B2B and B2C analysis for Olist using HBase, MapReduce, MySQL, Pig and Hive

Lalit Pathak  
*National college of ireland*   
*Dublin*   
x18110088@student.ncirl.ie

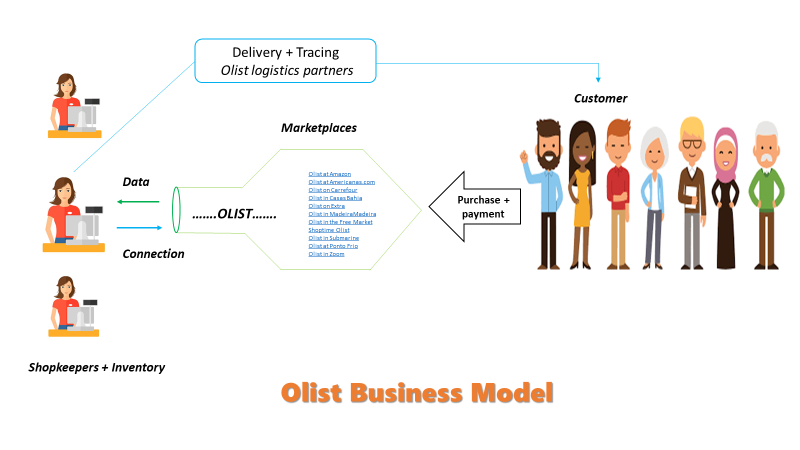
***Abstract— Big data is a massive amount of data and historically defined by volume, velocity, and variety. This data is useful to find patterns, trend, and association related to human behaviour and interaction. In retail industry marketing campaign is extremely important, and big data gives scientific elements for such a campaign. The retail industry is more dependent on data for applying various business strategies. With the availability of the abundant product, information shoppers compare and buy a product anywhere at any time. To increase profit and understand buying behaviour and mapping customer to retail product industry is heavily dependent on data. Olist is a large department store within the marketplace in brazil. Outline of the project is to provide a report to make business strategies for B2B2C model team of Olist. This project uses big data technologies to identify trend, pattern, and impact of location on sales. Publicly available Olist store dataset released by Olist used for analysis. Java-based MapReduce, hive, and pig used for analysis. Data stored in HDFS, MySQL and HBase. Tableau used for visualizations of the reports.***

*Research Question: --- How can big data technologies help to analyze B2B and B2C for retail industry to get the customer and business insight?*

Keywords—Retail industry, Big data technologies – Hbase, Hive, Pig, HDFS, MapReduce, MySQL

# **Introduction**

Olist is the company in the e-commerce segment. The main aim of the company is to connect small entrepreneurs or retails to sell their products in the most significant and big marketplaces in brazil. Olist connected to the number of retails, and all of them act as a distribution center. Once the order is placed on the Olist store through an e-commerce website. Invoice of product request sent to the retailer having the product in stock according to the closest location of the customer. Retail prepares the package and sends this safely to the customer. Below diagram shows the business model of the Olist.



***Fig 1: Olist business model***

From fig.1, there are two main segments of Olist business, and they are Retailers or merchants, and the other is customers.

In Olist B2B, merchants or retails have different incentive to operate in multiple marketplaces in brazil. Olist help them to

Manage their sales channel in different markets and manages orders and stock information. This project focused on B2B and B2C analysis for Olist.

   In the Report, section 2 defines the work related to retail analysis with big data by the previous author. Technologies used to build a framework is defined in section 3. Part 4 is an implementation which includes the process flow diagram, and script execution sequence, script log files. Results are visualizing in section 5. Section 6 concludes the paper.

# **Related Work**

With an increase in data, every industry focused on exploring the past data to get a competitive advantage. With an increase in volume and variety in the data manual analysis and some cases, dataset analysis is not possible using convolutional databases. Hadoop MapReduce, HBase, MySQL, Hive, Pig are useful in successfully fulfil the need for the process flow [3]. An interpretive framework of big data in the e-commerce industry is presented by [7]. This includes definitional aspects, distinctive, characteristics, business value and challenges of big data analytics in the e-commerce landscape. Big data analytics brings value to the e-commerce industry. In [8] author presents a predictive analysis in Retail using Big data. They showed how big data analysis helps to improve the retail business, but there is a barrier in big data analytics such as privacy of the data and scalability to the algorithm. Big data can be helpful to provide a personalized product recommendation to the customer. Inventory management is also possible through big data. Stock forecasting has been studied, which shows that with the help of historical sales data and seasonal sales data, it is possible to forecast the product demand.

    The rapid growth in the online retail markets embeds the need for high-end data processing with minimal consumption of time to predict the consumer's behaviour, that would turn into a successful sale. The traditional data mining approaches lack in doing such real-time intensive tasks. So, to perform this gigantic data management, a study proposed by [4] using Apriori-MapReduce (AMR) framework. [4] carried out the market basket analysis using AMR in conjunction with Hadoop based cloud architecture to insightfully understand customer behaviour. Another similar but theoretical study presented by [5] on strategical implementation of precision marketing in the retail industry. They illustrated that, with the embodiment of big data analytics like classification and regression, the precise model could be built to seek fruitful returns. The models like classification and regression tree (CART), differential regression analysis technique (DRA) and decision-making tree can granularize the customer profile to avoid the aversion of the valued customers. Thus, this data-driven marketing can build a sustainable model to consolidate retailers, suppliers, and customers.

[6] had applied big data analytics to evaluate the validity of the gravity model of the e-market firm in China named Taobao. They explored various business patterns and investigated the parameters that were affecting the amount of trade within the provinces and the cities, and customer buying behaviour was observed that largely depend on the reputation and the authentication of the sellers or retailers on the Taobao. The statistical analysis followed by regression analysis was performed.

# **TECHNOLOGIES INVOLVED**

**R Programming:** R Language is used for statistical computing and provide graphical support. R used for data, transformation and initial exploratory data analysis.

**Shell script:** Shell scripting is designed to run the Unix shell. The main purpose of shell scripting is the manipulation of data and program execution. The main aim of shell scripting for this project is as a wrapper to automate and integrate big data tools and execute the programs. Processing has been carried out with the “sed” command.

**MySQL:** MySQL is a Relational database management system (RDBMS) written in C ++ and C. Used in data warehousing, e-commerce, and logging application. Files with fixed length and know delimiter is stored in the MySQL database.

**Apache Sqoop:** Sqoop is developed in java used to load the data from RDBMS, CSV, TSV to Hadoop eco-system and vice versa. Sqoop is used to create table and load data from MySQL to Hive.

**Apache MapReduce:** MapReduce provides a benefit over for big data with scalability, flexibility, and speed. MapReduce is the backbone of all task performed by Hive, Pig. In this MapReduce task to filter data.

**Apache Hive:** This is a relational database for Big data environment. Tables and views are created for analysis in Apache Hive.

**Apache HBase:** It is a non-relational database system and like the big table is written in Java runs on top of HDFS. The data output of Pig, Hive, and MapReduce is loaded into HBase.

**Tableau:** Visualization of the data generated from Pig, MapReduce and Hive view.

# **PROCEUDRE AND METHODOLOGY**

This section describes data gathering, data pre-processing, Technologies in use and Implementation of Framework to

tackle Problem.

### **SOURCE OF DATA**

Business – Business(B2B) and Business – customer(B2C) data is downloaded from the Kaggle.[1] [2].

**I)  MARKETING FUNNEL BY OLIST (B2B)** contains information from of seller who sends filled-in requests of contact to sell their products on Olist platform. Data is from 1st June 2017 to Jun 1st, 2018. Dataset contain two files

olist\_closed\_deals\_dataset, olist\_marketing\_qualified\_leads\_dataset.

**II)** **BRAZILIAN E-COMMERCE PUBLIC DATASET BY OLIST (B2C).**

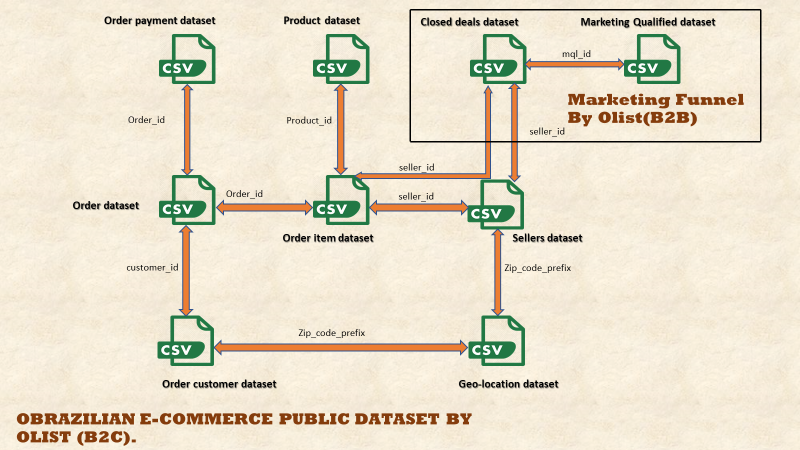
This dataset contains 100k orders from 2016-2018. This gives the library to view orders from different dimensions such as order status, price, payment and freight performance to customer location.

File contains below datasets:

olist\_customers\_dataset, olist\_geolocation\_dataset

olist\_order\_items\_dataset, olist\_order\_payments\_dataset,olist\_order\_reviews\_dataset,olist\_orders\_dataset,olist\_products\_dataset,olist\_sellers\_dataset,product\_category\_name\_translation

**Data Schema:**

****

***fig 2: Data schema created for B2B and B2C Olist Dataset.***

### **TECHNOLOGIES**

Cleaning: R studio

Big data: MapReduce, Apache Pig, Apache Hive,

shell script Sqoop, Apache Pig

Virtual machine: Cloudera QuickStart

Databases: MySQL, Hive, HBase

Visualization: Tableau

### **FRAMEWORK BUILD**

The architecture of the project consists of

• All B2B and B2C CSV formatted data in the R data frame to check the sample of the data, calculating the missing value and finding datatypes of the column in the dataset.

Splitting the date column MM/DD/YYYY HH:MM: SS into 4 different columns Month, Date, Year, Hours using string split function in R.

• Shell Script: Header and ‘ “ ” ’ removal from the file generated through R and created files with

m\_ prefix for loading data into Mysql and “h\_” to load data into Hive tables.

• Created table and data load into MySQL tables.

• With Sqoop command created and loaded MySQL data into Hive.

• Hive tables created for “h\_” files and loading data from files to these tables.

• Separate pig script and hive views are created for analysis and then call them in Shell Script.

