UNIVERSITI MALAYA

WQD7007

Big Data Management

Group Project

Customer Segmentation for Targeted Marketing in E-Commerce

Group 8 & Group 10

Name	**Matric Number**
Low Boon Kiat	17138399
Zhou Yao	S2177633
Wen Yukun	22108184
Du Pengfei	S2130205
Ong Yi Kwong	S2181260
Koh Rong Soon	22061463
Chunli Wang	22064827
Meng Guan	S2179731

Contents

1.0 Introduction

1.1 Background Study

The Southeast Asian region has witnessed a surge in e-commerce adoption, driven by increasing internet penetration, mobile device usage, and a burgeoning middle class. With a diverse landscape comprising countries such as Singapore, Malaysia, Indonesia, Thailand, Vietnam, and the Philippines, the e-commerce market in Southeast Asia has become a focal point for both local and international businesses. Among the myriad of e-commerce platforms in the region, Shopee has emerged as a dominant player. Launched in 2015 by Sea Limited, Shopee has gained popularity for its user-friendly interface, diverse product offerings, and strategic marketing campaigns. Operating in multiple Southeast Asian countries, Shopee has successfully navigated the unique challenges presented by each market.

In the fiercely competitive e-commerce landscape, retaining customers is integral to sustained success. Customer retention not only ensures a stable revenue stream but also fosters brand loyalty and positive word-of-mouth marketing. However, the e-commerce sector faces several challenges in maintaining customer retention. First, the crowded e-commerce space in Southeast Asia poses a challenge for platforms like Shopee to stand out and retain customers amidst numerous alternative platforms. Second, rapid shifts in consumer preferences and behaviours make it challenging for e-commerce platforms to adapt their strategies promptly, potentially leading to customer disengagement. Therefore, understanding the diverse needs, preferences, and behaviours of the customer base is paramount in developing effective marketing strategies. Customer segmentation allows e-commerce platforms to categorize their users into distinct groups, enabling personalized and targeted approaches for better engagement and satisfaction.

1.2 Problem Statement

Shopee, as an emerging player in the competitive e-commerce market, faces the challenge of effectively understanding its diverse user base to optimize marketing efforts, enhance user experience, and drive platform growth. Based on Measurable AI's data in 2023 Q1, Southeast Asian consumers have showcased steadfast loyalty to Shopee over the years with an average customer retention rate of ~49% (Sheng, 2023). There remains untapped potential for improvement. The challenge lies in accurately segmenting a diverse customer base into distinct groups based on their purchasing patterns, demographics, and interactions with the platform to allow for targeted marketing. To address the problem statement, we examine the following questions to provide crucial insights that can inform customer segmentation strategies.

<u>Segmentation by geographical location</u>

1. Which cities are generating the highest orders?
2. Which cities are generating the most revenue?

<u>Time-based segmentation</u>

3. What are the peak order timings?

4. What are the peak order days?

Product preference segmentation

5. What are the best and the worst performing product categories in terms of number of orders?
6. What are the best and the worst performing product categories in terms of total revenue generated?
7. What are the best and the worst performing product categories in terms of review scores?

Payment behaviour segmentation

8. What are the most popular payment methods?

Service-related segmentation

9. How long does it take for the products to be delivered?
10. Do longer delivery times lead to poor review scores?

1.3 Dataset Description

The dataset used in this project is the Brazilian E-Commerce Public Dataset by Olist available on Kaggle because we did not manage to find any suitable datasets on Shopee (Francisco, 2021). Olist is an e-commerce platform in Brazil that enables small and medium-sized businesses to sell their products on various marketplaces. The dataset provides information of 100,000 orders made across multiple marketplaces in Brazil from 2016 to 2018, encompassing various dimensions such as order status, pricing details, payment and freight performance, customer location, product attributes and customer reviews.

Table 1: List of attributes in the dataset.

olist_orders_dataset	
Attribute	**Description**
order_id	Unique identifier of an order.
customer_id	Key to customer dataset. Each order has a unique customer ID.
order_status	Order status (delivered, shipped, etc).
order_purchase_timestamp	Purchase timestamp.
order_approved_at	Payment approval timestamp.
order_delivered_carrier_date	Order posting timestamp.
order_delivered_customer_data	Actual order delivery date to the customer.
order_estimated_delivery_date	Estimated delivery date that was informed to customer.
olist_customers_dataset	
Attribute	**Description**
customer_id	Key to the orders dataset. Each order has a unique customer ID.
customer_unique_id	Unique identifier of a customer.

customer_zip_code_prefix	First five digits of customer zip code.
customer_city	Customer city name.
customer_state	Customer state name.
olist_order_items_dataset	
Attribute	**Description**
order_id	Unique identifier of an order.
order_item_id	Sequential number identifying number of items included in the same order.
product_id	Product unique identifier.
seller_id	Seller unique identifier.
shipping_limit_date	Seller shipping limit date for handling the order over to the logistic partner.
price	Item price.
freight_value	Item freight value.
olist_order_payments_dataset	
Attribute	**Description**
order_id	Unique identifier of an order.
payment_sequential	Identifier when customers pay with >1 payment methods.
payment_type	Method of payment chosen by customer.
payment_installments	Number of instalments chosen by customer.
payment_value	Transaction value.
olist_order_reviews_dataset	
Attribute	**Description**
review_id	Unique review identifier.
order_id	Unique identifier of an order.
review_score	Note ranging from 1 to 5 given by customer on a satisfaction survey.
review_comment_title	Comment title from the review.
review_comment_message	Comment message from the review.
review_creation_date	Date in which the satisfaction survey was sent.
review_answer_timestamp	Satisfaction survey answer timestamp.
olist_products_dataset	
Attribute	**Description**
product_id	Unique product identifier.
product_category_name	Root category of product.
product_name_length	Number of characters extracted from product name.
product_description_length	Number of characters extracted from product description.
product_photos_qty	Number of product published photos.
product_weight_g	Product weight measured in grams.
product_length_cm	Product length measured in centimetres.
product_height_cm	Product height measured in centimetres.
product_width_cm	Product width measured in centimetres.

2.0 Methodology

Figure 1: Big data pipeline for the project.

2.1 Data Preparation

- Download 6 dataset CSV files from Kaggle. Copy the CSV files from Downloads folder to the specified directory /var/lib/mysql-files/.
- Install MySQL on the Ubuntu platform.
- Specify MySQL database to be WQD7007.
- Use CREATE TABLE command to create six tables in the MySQL database, including olist_order_payments_dataset, olist_orders_dataset, olist_customers_dataset, olist_order_items_dataset, olist_order_reviews_dataset, and olist_products_dataset. Each table will accommodate data from the respective CSV files.
- Use LOAD DATA INFILE command to import data from CSV files into MySQL tables.
- Create a new table "olist_joined_dataset" to combine all the 6 MySQL tables by using CREATE TABLE command combined with LEFT JOIN and INNER JOIN commands.
- Perform a UNION ALL operation on a set of column names and add to the top of "olist_joined_dataset" table. This step ensures that the exported file has column headers.
- Export into a CSV file.
- "," and "Enter" in column "review_comment_message" are removed to avoid affecting the csv format.

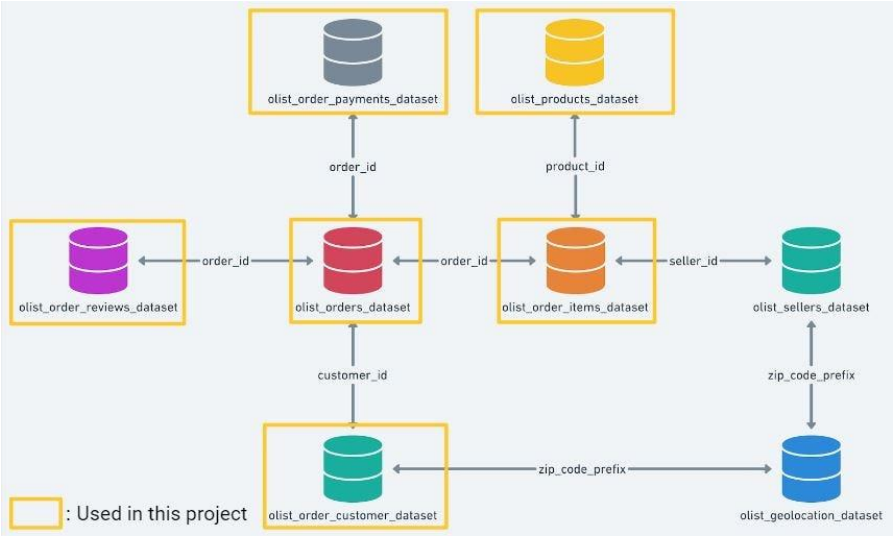


Figure 2: A schema that shows how six datasets are combined using common columns that serve as primary keys.

2.2 Data Storage

2.2.1 MySQL

- Install MySQL on the Ubuntu platform.

- Use the CREATE command in MySQL to create a table “olist_joined_dataset_2”.
- Use the LOAD DATA INFILE command to load data from joined dataset CSV.
- Retrieve information about MySQL table columns using the DESCRIBE command.
- Retrieve the first 10 rows from MySQL table using the SELECT command combined with the LIMIT command.

2.2.2 HBase

- Install HBase on the Ubuntu platform.
- Start the HDFS and YARN services and confirm the status with the "jps" command.
- Import the joined dataset CSV into HDFS.
- Create a data table named ‘OJD’ in HBase with 6 column families 'OrderInfo', 'ItemInfo', 'ProductInfo', 'CustomerInfo', 'PaymentInfo' and 'ReviewInfo'. Each column family represents a logical grouping of relevant data, which can be beneficial for efficient storage and retrieval. This organization follows the principle of grouping data that is likely to be accessed together, which can optimize read performance.
- Load the CSV file from HDFS into the ‘OJD’ table in HBase.
- View all the data in the HBase table.
- View the output of HBase table structure.
- View the first 10 rows of data in HBase table.
- Compare the output and configuration with MySQL.

2.2.3 Performance Comparison between MySQL and HBase

Table 2: Performance comparison between MySQL and HBase.

Comparison	Test Details	MySQL (Machine 1)	MySQL (Machine 3)	HBase (Machine 1)	HBase (Machine 2)
Data model (Data storage structure)	Define data table /column family structure	Adopt a table structure with rows and columns.		Adopt a non-relational model employing row keys, column families, and columns.	
Write performance (Speed of inserting data)	Insert 118311 sales records into a table	1.6249 s	4.7403 s	5.630 s	39.310 s
Read performance (Speed of querying data)	Retrieve the table structure, e.g. column, data type	0.0012 s	0.0190 s	0.1184 s	0.5400 s
	Retrieve all rows of a table	0.3696 s	0.5147 s	294.5066 s	751.1850 s
	Retrieve the first 10 rows of a table	0.0007 s	0.0011 s	0.0479 s	1.15 s

Note:

- Machine 1 is Intel(R) Core(TM) i7-1165G7 @ 2.80GHz, 16GB RAM; Ubuntu in Windows WSL environment.
- Machine 2 is Virtual Machine installed in 7th Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz 2.11 GHz, 8GB RAM; Cloudera environment (4GB RAM).
- Machine 3 is Virtual Machine installed in Intel(R) Core(TM) i7-1165G7 @ 2.80GHz, 16GB RAM; Ubuntu environment (4GB RAM). Equivalent to Machine 2.

2.3 Data Access

2.3.1 Hive

- Ubuntu and CentOS are used to setup Hive. The 'jps' command is used to verify the status once the HDFS and YARN services have been launched. The ‘hive’ command is performed on the terminal.
- In the CSV file, double quotes are used to enclose the dataset values. Therefore, quotation and separator characters are specified using the Serde Row Format.
- One disadvantage of CSV Serde is that, unless expressly stated as non-string columns, it regards all columns as strings. As a result, the "describe" command will display each column's data type as a string that may be converted for any necessary numerical calculations.
- To compare the query execution time depending on computational efficiency and hardware resource impact, an Orders Hive table is constructed on two distinct computers.
- Run the following command to load CSV data into the Orders Hive table:

```
LOAD DATA INPATH '/user/hdfs/data2.csv' INTO TABLE olist_order2;
```

Table 3 provides the query description and query execution time (sec) of the external table of Machine 1 and external table of Machine 2.

Table 3: Query performance of Hive.

No.	Query Description	Query Result	Machine 1	Machine 2
			Total Time (sec)	Total Time (sec)
1	Cities with the highest orders	Sao Paulo	37.329	40.528
2	Cities with the most revenue	Sao Paulo	37.537	36.615
3	The peak order timings	4pm	35.118	16.126
4	The peak order days	Monday	23.336	15.998
5	The best product categories of number of orders	Bed_bath_table	40.356	34.303
6	The worst product categories of number of orders	Security_and_services	35.252	35.182
7	The best product categories of total revenue generated	Bed_bath_table	41.348	34.844
8	The worst product categories of total revenue generated	Security_and_services	40.834	35.529

9	The best product categories of review scores	cds_dvds_musicals	39.625	36.935
10	The worst product categories of review scores	Security_and_services	38.024	34.889
11	The most popular payment methods	Credit Card	37.27	34.224
12	Product delivery time	12 days	17.844	16.433
13	Longer delivery times often result in lower review scores	Yes	18.837	16.799

Note:

- Machine 1 is Virtual Machine in ASUS TUF Gaming A15 R7 4800H @2.90GHz, 24GB RAM, Ubuntu Virtual Box (10GB RAM)
- Machine 2 is Virtual Machine in ASUS VivoBook 11th Gen Intel (R) Core (TM) i5-1135G7 @2.40GHz, 20GB RAM, Ubuntu Virtual Box (10GB RAM)

The external keyword in the CREATE TABLE command is used to construct the Hive external table, and HDFS is used to handle the data. The difference of execution time between Machine 1 and 2 is small may due the similarity of hardware.

2.3.2 Pig

- Ubuntu is used to setup Pig. The 'jps' command is used to verify the status once the HDFS and YARN services have been launched. The ‘pig’ command is performed on the terminal.
- For queries that involve Datetime, ToDate() function is used to convert variable format from string to Datetime.
- DUMP command is used to run the queries, and ToString(date,’EEE’) and GetHout() are used to get weekdays and hours.
- To compare the query execution time depending on computational efficiency and hardware resource impact, an Orders Pig table is constructed on two distinct computers.
- Run the following command to load csv data into pig:
  data4 = load '/user/hdfs/cleaned10.csv' using PigStorage(',');

Table 4 provides the query description and query execution time (sec) of the external table of Machine 1 and external table of Machine 2.

Table 4: Query performance of Pig.

