International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-2, May 2015
ISSN: 2395-3470
www.ijseas.com

Big Data Analytics on Cab Company's Customer Dataset Using Hive and Tableau

Dipesh Bhawnani[1], Ashish Sanwlani[2], Haresh Ahuja[3], Dimple Bohra[4]
[1,2,3,4] Vivekanand Education Society's Institute of Technology, Mumbai, India

Abstract

Project focuses on analyzing the cab company's customer dataset which will help company to analyze its frequent customers: so that the company can understand its customers and can provide different offers to them. Demand of cabs of particular type and at particular location and time, so that the company could make necessary arrangement of particular cab like small cabs, luxury cars, buses etc. We have analyzed the possible cancellations of cab booking by the customer using data obtained from the company. The goal is to reduce the cost incurred by the company as a result of cab cancellations made by the customer. Cab companies will be able to manage its vendors and drivers by providing them with up to date information about Customer cancellations. We have also analyzed travel and package type used by the customer. Tableau is used to connect hortonworks hive data source and the data is analyzed and shown in graphical format for better visualization and understanding.

Keywords: *Big Data, Hadoop, Hive, HiveQL, Tableau.*

1. Introduction

Big Data

There has been massive increase in volume of data in organizations. Now, organizations are discovering new ways to compete and win – transforming themselves to take advantage of the available information support [11].

Big data is the "buzz word" that represents large amount of data that cannot be handled by the traditional systems. Big data is often described by **3 V's**

1. Volume: Volume refers to large amount of data that is been generated every second.

2. Variety: Variety refers to the different types and formats of data.

3. Velocity: Velocity refers to speed with which the data is being generated [5].

Hadoop

To analyze large amount of the data Hadoop framework can be used. Hadoop is an open source framework for distributed storage and processing of large sets of data on commodity hardware. Hadoop enables businesses to quickly gain insight from massive amounts of structured and unstructured data. Numerous Apache Software Foundation projects make up the services required by an enterprise to deploy, integrate and work with Hadoop. Each project has been developed to deliver an explicit function and each has its own community of developers and individual release cycles [6].

For our project we have used a dataset which consists of more than 43,000 customer records and for analyzing this big data we have used a Hadoop subproject which is Apache Hive.

Hive & HiveQL

Hive is most suited for data warehouse applications, where relatively static data is analyzed, fast response times are not required, and when the data is not changing rapidly.

Hive provides a means of running MapReduce job through an SQL-like scripting language, called HiveQL, which can be applied towards summarization, querying, and analysis of large volumes of data.

HiveQL enables anyone already familiar with SQL to query the data. Hive makes it easier for developers to port SQL-based applications to Hadoop, compared with other Hadoop languages and tools.

However, like most SQL dialects, HiveQL does not conform to the ANSI SQL standard and it differs in various ways from the familiar SQL dialects provided by Oracle, MySQL, and SQL Server [10].

Beeswax

The Beeswax application enables you to perform queries on Apache Hive, a data warehousing system designed to work with Hadoop. You can create Hive tables, load data, run and manage Hive queries, and download the results in a Microsoft Office Excel worksheet file or a comma-separated values file [2].

Hive Configuration

Beeswax, the Hive user interface in Hue, uses your system's Hive installation and is compatible with Hive 0.7. Hive data is stored in the Hadoop Distributed File System(HDFS), typically in /user/hive/warehouse directory.

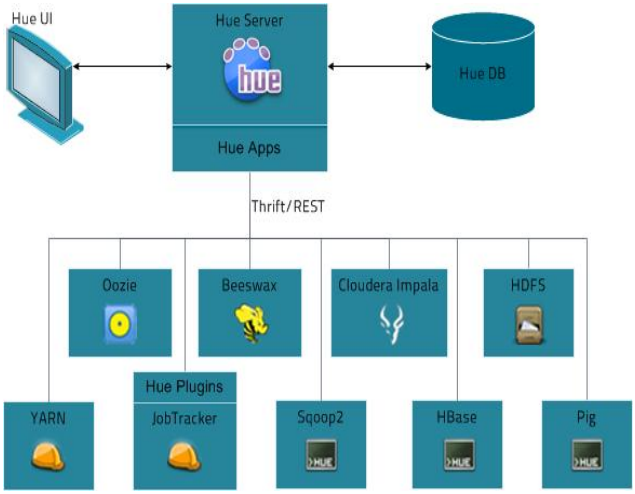

Fig 1: Hue Architecture

Tableau

Tableau is Business Intelligence and Analysis tool. *Tableau* can help anyone see and understand their data. Connect to almost any database, drag and drop to create visualizations, and share with a click. We can Connect to data and perform queries without writing a single line of code. We can measure data in petabytes stored in the cloud or in billions of rows, Tableau is built to work as fast as you do. It's self-service analytics, for everyone.