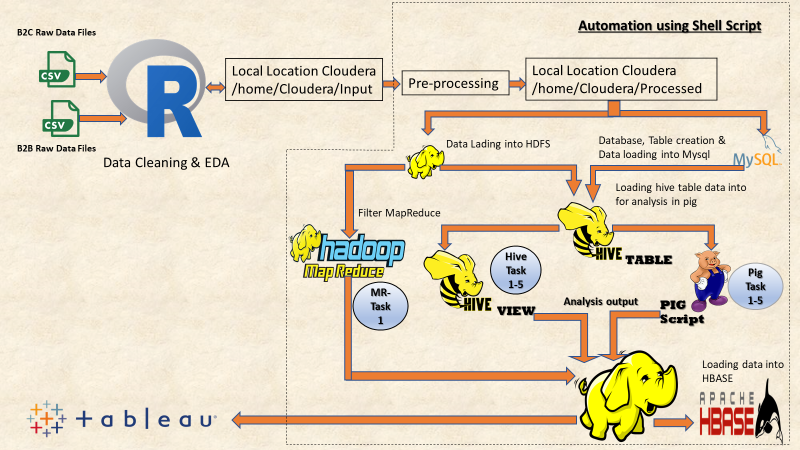
• HBase table created to store the result of Pig, Hive, and Map

Reduce tasks.

• Output: Hive\_task1, Hive\_task2, Hive\_task3, Hive\_task4, Hive\_task5, Pig\_task1, Pig\_task2, Pig\_task3, Pig\_task4, Pig\_task5, MapR\_Task1

### **IMPLEMENTAION**

#### **Process Flow Diagram:**

Process flow diagram defines the architecture of the project and showed how these all big data technologies used for the analysis of the data. As mentioned in the Implementation process cleaned and transform datasets are

***Fig 3: Process flow diagram showing loading of data***

loaded into Unix environment for further processing. Tables are created in MySQL and Hive from this dataset and data is loaded in the tables. Hive views are created for 5 different use cases and Pig Script is create for 5 different use cases. All this analysis is done on the table present in Olist database of Hive. Result of Pig, Hive, and MapReduce is stored into the HBase. This result is visualized through Tableau.

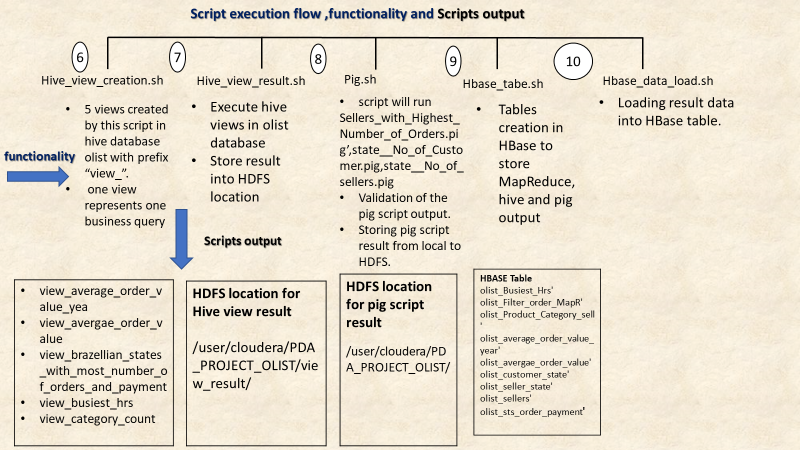
#### **ii) Script execution Sequence, Functionality and Result:**

This section explains working and implementation of each script. 10 sections for 10 different script is created and for evidence original script, log files and images are attached.

***Note: Please zoom small images***

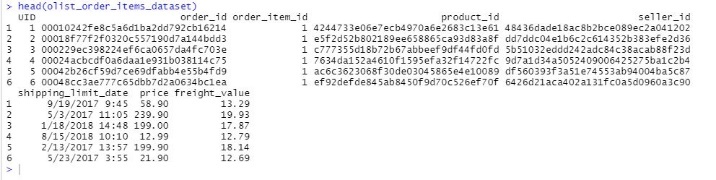


***Fig 4a: flow chart show Data load and analytics script execution***

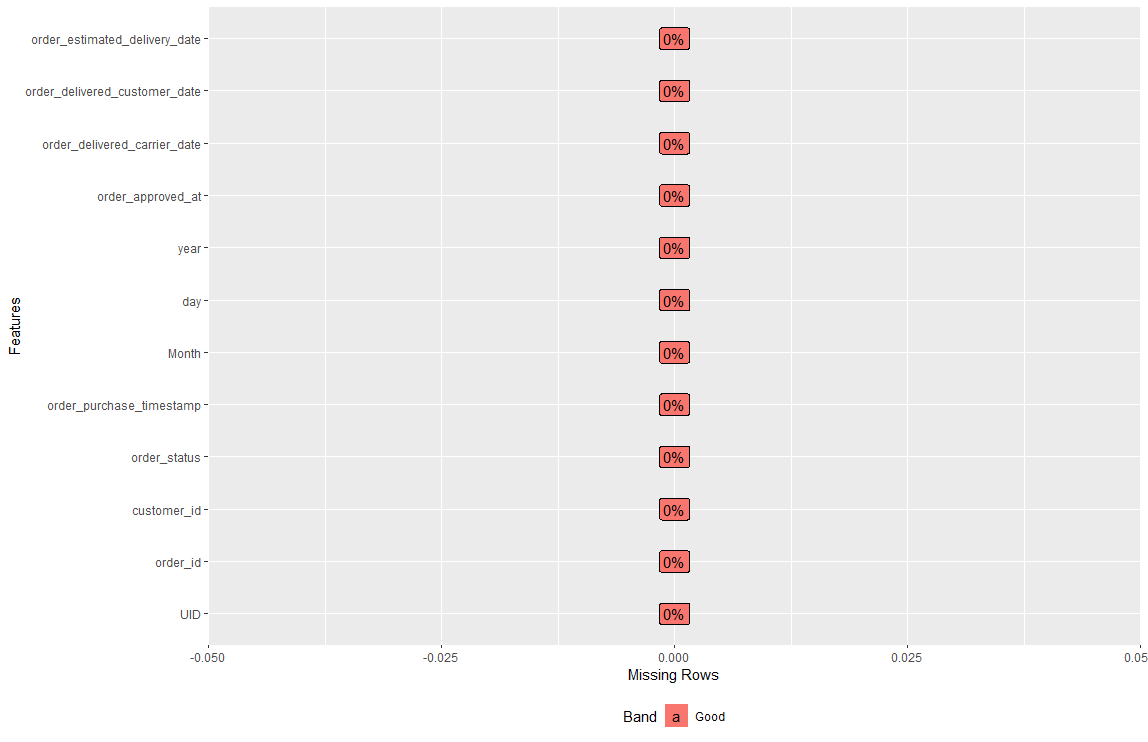


***Fig 4b: Defines the Data load and analytics execution***

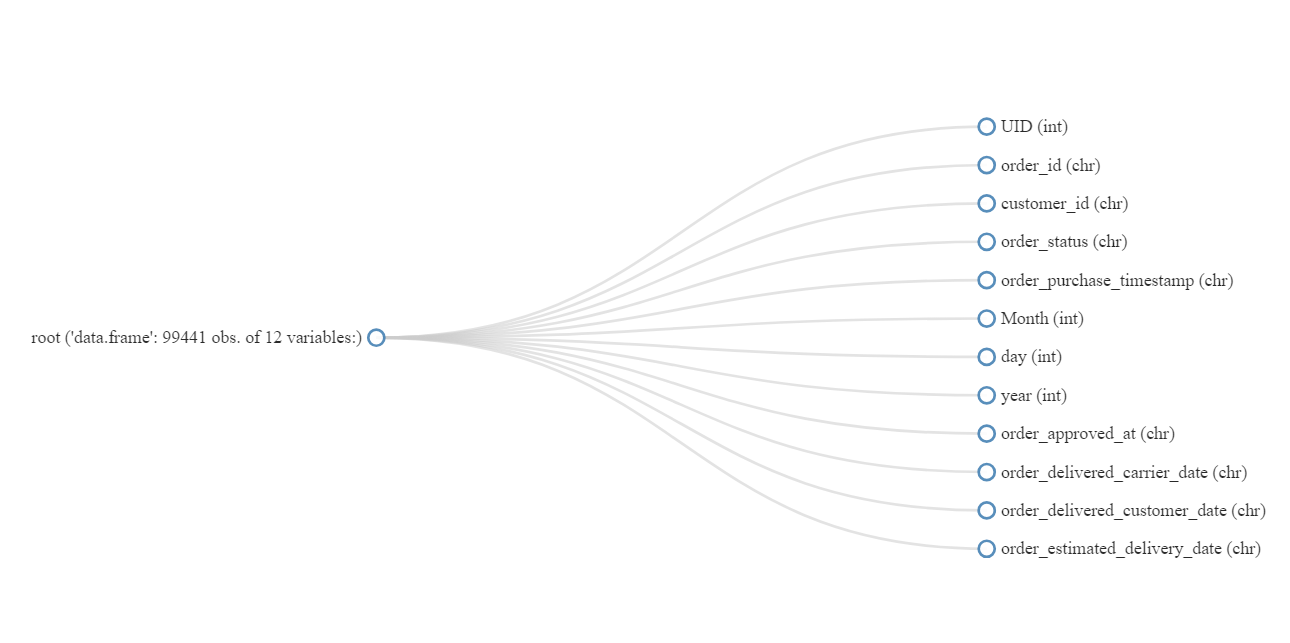
All scripts ae executed as per the process flow diagram. Before Shell script, R scripts are used for initial EDA and cleaning process. a screenshot shows the result of the R scripts. Data is loaded into the R data frame and sample data is checked***, Str plot*** function from library ***Data Explorer*** is used to plot the graph of the variable with the column and this is done for all data frame. Similarly, Missing values identified and removed from the dataset.



***Fig 5: Sample data of order***



***Fig 6: missing value plot for order dataset***



***Fig 7: fig shows structure of the order dataset.***

It is not possible to the attached result of all screenshot so attaching result for only ***orders dataset*** and attaching ***R result*** word file which includes evidence for missing value, sample data, and structure of all dataset and attached **R** code.



***Shell script explanation***

##### Pre-processing.sh

This script will remove the header and “‘” from the data set.

The output of the R script is used as input this script.

To load data into MySQL or Hive it is important to remove the header as it is considered as data. The output of R script contains header and it is removed by **“sed”** command of shell script.

Pre-processing.sh shell script file is attached input and result of this script is showed.