No.	**Query Description**	**Query Result**	**Machine 1**	**Machine 2**
			Total Time (sec)	**Total Time (sec)**
1	Cities with the highest orders	Sao Paulo	113	82
2	Cities with the most revenue	Sao Paulo	97	75
3	The peak order timings	4pm	31	21
4	The peak order days	Monday	41	17

5	The best product categories of number of orders	Bed_bath_table	95	79
6	The worst product categories of number of orders	Security_and_services	93	75
7	The best product categories of total revenue generated	Bed_bath_table	118	74
8	The worst product categories of total revenue generated	Security_and_services	97	71
9	The best product categories of review scores	cds_dvds_musicals	87	75
10	The worst product categories of review scores	Security_and_services	113	74
11	The most popular payment methods	Credit Card	83	53
12	Product delivery time	12 days	43	21
13	Longer delivery times often result in lower review scores	Yes	26	16

Note:

- Machine 1 is Virtual Machine in ASUS TUF Gaming A15 R7 4800H @2.90GHz, 24GB RAM, Ubuntu Virtual Box (10GB RAM)
- Machine 2 is Virtual Machine in ASUS VivoBook 11th Gen Intel (R) Core (TM) i5-1135G7 @2.40GHz, 20GB RAM, Ubuntu Virtual Box (10GB RAM)

The difference of execution time between Machine 1 and 2 is small may due the similarity of hardware.

3.0 Results and Discussion

3.1 Data Storage

MySQL and HBase employ distinct data models and storage structures. With a structured relational data model, MySQL is well-suited for handling CSV datasets that have a predefined and consistent schema. On the other hand, HBase adopts a non-relational model, utilizing row keys, column families, and columns. This schema-less approach is well-suited for handling vast amounts of sparse and distributed data, offering flexibility in the storage structure.

In terms of write performance, MySQL outperformed HBase, completing the task in just 3.1826 seconds (average value). HBase, on the other hand, falls significantly short of MySQL's efficiency, performing at 22.47 seconds. MySQL's advantage lies in its structured data handling with predefined schemas, allowing it to excel in write operations. In contrast, HBase's schema-less nature results in slower write performance when dealing with CSV data, possibly due to the additional processing overhead associated with its dynamic schema adaptability.

In terms of read performance, MySQL consistently outperformed HBase across all scenarios. For instance, retrieving table structure details in MySQL took a mere 0.0101 seconds. When it came to HBase's read performance, the same machine underperformed at 0.3292 seconds. In the context of retrieving data from a table, MySQL demonstrates remarkable efficiency, completing the task in just 0.0009 seconds. On the contrary, the same machine running HBase experiences a significant lag, requiring a substantial 0.5990 seconds. This stark contrast in performance can be attributed to the schema-less nature of HBase, which introduces complexities in querying extensive datasets.

To sum up, MySQL excels in scenarios where structured, relational data storage and retrieval are essential. Its predefined schema provides consistency for applications with well-defined data structures. HBase is designed for handling large volumes of unstructured data. Its schema-less flexibility may introduce complexities in managing structured data. Notably, both MySQL and HBase showed variation in read and write performance between Machine 1 and Machine 2 & 3 (Machine 3 is equivalent to Machine 2), suggesting that tool performance is influenced by hardware and configuration factors.

3.2 Data Access

When talking about big data processing and analytics, two popular tools are Hive and Pig. Hive is a data warehousing software for reading, writing, and managing large-scale datasets stored in the Hadoop file system. It provides a SQL-like query language called HiveQL that makes data querying and analysis more intuitive and easy to understand, especially for those familiar with SQL. Hive is suitable for scenarios where complex data aggregation, joins, and analysis are required. On the other hand, Pig is an advanced platform for creating complex MapReduce programs. Its language, Pig Latin, is a procedural language that gives users greater control over the processing of data. Pig is suitable for tasks like pipelining data streams and performing data transformations, and is great for scenarios that require fine-grained control over the data processing process.

The query results show that Hive and Pig exhibit different performances when executing the same 13 query statements. Hive demonstrates higher efficiency with an average time of only 31.97 seconds, while Pig's average time is 68.08 seconds. This difference could be due to the differences in architecture, execution, and optimization mechanisms between the two. Hive tends to make better use of its SQL-like query optimization, while Pig offers more flexibility and control, but may sacrifice performance in some cases. However, the choice between Hive or Pig depends on the specific application scenario, data characteristics, and user expertise.

4.0 Conclusion

The investigation into data storage and data access tools for the e-commerce dataset has provided valuable insights into the performance and suitability of different technologies. MySQL, with its structured relational database model, demonstrated superior read and write performance for CSV data due to its predefined schema, making it well-suited for scenarios

with well-defined data structures. On the other hand, HBase, with its schema-less and distributed storage approach, exhibited slower read and write performance likely due to additional processing overhead associated with its dynamic schema adaptability.

Furthermore, the exploration of data access tools revealed distinct characteristics between Hive and Pig. Hive, with its SQL-like query language (HiveQL), demonstrated higher efficiency in executing complex queries, making it suitable for scenarios requiring sophisticated data aggregation and analysis. Pig, with its procedural language (Pig Latin), provided greater control over data processing but exhibited slower query execution times on average. The choice between Hive and Pig depends on the specific requirements of the application, the nature of the data, and the level of control desired by users. Additionally, query execution performance is constrained by the capability of hardware resources, so it is necessary to ensure that sufficient hardware resources are available to support large-scale data processing.

In summary, the study underscores the importance of choosing the right technology stack based on the nature of the data, the intended use cases, and the desired trade-offs between structured and unstructured data processing. The findings can inform decision-making in optimizing marketing efforts, enhancing user experience, and driving growth for e-commerce platforms such as Shopee. Future directions of work include exploring advanced machine-learning techniques for more dynamic customer segmentation and improved personalization. Additionally, the integration of real-time data analytics for more accurate customer insights can help organizations make better decisions.

5.0 Reference

Chen, X., Hu, L., Liu, L., Chang, J., & Bone, D. L. (2017, June). Breaking down Hadoop distributed file systems data analytics tools: Apache Hive vs. Apache Pig vs. pivotal HWAQ. In 2017 IEEE 10th International Conference on Cloud Computing (CLOUD) (pp. 794-797). IEEE.

Francisco, M. (2021). Brazilian E-Commerce Public Dataset by OLIST. Kaggle. https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Preethi, R. A., & Elavarasi, J. (2017). Big data analytics using Hadoop tools—Apache Hive vs Apache Pig. International Journal of Emerging Technology in Computer Science & Electronics, 24(3).

Sheng, C. (2023). Understanding Southeast Asia e-commerce Shoppers: A Shopee Analysis | Data Insights - Measurable AI. Data Insights - Measurable AI. https://blog.measurable.ai/2023/07/25/understanding-southeast-asia-e-commerce-shoppers-a-shopee-analysis

6.0 Appendix

6.1 Data Pre-processing - MySQL

1. Change the permissions of the specified directory and its contents.

```
sudo chmod -R 755 /var/lib/mysql-files/
```

```
bklow@LAPTOP-M13O57I5:/var/lib$ sudo chmod -R 755 /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:/var/lib$ ls -l
total 132
drwxr-xr-x  2 root    root    4096 Dec 16 08:54 PackageKit
drwxr-xr-x  3 root    root    4096 Dec 16 08:52 apport
drwxr-xr-x  5 root    root    4096 Jan  8 20:00 apt
drwxr-xr-x  2 root    root    4096 Dec 16 08:54 command-not-found
drwxr-xr-x  2 root    root    4096 Dec 16 08:52 dbus
drwxr-xr-x  2 root    root    4096 Mar 24  2022 dhcp
drwxr-xr-x  7 root    root    4096 Jan  8 20:00 dpkg
drwxr-xr-x  2 root    root    4096 Jul  7  2023 git
drwxr-xr-x  2 root    root    4096 Jan  9 18:11 logrotate
drwxr-xr-x  2 root    root    4096 Nov 23 05:37 man-db
drwxr-xr-x  3 root    root    4096 Dec 20 13:36 mecab
drwxr-xr-x  2 root    root    4096 Apr 18  2022 misc
drwxr-xr-x  4 mongodb mongodb 4096 Jan  8 20:25 mongodb
drwx------  8 mysql   mysql   4096 Jan 11 09:09 mysql
drwxr-xr-x  2 mysql   mysql   4096 Jan 11 10:34 mysql-files
```

2. Copy the CSV files from Downloads folder to the specified directory.

```
sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_customers_dataset.csv
/var/lib/mysql-files/
sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_order_items_dataset.csv
/var/lib/mysql-files/
sudo cp
/mnt/c/Users/boonk/Downloads/Dataset/olist_order_payments_dataset.csv
/var/lib/mysql-files/
sudo cp
/mnt/c/Users/boonk/Downloads/Dataset/olist_order_reviews_dataset.csv
/var/lib/mysql-files/
sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_orders_dataset.csv
/var/lib/mysql-files/
sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_products_dataset.csv
/var/lib/mysql-files/
cd /var/lib/mysql-files/
ls -l
```

```
bklow@LAPTOP-M13O57I5:~$ sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_customers_dataset.csv /var/lib/mysql-files/
[sudo] password for bklow:
bklow@LAPTOP-M13O57I5:~$ sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_order_items_dataset.csv /var/lib/mysql-files/
[sudo] password for bklow:
bklow@LAPTOP-M13O57I5:~$ sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_order_payments_dataset.csv /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:~$ sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_order_reviews_dataset.csv /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:~$ sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_orders_dataset.csv /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:~$ sudo cp /mnt/c/Users/boonk/Downloads/Dataset/olist_products_dataset.csv /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:~$ cd /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:/var/lib/mysql-files$ ls -1
olist_customers_dataset.csv
olist_order_items_dataset.csv
olist_order_payments_dataset.csv
olist_order_reviews_dataset.csv
olist_orders_dataset.csv
olist_products_dataset.csv
```

3. Start the MySQL service and then access the MySQL command-line interface.

```
sudo systemctl start mysql
mysql -u root -p --local-infile=1
```

```
bklow@LAPTOP-M13O57I5:~$ sudo systemctl start mysql
bklow@LAPTOP-M13O57I5:~$ mysql -u root -p --local-infile=1
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 8
Server version: 8.0.35-0ubuntu0.22.04.1 (Ubuntu)

Copyright (c) 2000, 2023, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

4. Display all the databases available on the MySQL server and select a specific database.

```
SHOW databases;
USE WQD7007;
```

```
mysql> SHOW databases;
+--------------------+
| Database           |
+--------------------+
| WQD7007            |
| information_schema |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
5 rows in set (0.00 sec)

mysql> USE WQD7007;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
```

5. Create tables in MySQL database.

```
CREATE TABLE olist_customers_dataset (
    customer_id VARCHAR(255),
    customer_unique_id VARCHAR(255),
    customer_zip_code_prefix INT,
    customer_city VARCHAR(255),
    customer_state VARCHAR(255)
);
CREATE TABLE olist_order_items_dataset (
    order_id VARCHAR(255),
    order_item_id INT,
    product_id VARCHAR(255),
    seller_id VARCHAR(255),
    shipping_limit_date TIMESTAMP,
    price DECIMAL(10, 2),
    freight_value DECIMAL(10, 2)
);
CREATE TABLE olist_order_payments_dataset (
    order_id VARCHAR(255),
    payment_sequential INT,
    payment_type VARCHAR(255),
    payment_installments INT,
    payment_value DECIMAL(10, 2)
```

```
);
CREATE TABLE olist_order_reviews_dataset (
    review_id VARCHAR(255),
    order_id VARCHAR(255),
    review_score INT,
    review_comment_title VARCHAR(255),
    review_comment_message TEXT,
    review_creation_date TIMESTAMP,
    review_answer_timestamp TIMESTAMP
);
CREATE TABLE olist_orders_dataset (
    order_id VARCHAR(255),
    customer_id VARCHAR(255),
    order_status VARCHAR(255),
    order_purchase_timestamp TIMESTAMP,
    order_approved_at TIMESTAMP,
    order_delivered_carrier_date TIMESTAMP,
    order_delivered_customer_date TIMESTAMP,
    order_estimated_delivery_date TIMESTAMP
);
CREATE TABLE olist_products_dataset (
    product_id VARCHAR(255),
    product_category_name VARCHAR(255),
    product_name_length INT,
    product_description_length INT,
    product_photos_qty INT,
    product_weight_g INT,
    product_length_cm INT,
    product_height_cm INT,
    product_width_cm INT
);
```

```
mysql> CREATE TABLE olist_customers_dataset (
    ->     customer_id VARCHAR(255),
    ->     customer_unique_id VARCHAR(255),
    ->     customer_zip_code_prefix INT,
    ->     customer_city VARCHAR(255),
    ->     customer_state VARCHAR(255)
    -> );
Query OK, 0 rows affected (0.02 sec)

mysql> CREATE TABLE olist_order_items_dataset (
    ->     order_id VARCHAR(255),
    ->     order_item_id INT,
    ->     product_id VARCHAR(255),
    ->     seller_id VARCHAR(255),
    ->     shipping_limit_date TIMESTAMP,
    ->     price DECIMAL(10, 2),
    ->     freight_value DECIMAL(10, 2)
    -> );
Query OK, 0 rows affected (0.02 sec)

mysql> CREATE TABLE olist_order_payments_dataset (
    ->     order_id VARCHAR(255),
    paym    ->     payment_sequential INT,
    ->     payment_type VARCHAR(255),
    ->     payment_installments INT,
    ->     payment_value DECIMAL(10, 2)
    -> );
Query OK, 0 rows affected (0.02 sec)
```

```
mysql> CREATE TABLE olist_order_reviews_dataset (
    ->     review_id VARCHAR(255),
    ->     order_id VARCHAR(255),
    ->     review_score INT,
    ->     review_comment_title VARCHAR(255),
    ->     review_comment_message TEXT,
    ->     review_creation_date TIMESTAMP,
    ->     review_answer_timestamp TIMESTAMP
    -> );
Query OK, 0 rows affected (0.01 sec)

mysql> CREATE TABLE olist_orders_dataset (
    ->     order_id VARCHAR(255),
    ->     customer_id VARCHAR(255),
    ->     order_status VARCHAR(255),
    ->     order_purchase_timestamp TIMESTAMP,
    ->     order_approved_at TIMESTAMP,
    ->     order_delivered_carrier_date TIMESTAMP,
    ->     order_delivered_customer_date TIMESTAMP,
    ->     order_estimated_delivery_date TIMESTAMP
    -> );
Query OK, 0 rows affected (0.02 sec)

mysql> CREATE TABLE olist_products_dataset (
    ->     product_id VARCHAR(255),
    ->     product_category_name VARCHAR(255),
    ->     product_name_length INT,
    ->     product_description_length INT,
    ->     product_photos_qty INT,
    ->     product_weight_g INT,
    ->     product_length_cm INT,
    ->     product_height_cm INT,
    ->     product_width_cm INT
    -> );
Query OK, 0 rows affected (0.02 sec)
```

6. Display all tables in the MySQL database.

