Experimentation on Transaction Data for Online Retailer Dataset (Online Retailer Analysis)

J. Pradeep Kumar[1] L. Divya[2] B. Manvitha[3]
[1,2,3]Department of Information Technology
[1,2,3]MLRIT Hyderabad-5055, India

Abstract— Due to accelerated pace of economic globalization and increasing market competition, economic pressures and competition have lead enterprise managers to face the problems of choosing the right strategic decision-making policies. Recently, it has been recognized that precision marketing has become a key means of generating profit and is becoming increasingly important has customers become better inform about the products and their rights as customers. This paper considers a marketing problem where the supplier or manufacture provides different products for retail customers, of which some may sell well in the customer segments and some may not. Products that are not sold will be return back to the supplier. In recent years, decision-making problems have received much attention due to a wide range of real world applications. We are implementing a new approach using big data tools which can provide better decision making strategies and improve the online retail marketing.

Key words: Retailer Analysis, Cloudera, Hue, Impala, and VMware

I. INTRODUCTION

The term big data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies.

The urgency for collaborative research on big data topics is underscored by the U.S. federal government's recent $200 million funding initiative to support big data research. The remainder of this document is organized highlights the differences between traditional analytics and big data analytics, and briefly discusses tools used in big data analytics.

II. LITERATURE SURVEY

The term, big data has been coined to refer to the gargantuan bulk of data that cannot be dealt with by traditional data-handling techniques. Big data is still a novel concept, and in the following literature we intend to elaborate it in a palpable fashion. It commences with the concept of the subject in itself along with its properties and the two general approaches of dealing with it. The comprehensive study further goes on to elucidate the applications of big data in all diverse aspects of economy and being.

The utilization of big data analytics after integrating it with digital capabilities to secure business growth and its visualization to make it comprehensible to the technically apprenticed business analyzers has been discussed in depth. Aside this, the incorporation of big data in order to improve population health, for the betterment of finance, telecom industry, food industry and for fraud detection and sentiment analysis have been delineated. The challenges that are hindering the growth of big data analytics are accounted for in depth in the paper. This topic has been segregated into two arenas- one being the practical challenges faces whilst the other being the theoretical challenges. The hurdlers of securing the data and democratizing it have been elaborated amongst several others such as inability in finding sound data professionals in requited amounts and software that possess ability to process data at a high velocity. Through the article, the authors intend to decipher the notions in an intelligible manner embodying in text several use-cases and illustrations.

Every day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few such colossal amount of data that is being produced continuously is what can be coined as big data. Big data decodes previously untouched data to drive new insight that gets integrated into business operations.

III. EXISTING SYSTEM

Precision marketing offers personalized customer service and is used to help enterprises increase their profits by means of high-efficiency marketing. This paper presents a novel decision-making framework for precision marketing techniques. First, this study presents a trend model to accurately predict monthly supply quantity, second, it uses a RFM (Recency, Frequency and Monetary) model to select attributes to cluster customers into different groups, third, it uses CHAID decision tree and Pareto values to identify important attribute values to distinguish different customer groups, and finally, it creates different supply strategies targeting each customer group

IV. PROPOSED SYSTEM

The objective of the proposed precision-making framework is to help managers identify the potential characteristics of different customer categories and put forward appropriate precision.

V. REQUIREMENT ANALYSIS

After analyzing the requirements certain amount of tasks that have to be performed, the next step is to analyze the problem and understanding the context. There are majorly two phases that are present. The first activity in the phase is studying the existing system and the other is to understanding the requirements of the new system.

Understanding the properties and requirements of a system is more difficult and requires creative thinking and understanding of existing systems.

Requirements which we are using in this project include both Hardware and software requirements for the process.

1) Hardware Requirements

– Dual Quad-core CPU
– 8 GB RAM
2) Software Requirements
– MySQL
– VMware
– HDFS
– Hive
– Hue
– Impala
– Sqoop.

VI. IMPLEMENTATION

This paper considers a marketing problem where the supplier or manufacture provides different products for retail customers, of which some may sell well in some customer segments and some may not. Products that are not sild will be returned back to the supplier. In recent years, the decision-making problems have received much attention due to a wide range of real-world applications. We are implementing a new approach using big data tools which can provide better decision marking strategies and improve the online retail marketing.

A. Steps to implement big data

1) Entering into my sql

a) Step 1: Log into my sql by using mysql –u root –p command and the enter password.
b) Step 2: Create a database with some file name by using the below command create database onlineretailer;
c) Step 3: Then make use of the created database by using the below command

Use onlineretailer;

d) Step 4: We need to create a table in the database by giving the attributes along with the data type by using the below command

Create table retailer(voice_no float,stockcode float,description varchar(50),quantity float,invoicedata varchar(50),unitprice float,customerid float);

e) Step 5: To view the tables use the below

Show tables;

f) Step 6: We need to load the data into sql by using the below command

Load data local infile '/home/cloudera/Desktop/Online Retailer.csv' into table retailer fields terminated by ',' enclosed by "" lines terminated by '\r\n';

g) Step 7: Then exit from mysql by pressing cntrl+Z

2) Entering into Hive

a) Step 1: Enter into hive by giving hive as a command.
b) Step 2: To check the databases use the command