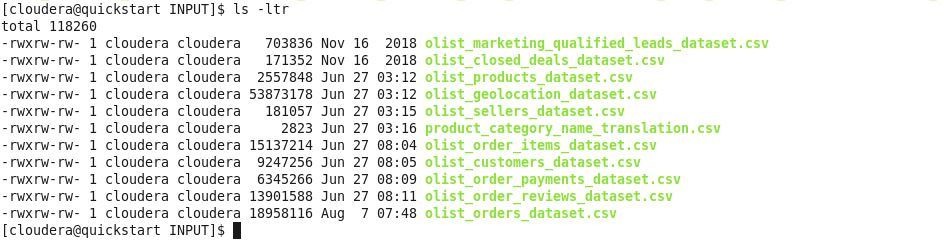


Fig 8: Input to files to pre-processing.sh



Fig 9: Shows output to pre-processing.sh

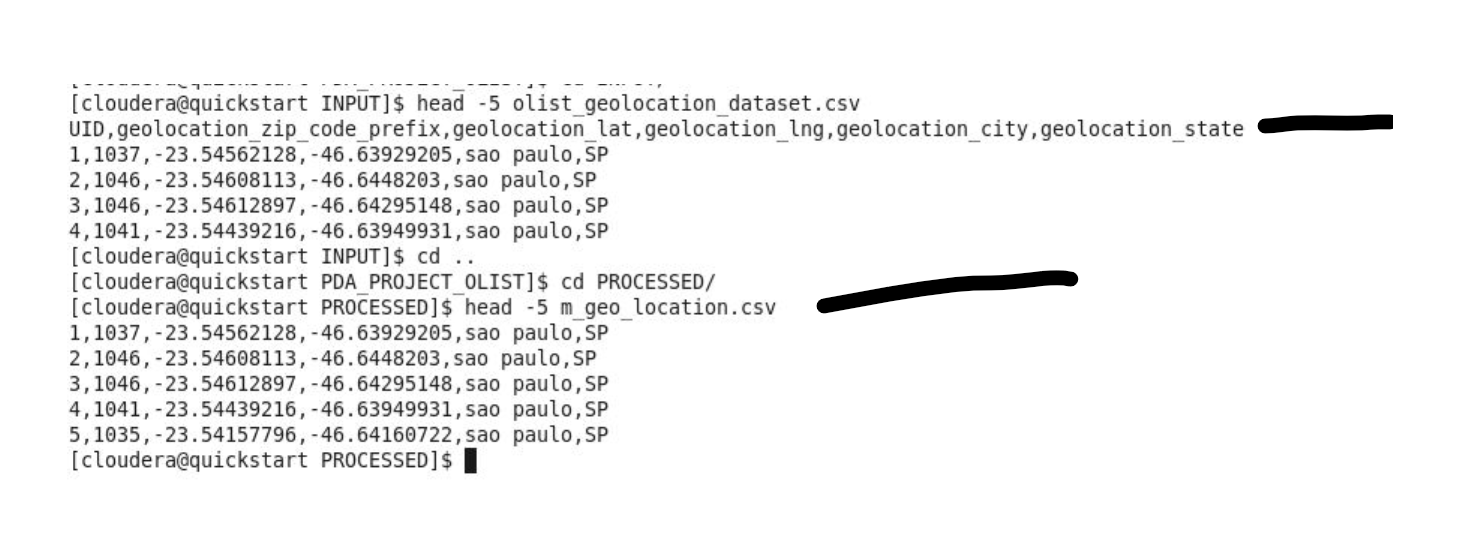


Fig: 10 output of Pre-processing.sh

##### Mysql.sh

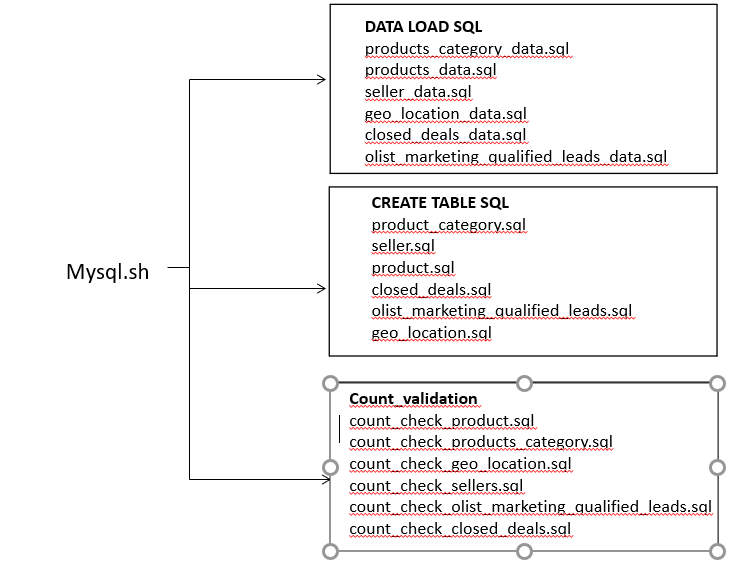
•        create database and table and into MySQL from

“m\_” Files

•        Load data into MySQL Table

•        Count verification of data

Fig shows structure of MySQL.sh .This file contains different files for above operation and shown in the figure. These .sql files with Mysql.sh and result is attached in this section.



**Fig. 11: Structure of MySQL file**

Below code explain the contain of each .sql file for sql table geo-location as example.

1. Table creation sql

Script to create geo location table

-- Create table Geo\_location

use Olist;

CREATE TABLE Olist.geo\_location

(

UID varchar(50) COLLATE utf8\_unicode\_ci,geolocation\_zip\_code\_prefix varchar(50) COLLATE utf8\_unicode\_ci, geolocation\_lat varchar(50) COLLATE utf8\_unicode\_ci,

geolocation\_lng varchar(50) COLLATE utf8\_unicode\_ci, geolocation\_city varchar(50) COLLATE utf8\_unicode\_ci,

geolocation\_state varchar(50) COLLATE utf8\_unicode\_ci);

1. Data load sql

Script geo location data .sql

LOAD DATA INFILE "/home/cloudera/PDA\_PROJECT\_OLIST/PROCESSED/m\_geo\_location.csv" INTO TABLE Olist.geo\_location COLUMNS TERMINATED BY ','

OPTIONALLY ENCLOSED BY '"' ESCAPED BY '"' LINES TERMINATED BY '\n' IGNORE 1 LINES;

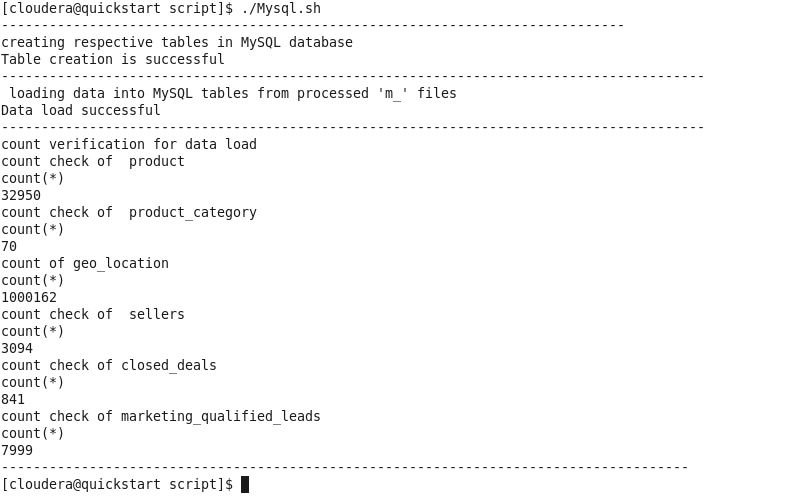
1. Count validation

select count (\*) from geo\_location;

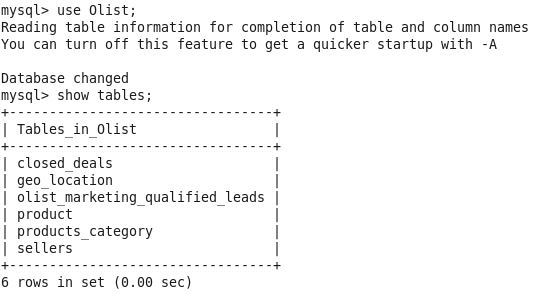
All Sql files are attached for evidence





******

*Fig 12: script execution log of Mysql.sh*

**

*Fig 13: Tables after execution of Mysql.sh script*

##### Sqoop.sh

This file will load Tables of MySQL database Olist generated n previous script into Hive database with data and schema. It is not possible to take a screenshot of Sqoop script, so log file is attached for evidence which consist of Table creation logs and Count validation.



Fig14 Sqoop.sh script

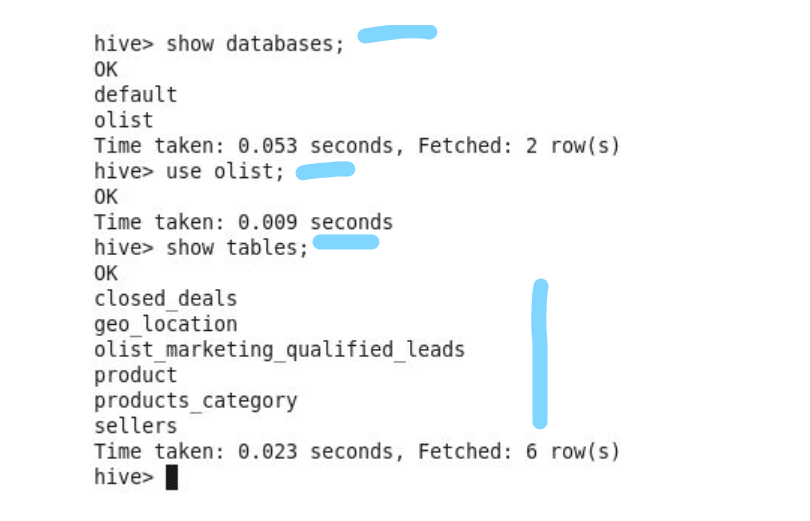


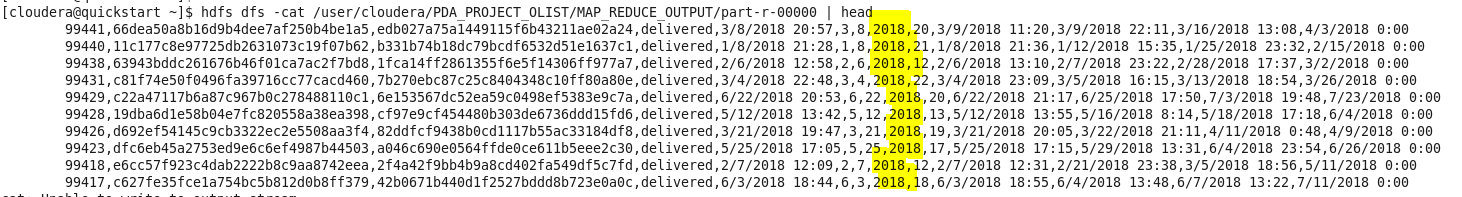
Fig 15: Sqoop output showing created database olist and tables.