```
SHOW tables;
```

```
mysql> SHOW tables;
+------------------------------+
| Tables_in_WQD7007            |
+------------------------------+
| churn                        |
| olist_customers_dataset      |
| olist_order_items_dataset    |
| olist_order_payments_dataset |
| olist_order_reviews_dataset  |
| olist_orders_dataset         |
| olist_products_dataset       |
+------------------------------+
7 rows in set (0.00 sec)
```

7. Retrieve the current directory where the MySQL server can read and write files.

```
SHOW VARIABLES LIKE 'secure_file_priv';
```

```
mysql> SHOW VARIABLES LIKE 'secure_file_priv';
+------------------+-----------------------+
| Variable_name    | Value                 |
+------------------+-----------------------+
| secure_file_priv | /var/lib/mysql-files/ |
+------------------+-----------------------+
1 row in set (0.01 sec)
```

8. Gives the MySQL user account 'root'@'localhost' the FILE privilege on all databases and tables (***.***).

```
GRANT FILE ON *.* TO 'root'@'localhost';
```

```
mysql> GRANT FILE ON *.* TO 'root'@'localhost';
Query OK, 0 rows affected (0.01 sec)
```

9. Allows data to be loaded from a local file into a MySQL table.

```
SET GLOBAL local_infile=1;
SHOW GLOBAL VARIABLES LIKE 'local_infile';
```

```
mysql> SET GLOBAL local_infile=1;
Query OK, 0 rows affected (0.00 sec)

mysql> SHOW GLOBAL VARIABLES LIKE 'local_infile';
+---------------+-------+
| Variable_name | Value |
+---------------+-------+
| local_infile  | ON    |
+---------------+-------+
1 row in set (0.00 sec)
```

10. Change the ownership of the /var/lib/mysql-files/ directory to the user ‘mysql. Change the permissions of directory and its contents.

```
sudo chown -R mysql:mysql /var/lib/mysql-files/
sudo chmod -R 755 /var/lib/mysql-files/
ls -ld /var/lib/mysql-files/
```

```
bklow@LAPTOP-M13O57I5:~$ sudo chown -R mysql:mysql /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:~$ sudo chmod -R 755 /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:~$ ls -ld /var/lib/mysql-files/
drwxr-xr-x 2 mysql mysql 4096 Jan 11 09:13 /var/lib/mysql-files/
```

11. Import data from CSV files into MySQL tables.

```
LOAD DATA INFILE '/var/lib/mysql-files/olist_customers_dataset.csv' INTO
TABLE olist_customers_dataset FIELDS TERMINATED BY ',' ENCLOSED BY '"'
LINES TERMINATED BY '\n' IGNORE 1 ROWS;
LOAD DATA LOCAL INFILE '/var/lib/mysql-files/olist_order_items_dataset.csv'
INTO TABLE olist_order_items_dataset FIELDS TERMINATED BY ',' ENCLOSED BY
'"' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
LOAD DATA LOCAL INFILE '/var/lib/mysql-
files/olist_order_payments_dataset.csv' INTO TABLE
olist_order_payments_dataset FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES
TERMINATED BY '\n' IGNORE 1 ROWS;
LOAD DATA LOCAL INFILE '/var/lib/mysql-
files/olist_order_reviews_dataset.csv' INTO TABLE
olist_order_reviews_dataset FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES
TERMINATED BY '\n' IGNORE 1 ROWS;
LOAD DATA LOCAL INFILE '/var/lib/mysql-files/olist_orders_dataset.csv' INTO
TABLE olist_orders_dataset FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES
TERMINATED BY '\n' IGNORE 1 ROWS;
LOAD DATA LOCAL INFILE '/var/lib/mysql-files/olist_products_dataset.csv'
INTO TABLE olist_products_dataset FIELDS TERMINATED BY ',' ENCLOSED BY '"'
LINES TERMINATED BY '\n' IGNORE 1 ROWS;
```

```
mysql> LOAD DATA INFILE '/var/lib/mysql-files/olist_customers_dataset.csv' INTO TABLE olist_customers_dataset FIELDS TERMINATED BY ',' ENCLOSED
BY '"' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
Query OK, 99441 rows affected (0.51 sec)
Records: 99441  Deleted: 0  Skipped: 0  Warnings: 0

mysql> LOAD DATA LOCAL INFILE '/var/lib/mysql-files/olist_order_items_dataset.csv' INTO TABLE olist_order_items_dataset FIELDS TERMINATED BY ','
 ENCLOSED BY '"' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
Query OK, 112650 rows affected (0.74 sec)
Records: 112650  Deleted: 0  Skipped: 0  Warnings: 0

mysql> LOAD DATA LOCAL INFILE '/var/lib/mysql-files/olist_order_payments_dataset.csv' INTO TABLE olist_order_payments_dataset FIELDS TERMINATED
BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
Query OK, 103886 rows affected (0.46 sec)
Records: 103886  Deleted: 0  Skipped: 0  Warnings: 0
mysql> LOAD DATA LOCAL INFILE '/var/lib/mysql-files/olist_order_reviews_dataset.csv' INTO TABLE olist_order_reviews_dataset FIELDS TERMINATED BY
 ',' ENCLOSED BY '"' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
Query OK, 99223 rows affected, 65535 warnings (0.74 sec)
Records: 99223  Deleted: 0  Skipped: 0  Warnings: 99225

mysql> LOAD DATA LOCAL INFILE '/var/lib/mysql-files/olist_orders_dataset.csv' INTO TABLE olist_orders_dataset FIELDS TERMINATED BY ',' ENCLOSED
BY '"' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
Query OK, 99441 rows affected, 4908 warnings (0.75 sec)
Records: 99441  Deleted: 0  Skipped: 0  Warnings: 4908

mysql> LOAD DATA LOCAL INFILE '/var/lib/mysql-files/olist_products_dataset.csv' INTO TABLE olist_products_dataset FIELDS TERMINATED BY ',' ENCLO
SED BY '"' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
Query OK, 32951 rows affected, 1838 warnings (0.21 sec)
Records: 32951  Deleted: 0  Skipped: 0  Warnings: 1838
```

12. Retrieve the first 10 rows from the MySQL tables.

```
SELECT * FROM olist_customers_dataset LIMIT 10;
SELECT * FROM olist_order_items_dataset LIMIT 10;
SELECT * FROM olist_order_payments_dataset LIMIT 10;
SELECT * FROM olist_order_reviews_dataset LIMIT 10;
SELECT * FROM olist_orders_dataset LIMIT 10;
SELECT * FROM olist_products_dataset LIMIT 10;
```

```
mysql> SELECT * FROM olist_customers_dataset LIMIT 10;
+----------------------------------+----------------------------------+--------------------------+---------------------+----------------+
| customer_id                      | customer_unique_id               | customer_zip_code_prefix | customer_city       | customer_state |
+----------------------------------+----------------------------------+--------------------------+---------------------+----------------+
| 06b8999e2fba1a1fbc88172c00ba8bc7 | 861eff4711a542e4b93843c6dd7febb0 |                    14409 | franca              | SP             |
| 18955e83d337fd6b2def6b18a428ac77 | 290c77bc529b7ac935b93aa66c333dc3 |                     9790 | sao bernardo do campo | SP           |
| 4e7b3e00288586ebd08712fdd0374a03 | 060e732b5b29e8181a18229c7b0b2b5e |                     1151 | sao paulo           | SP             |
| b2b6027bc5c5109e529d4dc6358b12c3 | 259dac757896d24d7702b9acbbff3f3c |                     8775 | mogi das cruzes     | SP             |
| 4f2d8ab171c80ec8364f7c12e35b23ad | 345ecd01c38d18a9036ed96c73b8d066 |                    13056 | campinas            | SP             |
| 879864dab9bc30475222c92c82e1212b8 | 4c93744516667ad3b8f1fb645a3116a4 |                    89254 | jaragua do sul      | SC             |
| fd826e7cf63160e536e0908c76c3f441 | addec96d2e059c80c30fe6871d30d177 |                     4534 | sao paulo           | SP             |
| 5e274e7a0c3809e14aba7ad5aae0d407 | 57b2a98a409812fe9618067b6b8ebe4f |                    35182 | timoteo             | MG             |
| 5adf08e34b2e993982a47070956c5c65 | 1175e95fb47ddff9de6b2b06188f7e0d |                    81560 | curitiba            | PR             |
| 4b7139f34592b3a31687243a302fa75b | 9afe194fb833f79e300e37e580171f22 |                    30575 | belo horizonte      | MG             |
+----------------------------------+----------------------------------+--------------------------+---------------------+----------------+
10 rows in set (0.00 sec)
```

```
mysql> SELECT * FROM olist_order_items_dataset LIMIT 10;
+----------------------------------+---------------+----------------------------------+----------------------------------+---------------------+---------+---------------+
| order_id                         | order_item_id | product_id                       | seller_id                        | shipping_limit_date | price   | freight_value |
+----------------------------------+---------------+----------------------------------+----------------------------------+---------------------+---------+---------------+
| 00010242fe8c5a6d1ba2dd792cb16214 |             1 | 4244733e06e7ecb4970a6e2683c13e61 | 48436dade18ac8b2bce089ec2a041202 | 2017-09-19 09:45:35 |  58.90  |         13.29 |
| 00018f77f2f0320c557190d7a144bdd3 |             1 | e5f2d52b802189ee658865ca93d83a8f | dd7ddc04e1b6c2c614352b383efe2d36 | 2017-05-03 11:05:13 | 239.90  |         19.93 |
| 000229ec398224ef6ca0657da4fc703e |             1 | c7773555d18b72b67abbeef9df44fd0fd | 5b51032eddd242adc84c38acab88f23d | 2018-01-18 14:48:30 | 199.00  |         17.87 |
| 00024acbcdf0a6daa1e931b038114c75 |             1 | 7634da152a4610f1595efa32f14722fc | 9d7a1d34a5052409006425275ba1c2b4 | 2018-08-15 10:10:18 |  12.99  |         12.79 |
| 00042b26cf59d7ce69dfabb4e55b4fd9 |             1 | ac6c3623068f30de03045865e4e10089 | df560393f3a51e74553ab94004ba5c87 | 2017-02-13 13:57:51 | 199.90  |         18.14 |
| 00048cc3ae777c65dbb7d2a0634bc1ea |             1 | ef92defde845ab8450f9d70c526ef70f | 6426d21aca402a131fc0a5d0960a3c90 | 2017-05-23 03:55:27 |  21.90  |         12.69 |
| 00054e8431b9d7675808bcb819fb4a32 |             1 | 8d4f2bb7e93e6710a28f34fa83ee7d28 | 7040e82f899a04d1b434b795a43b4617 | 2017-12-14 12:10:31 |  19.90  |         11.85 |
| 000576fe39319847cbb9d288c5617fa6 |             1 | 557d850972a7d6f792fd18ae1400d9b6 | 5996cddab893a4652a15592fb58ab8db | 2018-07-10 12:30:45 | 810.00  |         70.75 |
| 0005a1a1728c9d785b8e2b08b904576c |             1 | 310ae3c140ff94b03219ad0adc3c778f | a416b6a846a11724393025641d4edd5e | 2018-03-26 18:31:29 | 145.95  |         11.65 |
| 0005f50442cb953dcd1d21e1fb923495 |             1 | 4535b0e1091c278dfd193e5a1d63b39f | ba143b05f0110f0dc71ad71b4466ce92 | 2018-07-06 14:10:56 |  53.99  |         11.40 |
+----------------------------------+---------------+----------------------------------+----------------------------------+---------------------+---------+---------------+
10 rows in set (0.00 sec)
```

```
mysql> SELECT * FROM olist_order_payments_dataset LIMIT 10;
+----------------------------------+--------------------+--------------+----------------------+---------------+
| order_id                         | payment_sequential | payment_type | payment_installments | payment_value |
+----------------------------------+--------------------+--------------+----------------------+---------------+
| b81ef226f3fe1789b1e8b2acac839d17 |                  1 | credit_card  |                    8 |         99.33 |
| a9810da82917af2d9aefd1278f1dcfa0 |                  1 | credit_card  |                    1 |         24.39 |
| 25e8ea4e93396b6fa0d3dd708e76c1bd |                  1 | credit_card  |                    1 |         65.71 |
| ba78997921bbcdc1373bb41e913ab953 |                  1 | credit_card  |                    8 |        107.78 |
| 42fdf880ba16b47b59251dd489d4441a |                  1 | credit_card  |                    2 |        128.45 |
| 298fcdf1f73eb413e4d26d01b25bc1cd |                  1 | credit_card  |                    2 |         96.12 |
| 771ee386b001f06208a7419e4fc1bbd7 |                  1 | credit_card  |                    1 |         81.16 |
| 3d7239c394a212faae122962df514ac7 |                  1 | credit_card  |                    3 |         51.84 |
| 1f78449c87a54faf9e96e88ba1491fa9 |                  1 | credit_card  |                    6 |        341.09 |
| 0573b5e23cbd798006520e1d5b4c6714 |                  1 | boleto       |                    1 |         51.95 |
+----------------------------------+--------------------+--------------+----------------------+---------------+
10 rows in set (0.00 sec)
```

```
mysql> SELECT * FROM olist_order_reviews_dataset LIMIT 10;
+----------------------------------+----------------------------------+--------------+----------------------+---------------------------+----------------------------+
| review_id                        | order_id                         | review_score | review_comment_title | review_comment_message    | review_creation_date       | review_answer_timestamp    |
+----------------------------------+----------------------------------+--------------+----------------------+---------------------------+----------------------------+----------------------------+
| 7bc2406110b926393aa56f80a40eba40 | 73fc7af87114b39712e6da79b0a377eb |            4 |                      |                           | 2018-01-18 00:00:00        | 2018-01-18 21:46:59        |
| 80e641a11e56f04c1ad469d5645fdfde | a548910a1c6147796b98fdf73dbeba33 |            5 |                      |                           | 2018-03-10 00:00:00        | 2018-03-11 03:05:13        |
| 228ce5500dc1d8e020d8d1322874b6f0 | f9e4b658b201a9f2ecdecbb34bed034b |            5 |                      |                           | 2018-02-17 00:00:00        | 2018-02-18 14:36:24        |
| e64fb393e7b32834bb789ff8bb30750e | 658677c97b385a9be170737859d3511b |            5 |                      | Recebí bem antes do prazo estipulado. | 2017-04-21 00:00:00        | 2017-04-21 22:02:06        |
| f7c4243c7fe1938f181bec41a392bdeb | 8e6bfb81e283fa7e4f11123a3fb894f1 |            5 |                      | Parabéns lojas lannister adorei comprar pela Internet seguro e prático Parabéns a todos feliz Páscoa | 2018-03-01 00:00:00        | 2018-03-02 10:26:53        |
| 15197aa66ff4d0650b5434f1b46cda19 | b18dcdf73be66366873cd26c5724d1dc |            1 |                      |                           | 2018-04-13 00:00:00        | 2018-04-16 00:39:37        |
| 07f9bee5d1b850860defd761afa7ff16 | e48aa0d2dcec3a2e87348811bcfdf22b |            5 |                      |                           | 2017-07-16 00:00:00        | 2017-07-18 19:30:34        |
| 7c6400515c67679fbee952a7525281ef | c31a859e34e3adac22f376954e19b39d |            5 |                      |                           | 2018-08-14 00:00:00        | 2018-08-14 21:36:06        |
| a3f6f7f6f433de0aefbb97da197c554c | 9c2114ac970e84273583ab523dfafd09b |            5 |                      |                           | 2017-05-17 00:00:00        | 2017-05-18 12:05:37        |
| 8670d52e15e00043ae7de4c01cc2fe06 | b9bf720beb4ab3728760088589c62129 |            4 | recomendo            | aparelho eficiente. no site a marca do aparelho esta impresso como 3desinfector e ao chegar esta com outro nome...atualizar com a marca correta uma vez que é o mesmo aparelho | 2018-05-22 00:00:00        | 2018-05-23 16:45:47        |
+----------------------------------+----------------------------------+--------------+----------------------+---------------------------+----------------------------+----------------------------+
10 rows in set (0.00 sec)
```