Show databases;

c) Step3: Then create a staging table as below

```
CREATE TABLE retailer_staging
(voice_no float,
Stockcode float,
Description varchar(50),
Quantity float,
Invoicedate varchar(50),
Unitprice float,
Customerid float,
Country varchar(50))
Rowformat Delimited Fields Terminated By ',';
```

d) Step 4: Later on load the data into staging table by using the below command

Load data local inpath'/home/cloudera/Desktop/Online Retailer.csv' into table retailer_staging;

e) Step 5: Create a production table as below

```
CREATE TABLE retailer_production1
(voice_no float,
Stockcode float,
description varchar(50),
quantity float,
invoicedate float,
unitprice float,
customerid float,
country string)
CLUSTERED BY(country) INTO 20 BUCKETS;
```

f) Step 6: Set the hive environment as follows

```
SET hive.exec.dynamic.patition.mode = nonstrict;
Set hive.enforce.bucketing = true;
```

g) Step 7: INERT OVERWRITE TABLE retailer_production1 select *from retailer_staging;

B. Entering into Impala

a) Step 1: Enter into impala by giving impala-shell command
b) Step 2: Select count(*)from retailer;

Here retailer is the table name.

Fig. 1: execution of above queries

The above fig: 1 shows the execution of above queries

VII. RESULTS

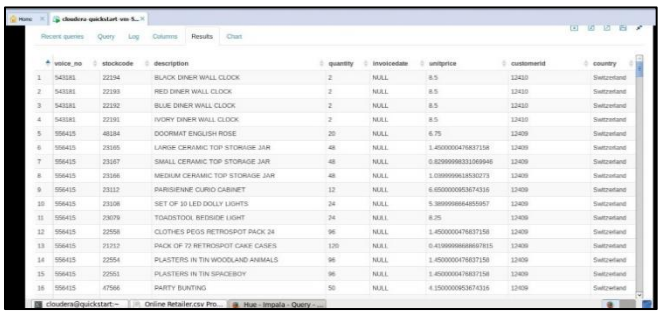

Fig. 2: Results

The above Fig 2 shows the results that were obtained from impala.

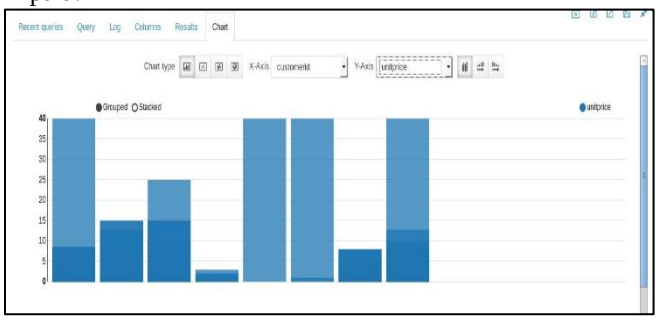

Fig. 3: Results

The above Fig. 3 shows the bar charts that ware obtained as a result in impala. Here we can find these bar graphs were drawn by taking customer id on the X-Axis and unit price on the Y-Axis.

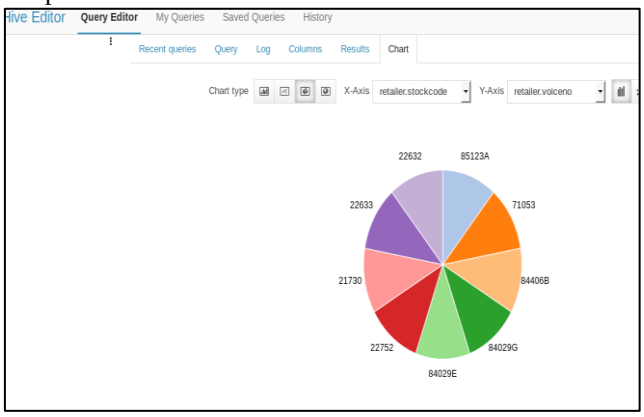

Fig. 5: Results

The above Fig 5 shows the results in pie chart.

VIII. FUTURE ENHANCEMENT

We can extend this idea by using different tools of big data like pig ,tableau, Ooize etc. Pig can execute its Hadoop jobs in MapReduce, Apache Tez. Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages such as Map Reduce, Pig and Hive then intelligently link them to one another. By using these tools we can process the same data in a different way.

IX. CONCLUSION

In this paper we have used many tools like hive, sqoop, impala. The processing of data varies for different tools. To process the data hive takes much time compared to impala. If you run a query on hive there is start time overhead on queries run on map reduce but not on impala. Hive is fault tolerant where as impala is not.

REFERENCES

[1] https://www.edureka.co/big-data-and-hadoop?utm_source=googlesearch&utm_medium_term=big%20hadoop&utm_campaign=Big-Data-hadoop-Search-IN-New&gclid=Cj0KEQjw5sHHBRDg5IK6k938j_IBEiQARZBJWtWxfWjwCWCgpthTDas2WLW1wPbTW1TK9guA2thTZEaAoe-8P8HAQ
[2] https://www.simplilearn.com/big-data-and-analytics/big-data-and-hadoop-training?utm_source=google&utm_medium=cpc&utm_content=lvc&utm_term=big%20data%20hadoop&utm_campaign=search-bigdata-lvc-ind-fpass&gclid=Cj0KEQjw5sHHBRDg5IK6k938j_IBEiQARZBJWhUO1L8gILnNwmsLYgNHx5Gye6JvS1UsY0o2GKk7-PcaAKMx8P8HAQ
[3] http://searchcloudcomputing.techtarget.com/definition/hadoop