##### MapReduce\_filter.sh

* Creation of HDFS directory
* Loading file from local to HDFS for Hadoop MapReduce processing
* Processing HDFS data with jar file (Filtering program to keep data of the year 2018 and remove all other data ) created through eclipse.
* Storing result on HDFS output directory

Attach MapReduce script contains all this functionality



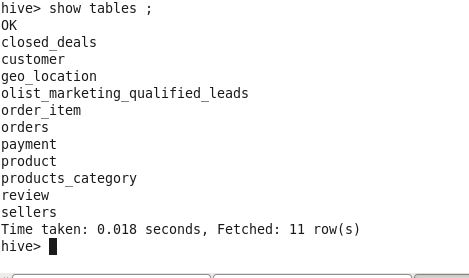


***Fig 16: Map reduce filter output on HDFS***

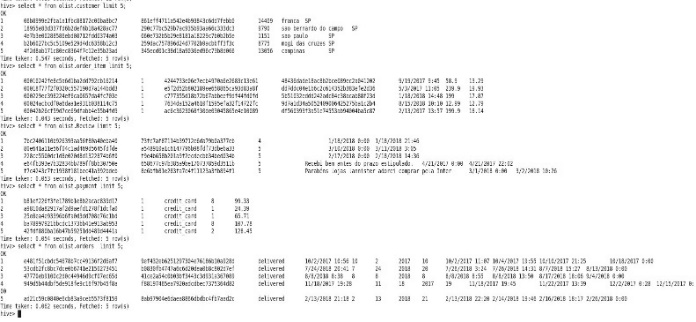
##### hive\_table\_creation.sh

creation of the schema on the **“h\_”** files stored by pre-**processing.sh** into a Hive and load the data from each file and finally count verification of each file.

******

***Fig 17: hive tables in database olist***

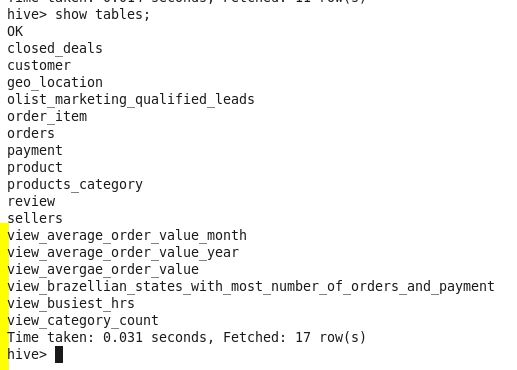
******

***Fig18: data validation for tables created by hive table creation script***

##### Hive\_view\_creation.sh

Hive views are created for analysis. Total 5 views are created for different use- case by joining tables in database olist. All view creation query process automated through Hive\_view\_creation.sh script.



***Use case 1:***

***Brazellian\_States\_With\_Most\_number\_of\_Orders\_and\_payment.sh***

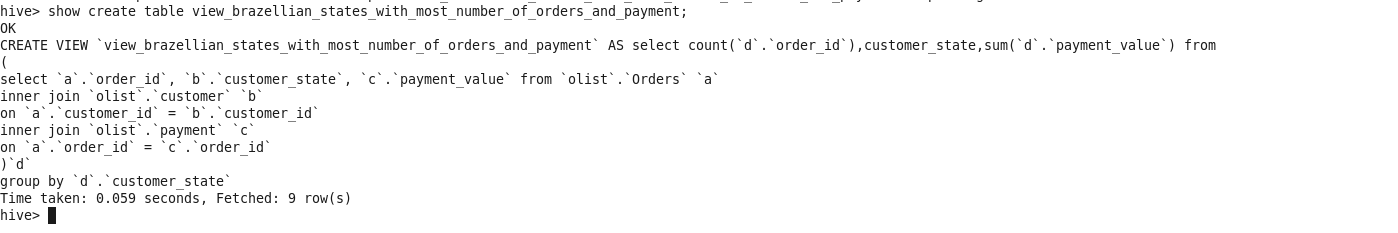
*Table used: Orders,* *customer,* *payment*

*Join condition:*

*Order and customer = customer\_id*

*Order and payment = order\_id*

***Query:***



***Use case 2***

***Identifying Average order value per month for brazil***

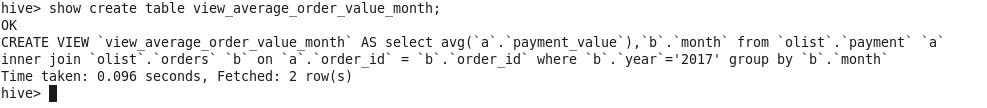
***For year 2017***

*Table used: payment, orders*

*Join condition:*

*Order and payment = order\_id*

***Query:***



***Use case 3***

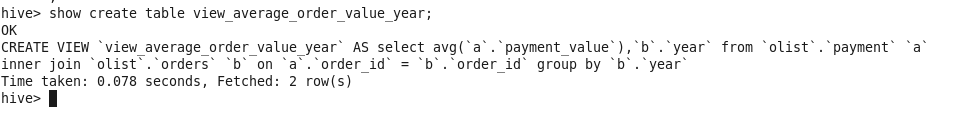
***Identification of average order value for each year***

*Table used: payment, orders*

*Join condition:*

*Order and payment = order\_id*

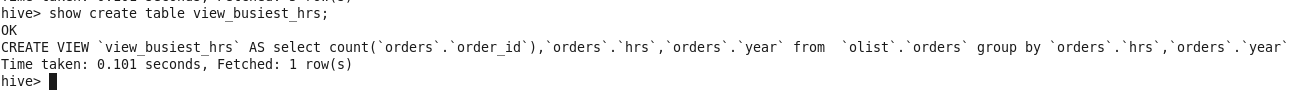
***Query:***



***Use Case 4***

***Identification of Busiest hours from the olist dataset***

*Table used: orders*



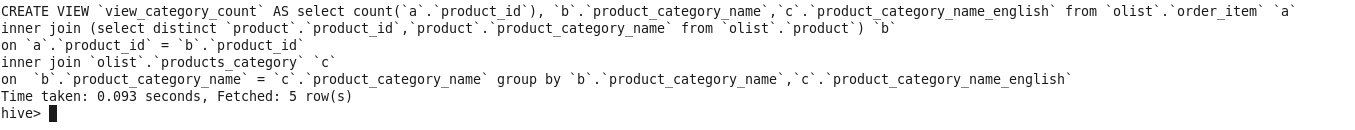
***Use case 5***

***Identification of product sell of category as per orders***

*Table used: order\_item, product, products\_category*

*Join condition:*

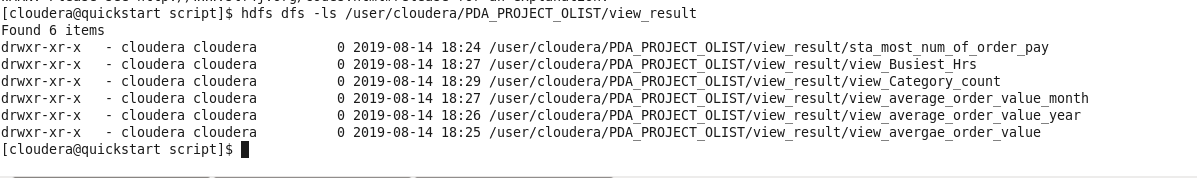
*order\_item and product = order\_id*



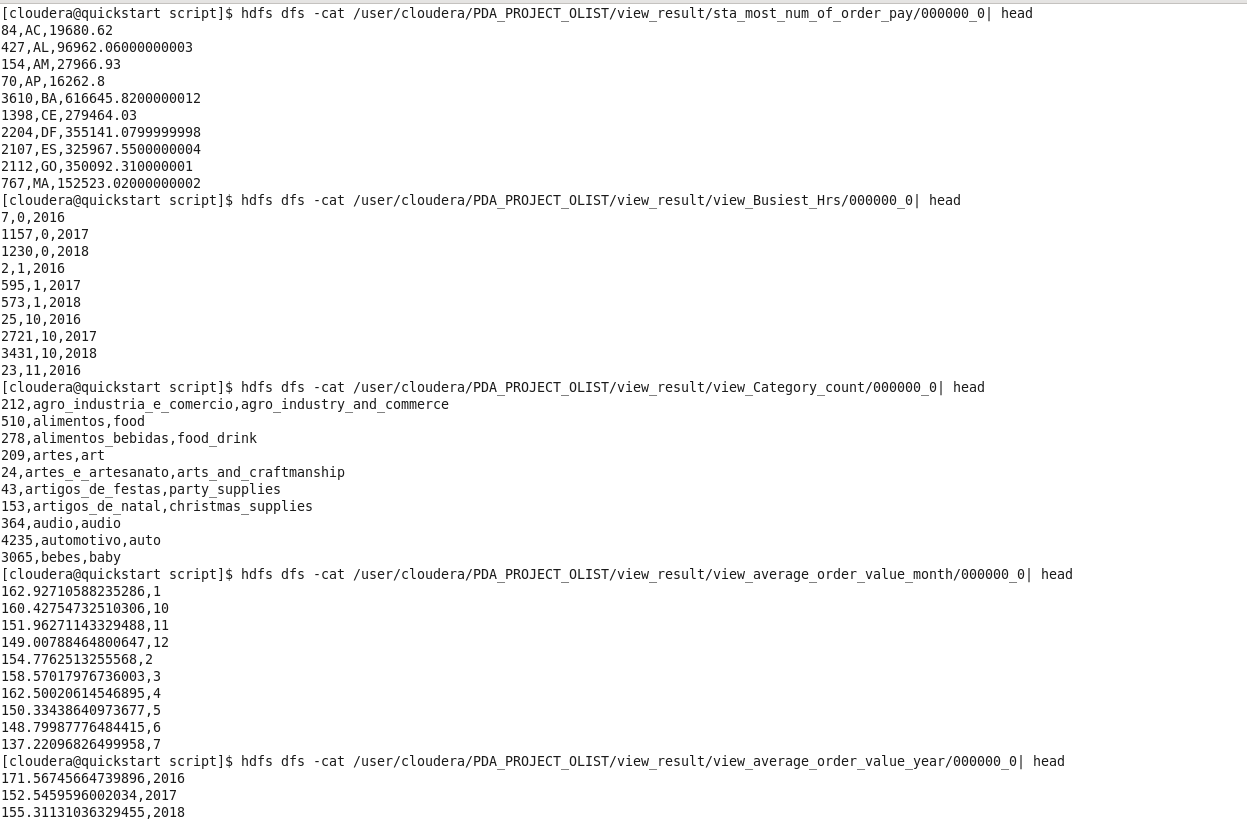
##### Hive\_view\_result.sh

Views created by hive view creation.sh is executed with this script and stored results of each hive in a different location on HDFS. This is because after processing with the hive query results generated with the file name **“000000-0”. If we stored result of all views in same output directory it will replace the previous result.** So Each **use case** result stored in a separate location on HDFS.