```
mysql> SELECT * FROM olist_orders_dataset LIMIT 10;
+----------------------------------+----------------------------------+--------------+--------------------------+---------------------+---------------------------+-----------------------------+-----------------------------+
| order_id                         | customer_id                      | order_status | order_purchase_timestamp | order_approved_at   | order_delivered_carrier_date | order_delivered_customer_date | order_estimated_delivery_date |
+----------------------------------+----------------------------------+--------------+--------------------------+---------------------+---------------------------+-----------------------------+-----------------------------+
| e481f51cbdc54678b7cc49136f2d6af7 | 9ef432eb6251297304e76186b10a928d | delivered    | 2017-10-02 10:56:33      | 2017-10-02 11:07:15 | 2017-10-04 19:55:00       | 2017-10-10 21:25:13         | 2017-10-18 00:00:00         |
| 53cdb2fc8bc7dce0b6741e2150273451 | b0830fb4747a6c6d20dea0b8c802d7ef | delivered    | 2018-07-24 20:41:37      | 2018-07-26 03:24:27 | 2018-07-26 14:31:00       | 2018-08-07 15:27:45         | 2018-08-13 00:00:00         |
| 47770eb9100c2d0c44946d9cf07ec65d | 41ce2a54c0b03bf3443c3d931a367089 | delivered    | 2018-08-08 08:38:49      | 2018-08-08 08:55:23 | 2018-08-08 13:50:00       | 2018-08-17 18:06:29         | 2018-09-04 00:00:00         |
| 949d5b44dbf5de918fe9c16f97b0545a | f88197465ea7920adcdbec7375364d82 | delivered    | 2017-11-18 19:28:06      | 2017-11-18 19:45:59 | 2017-11-22 13:39:59       | 2017-12-02 00:28:42         | 2017-12-15 00:00:00         |
| ad21c59c0840e6cb83a9ceb5573f8159 | 8ab97904e6daea8866dbdbc4fb7aad2c | delivered    | 2018-02-13 21:18:39      | 2018-02-13 22:20:29 | 2018-02-14 19:46:34       | 2018-02-16 18:17:02         | 2018-02-26 00:00:00         |
| a4591c265e18cb1dcee52889e2d8acc3 | 503740e9ca751ccdda7ba28e9ab8f608 | delivered    | 2017-07-09 21:57:05      | 2017-07-09 22:10:13 | 2017-07-11 14:58:04       | 2017-07-26 10:57:55         | 2017-08-01 00:00:00         |
| 136cce7faa42fdb2cefd53fdc79a6098 | ed0271e0b7da060a393796590e7b737a | invoiced     | 2017-04-11 12:22:08      | 2017-04-13 13:25:17 | 0000-00-00 00:00:00       | 0000-00-00 00:00:00         | 2017-05-09 00:00:00         |
| 6514b8ad8028c9f2cc2374ded245783f | 9bdf08b4b3b52b5526ff42d37d47f222 | delivered    | 2017-05-16 13:10:30      | 2017-05-16 13:22:11 | 2017-05-22 10:07:46       | 2017-05-26 12:55:51         | 2017-06-07 00:00:00         |
| 76c6e866289321a7c93b82b54852dc33 | f54a9f0e6b351c431402b8461eaa51999 | delivered    | 2017-01-23 18:29:09      | 2017-01-25 02:50:47 | 2017-01-26 14:16:31       | 2017-02-02 14:08:10         | 2017-03-06 00:00:00         |
| e69bfb5eb88e0ed6a785585b27e16dbf | 31ad1d1b63eb9962463f764d4e6e0c9d | delivered    | 2017-07-29 11:55:02      | 2017-07-29 12:05:32 | 2017-08-10 19:45:24       | 2017-08-16 17:14:30         | 2017-08-23 00:00:00         |
+----------------------------------+----------------------------------+--------------+--------------------------+---------------------+---------------------------+-----------------------------+-----------------------------+
10 rows in set (0.00 sec)
```

```
mysql> SELECT * FROM olist_products_dataset LIMIT 10;
+----------------------------------+--------------------------+--------------------+--------------------------+-------------------+----------------+--------------------+---------------------+--------------------+
| product_id                       | product_category_name    | product_name_length | product_description_length | product_photos_qty | product_weight_g | product_length_cm | product_height_cm | product_width_cm |
+----------------------------------+--------------------------+--------------------+--------------------------+-------------------+----------------+--------------------+---------------------+--------------------+
| 1e9e8ef04dbcff4541ed26657ea517e5 | perfumaria               |                 40 |                      287 |                 1 |            225 |                 16 |                  10 |                 14 |
| 3aa071139cb16b67ca9e5dea641aaa2f | artes                    |                 44 |                      276 |                 1 |           1000 |                 30 |                  18 |                 20 |
| 96bd76ec8810374ed1b65e291975717f | esporte_lazer            |                 46 |                      250 |                 1 |            154 |                 18 |                   9 |                 15 |
| cef67bcfe19066a932b7673e239eb23d | bebes                    |                 27 |                      261 |                 1 |            371 |                 26 |                   4 |                 26 |
| 9dc1a7de274444849c219cff195d0b71 | utilidades_domesticas    |                 37 |                      402 |                 4 |            625 |                 20 |                  17 |                 13 |
| 41d3672d4792049fa1779bb35283ed13 | instrumentos_musicais    |                 60 |                      745 |                 1 |            200 |                 38 |                   5 |                 11 |
| 732bd381ad09e530fe0a5f457d81becb | cool_stuff               |                 56 |                     1272 |                 4 |          18350 |                 70 |                  24 |                 44 |
| 2548af3e6e77a690cf3eb6368e9ab61e | moveis_decoracao         |                 56 |                      184 |                 2 |            900 |                 40 |                   8 |                 40 |
| 37cc742be07708b53a98702e77a21a02 | eletrodomesticos         |                 57 |                      163 |                 1 |            400 |                 27 |                  13 |                 17 |
| 8c92109888e8cdf9d66dc7e463025574 | brinquedos               |                 36 |                     1156 |                 1 |            600 |                 17 |                  10 |                 12 |
+----------------------------------+--------------------------+--------------------+--------------------------+-------------------+----------------+--------------------+---------------------+--------------------+
10 rows in set (0.00 sec)
```

13. Retrieve information about the columns of MySQL tables.

```
DESCRIBE olist_customers_dataset;
DESCRIBE olist_order_items_dataset;
DESCRIBE olist_order_payments_dataset;
DESCRIBE olist_order_reviews_dataset;
DESCRIBE olist_orders_dataset;
DESCRIBE olist_products_dataset;
```

```
mysql> DESCRIBE olist_customers_dataset;
+--------------------------+--------------+------+-----+---------+-------+
| Field                    | Type         | Null | Key | Default | Extra |
+--------------------------+--------------+------+-----+---------+-------+
| customer_id              | varchar(255) | YES  |     | NULL    |       |
| customer_unique_id       | varchar(255) | YES  |     | NULL    |       |
| customer_zip_code_prefix | int          | YES  |     | NULL    |       |
| customer_city            | varchar(255) | YES  |     | NULL    |       |
| customer_state           | varchar(255) | YES  |     | NULL    |       |
+--------------------------+--------------+------+-----+---------+-------+
5 rows in set (0.00 sec)

mysql> DESCRIBE olist_order_items_dataset;
+--------------------+---------------+------+-----+---------+-------+
| Field              | Type          | Null | Key | Default | Extra |
+--------------------+---------------+------+-----+---------+-------+
| order_id           | varchar(255)  | YES  |     | NULL    |       |
| order_item_id      | int           | YES  |     | NULL    |       |
| product_id         | varchar(255)  | YES  |     | NULL    |       |
| seller_id          | varchar(255)  | YES  |     | NULL    |       |
| shipping_limit_date | timestamp     | YES  |     | NULL    |       |
| price              | decimal(10,2) | YES  |     | NULL    |       |
| freight_value      | decimal(10,2) | YES  |     | NULL    |       |
+--------------------+---------------+------+-----+---------+-------+
7 rows in set (0.00 sec)

mysql> DESCRIBE olist_order_payments_dataset;
+----------------------+---------------+------+-----+---------+-------+
| Field                | Type          | Null | Key | Default | Extra |
+----------------------+---------------+------+-----+---------+-------+
| order_id             | varchar(255)  | YES  |     | NULL    |       |
| payment_sequential   | int           | YES  |     | NULL    |       |
| payment_type         | varchar(255)  | YES  |     | NULL    |       |
| payment_installments | int           | YES  |     | NULL    |       |
| payment_value        | decimal(10,2) | YES  |     | NULL    |       |
+----------------------+---------------+------+-----+---------+-------+
5 rows in set (0.00 sec)

mysql> DESCRIBE olist_order_reviews_dataset;
+-------------------------+--------------+------+-----+---------+-------+
| Field                   | Type         | Null | Key | Default | Extra |
+-------------------------+--------------+------+-----+---------+-------+
| review_id               | varchar(255) | YES  |     | NULL    |       |
| order_id                | varchar(255) | YES  |     | NULL    |       |
| review_score            | int          | YES  |     | NULL    |       |
| review_comment_title    | varchar(255) | YES  |     | NULL    |       |
| review_comment_message  | text         | YES  |     | NULL    |       |
| review_creation_date    | timestamp    | YES  |     | NULL    |       |
| review_answer_timestamp | timestamp    | YES  |     | NULL    |       |
+-------------------------+--------------+------+-----+---------+-------+
7 rows in set (0.00 sec)

mysql> DESCRIBE olist_orders_dataset;
+-------------------------------+--------------+------+-----+---------+-------+
| Field                         | Type         | Null | Key | Default | Extra |
+-------------------------------+--------------+------+-----+---------+-------+
| order_id                      | varchar(255) | YES  |     | NULL    |       |
| customer_id                   | varchar(255) | YES  |     | NULL    |       |
| order_status                  | varchar(255) | YES  |     | NULL    |       |
| order_purchase_timestamp      | timestamp    | YES  |     | NULL    |       |
| order_approved_at             | timestamp    | YES  |     | NULL    |       |
| order_delivered_carrier_date  | timestamp    | YES  |     | NULL    |       |
| order_delivered_customer_date | timestamp    | YES  |     | NULL    |       |
| order_estimated_delivery_date | timestamp    | YES  |     | NULL    |       |
+-------------------------------+--------------+------+-----+---------+-------+
8 rows in set (0.00 sec)
```

```
mysql> DESCRIBE olist_products_dataset;
+-----------------------------+-------------+------+-----+---------+-------+
| Field                       | Type        | Null | Key | Default | Extra |
+-----------------------------+-------------+------+-----+---------+-------+
| product_id                  | varchar(255) | YES  |     | NULL    |       |
| product_category_name       | varchar(255) | YES  |     | NULL    |       |
| product_name_length         | int         | YES  |     | NULL    |       |
| product_description_length  | int         | YES  |     | NULL    |       |
| product_photos_qty          | int         | YES  |     | NULL    |       |
| product_weight_g            | int         | YES  |     | NULL    |       |
| product_length_cm           | int         | YES  |     | NULL    |       |
| product_height_cm           | int         | YES  |     | NULL    |       |
| product_width_cm            | int         | YES  |     | NULL    |       |
+-----------------------------+-------------+------+-----+---------+-------+
9 rows in set (0.00 sec)
```

14. Create indexes on specific column of the MySQL tables.

```
CREATE INDEX idx_order_id_olist_orders ON olist_orders_dataset(order_id);
CREATE INDEX idx_customer_id_olist_customers ON
olist_customers_dataset(customer_id);
CREATE INDEX idx_product_id_olist_products ON
olist_products_dataset(product_id);
CREATE INDEX idx_order_id_olist_order_items ON
olist_order_items_dataset(order_id);
CREATE INDEX idx_order_id_olist_order_reviews ON
olist_order_reviews_dataset(order_id);
CREATE INDEX idx_order_id_olist_order_payments ON
olist_order_payments_dataset(order_id);
```

```
mysql> CREATE INDEX idx_order_id_olist_orders ON olist_orders_dataset(order_id);
Query OK, 0 rows affected (0.41 sec)
Records: 0  Duplicates: 0  Warnings: 0

mysql> CREATE INDEX idx_customer_id_olist_customers ON olist_customers_dataset(customer_id);
Query OK, 0 rows affected (0.33 sec)
Records: 0  Duplicates: 0  Warnings: 0

mysql> CREATE INDEX idx_product_id_olist_products ON olist_products_dataset(product_id);
Query OK, 0 rows affected (0.12 sec)
Records: 0  Duplicates: 0  Warnings: 0

mysql> CREATE INDEX idx_order_id_olist_order_items ON olist_order_items_dataset(order_id);
Query OK, 0 rows affected (0.36 sec)
Records: 0  Duplicates: 0  Warnings: 0

mysql> CREATE INDEX idx_order_id_olist_order_reviews ON olist_order_reviews_dataset(order_id);
Query OK, 0 rows affected (0.33 sec)
Records: 0  Duplicates: 0  Warnings: 0

mysql> CREATE INDEX idx_order_id_olist_order_payments ON olist_order_payments_dataset(order_id);
Query OK, 0 rows affected (0.33 sec)
Records: 0  Duplicates: 0  Warnings: 0
```

15. Set the SQL mode of the MySQL server temporarily to allow invalid dates.

```
SET sql_mode = 'ALLOW_INVALID_DATES';
```

```
mysql> SET sql_mode = 'ALLOW_INVALID_DATES';
Query OK, 0 rows affected (0.00 sec)
```

16. Create a new table by joining several MySQL tables.

