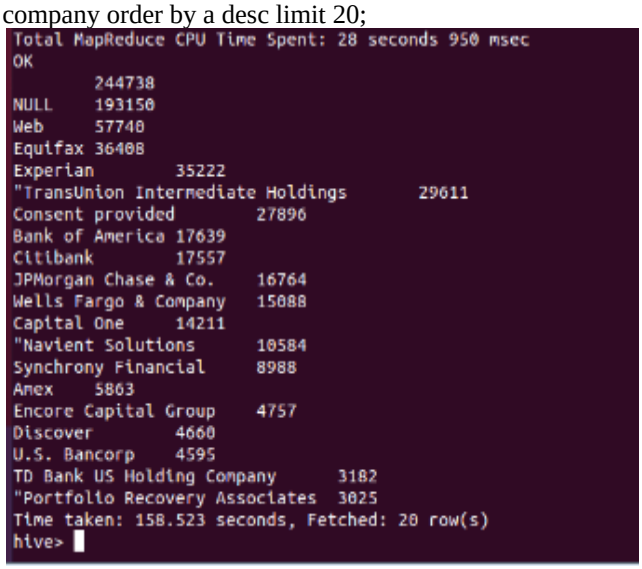

Fig 8. Query-3 result using hive

Query-4. Top twenty Issue along with company

Using Pig:-

V = Load the data set using Pig Storage;
W = foreach V 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

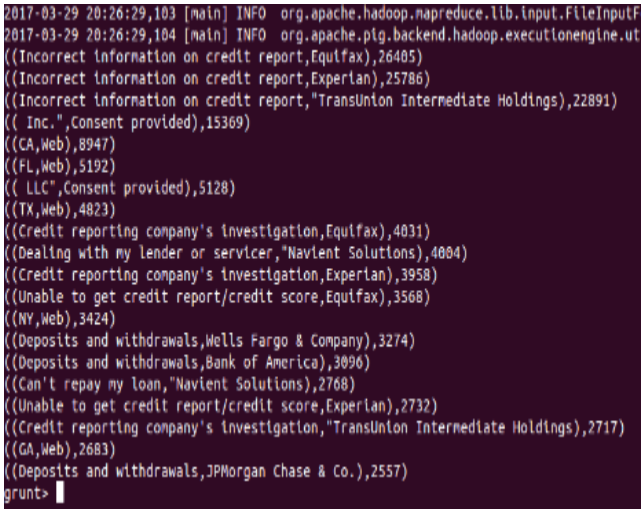

Fig 9. Query-4 result using pig

Using Hive:-

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

create table new1 as select issue,company from complaints 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;