*** ***



***Fig 19: Hive result stored on HDFS location.***



***Fig 20: Data verification of the view results on HDFS directory***

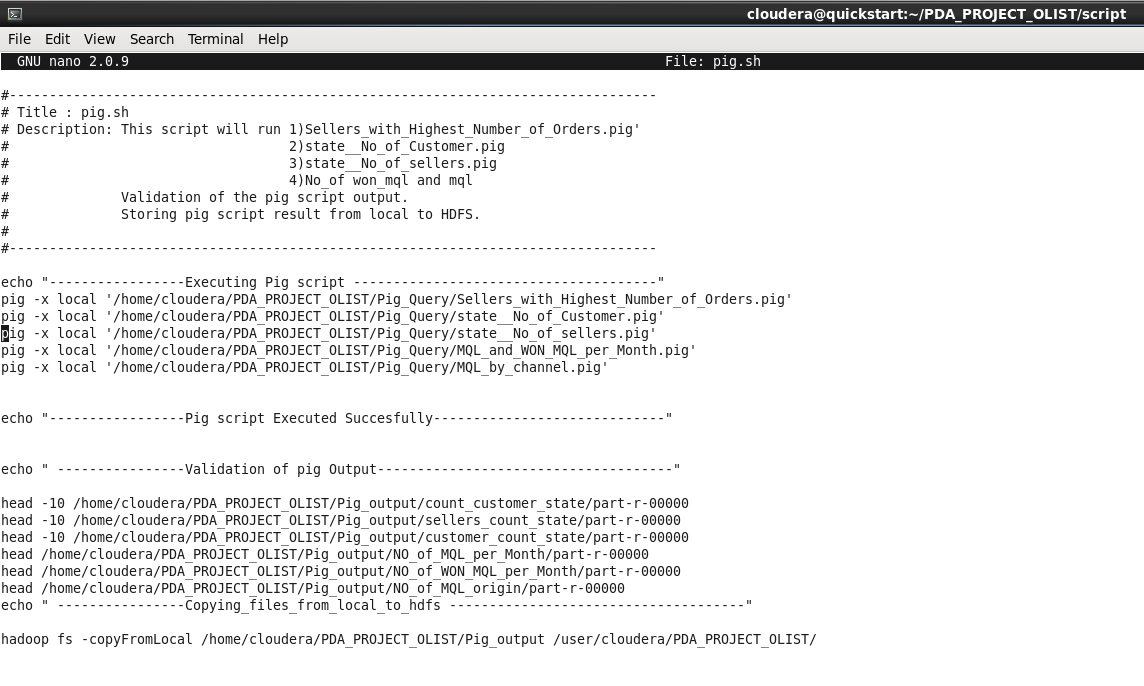
##### Pig.sh

Pig scrips are created for analysis purpose. Total 5 scripts are created for 5 different use cases. Data for pig analysis is loaded from Hive table directly in pig through ***HCatLoader***. all 5 scripts are call inside the shell script to automate the process.

**Functionality**

* Execution of pig scripts
* Validation of the pig script output.
* Storing pig script result from local to HDFS.

Pig.sh



Below is description of each pig script used in shell script pig.sh

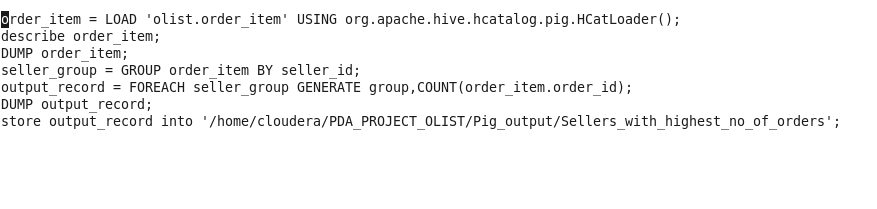
***Use case 1***

***Identification of the sellers having highest number of orders.***

*Pig script name: Sellers\_with\_Highest\_Number\_of\_Orders.pig*

*Table used: order\_item,*

***Query***



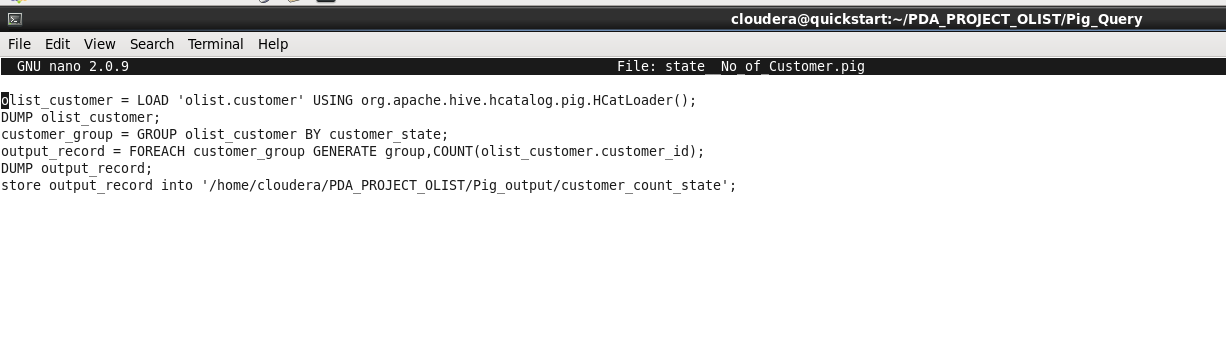
***Use case 2***

***Geo-location analysis to identify of customer in each State***

*Pig script name:* *state\_\_No\_of\_Customer.pig*

*Table used: olist.customer*

***Query***

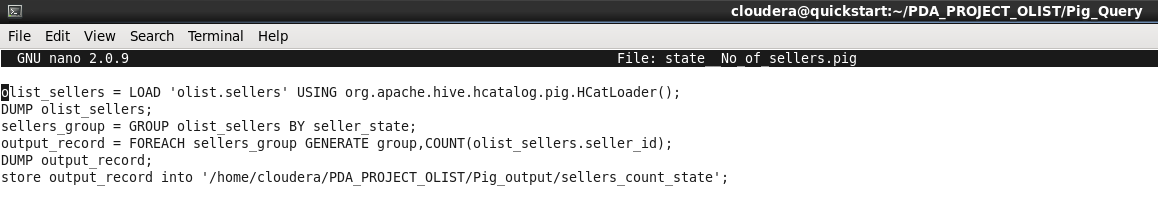


***Use case 3***

***Geo-location analysis to identify of customer in each State***

*Pig script name:* *state\_\_No\_of\_sellers.pig*

*Table used: olist.sellers*



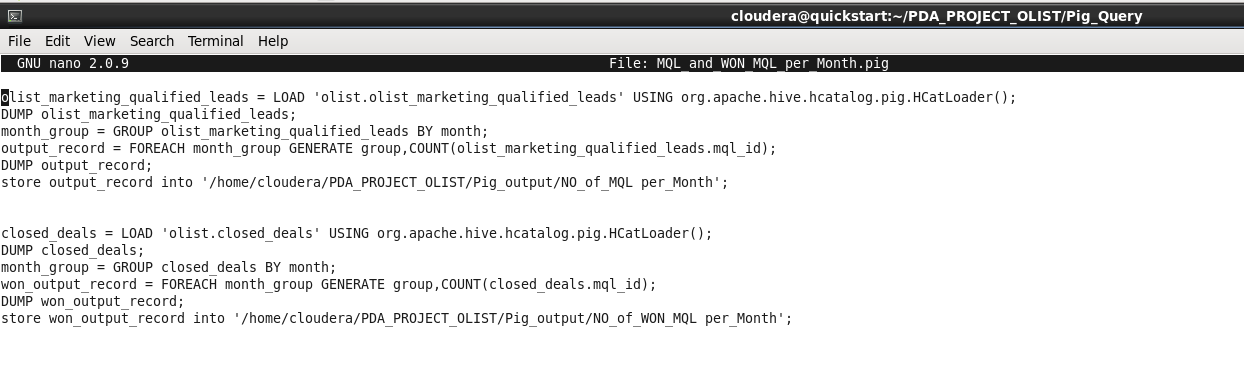
***Use case 4***

***This is for business analysis where request of total marketing qualified lead (MQL) and corresponding Won MQL*** ***are identified.***