We can connect Tableau to various data sources such as Hortonworks Hadoop Hive, MySQL, IBM DB2, SAP HANA, and many more [3].

2. Relevance of the Project

San Francisco-based radio cab service provider Uber started operations in India recently. While one can book Uber's luxury taxi service using a smartphone app, most Indian radio cab players, too, are using similar technology, making use of big data to understand the future demand. Companies such as Meru Cabs and OlaCabs among radio cabs and bus ticket booking provider Redbus have also been using big data analytics.

By using this data and based on everyday predictability of data, we try to enhance the number of trips that one cab makes per day. About a year-and-a-half- ago, the average trip made by a taxi driver at Meru was four per day. This has improved to 5.8 trips now. Radio cab firms are also collecting data on traffic situation, road condition and speed, which they say help them offer better services to the customer [4].

3. System Block Diagram

Using Hive Beeswax UI we create the table and loaded the input dataset and written the queries using HiveQL to analyze the data. The query results are shown in the form of tables. Visualization is done in Tableau by connecting Hive data source with the help of Microsoft Hive ODBC Driver.

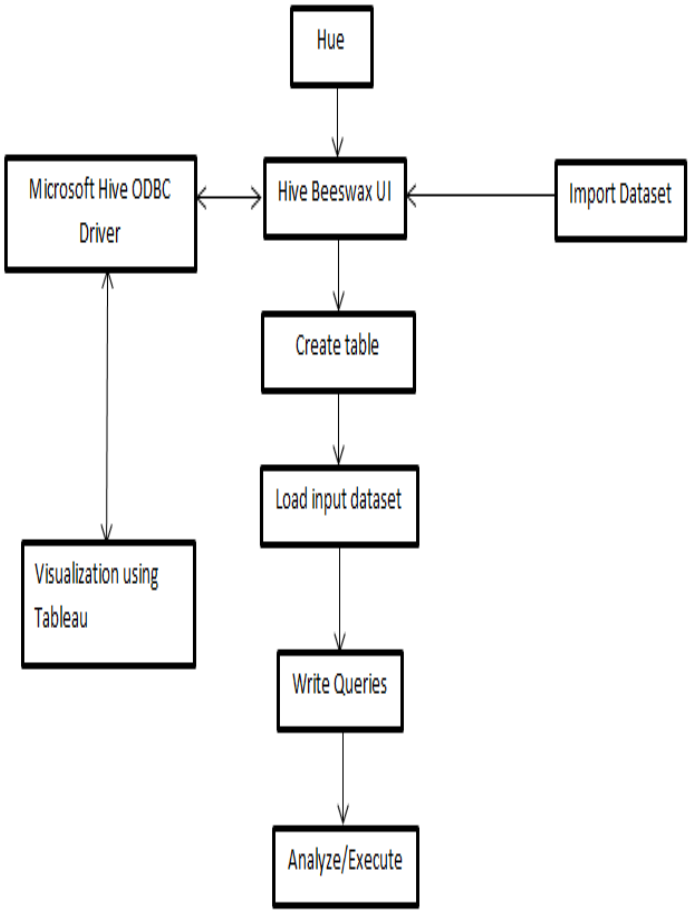

Fig. 2 System Block Diagram

International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-2, May 2015
ISSN: 2395-3470
www.ijseas.com

4. Implementation

4.1 Input Dataset

Cabs.csv (over 43,000 records)

1	id	user_id	vehicle_model_id	package_id	travel_type_id	from_area_id	to_area_id	from_city_id	to_city_
2	132512	22177	28	NULL	2	83	448	NULL	NULL
3	132513	21413	12	NULL	2	1010	540	NULL	NULL
4	132514	22178	12	NULL	2	1301	1034	NULL	NULL
5	132515	13034	12	NULL	2	768	398	NULL	NULL
6	132517	22180	12	NULL	2	1365	849	NULL	NULL
7	132518	17712	12	NULL	2	1021	1323	NULL	NULL
8	132519	22172	12	NULL	1	571	NULL	15	1
9	132520	22181	12	NULL	2	1192	832	NULL	NULL
10	132521	22182	65	2	3	448	NULL	NULL	NULL
11	132522	22184	12	NULL	2	516	376	NULL	NULL
12	132523	4941	12	NULL	2	150	776	NULL	NULL
13	132524	17037	12	NULL	2	455	1330	NULL	NULL
14	132525	22185	12	NULL	2	1166	1328	NULL	NULL
15	132526	22186	12	NULL	2	793	590	NULL	NULL
16	132527	16	28	NULL	2	1063	58	NULL	NULL
17	132528	15097	12	NULL	2	1102	292	NULL	NULL
18	132530	761	12	NULL	2	814	393	NULL	NULL
19	132531	22189	24	NULL	1	1383	NULL	NULL	
20	132533	17226	12	NULL	2	1353	585	NULL	NULL
21	132534	868	12	NULL	2	297	212	NULL	NULL
22	132535	22190	87	2	3	471	NULL	NULL	NULL
23	132536	16	28	NULL	2	58	1063	NULL	NULL
24	132537	21558	12	NULL	2	540	409	NULL	NULL
25	132538	21995	12	1	3	1286	NULL	NULL	NULL
26	132539	22192	54	NULL	1	515	NULL	15	
27	132540	22193	12	NULL	2	142	393	NULL	NULL
28	132541	20430	12	NULL	2	393	61	NULL	NULL