```
CREATE TABLE olist_joined_dataset AS
SELECT
o.*,
i.order_item_id,
i.product_id,
i.seller_id,
i.shipping_limit_date,
i.price,
i.freight_value,
```

```
pr.product_category_name,
pr.product_name_length,
pr.product_description_length,
pr.product_photos_qty,
pr.product_weight_g,
pr.product_length_cm,
pr.product_height_cm,
pr.product_width_cm,
p.payment_sequential,
p.payment_type,
p.payment_installments,
p.payment_value,
r.review_id,
r.review_score,
r.review_comment_title,
r.review_comment_message,
r.review_creation_date,
r.review_answer_timestamp,
c.customer_unique_id,
c.customer_zip_code_prefix,
c.customer_city,
c.customer_state
FROM olist_orders_dataset AS o
LEFT JOIN olist_order_items_dataset AS i ON o.order_id = i.order_id
INNER JOIN olist_products_dataset AS pr ON i.product_id = pr.product_id
LEFT JOIN olist_order_payments_dataset AS p ON o.order_id = p.order_id
LEFT JOIN olist_order_reviews_dataset AS r ON o.order_id = r.order_id
LEFT JOIN olist_customers_dataset AS c ON o.customer_id = c.customer_id;
```

```
mysql> CREATE TABLE olist_joined_dataset AS
    -> SELECT
    -> o.*,
    -> i.order_item_id,
    -> i.product_id,
    -> i.seller_id,
    -> i.shipping_limit_date,
    -> i.price,
    -> i.freight_value,
    -> pr.product_category_name,
    -> pr.product_name_length,
    -> pr.product_description_length,
    -> pr.product_photos_qty,
    -> pr.product_weight_g,
    -> pr.product_length_cm,
    -> pr.product_height_cm,
    -> pr.product_width_cm,
    -> p.payment_sequential,
    -> p.payment_type,
    -> p.payment_installments,
    -> p.payment_value,
    -> r.review_id,
    -> r.review_score,
    -> r.review_comment_title,
    -> r.review_comment_message,
    -> r.review_creation_date,
    -> r.review_answer_timestamp,
    -> c.customer_unique_id,
    -> c.customer_zip_code_prefix,
    -> c.customer_city,
    -> c.customer_state
    -> FROM olist_orders_dataset AS o
    -> LEFT JOIN olist_order_items_dataset AS i ON o.order_id = i.order_id
    -> INNER JOIN olist_products_dataset AS pr ON i.product_id = pr.product_id
    -> LEFT JOIN olist_order_payments_dataset AS p ON o.order_id = p.order_id
    -> LEFT JOIN olist_order_reviews_dataset AS r ON o.order_id = r.order_id
    -> LEFT JOIN olist_customers_dataset AS c ON o.customer_id = c.customer_id;
Query OK, 118310 rows affected (4.67 sec)
```

17. Retrieve information about the columns of MySQL table.

```
DESCRIBE olist_joined_dataset;
```

```
mysql> DESCRIBE olist_joined_dataset;
+-------------------------------+---------------+------+-----+---------+-------+
| Field                         | Type          | Null | Key | Default | Extra |
+-------------------------------+---------------+------+-----+---------+-------+
| order_id                      | varchar(255)  | YES  |     | NULL    |       |
| customer_id                   | varchar(255)  | YES  |     | NULL    |       |
| order_status                  | varchar(255)  | YES  |     | NULL    |       |
| order_purchase_timestamp      | timestamp     | YES  |     | NULL    |       |
| order_approved_at             | timestamp     | YES  |     | NULL    |       |
| order_delivered_carrier_date  | timestamp     | YES  |     | NULL    |       |
| order_delivered_customer_date | timestamp     | YES  |     | NULL    |       |
| order_estimated_delivery_date | timestamp     | YES  |     | NULL    |       |
| order_item_id                 | int           | YES  |     | NULL    |       |
| product_id                    | varchar(255)  | YES  |     | NULL    |       |
| seller_id                     | varchar(255)  | YES  |     | NULL    |       |
| shipping_limit_date           | timestamp     | YES  |     | NULL    |       |
| price                         | decimal(10,2) | YES  |     | NULL    |       |
| freight_value                 | decimal(10,2) | YES  |     | NULL    |       |
| product_category_name         | varchar(255)  | YES  |     | NULL    |       |
| product_name_length           | int           | YES  |     | NULL    |       |
| product_description_length    | int           | YES  |     | NULL    |       |
| product_photos_qty            | int           | YES  |     | NULL    |       |
| product_weight_g              | int           | YES  |     | NULL    |       |
| product_length_cm             | int           | YES  |     | NULL    |       |
| product_height_cm             | int           | YES  |     | NULL    |       |
| product_width_cm              | int           | YES  |     | NULL    |       |
| payment_sequential            | int           | YES  |     | NULL    |       |
| payment_type                  | varchar(255)  | YES  |     | NULL    |       |
| payment_installments          | int           | YES  |     | NULL    |       |
| payment_value                 | decimal(10,2) | YES  |     | NULL    |       |
| review_id                     | varchar(255)  | YES  |     | NULL    |       |
| review_score                  | int           | YES  |     | NULL    |       |
| review_comment_title          | varchar(255)  | YES  |     | NULL    |       |
| review_comment_message        | text          | YES  |     | NULL    |       |
| review_creation_date          | timestamp     | YES  |     | NULL    |       |
| review_answer_timestamp       | timestamp     | YES  |     | NULL    |       |
| customer_unique_id            | varchar(255)  | YES  |     | NULL    |       |
| customer_zip_code_prefix      | int           | YES  |     | NULL    |       |
| customer_city                 | varchar(255)  | YES  |     | NULL    |       |
| customer_state                | varchar(255)  | YES  |     | NULL    |       |
+-------------------------------+---------------+------+-----+---------+-------+
36 rows in set (0.01 sec)
```

18. Start all the Hadoop daemons.

```
./start-all.sh
jps
```

```
bklow@LAPTOP-M13O57I5:~$ cd hadoop/sbin
bklow@LAPTOP-M13O57I5:~/hadoop/sbin$ ./start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
bklow@localhost's password:
localhost: starting namenode, logging to /home/bklow/hadoop/logs/hadoop-bklow-namenode-LAPTOP-M13O57I5.out
bklow@localhost's password:
localhost: starting datanode, logging to /home/bklow/hadoop/logs/hadoop-bklow-datanode-LAPTOP-M13O57I5.out
Starting secondary namenodes [0.0.0.0]
bklow@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/bklow/hadoop/logs/hadoop-bklow-secondarynamenode-LAPTOP-M13O57I5.out
starting yarn daemons
starting resourcemanager, logging to /home/bklow/hadoop/logs/yarn-bklow-resourcemanager-LAPTOP-M13O57I5.out
bklow@localhost's password:
localhost: starting nodemanager, logging to /home/bklow/hadoop/logs/yarn-bklow-nodemanager-LAPTOP-M13O57I5.out
bklow@LAPTOP-M13O57I5:~/hadoop/sbin$ jps
1784 NodeManager
1482 ResourceManager
1915 Jps
988 NameNode
1325 SecondaryNameNode
1135 DataNode
```

19. Import data from a MySQL database to HDFS using Sqoop.

```
./sqoop import -connect jdbc:mysql://localhost/WQD7007 -username root -password root -table olist_joined_dataset -m 1 -target-dir /user/hdfs/olist
```

```
bklow@LAPTOP-M13O57I5:~/sqoop/bin$ ./sqoop import -connect jdbc:mysql://localhost/WQD7007 -username root -password root -table olist_joined_dataset -m 1 -target-dir /user/hdfs/olist
```

```
24/01/11 10:21:21 INFO mapreduce.Job: Job job_1704937930029_0001 completed successfully
24/01/11 10:21:21 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=219077
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=60607158
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=8810
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4405
                Total vcore-milliseconds taken by all map tasks=4405
                Total megabyte-milliseconds taken by all map tasks=1127680
        Map-Reduce Framework
                Map input records=118310
                Map output records=118310
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=252
                CPU time spent (ms)=7500
                Physical memory (bytes) snapshot=251318272
                Virtual memory (bytes) snapshot=1910280192
                Total committed heap usage (bytes)=126877696
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=60607158
24/01/11 10:21:21 INFO mapreduce.ImportJobBase: Transferred 57.7995 MB in 19.6539 seconds (2.9409 MB/sec)
24/01/11 10:21:21 INFO mapreduce.ImportJobBase: Retrieved 118310 records.
```

20. List the contents of HDFS directory and view the first 10 lines of the file.

```
./hdfs dfs -ls /user/hdfs/olist/
./hdfs dfs -cat /user/hdfs/olist/part-m-00000 | head -n 10
```

```
bklow@LAPTOP-M13O57I5:~/hadoop/bin$ ./hdfs dfs -ls /user/hdfs/olist/
Found 2 items
-rw-r--r--   1 bklow supergroup          0 2024-01-11 10:21 /user/hdfs/olist/_SUCCESS
-rw-r--r--   1 bklow supergroup   60607158 2024-01-11 10:21 /user/hdfs/olist/part-m-00000
bklow@LAPTOP-M13O57I5:~/hadoop/bin$ ./hdfs dfs -cat /user/hdfs/olist/part-m-00000 | head -n 10
e481f51cbdc54678b7cc49136f2d6af7,9ef432eb6251297304e76186b10a928d,delivered,2017-10-02 10:56:33.0,2017-10-02 11:07:15.0,2017-10-04 19:55:00.0,2017-10-10 21:25:13.0,2017-10-18 00:00:00.0,1,87285b34884572647811a353c7ac498a,3504c0cb71d7fa48d967e0e4c94d59d9,2017-10-06 11:07:15.0,29.99,8.72,utilidades_domesticas,40,268,4,500,19,8,13,1,credit_card,1,18.12,a54f0611adc9ed256b57ede6b6eb5114,4,,Não testei o produto ainda, mas ele veio correto e em boas condições. Apenas a caixa que veio bem amassada e danificada, o que ficará chato, pois se trata de um presente.,2017-10-11 00:00:00.0,2017-10-12 03:43:48.0,7c396fd4830fd04220f754e42b4e5bff,3149,sao paulo,SP
e481f51cbdc54678b7cc49136f2d6af7,9ef432eb6251297304e76186b10a928d,delivered,2017-10-02 10:56:33.0,2017-10-02 11:07:15.0,2017-10-04 19:55:00.0,2017-10-10 21:25:13.0,2017-10-18 00:00:00.0,1,87285b34884572647811a353c7ac498a,3504c0cb71d7fa48d967e0e4c94d59d9,2017-10-06 11:07:15.0,29.99,8.72,utilidades_domesticas,40,268,4,500,19,8,13,3,voucher,1,2.00,a54f0611adc9ed256b57ede6b6eb5114,4,,Não testei o produto ainda, mas ele veio correto e em boas condições. Apenas a caixa que veio bem amassada e danificada, o que ficará chato, pois se trata de um presente.,2017-10-11 00:00:00.0,2017-10-12 03:43:48.0,7c396fd4830fd04220f754e42b4e5bff,3149,sao paulo,SP
e481f51cbdc54678b7cc49136f2d6af7,9ef432eb6251297304e76186b10a928d,delivered,2017-10-02 10:56:33.0,2017-10-02 11:07:15.0,2017-10-04 19:55:00.0,2017-10-10 21:25:13.0,2017-10-18 00:00:00.0,1,87285b34884572647811a353c7ac498a,3504c0cb71d7fa48d967e0e4c94d59d9,2017-10-06 11:07:15.0,29.99,8.72,utilidades_domesticas,40,268,4,500,19,8,13,2,voucher,1,18.59,a54f0611adc9ed256b57ede6b6eb5114,4,,Não testei o produto ainda, mas ele veio correto e em boas condições. Apenas a caixa que veio bem amassada e danificada, o que ficará chato, pois se trata de um presente.,2017-10-11 00:00:00.0,2017-10-12 03:43:48.0,7c396fd4830fd04220f754e42b4e5bff,3149,sao paulo,SP
53cdb2fc8bc7dce0b6741e2150273451,b0830fb4747a6c6d20dea0b8c802d7ef,delivered,2018-07-24 20:41:37.0,2018-07-26 03:24:27.0,2018-07-26 14:31:00.0,2018-08-07 15:27:45.0,2018-08-13 00:00:00.0,1,595fac2a385ac33a80bd5114aec74eb8,289cdb325fb7e7f891c38608bf9e0962,2018-07-30 03:24:27.0,118.70,22.76,perfumaria,29,178,1,400,19,13,19,1,boleto,1,141.46,8d5266042046a06655c8db133d120ba5,4,Muito boa a loja,Muito bom o produto.,2018-08-08 00:00:00.0,2018-08-08 18:37:50.0,af07308b275d755c9edb36a90c618231,47813,barreiras,BA
47770eb9100c2d0c44946d9cf07ec65d,41ce2a54c0b03bf3443c3d931a367089,delivered,2018-08-08 08:38:49.0,2018-08-08 08:55:23.0,2018-08-08 13:50:00.0,2018-08-17 18:06:29.0,2018-09-04 00:00:00.0,1,aa4383b373c6aca5d8797843e5594415,4869f7a5dfa277a7dca6462dcf3b52b2,2018-08-13 08:55:23.0,159.90,19.22,automotivo,46,232,1,420,24,19,21,1,credit_card,3,179.12,e73b67b67587f7644d5bd1a52deb1b01,5,,,2018-08-18 00:00:00.0,2018-08-22 19:07:58.0,3a653a41f6f9fc3d2a113cf8398680e8,75265,vianopolis,GO
949d5b44dbf5de918fe9c16f97b45f8a,f88197465ea7920adcdbec7375364d82,delivered,2017-11-18 19:28:06.0,2017-11-18 19:45:59.0,2017-11-22 13:39:59.0,2017-12-02 00:28:42.0,2017-12-15 00:00:00.0,1,d0b61bfb1de832b15ba9d266ca96e5b0,66922902710d126a0e7d26b0e3805106,2017-11-23 19:45:59.0,45.00,27.20,pet_shop,59,468,3,450,30,10,20,1,credit_card,1,72.20,359d03e676b3c069f62cadba8dd3f6e8,5,,O produto foi exatamente o que eu esperava e estava descrito no site e chegou bem antes da data prevista.,2017-12-03 00:00:00.0,2017-12-05 19:21:58.0,7c142cf63193a1473d2e66489a9ae977,59296,sao goncalo do amarante,RN
ad21c59c0840e6cb83a9ceb5573f8159,8ab97904e6daea8866dbdbc4fb7aad2c,delivered,2018-02-13 21:18:39.0,2018-02-13 22:20:29.0,2018-02-14 19:46:34.0,2018-02-16 18:17:02.0,2018-02-26 00:00:00.0,1,65266b2da20d04dbe00c5c2d3bb7859e,2c9e548be18521d1c43cde1c582c6de8,2018-02-19 20:31:37.0,19.90,8.72,papelaria,38,316,4,250,51,15,15,1,credit_card,1,28.62,e50934924e227544ba8246aeb3770dd4,5,,,2018-02-17 00:00:00.0,2018-02-18 13:02:51.0,72632f0f9dd73dfee390c9b22eb56dd6,9195,santo andre,SP
a4591c265e18cb1dcee52889e2d8acc3,503740e9ca751ccdda7ba28e9ab8f608,delivered,2017-07-09 21:57:05.0,2017-07-09 22:10:13.0,2017-07-11 14:58:04.0,2017-07-26 10:57:55.0,2017-08-01 00:00:00.0,1,060cb19345d90064d1015407193c233d,8581055ce74af1dabab164fdbd55a40de,2017-07-13 22:10:13.0,147.90,27.36,automotivo,49,608,1,7150,65,10,65,1,credit_card,6,175.26,89b738e70a1ce346db29a20fb2910161,4,,,2017-07-27 00:00:00.0,2017-07-27 22:48:30.0,80bb27c7c16e8f973207a5086ab329e2,86320,congonhinhas,PR
136cce7faa42fdb2cefd53fdc79a6098,ed0271e0b7da060a393796590e7b737a,invoiced,2017-04-11 12:22:08.0,2017-04-13 13:25:17.0,null,null,2017-05-09 00:00:00.0,1,a1804276d99411ac0733cfd409f5206eb,dc8798cbf453b7e0f98745e396cc5616,2017-04-19 13:25:17.0,49.90,16.05,,0,0,0,600,35,35,15,1,credit_card,1,65.95,e07549ef5311abcc92ba1784b093fb56,2,,fiquei triste por n ter me atendido.,2017-05-13 00:00:00.0,2017-05-13 20:25:42.0,36edbb3fb164b1f16485364b6fb04c73,98900,santa rosa,RS
6514b8ad8028c9f2cc2374ded245783f,9bdf08b4b3b52b5526ff42d37d47f222,delivered,2017-05-16 13:10:30.0,2017-05-16 13:22:11.0,2017-05-22 10:07:46.0,2017-05-26 12:55:51.0,2017-06-07 00:00:00.0,1,4520766ec412348b8d4caa5e8a18c464,16090f2ca825584b5a147ab24aa30c86,2017-05-22 13:22:11.0,59.99,15.17,automotivo,59,956,1,50,16,16,17,1,credit_card,3,75.16,07d67dd06ed5f88bef11ef6b464e79ae,5,,,2017-05-27 00:00:00.0,2017-05-28 02:59:57.0,932afa1e708222e5821dac9cd5db4cae,26525,nilopolis,RJ
cat: Unable to write to output stream.
```

21. Perform a UNION ALL operation on a set of column names followed by selecting all columns from “olist_joined_dataset” table. Export the result into a CSV file.