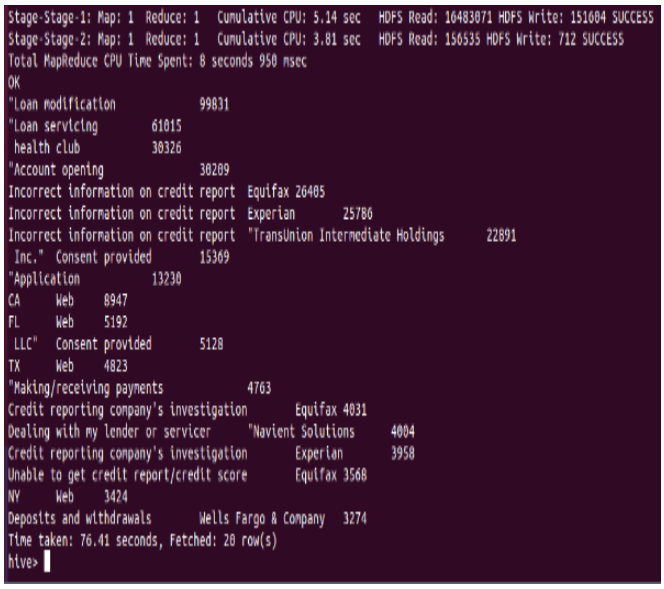

Fig 10. Query-4 result using hive

VIII. 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 also introduced hive which is more useful as compared to pig on analysis of .csv datasets. We can say that hive perform faster as compared to pig on the basis of various parameters, also the above query results demonstrate that the execution time taken by hive is very less as compared to pig. And the mapreduce jobs generated by hive are less as compared to pig whereby the execution time is less in hive. Another benefit of using hive is number of lines of code, which are more in pig but in hive only one line of query is sufficient. Another parameter is load over mr-jobhistory server, there is much load over history server when we execute pig scripts because there is more switching between the alias in pig whereas hive imparts less switching thereby reducing the load on mr-jobhistory server. The experimental results are shown below-

Execution time taken (in min.)	Pig	Hive
Query-1	7	3.64
Query-2	5	3.18
Query-3	4	2.65
Query-4	19	6

Table 1. Execution time taken by hive and pig

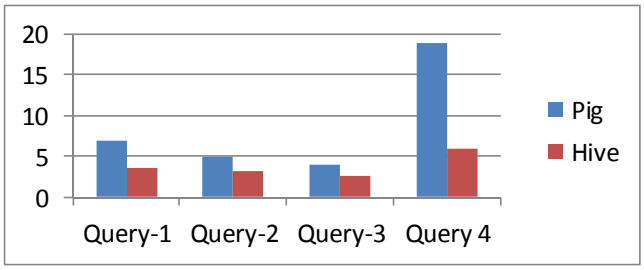

Fig 10. Execution time taken by hive and pig

No. of jobs launched	Pig	Hive
Query-1	4	2
Query-2	4	2
Query-3	4	2
Query-4	9	5

Table 2. No. of jobs launched by pig and hive

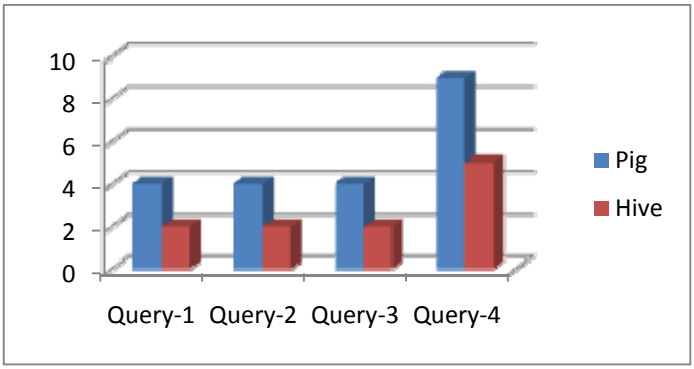

Fig 11. No. of jobs launched by pig and hive

Query executed by	Execution time taken (in min.)
Pig with mr-history server	4
Pig without mr-history server	16
Hive with mr-history server	2.65
Hive withour mr-history server	4.33

Table 3. Query executed w.r.t mr-history server

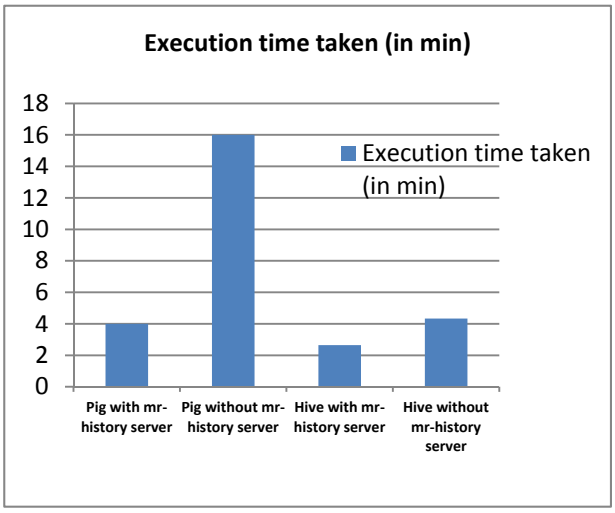

Fig 12. Query executed w.r.t mr-history server

VI. CONCLUSION

Hadoop Mapreduce is now a popular choice for performing large-scale data analytics. Bigdata analytics using pig and hive sheds light on significant issues faced by consumers and helps the institutions or corporations to rectify these issues, provide proper satisfaction to the consumers, improvement in services, to keep check on issues and to build up good will in the market. On the other hand, it provides consumers to distinguish properly among the institutions and make the service provider selection vigorously. 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