4.2 Parameters

1. **id-** Booking ID
2. **user_id** - ID of the customer (based on mobile number)
3. **vehicle_model_id -** Vehicle model type.
4. **package_id** - Type of package (1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs & 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms)
5. **travel_type_id** - type of travel (1=long distance, 2= point to point, 3= hourly rental).
6. **from_area_id** - unique identifier of area. Applicable only for point-to-point travel and packages
7. **from_city_id -** unique identifier of city
8. **to_city_id -** unique identifier of city (only for intercity)
9. **month–** booking month
10. **year –** year in which booking was done
11. **hour –** hour of trip start
12. **online_booking** - if booking was done on desktop website
13. **mobile_site_booking** - if booking was done on mobile website
14. **Car_Cancellation** - whether the booking was canceled (1) or not (0)

4.3 Querying using HiveQL

Some of the queries are listed here

Frequent Customers

```
select user_id,count(*) as frequent_customers
from cab_customers
group by user_id
having count(user_id)>10;
```

Total bookings and cancellations done by each customers

```
select user_id,count(*) as total_bookings,
sum(Car_Cancellation) as total_cancellations
from cab_customers
group by user_id
having count(user_id)>1;
```

Demand of cabs and number of cancelations in each month

```
select month,count(id) as total_bookings ,sum(Car_Ca
ncellation) as total_cancellations

from cab_customers

where month>=1 and month<=12

group by month;
```

Similarly we have analyzed for travel type and package type used by the customer, Total number of bookings done via mobile site and desktop site, Type of cab more booked and cancelled.

4.4 Results

Fig 3 shows the frequent customers of the company those who have booked cabs more than 10 times.

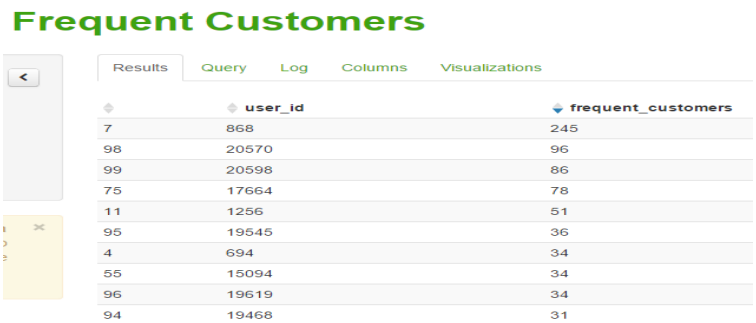

Fig 3: Frequent Customers

Fig 4 shows the total number of bookings and cancellations done by each customer.

Total bookings and cancellations done by each c

Results | Query | Log | Columns | Visualizations

	user_id	total_bookings	total_cancellations
70	868	245	3
46	694	34	0
12	97	23	0
96	1154	20	1
61	819	17	0
82	998	17	0
8	74	15	0
15	115	15	0
47	704	14	0
13	99	13	0

Fig 4: Total bookings and cancellations done by each customer

Fig 5 shows the demand of cabs and the number of cancellations done in each month.

emand of cabs and number of cancelations in e

Results | Query | Log | Columns | Visualizations

	month	total_bookings	total_cancellations
0	1	2852	34
1	2	2890	69
2	3	2790	59
3	4	3119	141
4	5	4296	520
5	6	4392	310
6	7	4907	109
7	8	5445	245
8	9	4736	401
9	10	4814	650

Fig 5: Demand of cabs and number of cancelations in each month

4.5 Visualization Using Tableau

Fig 6 shows the visualization of the total number of bookings and cancellations done by each customer.