```
SELECT
        'order_id',
        'customer_id',
        'order_status',
        'order_purchase_timestamp',
        'order_approved_at',
```

```
      'order_delivered_carrier_date',
      'order_delivered_customer_date',
      'order_estimated_delivery_date',
      'order_item_id',
      'product_id',
      'seller_id',
      'shipping_limit_date',
      'price',
      'freight_value',
      'product_category_name',
      'product_name_length',
      'product_description_length',
      'product_photos_qty',
      'product_weight_g',
      'product_length_cm',
      'product_height_cm',
      'product_width_cm',
      'payment_sequential',
      'payment_type',
      'payment_installments',
      'payment_value',
      'review_id',
      'review_score',
      'review_comment_title',
      'review_comment_message',
      'review_creation_date',
      'review_answer_timestamp',
      'customer_unique_id',
      'customer_zip_code_prefix',
      'customer_city',
      'customer_state'
UNION ALL
SELECT *
    FROM olist_joined_dataset
    INTO OUTFILE '/var/lib/mysql-files/olist_joined_dataset.csv'
    FIELDS TERMINATED BY ','
    ENCLOSED BY '"'
    LINES TERMINATED BY '\n';
```

```
mysql> SELECT
    -> 'order_id',
    -> 'customer_id',
order_p    -> 'order_status',
    -> 'order_purchase_timestamp',
    -> 'order_approved_at',
    -> 'order_delivered_carrier_date',
    -> 'order_delivered_customer_date',
    -> 'order_estimated_delivery_date',
    -> 'order_item_id',
    -> 'product_id',
    -> 'seller_id',
    -> 'shipping_limit_date',
    -> 'price',
    -> 'freight_value',
    -> 'product_category_name',
    -> 'product_name_length',
    -> 'product_description_length',
    -> 'product_photos_qty',
    -> 'product_weight_g',
    -> 'product_length_cm',
    -> 'product_height_cm',
    -> 'product_width_cm',
    -> 'payment_sequential',
    -> 'payment_type',
    -> 'payment_installments',
'payment    -> 'payment_value',
    -> 'review_id',
    -> 'review_score',
    -> 'review_comment_title',
    -> 'review_comment_message',
zip_co    -> 'review_creation_date',
    -> 'review_answer_timestamp',
    -> 'customer_unique_id',
    -> 'customer_zip_code_prefix',
    -> 'customer_city',
    -> 'customer_state'
    -> UNION ALL
    -> SELECT *
    ->     FROM olist_joined_dataset
    ->     INTO OUTFILE '/var/lib/mysql-files/olist_joined_dataset.csv'
   FIELDS TERMINATED BY ','
    ENCLOSED BY     ->     FIELDS TERMINATED BY ','
'"'
       ->     ENCLOSED BY '"'
 LINE    ->     LINES TERMINATED BY '\n';
Query OK, 118311 rows affected (0.98 sec)
```

22. Change permission of the CSV file.

```
sudo chmod 755 /var/lib/mysql-files/olist_joined_dataset.csv
```

```
bklow@LAPTOP-M13O57I5:/var/lib/mysql-files$ cd /var/lib/mysql-files/
bklow@LAPTOP-M13O57I5:/var/lib/mysql-files$ ls -l
total 128948
-rwxr-xr-x 1 mysql mysql  9033957 Jan  9 18:21 olist_customers_dataset.csv
-rw-r----- 1 mysql mysql 67286550 Jan 11 10:34 olist_joined_dataset.csv
-rwxr-xr-x 1 mysql mysql 15438671 Jan 11 09:12 olist_order_items_dataset.csv
-rwxr-xr-x 1 mysql mysql  5777138 Jan 11 09:13 olist_order_payments_dataset.csv
-rwxr-xr-x 1 mysql mysql 14451670 Jan 11 09:13 olist_order_reviews_dataset.csv
-rwxr-xr-x 1 mysql mysql 17654914 Jan 11 09:13 olist_orders_dataset.csv
-rwxr-xr-x 1 mysql mysql  2379446 Jan 11 09:13 olist_products_dataset.csv
bklow@LAPTOP-M13O57I5:/var/lib/mysql-files$ sudo chmod 755 /var/lib/mysql-files/olist_joined_dataset.csv
bklow@LAPTOP-M13O57I5:/var/lib/mysql-files$ ls -l
total 128948
-rwxr-xr-x 1 mysql mysql  9033957 Jan  9 18:21 olist_customers_dataset.csv
-rwxr-xr-x 1 mysql mysql 67286550 Jan 11 10:34 olist_joined_dataset.csv
-rwxr-xr-x 1 mysql mysql 15438671 Jan 11 09:12 olist_order_items_dataset.csv
-rwxr-xr-x 1 mysql mysql  5777138 Jan 11 09:13 olist_order_payments_dataset.csv
-rwxr-xr-x 1 mysql mysql 14451670 Jan 11 09:13 olist_order_reviews_dataset.csv
-rwxr-xr-x 1 mysql mysql 17654914 Jan 11 09:13 olist_orders_dataset.csv
-rwxr-xr-x 1 mysql mysql  2379446 Jan 11 09:13 olist_products_dataset.csv
```

23. Copy the CSV file to Downloads folder.

```
sudo cp /var/lib/mysql-files/olist_joined_dataset.csv /mnt/c/Users/boonk/Downloads
ls /mnt/c/Users/boonk/Downloads/olist_joined_dataset.csv
```

```
bklow@LAPTOP-M13O57I5:/var/lib/mysql-files$ sudo cp /var/lib/mysql-files/olist_joined_dataset.csv /mnt/c/Users/boonk/Downloads
bklow@LAPTOP-M13O57I5:/var/lib/mysql-files$ ls /mnt/c/Users/boonk/Downloads/olist_joined_dataset.csv
/mnt/c/Users/boonk/Downloads/olist_joined_dataset.csv
```

6.2 Data Storage - MySQL

Running MySQL in Machine 1 – Ubuntu in Local Machine (16 GB RAM)

To compare performance of data storage tools, we create a new table named "olist_joined_dataset_2" in the MySQL database and load data from "olist_joined_dataset.csv" into the new table, despite having a MySQL table named "olist_joined_dataset" containing the same data when we were performing data pre-processing.

1. Create tables in MySQL database.

```
CREATE TABLE olist_joined_dataset_2 (
    order_id VARCHAR(255),
    customer_id VARCHAR(255),
    order_status VARCHAR(255),
    order_purchase_timestamp TIMESTAMP,
    order_approved_at TIMESTAMP,
    order_delivered_carrier_date TIMESTAMP,
    order_delivered_customer_date TIMESTAMP,
    order_estimated_delivery_date TIMESTAMP,
    order_item_id INT,
    product_id VARCHAR(255),
    seller_id VARCHAR(255),
    shipping_limit_date TIMESTAMP,
    price DECIMAL(10, 2),
    freight_value DECIMAL(10, 2),
    product_category_name VARCHAR(255),
    product_name_length INT,
    product_description_length INT,
    product_photos_qty INT,
    product_weight_g INT,
    product_length_cm INT,
    product_height_cm INT,
    product_width_cm INT,
    payment_sequential INT,
    payment_type VARCHAR(255),
    payment_installments INT,
    payment_value DECIMAL(10, 2),
    review_id VARCHAR(255),
    review_score INT,
    review_comment_title VARCHAR(255),
    review_comment_message TEXT,
    review_creation_date TIMESTAMP,
    review_answer_timestamp TIMESTAMP,
    customer_unique_id VARCHAR(255),
    customer_zip_code_prefix INT,
    customer_city VARCHAR(255),
    customer_state VARCHAR(255)
);
```

```
mysql> CREATE TABLE olist_joined_dataset_2 (
    ->     order_id VARCHAR(255),
    ->     customer_id VARCHAR(255),
    ->     order_status VARCHAR(255),
    ->     order_purchase_timestamp TIMESTAMP,
    ->     order_approved_at TIMESTAMP,
    ->     order_delivered_carrier_date TIMESTAMP,
    ->     order_delivered_customer_date TIMESTAMP,
    ->     order_estimated_delivery_date TIMESTAMP,
    ->     order_item_id INT,
    ->     product_id VARCHAR(255),
    ->     seller_id VARCHAR(255),
    ->     shipping_limit_date TIMESTAMP,
    ->     price DECIMAL(10, 2),
    ->     freight_value DECIMAL(10, 2),
    ->     product_category_name VARCHAR(255),
    ->     product_name_length INT,
    ->     product_description_length INT,
    ->     product_photos_qty INT,
    ->     product_weight_g INT,
    ->     product_length_cm INT,
    ->     product_height_cm INT,
    ->     product_width_cm INT,
    ->     payment_sequential INT,
    ->     payment_type VARCHAR(255),
    ->     payment_installments INT,
    ->     payment_value DECIMAL(10, 2),
    ->     review_id VARCHAR(255),
    ->     review_score INT,
    ->     review_comment_title VARCHAR(255),
    ->     review_comment_message TEXT,
    ->     review_creation_date TIMESTAMP,
    ->     review_answer_timestamp TIMESTAMP,
    ->     customer_unique_id VARCHAR(255),
    ->     customer_zip_code_prefix INT,
    ->     customer_city VARCHAR(255),
    ->     customer_state VARCHAR(255)
    -> );
Query OK, 0 rows affected (0.02 sec)
```

2. Import data from CSV file into MySQL table.

```
LOAD DATA INFILE '/var/lib/mysql-files/olist_joined_dataset.csv' INTO TABLE
olist_joined_dataset_2 FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES
TERMINATED BY '\n' IGNORE 1 ROWS;
```

```
mysql> LOAD DATA INFILE '/var/lib/mysql-files/olist_joined_dataset.csv' INTO TABLE olist_joined_dataset_2 FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n
' IGNORE 1 ROWS;
Query OK, 118310 rows affected (1.62 sec)
Records: 118310  Deleted: 0  Skipped: 0  Warnings: 0

mysql> SHOW PROFILES;
+----------+------------+--------------------------------------------------
---------------------------------------------------+
| Query_ID | Duration   | Query
                                                   |
+----------+------------+--------------------------------------------------
---------------------------------------------------+
|        1 | 0.01393800 | CREATE TABLE olist_joined_dataset_2 (
    order_id VARCHAR(255),
    customer_id VARCHAR(255),
    order_status VARCHAR(255),
    order_purchase_timestamp TIMESTAMP,
    order_approved_at TIMESTAMP,
    order_delivered_carrier_date TIMESTAMP,
    order_delivered_customer_date TIMESTAMP,
    order_es |
|        2 | 1.62492825 | LOAD DATA INFILE '/var/lib/mysql-files/olist_joined_dataset.csv' INTO TABLE olist_joined_dataset_2 FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINE
S TERMINATED BY '\n' IGNORE 1 ROWS
                                                   |
+----------+------------+--------------------------------------------------
---------------------------------------------------+
2 rows in set, 1 warning (0.00 sec)
```

3. Retrieve information about the columns of MySQL tables.