[01] Arushi Jain, Vishal Bhatnagar, "Crime Data Analysis Using Pig with Hadoop" in International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA, in ELSEVIER 2015.
[02] Kamalpreet Singh, Ravinder Kaur, "Hadoop: Addressing Challenges of Big Data" in 2014 IEEE International Advance Computing Conference (IACC).
[03] Sarmistha Agasti, Partha Pratim Sengupta, "Business regulation for consumer welfare consumer protection in India, related to marketing" in 2014 2nd International Conference on Business and Information Management (ICBIM), IEEE.
[04] http://hadoop.apache.org/
[05] https://pig.apache.org/
[06] https://hive.apache.org/
[07] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, "Big Data and Advanced Analytics Tools",IEEE 2016, in Symposium on Colossal Data Analysis and Networking (CDAN).
[08] Jurmo Mehine, Satish Srirama, Pelle Jakovits "Large Scale Data Analysis Using Apache Pig"
[09] Dave Jaffe "Three Approaches to Data Analysis with Hadoop".
[10] China E-Commerce Data Monitoring Report in 2015 (on)[J]. http://wwwl 00eccn/ztl2015sndbg/, 20 IS.
[11] Wu I-L, Huang C-YAnalysing Complaint Intentions in Online Shopping: The Antecedents of Justice and Technology Use and the Mediator of Customer Satisfaction[J]. Behaviour & Information Technology, 20 15, 34(1): 69-80.
[12] Yilmaz C, Varnali K, Kasnakoglu B T.How Do Firms Benefit from Customer Complaints?[J]. Journal of Business Research, 20 16, 69(2): 944-955.
[13] Knox G, Van Oest R.Customer Complaints and Recovery Eflectiveness: A Customer Base Approach[J]. Journal of Marketing, 20 14, 78(5): 42-57.
[14] Day R L, Grabicke K, Schaetzle T, et al.The Hidden Agenda of Consumer Complaining[J]. Journal of Retailing, 1981, 57(3): 86.
[15] Chang C C, Chin Y c.Comparing Consumer Complaint Responses to Online and Offline Environment[J]. Internet Research, 20 1 1, 2 1(2): 124-137.
[16] Tax S S, Brown S W, Chandrashekaran M.Customer Evaluations of Service Complaint Experiences: Implications for Relationship Marketing[J]. Journal of Marketing, 1998, 62(2): 60-76.
[17] Homburg C, Fiirst A.How Organizational Complaint Handling Drives Customer Loyalty: An Analysis of the Mechanistic and the Organic Approach[J].Journal of Marketing,2005, 69(3): 95- 1 14.
[18] Dick a S, Basu K.Customer Loyalty: Toward an Integrated Conceptual Framework[J]. Journal of the Academy of Marketing Science, 1994, 22(2): 99- 1 13.
[19] Pizzutti C, Fernandes D.Effect of Recovery Efforts on Consumer Trust and Loyalty in E-Tail: A Contingency Model[J]. International Journal of Electronic Commerce, 20 10, 14(4): 127-160.
[20] Uruefia A, Hidalgo A.Successflil Loyalty in E-Complaints: Fsqca and Structural Equation Modeling Analyses[J]. Journal of Business Research, 2015.
[21] Rothenberger S, Grewal 0, Iyer G R.Understanding the Role of Complaint Handling on Consumer Loyalty in Service Relationships[J]. Journal of Relationship Marketing, 2008, 7(4): 359-376.
[22] Hadoop Wiki Website, Apache, http://wiki.apache.org/hadoop