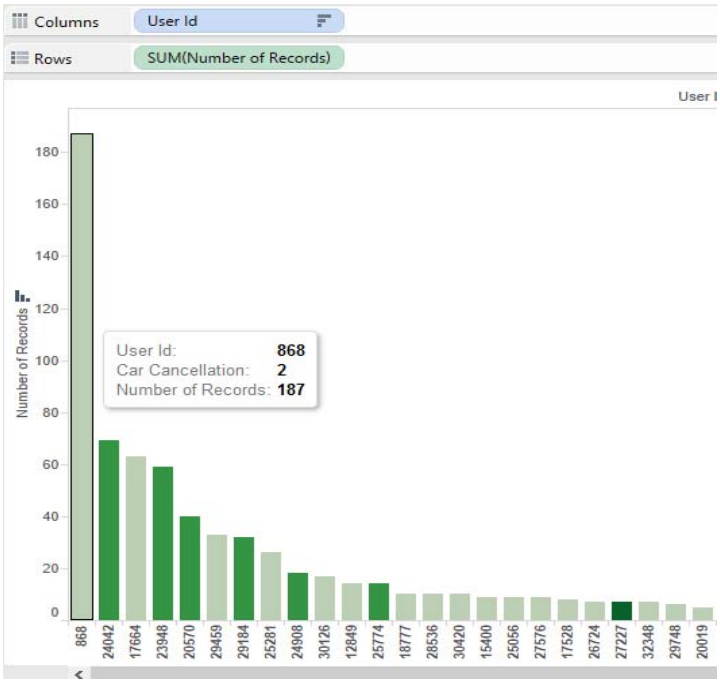

Fig 6: Total bookings and cancellations done by each customer

Fig 7 shows the visualization for demand of cabs and number of cancellations in each month.

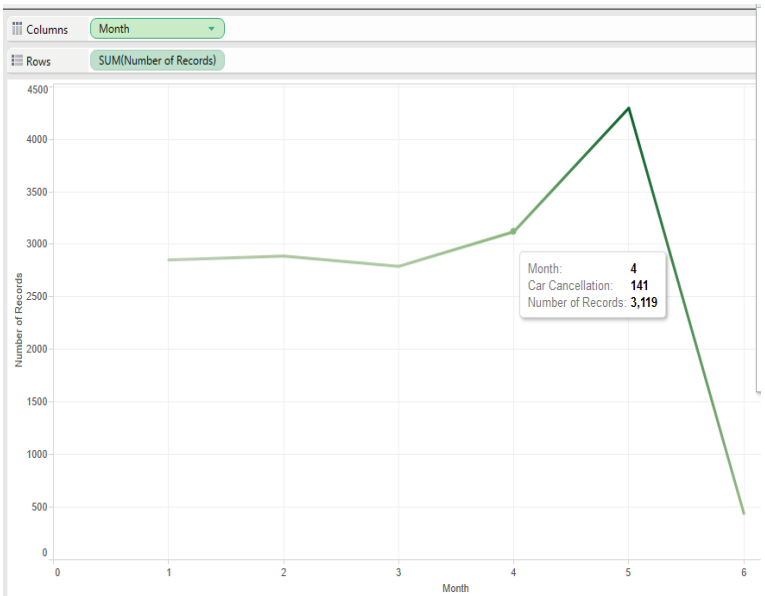

Fig 7: Demand of cabs and number of cancelations in each month

5. Conclusions

The project helps Cab Company to analyze its customers which will help Company to find meaningful patterns from large amount of data and can increase revenue by taking better and faster decisions.

Acknowledgments

This idea would not have been possible without noteworthy contribution of Assistant Prof. Dimple Bohra who inspired us for making this project the way it is and helps us in understanding the concept of big data in detail.

References

[1] http://www.slideshare.net/zanorte/big-data-analytics-2013
[2]http://www.computerworld.com/s/article/9227146/Yahoo_launches_big_data_analytics_tool_for_online_advertiser
[3] http://www.tableau.com
[4]http://www.businessstandard.com/article/technology/big-data-analytics-changes-the-rules-of-road-travel-113091000588_1.html
[5] http://blog.sqlauthority.com/2013/10/29/big-data-final-wrap-and-what-next-day-21-of-21/
[6] http://hortonworks.com/hadoop/
[7] http://en.wikipedia.org/wiki/Hortonworks
[9] http://hadoop.apache.org/
[10] Edward Capriolo,Dean Wampler,Jason Rutherglen, “Programming Hive”, O’REILLY, 2012
[11] Ashish Thusoo, et al., “Hive-A Petabyte Scale Data Warehousing Using Hadoop”, in ICDE Conference, Year: 2010, Page(s): 996-1005

Dipesh Bhawnani B.E., Student, Fourth year Computer Engineering in Vivekanand Education Society’s Institute of Technology, Mumbai

Ashish Sanwlani B.E., Student, Fourth year Computer Engineering in Vivekanand Education Society’s Institute of Technology, Mumbai

Haresh Ahuja B.E., Student, Fourth year Computer Engineering in Vivekanand Education Society’s Institute of Technology, Mumbai

Dimple Bohra M.E., Assistant Professor, Department of Computer Engineering in Vivekanand Education Society’s Institute of Technology, Mumbai