```
DESCRIBE olist_joined_dataset_2;
```

```
mysql> DESCRIBE olist_joined_dataset_2;
+--------------------------------+---------------+------+-----+---------+-------+
| Field                          | Type          | Null | Key | Default | Extra |
+--------------------------------+---------------+------+-----+---------+-------+
| order_id                       | varchar(255)  | YES  |     | NULL    |       |
| customer_id                    | varchar(255)  | YES  |     | NULL    |       |
| order_status                   | varchar(255)  | YES  |     | NULL    |       |
| order_purchase_timestamp       | timestamp     | YES  |     | NULL    |       |
| order_approved_at              | timestamp     | YES  |     | NULL    |       |
| order_delivered_carrier_date   | timestamp     | YES  |     | NULL    |       |
| order_delivered_customer_date  | timestamp     | YES  |     | NULL    |       |
| order_estimated_delivery_date  | timestamp     | YES  |     | NULL    |       |
| order_item_id                  | int           | YES  |     | NULL    |       |
| product_id                     | varchar(255)  | YES  |     | NULL    |       |
| seller_id                      | varchar(255)  | YES  |     | NULL    |       |
| shipping_limit_date            | timestamp     | YES  |     | NULL    |       |
| price                          | decimal(10,2) | YES  |     | NULL    |       |
| freight_value                  | decimal(10,2) | YES  |     | NULL    |       |
| product_category_name          | varchar(255)  | YES  |     | NULL    |       |
| product_name_length            | int           | YES  |     | NULL    |       |
| product_description_length     | int           | YES  |     | NULL    |       |
| product_photos_qty             | int           | YES  |     | NULL    |       |
| product_weight_g               | int           | YES  |     | NULL    |       |
| product_length_cm              | int           | YES  |     | NULL    |       |
| product_height_cm              | int           | YES  |     | NULL    |       |
| product_width_cm               | int           | YES  |     | NULL    |       |
| payment_sequential             | int           | YES  |     | NULL    |       |
| payment_type                   | varchar(255)  | YES  |     | NULL    |       |
| payment_installments           | int           | YES  |     | NULL    |       |
| payment_value                  | decimal(10,2) | YES  |     | NULL    |       |
| review_id                      | varchar(255)  | YES  |     | NULL    |       |
| review_score                   | int           | YES  |     | NULL    |       |
| review_comment_title           | varchar(255)  | YES  |     | NULL    |       |
| review_comment_message         | text          | YES  |     | NULL    |       |
| review_creation_date           | timestamp     | YES  |     | NULL    |       |
| review_answer_timestamp        | timestamp     | YES  |     | NULL    |       |
| customer_unique_id             | varchar(255)  | YES  |     | NULL    |       |
| customer_zip_code_prefix       | int           | YES  |     | NULL    |       |
| customer_city                  | varchar(255)  | YES  |     | NULL    |       |
| customer_state                 | varchar(255)  | YES  |     | NULL    |       |
+--------------------------------+---------------+------+-----+---------+-------+
36 rows in set (0.00 sec)

mysql> SHOW PROFILES;
+----------+------------+-----------------------------------+
| Query_ID | Duration   | Query                             |
+----------+------------+-----------------------------------+
|        1 | 0.00118800 | DESCRIBE olist_joined_dataset_2 |
+----------+------------+-----------------------------------+
1 row in set, 1 warning (0.00 sec)
```

4. Retrieve the first 10 rows from the MySQL table.

```
SELECT * FROM olist_joined_dataset_2 LIMIT 10;
```

```
mysql> SHOW PROFILES;
+----------+------------+-----------------------------------------------+
| Query_ID | Duration   | Query                                         |
+----------+------------+-----------------------------------------------+
|        1 | 0.00118800 | DESCRIBE olist_joined_dataset_2               |
|        2 | 0.00065350 | SELECT * FROM olist_joined_dataset_2 LIMIT 10 |
+----------+------------+-----------------------------------------------+
2 rows in set, 1 warning (0.00 sec)
```

5. Retrieve all rows from the MySQL table.

```
SELECT * FROM olist_joined_dataset_2;
```

```
mysql> SELECT * FROM olist_joined_dataset_2;

1d36f14b6053 | 1caf283236cd69af44cbc09a0a1e7d32 | 2017-10-10 20:07:14 |   42.11 |        0.80 | brinquedos |   19 |   328 |   3 |   600 |   48 |
      10 |       48 |      2 | voucher     |      1 |      64.37 | ebd75732b5804e934123d11ec1f11db0 |      5 |  |
                                                                                      | 2017-10-21 00:00:00 | 2017-10-23 14:48:40 | c716cf2b5b86fb24257c

ffe9e7969df8 |      78048 | cuiaba        | MT     |
| 9115838be804184b91f5c00f6f49f92d | da2124f134f5dfbce9d06f29bdb6c308 | delivered | 2017-10-04 19:57:37 | 2017-10-04 20:07:14 | 2017-10-05 16:52:52 | 2017-10-20 20:25:45 | 2017-11-07 00:00:00 |      2 | 49d2e2460386273b195e
7e59b43587c3 | 1caf283236cd69af44cbc09a0a1e7d32 | 2017-10-10 20:07:14 |   26.90 |       36.98 | brinquedos |   40 |   180 |   3 |   750 |   26 |
      15 |       26 |      2 | voucher     |      1 |      64.37 | ebd75732b5804e934123d11ec1f11db0 |      5 |  |
                                                                                      | 2017-10-21 00:00:00 | 2017-10-23 14:48:40 | c716cf2b5b86fb24257c

ffe9e7969df8 |      78048 | cuiaba        | MT     |
| aa04ef5214580b06b10e2a378300db44 | f01a6bfcc730456317e4081fe0c9940e | delivered | 2017-01-27 00:30:03 | 2017-01-27 01:05:25 | 2017-01-30 11:40:16 | 2017-02-07 13:15:25 | 2017-03-17 00:00:00 |      1 | 9fc063fd34fed29ccc57
b7f8e8d03388 | ccc4bbb5f32a6ab2b7066a4130f114e3 | 2017-02-03 00:30:03 |  370.00 |       19.43 | beleza_saude |   48 |   657 |   1 |   750 |   38 |
      12 |       25 |      2 | voucher     |      1 |     250.00 | e8995c053d3db2d9c07407efe7de52dd |      5 |  |
                                                                                      | 2017-02-08 00:00:00 | 2017-02-11 12:37:36 | e03dbdf5e56c96b106d8

115ac336f47f |      35502 | divinopolis   | MG     |
| aa04ef5214580b06b10e2a378300db44 | f01a6bfcc730456317e4081fe0c9940e | delivered | 2017-01-27 00:30:03 | 2017-01-27 01:05:25 | 2017-01-30 11:40:16 | 2017-02-07 13:15:25 | 2017-03-17 00:00:00 |      1 | 9fc063fd34fed29ccc57
b7f8e8d03388 | ccc4bbb5f32a6ab2b7066a4130f114e3 | 2017-02-03 00:30:03 |  370.00 |       19.43 | beleza_saude |   48 |   657 |   1 |   750 |   38 |
      12 |       25 |      1 | credit_card |      5 |     139.43 | e8995c053d3db2d9c07407efe7de52dd |      5 |  |
                                                                                      | 2017-02-08 00:00:00 | 2017-02-11 12:37:36 | e03dbdf5e56c96b106d8

115ac336f47f |      35502 | divinopolis   | MG     |
| 880675dff2150932f1601e1c07eadeeb | 47cd45a6ac7b9fb16537df2ccffeb5ac | delivered | 2017-02-23 09:05:12 | 2017-02-23 09:15:11 | 2017-03-01 10:22:52 | 2017-03-06 11:08:08 | 2017-03-22 00:00:00 |      1 | ea73128566d1b082e510
1ce46f8107c7 | 391fc6631aebcf3004804e51b40bcf1e | 2017-02-27 09:05:12 |  139.90 |       16.09 | moveis_decoracao |   63 |   254 |   2 |  2500 |   49 |
      13 |       41 |      1 | credit_card |      3 |     155.99 | 96e8f371a3983122b739944537e15525 |      5 |  |
                                                                                      | 2017-03-07 00:00:00 | 2017-03-11 15:42:41 | 831ce3f1bacbd424fc4e

38fbd4d66d29 |       5127 | sao paulo     | SP     |
| 9c5dedf39a927c1b2549525ed64a053c | 39bd1228ee8140590ac3aca26f2dfe00 | delivered | 2017-03-09 09:54:05 | 2017-03-09 09:54:05 | 2017-03-10 11:18:03 | 2017-03-17 15:08:01 | 2017-03-28 00:00:00 |      1 | ac35486adb7b02598c18
2c2ff2e05254 | e24fc9fcd865784fb25705606fe3dfe7 | 2017-03-15 09:54:05 |   72.00 |       13.08 | beleza_saude |   50 |  1517 |   1 |  1175 |   22 |
      13 |       18 |      1 | credit_card |      3 |      85.08 | e262b3f92d1ce917aa412a9406cf61a6 |      5 |  |
                                                                                      | 2017-03-22 00:00:00 | 2017-03-23 11:02:08 | 6359f309b166b0196dbf

7ad2ac62bb5a |      12209 | sao jose dos campos | SP     |
| 63943bddc261676b46f01ca7ac2f7bd8 | 1fca14ff2861355f6e5f14306ff977a7 | delivered | 2018-02-06 12:58:58 | 2018-02-06 13:10:37 | 2018-02-07 23:22:42 | 2018-02-28 17:37:56 | 2018-03-02 00:00:00 |      1 | f1d4ce8c6dd66c47bbaa
8c6781c2a923 | 1f9ab4708f3056ede07124aad39a2554 | 2018-02-12 13:10:37 |  174.90 |       20.10 | bebes |   52 |   828 |   4 |  4950 |   40 |
      10 |       40 |      1 | credit_card |      3 |     195.00 | 29bb71b2760d0f876dfa178a76bc4734 |      4 | So uma peça que veio rachado mas tudo bem rs
                                                                                      | 2018-03-01 00:00:00 | 2018-03-02 17:50:01 | da62f9e57a76d978d02a

b5362c509660 |      11722 | praia grande  | SP     |
| 83c1379a015df1e13d02aae0204711ab | 1aa71eb042121263aafbe80c1b562c9c | delivered | 2017-08-27 14:46:43 | 2017-08-27 15:04:16 | 2017-08-28 20:52:26 | 2017-09-21 11:24:17 | 2017-09-27 00:00:00 |      1 | b80910977a37536adedd
d63663f916ad | d50d79cb34e38265a8649c383dcffd48 | 2017-09-05 15:04:16 |  205.99 |       65.02 | eletrodomesticos_2 |   51 |   500 |   2 | 13300 |   32 |
      90 |       22 |      1 | credit_card |      5 |     271.01 | 371579771219f6db2d830d50805977bb |      5 | Foi entregue antes do prazo.
                                                                                      | 2017-09-22 00:00:00 | 2017-09-22 23:10:57 | 737520a9aad80b3fbbda

d19b66b37b30 |      45920 | nova vicosa   | BA     |
| 11c177c8e97725db2631073c19f07b62 | b331b74b18dc79bcdf6532d51e1637c1 | delivered | 2018-01-08 21:28:27 | 2018-01-08 21:36:21 | 2018-01-12 15:35:03 | 2018-01-25 23:32:54 | 2018-02-15 00:00:00 |      1 | d1c427060a0f73f6b889
a5c7c61f2ac4 | a1043bafd471dff536d0c462352beb48 | 2018-01-12 21:36:21 |  179.99 |       40.59 | informatica_acessorios |   59 |  1893 |   1 |  6550 |   20 |
      20 |       20 |      1 | credit_card |      4 |     441.16 | 8ab6855b9fe9b812cd03a480a25058a1 |      2 | Foi entregue somente 1. Quero saber do outro produto.
                                                                                      | 2018-01-26 00:00:00 | 2018-01-27 09:16:56 | 5097a5312c8b157bb7be

58ae360ef43c |      28685 | japuiba       | RJ     |
| 11c177c8e97725db2631073c19f07b62 | b331b74b18dc79bcdf6532d51e1637c1 | delivered | 2018-01-08 21:28:27 | 2018-01-08 21:36:21 | 2018-01-12 15:35:03 | 2018-01-25 23:32:54 | 2018-02-15 00:00:00 |      2 | d1c427060a0f73f6b889
a5c7c61f2ac4 | a1043bafd471dff536d0c462352beb48 | 2018-01-12 21:36:21 |  179.99 |       40.59 | informatica_acessorios |   59 |  1893 |   1 |  6550 |   20 |
      20 |       20 |      1 | credit_card |      4 |     441.16 | 8ab6855b9fe9b812cd03a480a25058a1 |      2 | Foi entregue somente 1. Quero saber do outro produto.
                                                                                      | 2018-01-26 00:00:00 | 2018-01-27 09:16:56 | 5097a5312c8b157bb7be

58ae360ef43c |      28685 | japuiba       | RJ     |
| 66dea50a8b16d9b4dee7af250b4be1a5 | edb027a75a1449115f6b43211ae02a24 | delivered | 2018-03-08 20:57:30 | 2018-03-09 11:20:28 | 2018-03-09 22:11:59 | 2018-03-16 13:08:30 | 2018-04-03 00:00:00 |      1 | 006619bbed68b000c8ba
3f8725d5409e | ececbfcff9804a2d6b40f589df8eef2b | 2018-03-15 10:55:42 |   68.50 |       18.36 | beleza_saude |   45 |   569 |   1 |   150 |   16 |
       7 |       15 |      1 | debit_card  |      1 |      86.86 | dc9c59b4688062c25758c2be4cafc523 |      5 |  |
                                                                                      | 2018-03-17 00:00:00 | 2018-03-17 16:33:31 | 60350aa974b26ff12caa

d89e55993bd6 |      83750 | lapa          | PR     |
+--------------+------------+---------------+--------+--------------------+---------------------+----------------------+
---------------+-----------------+------------+------------------+-------------------------------------------------------
---------+----------+------------+---------------+-------------+-------------+---------------------------------------+--
--------------------------------------------------------------------------------+------------+----------------+---------
---------------+---------------+-----------------+--------+

118310 rows in set (0.37 sec)
```

```
mysql> SHOW PROFILES;
+----------+------------+----------------------------------------------+
| Query_ID | Duration   | Query                                        |
+----------+------------+----------------------------------------------+
|        1 | 0.00118800 | DESCRIBE olist_joined_dataset_2              |
|        2 | 0.00065350 | SELECT * FROM olist_joined_dataset_2 LIMIT 10 |
|        3 | 0.36963050 | SELECT * FROM olist_joined_dataset_2         |
+----------+------------+----------------------------------------------+
3 rows in set, 1 warning (0.01 sec)
```

Running MySQL in Machine 3 –Virtual Machine, UBuntu (4 GB RAM)

1. Create tables in MySQL database.