*Pig script used: MQL\_and\_WON\_MQL\_per\_Month.pig*

***T****able used:* *olistmarketing\_qualified\_leads*

*closed\_deals*

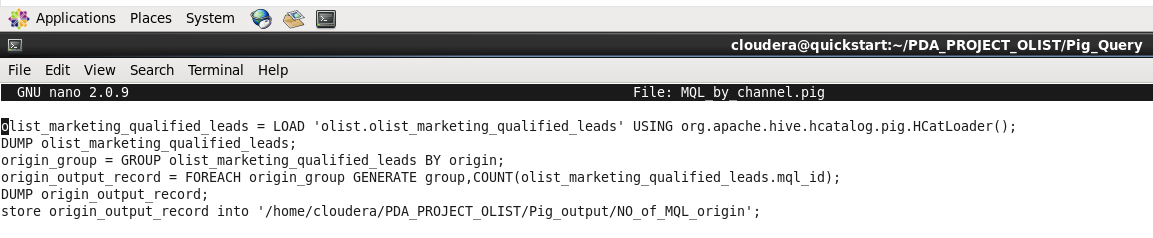


***Use case 5***

***Analysis of Channels through which MQLs are requested***

*Pig script used: MQL\_by\_channel.pig*

*Table used: olist\_marketing\_qualified\_leads*

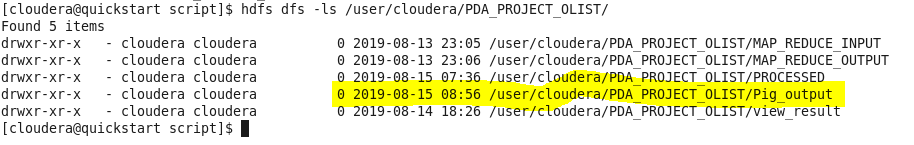


***Below files are shell scrip pig.sh which will execute all 5 pig scripts and result showed***

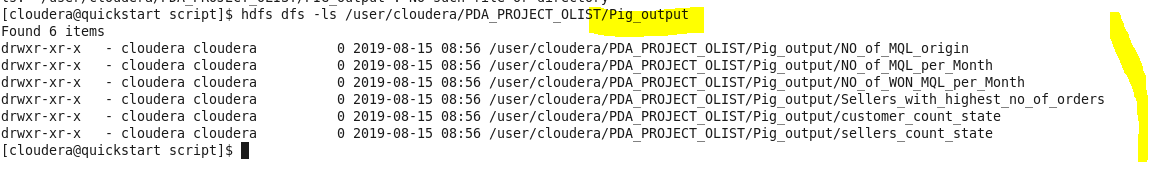
******

******

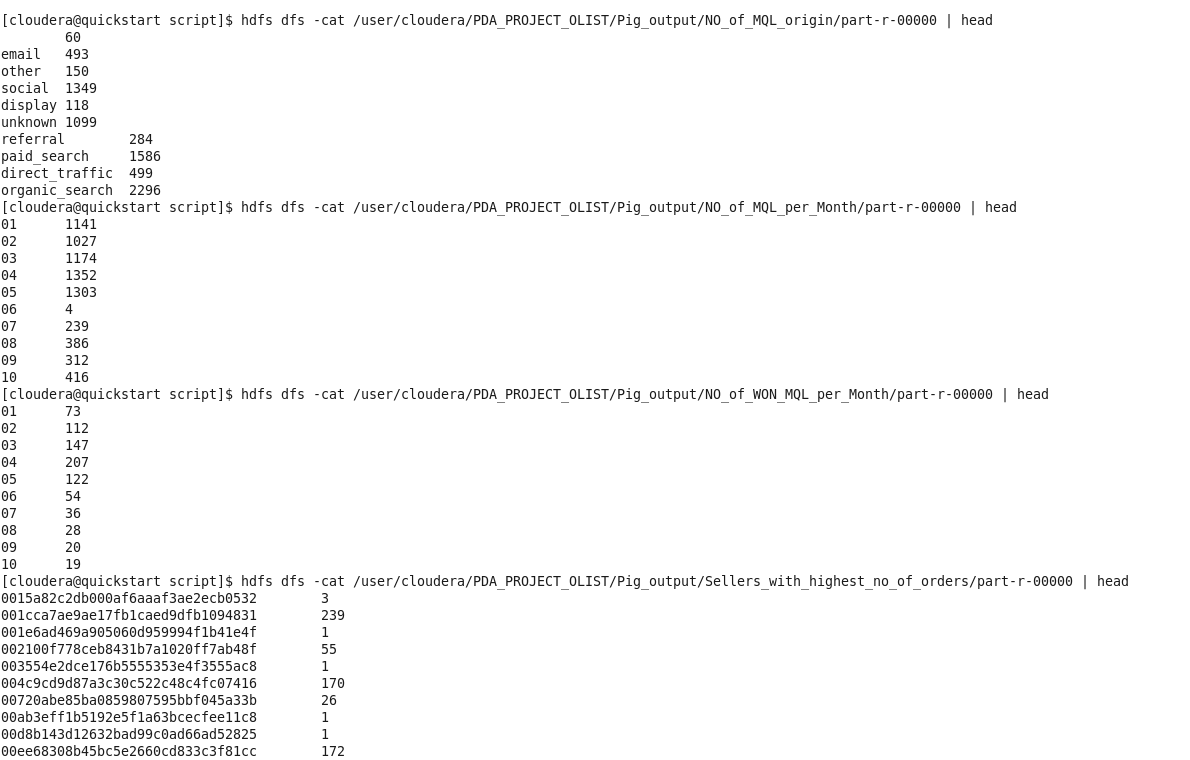
Pig output directory on HDFS



Output of Each Use\_case stored in differnent folder on HDFS



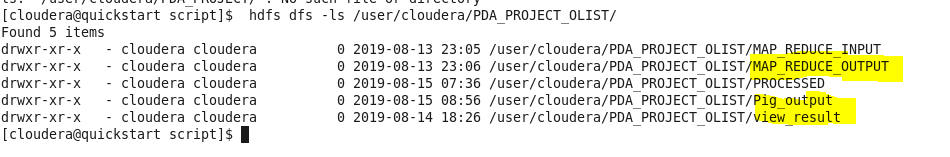
***Data validation of Pig Script***





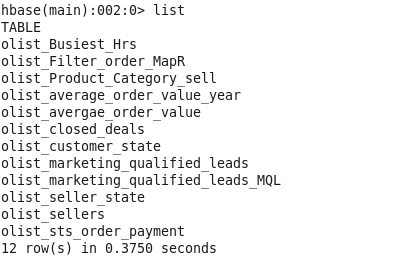
##### Hbase\_tabe.sh

Tables are created to store generated output files of Mapreduce, Hive and Pig.





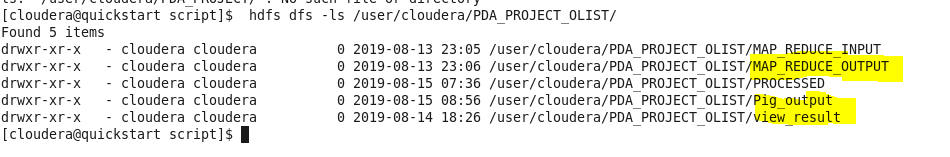
Tables created through hbase\_table\_cre.sh



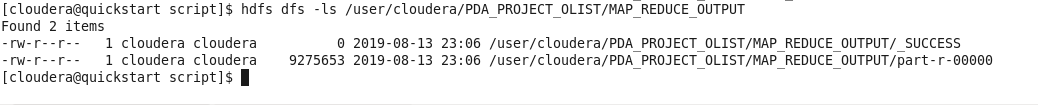
##### **Hbase\_data\_load.sh**

Script is used to load the tables created by hbase table Cre script from the HDFS location.

HDFS location of results generated by Pig, Hive and MapReduce

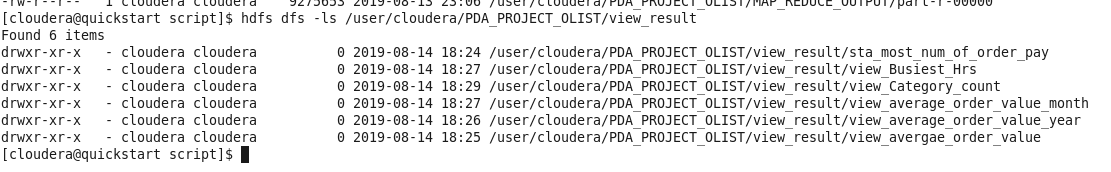


**MapReduce analysis output location**



**Hive view output**

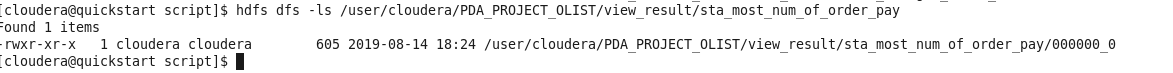
Each use case result is stored in respective folder during the execution of Hive view script and the location is below.



*Example*

***Hive use case***: Identification of busiest hrs.

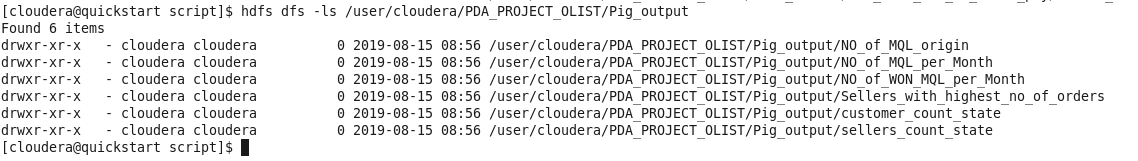
result file is generated with 000000-0 in folder



Similarly, all files are stored in each subfolder for respective use case

**Pig script output**

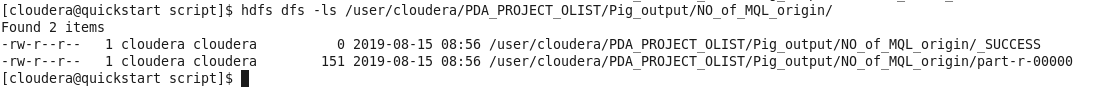
Each use case result is stored in respective folder during the execution of Pig script and the location is below



*Example*

***Pig use case***: Channels through which MQLs are generated

result file is generated with part -r- 00000 in folder.



Similarly, all result files are stored in each subfolder for respective use case These files are then loaded into HBase table.

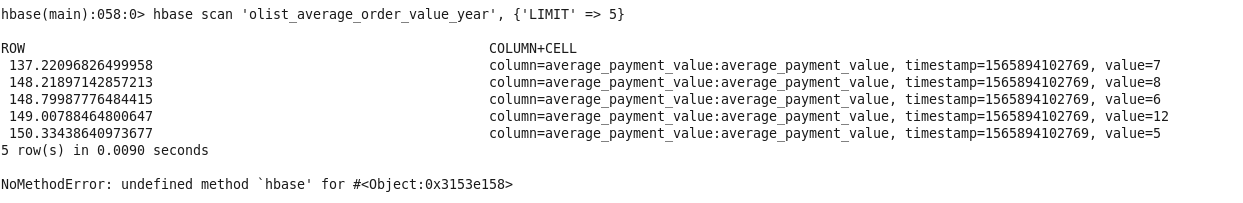
**Hbase data load script and result log file**

****

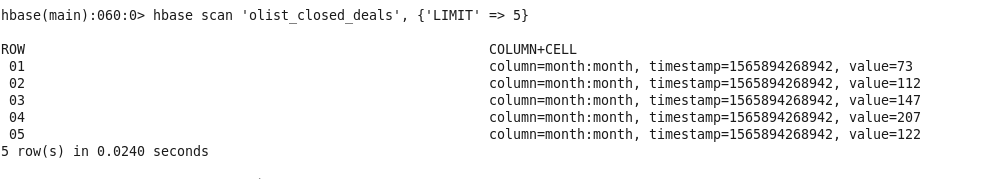
Log file contains evidence of data load for each table attaching screenshot of few Hbase tables with limit 5, showing data load successful.

**Data load check**

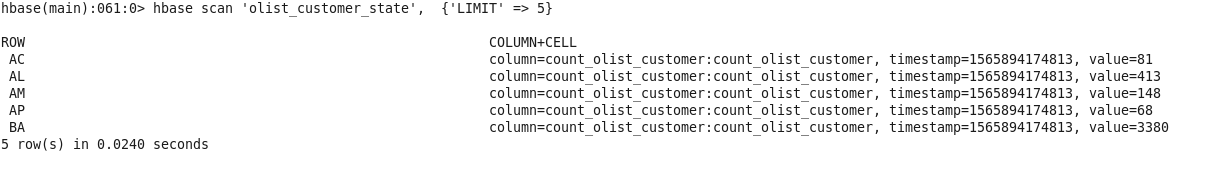
**HBase table olist\_average\_order\_value\_year**



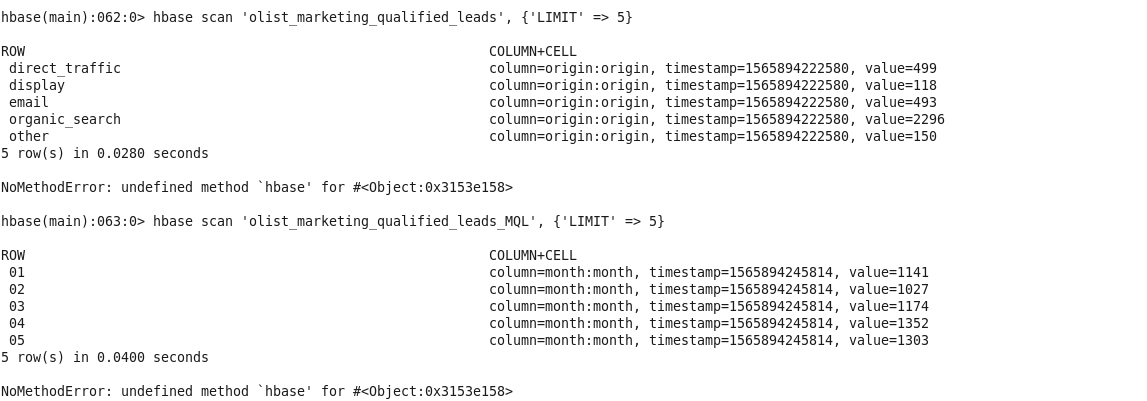
**HBase table: Closed deal:**



**HBase Table: Olist customer state**



**HBase Table: Marketing qualified lead and Marketing qualified lead**



##### **Master.sh**

This script call functionality of all shell script into single script. This avoid manual execution of each script and all script can execute on single click.

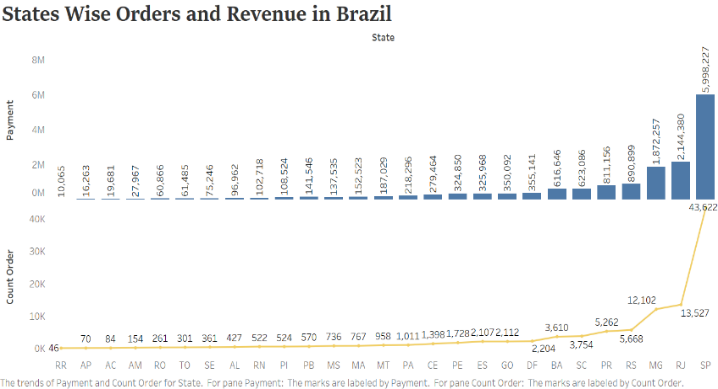


# VISUALIZATION TASKS (OR) USE CASES

### Business to customer Analysis

#### case I

***Identification of Brazilian state with orders and revenue of the state:***

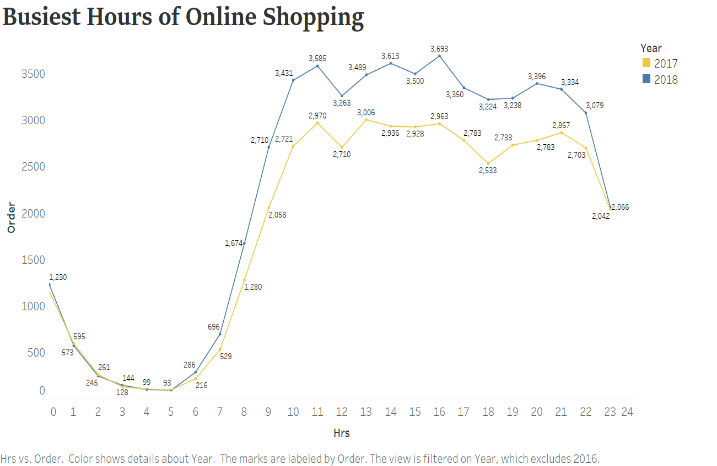


***Fig: showing distribution of orders and the revenue generated from the state for Olist.***

Inventory management is essential for sellers. Olist is the platform which provides merchants to connect to the marketplaces in brazil. From the graph, it is clear that only a few states of brazil are generating 9o% of the revenue as maximum orders made from those states. So it is essential to provide more details in order Fulfillment centers in these states for better customer experience.

#### case II:

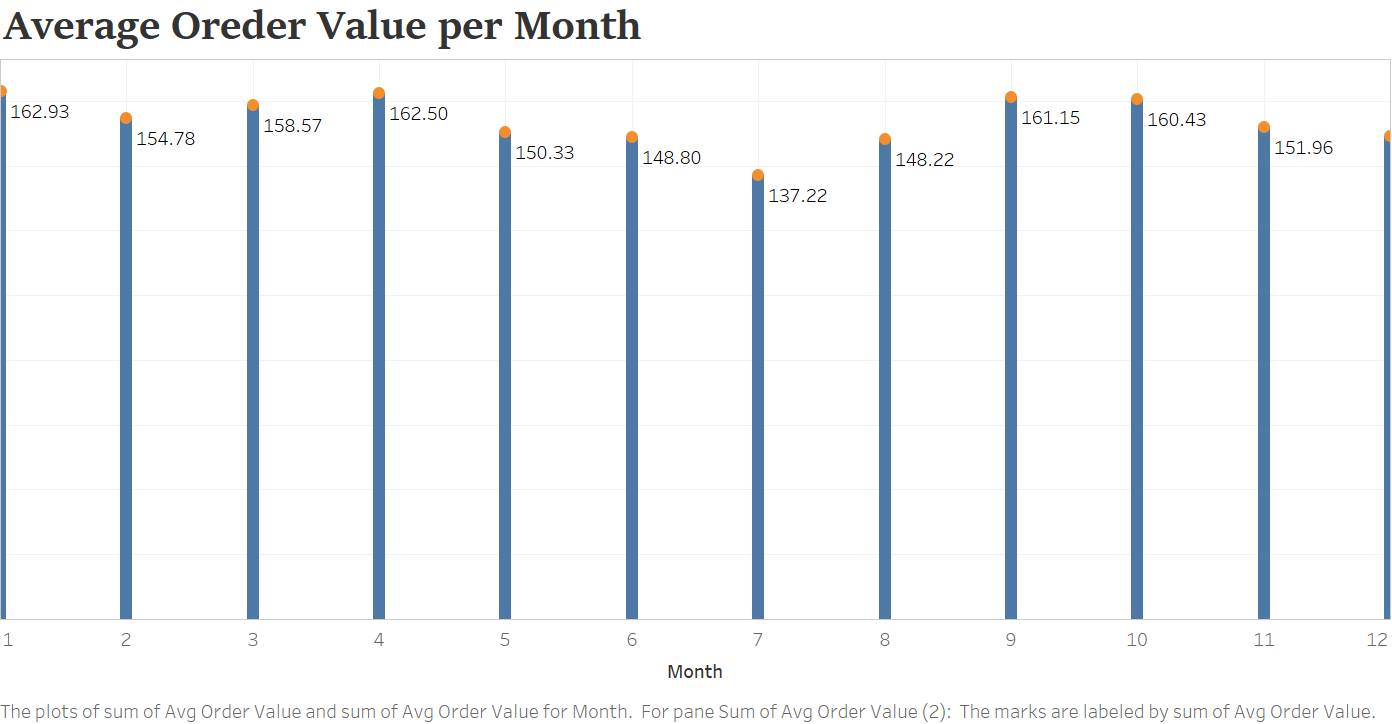
***Identification of Busiest hours of online shopping***



Consumer behaviour is essential for the e-commerce business, and it is necessary to identify when do people shop online. The trend of shopping identified with the help of data of 2017-18. X-axis showed hrs. And the y-axis is order data. The average amount of the orders is placed in between 10 am to 9 pm. The peak time for online shopping is 11 am and 4 pm. There is a various reason such as age, pc knowledge. Most of the time people with age between 35- 45 place the orders in between 11 am – 4 pm.

#### Case III:

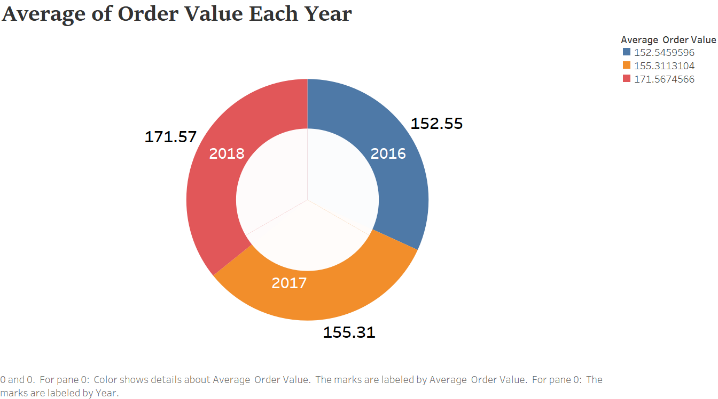
***Identifying Average order value per month for brazil***



All data-driven business need average order value metric, which is essential to understand the profit and scaling the revenue growth. Knowing the amount spend on the product allows Olist to define the marketing strategies and price planning as average order value shows the average amount of money spend on the product. The graph shows the average amount of spending on a product, and Average cost decreased in the 7th month. Average monthly value dropped from 5th – 7th month and then increased, a pattern repeated for 9th -12th month again.

#### Case IV:

***Identification of average order value for each year***

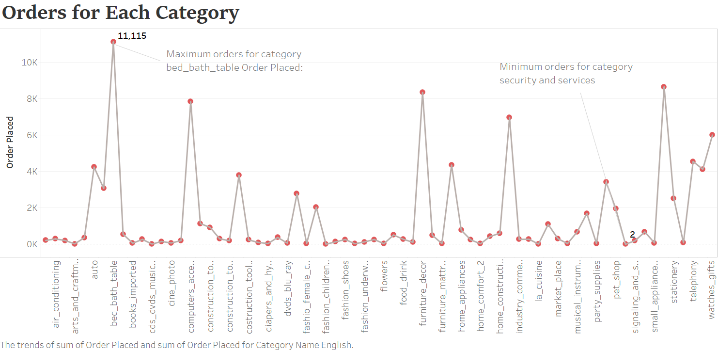


The previous case showed the monthly average order value.

In this case, the yearly average order value calculated, which showed that average order value (AOV) in 2018 is maximum, and in the year 2016, it is minimum. This shows average order values increase exponentially.

#### Case V

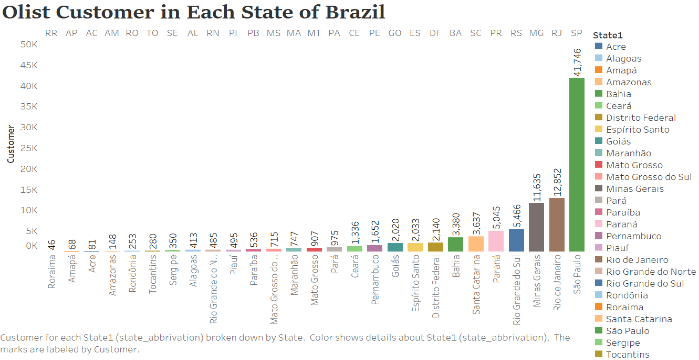
***Identification of product sell of category as per orders***



For inventory management, it is essential to understand selling frequency of the product and this analysis showed which product has the maximum number of the orders. From the graph product category, bed bath table has maximum purchase by the buyer and signal and security has only two orders from the year 2016-18.

#### Case VI:

***Geo-location analysis to identify of customer in each State***



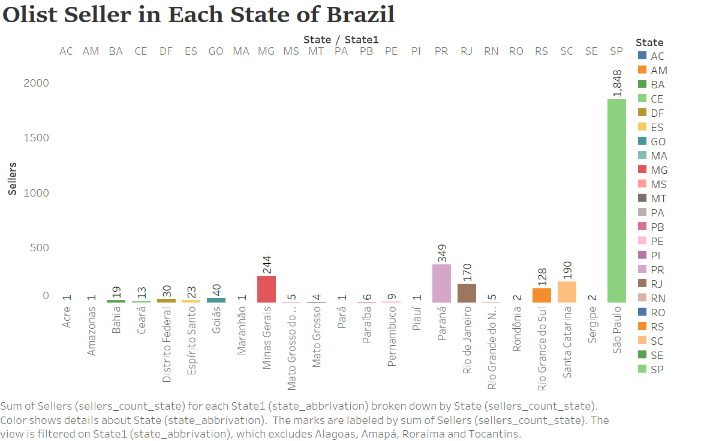
In the e-commerce industry, customer location is essential as a lot of things such as the delivery of the product, an incentive of the sellers, and managing human resource based on customer location. This analysis showed the customers in each state. This is a useful insight to make pricing strategies and business strategies. Understand of the place is vital for the e-commerce industry as it shows the offline environment and referral to the new customer With the increase in the mobile shopping location is extremely important. This chart shows the analysis of the customer location according to the state. Sao paulo contains around 90% of the overall customer. State Roraima has minimum Olist shopper. This graph also helpful for the Olist team to increase the customers according to the location.

## **Business2 Business**

#### Case VII

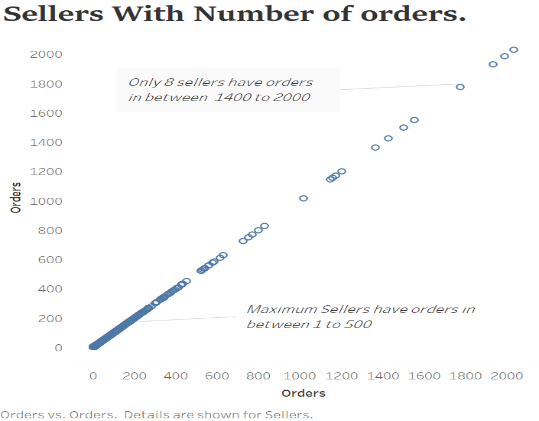
***Geo-location analysis to identify of sellers in each State***

Geo-location analysis to identify sellers in each State has done in the previous case. Olist provides a platform to the small seller for selling their product. Customer location analysis is useful to decide the seller’s requirement for the Olist team. Maximum sellers required where the density of customer is maximum. To sale, the product on the Olist platform seller must send the request through the various channels available. Sales development representative (SDR) decide whether to close the deal (accept the lead) or lose the deal (lose the lead) after consulting with MQL. Once the SDR allows MQLs application, lead become a seller.

Below analysis shows the distribution of seller is maximum where the frequency of the customer is the high customer. X-axis shows the

#### Case VIII:

***Identification of sellers with highest Number of orders***



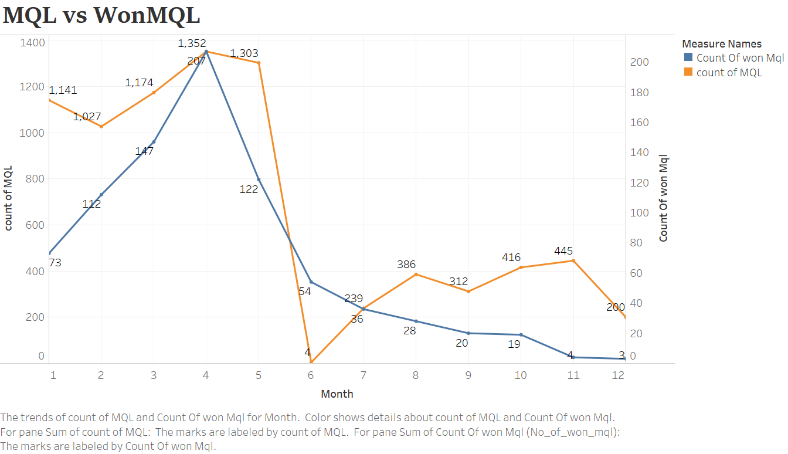
Sellers have a different percentage of incentives for sailing their product in Brazilian marketplaces. Olist provides inventory management and order management to these sellers. From their perspective, they want to increase the number of merchants through Olist. From Olist, it is essential to have a faithful and valid seller who can fulfil the customer order without any issue. Olist can also provide extraordinary benefits to sellers having maximum order and encourage them to sell their product through Olist. This way, they can also grow their own business to maintain a healthy relationship with the sellers. The graph shows the sellers with the maximum number of orders in each state.

Maximum Olist sellers have orders in between 1- 500 order and only eight sellers have order between 1400 – 2000. In this way, olist can categories seller according to purchase.

#### Case IX:

***Requests marketing qualified lead (MQL) and corresponding Won MQL are plotted.***

MQL is a marketing qualified lead request Olist to sell their product. An executive from Olist after consulting with MQL decide whether to accept MQL or reject MQL. This analysis shows a graph of the comparison of accepted MQL represented by Won MQL Deal with the requested MQLs. Data use for this analysis is from Jun. 1st2017 and Jun 1st, 2018



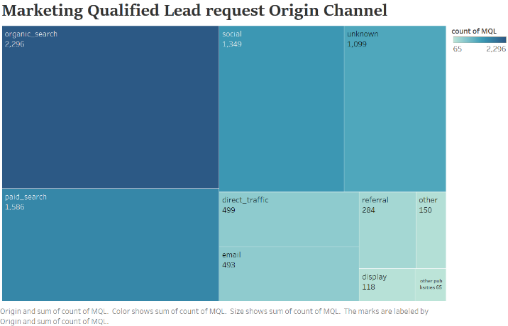
The graph shows the No of MQL requested by orange line and Won MQL by blue. In the month of 4th more than 200 MQLs accepted by Olist. from the graph there are a lot of request by the MQLs to Olist in the first five months of the year and decreases suddenly in the month

#### Case X:

Marketing qualified lead can generate their request through different channels provided by Olist. One MQL can use multiple channels to request Olist. This analysis focused on the identification of the channel through which maximum requests generated. In the dataset there is distinct MQL id it means the request not made through multiple channels.

The graph is shown in the figure, and it showed that there are a total of 10 categories which are used by Olist and among these categories maximum request is generated through organic search

Which are around 2300 and minimum offer made by other publicities 65.



# programming and data handling challenges

### order dataset contains purchase column in timestamp and for purpose of the project we need three different columns month, date, year and hrs. Creation of these four columns by Spliting timestamp column is tried in hive view directly using str-split function but this coumn consis of two delimiter “/” for date, month, year and “ ” for hrs and second

sample data 7/24/2018 20:41

This s handled in R after completion of EDA

### Hbase data load : Output of hive view is comma separated and output of pig script is Tab separated.

While loading the data output of MapReduce for pig task showed read bytes =”xxxx” , write bytes “ ” and bad lines =”xxxx” and this problem is solved by using

For hive: Dimporttsv.separator=','

For pig: Dimporttsv.separator=' '

### 

# **Conclusion**

In this study, we have outlined the analysis of B2B and B2C for retail data analysis. B2C analysis focused on the orders, Revenue, average order value in month & year, shopping hours and product sale. B2B focused on the

The seller who is selling the product on Olist platform and categorizing the sellers according to maximum sale and location. Request by MQL are also analyzed and compare with Won MQL. This helps to convert the MQLs into Seller for Olist team. Big data architecture is used in this case Three main patterns used in this analysis which is widely accepted in big data analytics project, and they are

*Input-Map-Reduce-Output*

*Input-Map-output -*

*Input-Multiple Maps-Reduce-output*

Database use in the projects is suitable for the available data set. Hive and Pig query both successfully used for analysis purpose. The shell script is successfully implemented as the wrapper script to automate various operation. Sales prediction can be implemented using the spark with the available framework.

**REFERENCES:**

**[1]** “Brazilian E-Commerce Public Dataset by Olist.” [Online].Available: https://kaggle.com/olistbr/brazilian-ecommerce. [Accessed: 15-Aug-2019].

**[2]** “Marketing Funnel by Olist.” [Online]. Available: https://kaggle.com/olistbr/marketing-funnel-olist. [Accessed: 15-Aug-2019].

**[3]** H. Manjunatha and R. Mohanasundaram, “BRNADS: Big data real-time node anomaly detection in social networks,” *2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 929–932, 2018.

**[4]** N. Verma and J. Singh, “An intelligent approach to Big Data analytics for sustainable retail environment using Apriori-MapReduce framework,” *Industrial Management & Data Systems*, vol. 117, no. 7, pp. 1503–1520, Aug. 2017.

**[5]** G. Zhu and X. Gao, “The Digital Sales Transformation Featured by Precise Retail Marketing Strategy,” *Expert Journal of Marketing*, Volume 7, Issue 1, pp. 72-76, 2019.

**[6]** X. Ye, Z. Lian, B. She, and S. Kudva, “Spatial and big data analytics of E-market transaction in China,” *GeoJournal,* Jan. 2019.

**[7]** S. Akter and S. F. Wamba, “Big data analytics in E-commerce: a systematic review and agenda for future research,” *Electron Markets,* vol. 26, no. 2, pp. 173–194, May 2016.

**[8]** S. Chandramana, “Retail Analytics: Driving Success in Retail Industry with Business Analytics,” vol. Volume 7, Aug. 2017.