```
CREATE TABLE olist_joined_dataset_2 (
    order_id VARCHAR(255),
    customer_id VARCHAR(255),
    order_status VARCHAR(255),
    order_purchase_timestamp TIMESTAMP,
    order_approved_at TIMESTAMP,
    order_delivered_carrier_date TIMESTAMP,
    order_delivered_customer_date TIMESTAMP,
    order_estimated_delivery_date TIMESTAMP,
    order_item_id INT,
    product_id VARCHAR(255),
    seller_id VARCHAR(255),
    shipping_limit_date TIMESTAMP,
    price DECIMAL(10, 2),
    freight_value DECIMAL(10, 2),
    product_category_name VARCHAR(255),
    product_name_length INT,
    product_description_length INT,
    product_photos_qty INT,
    product_weight_g INT,
    product_length_cm INT,
    product_height_cm INT,
    product_width_cm INT,
    payment_sequential INT,
    payment_type VARCHAR(255),
    payment_installments INT,
    payment_value DECIMAL(10, 2),
    review_id VARCHAR(255),
    review_score INT,
    review_comment_title VARCHAR(255),
    review_comment_message TEXT,
    review_creation_date TIMESTAMP,
    review_answer_timestamp TIMESTAMP,
    customer_unique_id VARCHAR(255),
    customer_zip_code_prefix INT,
    customer_city VARCHAR(255),
    customer_state VARCHAR(255)
);
```

```
mysql> CREATE TABLE olist_joined_dataset_2 (
    ->     order_id VARCHAR(255),
    ->     customer_id VARCHAR(255),
    ->     order_status VARCHAR(255),
    ->     order_purchase_timestamp TIMESTAMP,
    ->     order_approved_at TIMESTAMP,
    ->     order_delivered_carrier_date TIMESTAMP,
    ->     order_delivered_customer_date TIMESTAMP,
    ->     order_estimated_delivery_date TIMESTAMP,
    ->     order_item_id INT,
    ->     product_id VARCHAR(255),
    ->     seller_id VARCHAR(255),
    ->     shipping_limit_date TIMESTAMP,
    ->     price DECIMAL(10, 2),
    ->     freight_value DECIMAL(10, 2),
    ->     product_category_name VARCHAR(255),
    ->     product_name_length INT,
    ->     product_description_length INT,
    ->     product_photos_qty INT,
    ->     product_weight_g INT,
    ->     product_length_cm INT,
    ->     product_height_cm INT,
    ->     product_width_cm INT,
    ->     payment_sequential INT,
    ->     payment_type VARCHAR(255),
    ->     payment_installments INT,
    ->     payment_value DECIMAL(10, 2),
    ->     review_id VARCHAR(255),
    ->     review_score INT,
    ->     review_comment_title VARCHAR(255),
    ->     review_comment_message TEXT,
    ->     review_creation_date TIMESTAMP,
    ->     review_answer_timestamp TIMESTAMP,
    ->     customer_unique_id VARCHAR(255),
    ->     customer_zip_code_prefix INT,
    ->     customer_city VARCHAR(255),
    ->     customer_state VARCHAR(255)
    -> );
Query OK, 0 rows affected (0.05 sec)
```

2. Import data from CSV file into MySQL table.

```
LOAD DATA INFILE '/var/lib/mysql-files/olist_joined_dataset.csv' INTO TABLE
olist_joined_dataset_2 FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES
TERMINATED BY '\n' IGNORE 1 ROWS;
```

```
mysql> set profiling = 1;
Query OK, 0 rows affected, 1 warning (0.01 sec)

mysql> LOAD DATA INFILE '/var/lib/mysql-files/olist_joined_dataset.csv' INTO TABLE olist_joined_dataset_2 FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
Query OK, 118310 rows affected (4.74 sec)
Records: 118310  Deleted: 0  Skipped: 0  Warnings: 0

mysql> SHOW PROFILES;
+----------+------------+-----------------------------------------------------------------------------------------------------------------------------------------
-+
| Query_ID | Duration   | Query
 |
+----------+------------+-----------------------------------------------------------------------------------------------------------------------------------------
-+
|        1 | 4.74025750 | LOAD DATA INFILE '/var/lib/mysql-files/olist_joined_dataset.csv' INTO TABLE olist_joined_dataset_2 FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\
 |
+----------+------------+-----------------------------------------------------------------------------------------------------------------------------------------
-+
1 row in set, 1 warning (0.00 sec)
```

3. Retrieve information about the columns of MySQL tables.

```
DESCRIBE olist_joined_dataset_2;
```

```
mysql> DESCRIBE olist_joined_dataset_2;
+-------------------------------+---------------+------+-----+---------+-------+
| Field                         | Type          | Null | Key | Default | Extra |
+-------------------------------+---------------+------+-----+---------+-------+
| order_id                      | varchar(255)  | YES  |     | NULL    |       |
| customer_id                   | varchar(255)  | YES  |     | NULL    |       |
| order_status                  | varchar(255)  | YES  |     | NULL    |       |
| order_purchase_timestamp      | timestamp     | YES  |     | NULL    |       |
| order_approved_at             | timestamp     | YES  |     | NULL    |       |
| order_delivered_carrier_date  | timestamp     | YES  |     | NULL    |       |
| order_delivered_customer_date | timestamp     | YES  |     | NULL    |       |
| order_estimated_delivery_date | timestamp     | YES  |     | NULL    |       |
| order_item_id                 | int           | YES  |     | NULL    |       |
| product_id                    | varchar(255)  | YES  |     | NULL    |       |
| seller_id                     | varchar(255)  | YES  |     | NULL    |       |
| shipping_limit_date           | timestamp     | YES  |     | NULL    |       |
| price                         | decimal(10,2) | YES  |     | NULL    |       |
| freight_value                 | decimal(10,2) | YES  |     | NULL    |       |
| product_category_name         | varchar(255)  | YES  |     | NULL    |       |
| product_name_length           | int           | YES  |     | NULL    |       |
| product_description_length    | int           | YES  |     | NULL    |       |
| product_photos_qty            | int           | YES  |     | NULL    |       |
| product_weight_g              | int           | YES  |     | NULL    |       |
| product_length_cm             | int           | YES  |     | NULL    |       |
| product_height_cm             | int           | YES  |     | NULL    |       |
| product_width_cm              | int           | YES  |     | NULL    |       |
| payment_sequential            | int           | YES  |     | NULL    |       |
| payment_type                  | varchar(255)  | YES  |     | NULL    |       |
| payment_installments          | int           | YES  |     | NULL    |       |
| payment_value                 | decimal(10,2) | YES  |     | NULL    |       |
| review_id                     | varchar(255)  | YES  |     | NULL    |       |
| review_score                  | int           | YES  |     | NULL    |       |
| review_comment_title          | varchar(255)  | YES  |     | NULL    |       |
| review_comment_message        | text          | YES  |     | NULL    |       |
| review_creation_date          | timestamp     | YES  |     | NULL    |       |
| review_answer_timestamp       | timestamp     | YES  |     | NULL    |       |
| customer_unique_id            | varchar(255)  | YES  |     | NULL    |       |
| customer_zip_code_prefix      | int           | YES  |     | NULL    |       |
| customer_city                 | varchar(255)  | YES  |     | NULL    |       |
| customer_state                | varchar(255)  | YES  |     | NULL    |       |
+-------------------------------+---------------+------+-----+---------+-------+
36 rows in set (0.02 sec)
```

```
mysql> SHOW PROFILES;
+----------+------------+------------------------------------------------------------------------------------------------------
-+
| Query_ID | Duration   | Query
 |
+----------+------------+------------------------------------------------------------------------------------------------------
-+
|        1 | 4.74025750 | LOAD DATA INFILE '/var/lib/mysql-files/olist_joined_dataset.csv' INTO TABLE olist_joined_dataset_2 FIELDS TERMINATED BY ',' ENC
 |
|        2 | 0.01899300 | DESCRIBE olist_joined_dataset_2
 |
+----------+------------+------------------------------------------------------------------------------------------------------
-+
2 rows in set, 1 warning (0.00 sec)
```

4. Retrieve the first 10 rows from the MySQL table.

```
SELECT * FROM olist_joined_dataset_2 LIMIT 10;
```

```
mysql> SELECT * FROM olist_joined_dataset_2 LIMIT 10;
| order_id                         | customer_id                      | order_status | order_purchase_timestamp | order_approved_at   | order_delivered_carrier_date | order_delivered_customer_date | order
_estimated_delivery_date | order_item_id | product_id                       | seller_id                        | shipping_limit_date | price  | freight_value | product_category_name | product_name_length
| product_description_length | product_photos_qty | product_weight_g | product_length_cm | product_height_cm | product_width_cm | payment_sequential | payment_type | payment_installments | payment_value |
 review_id                        | review_score | review_comment_title | review_comment_message
                        | review_creation_date | review_answer_timestamp | customer_unique_id               | customer_zip_code_prefix | customer_city          | customer_state |
| e481f51cbdc54678b7cc49136f2d6af7 | 9ef432eb6251297304e76186b10a928d | delivered    | 2017-10-02 10:56:33      | 2017-10-02 11:07:15 | 2017-10-04 19:55:00          | 2017-10-10 21:25:13           | 2017-
10-18 00:00:00           |             1 | 87285b34884572647811a353c7ac498a | 3504c0cb71d7fa48d967e0e4c94d59d9 | 2017-10-06 11:07:15 |  29.99 |          8.72 | utilidades_domesticas |                  40
|                        268 |                  4 |              500 |                19 |                 8 |               13 |                  1 | credit_card  |                    1 |         18.12 |
 a54f0611adc9ed256b57ede6b6eb5114 |            4 |                      | Não testei o produto ainda, mas ele veio correto e em boas condições. Apenas a caixa que veio bem amassada e danificada, o que fic
ará chato, pois se trata de um presente.    | 2017-10-11 00:00:00  | 2017-10-12 03:43:48     | 7c396fd4830fd04220f754e42b4e5bff |                     3149 | sao paulo              | SP             |
| e481f51cbdc54678b7cc49136f2d6af7 | 9ef432eb6251297304e76186b10a928d | delivered    | 2017-10-02 10:56:33      | 2017-10-02 11:07:15 | 2017-10-04 19:55:00          | 2017-10-10 21:25:13           | 2017-
10-18 00:00:00           |             1 | 87285b34884572647811a353c7ac498a | 3504c0cb71d7fa48d967e0e4c94d59d9 | 2017-10-06 11:07:15 |  29.99 |          8.72 | utilidades_domesticas |                  40
|                        268 |                  4 |              500 |                19 |                 8 |               13 |                  3 | voucher      |                    1 |          2.00 |
 a54f0611adc9ed256b57ede6b6eb5114 |            4 |                      | Não testei o produto ainda, mas ele veio correto e em boas condições. Apenas a caixa que veio bem amassada e danificada, o que fic
ará chato, pois se trata de um presente.    | 2017-10-11 00:00:00  | 2017-10-12 03:43:48     | 7c396fd4830fd04220f754e42b4e5bff |                     3149 | sao paulo              | SP             |
| e481f51cbdc54678b7cc49136f2d6af7 | 9ef432eb6251297304e76186b10a928d | delivered    | 2017-10-02 10:56:33      | 2017-10-02 11:07:15 | 2017-10-04 19:55:00          | 2017-10-10 21:25:13           | 2017-
10-18 00:00:00           |             1 | 87285b34884572647811a353c7ac498a | 3504c0cb71d7fa48d967e0e4c94d59d9 | 2017-10-06 11:07:15 |  29.99 |          8.72 | utilidades_domesticas |                  40
|                        268 |                  4 |              500 |                19 |                 8 |               13 |                  2 | voucher      |                    1 |         18.59 |
 a54f0611adc9ed256b57ede6b6eb5114 |            4 |                      | Não testei o produto ainda, mas ele veio correto e em boas condições. Apenas a caixa que veio bem amassada e danificada, o que fic
ará chato, pois se trata de um presente.    | 2017-10-11 00:00:00  | 2017-10-12 03:43:48     | 7c396fd4830fd04220f754e42b4e5bff |                     3149 | sao paulo              | SP             |
| 53cdb2fc8bc7dce0b6741e2150273451 | b0830fb4747a6c6d20dea0b8c802d7ef | delivered    | 2018-07-24 20:41:37      | 2018-07-26 03:24:27 | 2018-07-26 14:31:00          | 2018-08-07 15:27:45           | 2018-
08-13 00:00:00           |             1 | 595fac2a385ac33a80bd5114aec74eb8 | 289cdb325fb7e7f891c38608bf9e0962 | 2018-07-30 03:24:27 | 118.70 |         22.76 | perfumaria            |                  29
|                        178 |                  1 |              400 |                19 |                13 |               19 |                  1 | boleto       |                    1 |        141.46 |
 8d5266042046a06655c8db133d120ba5 |            4 | Muito boa a loja     | Muito bom o produto.
                                            | 2018-08-08 00:00:00  | 2018-08-08 18:37:50     | af07308b275d755c9edb36a90c618231 |                    47813 | barreiras              | BA             |
| 47770eb9100c2d0c44946d9cf07ec65d | 41ce2a54c0b03bf3443c3d931a367089 | delivered    | 2018-08-08 08:38:49      | 2018-08-08 08:55:23 | 2018-08-08 13:50:00          | 2018-08-17 18:06:29           | 2018-
09-04 00:00:00           |             1 | aa4383b373c6aca5d8797843e5594415 | 4869f7a5dfa277a7dca6462dcf3b52b2 | 2018-08-13 08:55:23 | 159.90 |         19.22 | automotivo            |                  46
|                        232 |                  1 |              420 |                24 |                19 |               21 |                  1 | credit_card  |                    3 |        179.12 |
 e73b67b67587f7644d5bd1a52deb1b01 |            5 |                      |
                                            | 2018-08-18 00:00:00  | 2018-08-22 19:07:58     | 3a653a41f6f9fc3d2a113cf8398680e8 |                    75265 | vianopolis             | GO             |
| 949d5b44dbf5de918fe9c16f97b45f8a | f88197465ea7920adcdbec7375364d82 | delivered    | 2017-11-18 19:28:06      | 2017-11-18 19:45:59 | 2017-11-22 13:39:59          | 2017-12-02 00:28:42           | 2017-
12-15 00:00:00           |             1 | d0b61bfb1de832b15ba9d266ca96e5b0 | 66922902710d126a0e7d26b0e3805106 | 2017-11-23 19:45:59 |  45.00 |         27.20 | pet_shop              |                  59
|                        468 |                  3 |              450 |                30 |                10 |               20 |                  1 | credit_card  |                    1 |         72.20 |
```

```
|                        956 |                  1 |               50 |                16 |                16 |               17 |                  1 | credit_card  |                    3 |         75.16 |
 07d67dd06ed5f88bef11ef6b464e79ae |            5 |                      |
                                            | 2017-05-27 00:00:00  | 2017-05-28 02:59:57     | 932afa1e708222e5821dac9cd5db4cae |                    26525 | nilopolis              | RJ             |
10 rows in set (0.00 sec)

mysql> SHOW PROFILES;
+----------+------------+------------------------------------------------------------------------------------------------------
-+
| Query_ID | Duration   | Query
 |
+----------+------------+------------------------------------------------------------------------------------------------------
-+
|        1 | 4.74025750 | LOAD DATA INFILE '/var/lib/mysql-files/olist_joined_dataset.csv' INTO TABLE olist_joined_dataset_2 FIELDS TERMINATED BY ',' ENCLOSED BY '"' LINES TERMINATED BY '\n' IGNORE 1 ROWS
 |
|        2 | 0.01899300 | DESCRIBE olist_joined_dataset_2
 |
|        3 | 0.00107275 | SELECT * FROM olist_joined_dataset_2 LIMIT 10
 |
+----------+------------+------------------------------------------------------------------------------------------------------
-+
3 rows in set, 1 warning (0.02 sec)
```

5. Retrieve all rows from the MySQL table.

```
SELECT * FROM olist_joined_dataset_2;
```

6.3 Data Storage – HBase

Running HBase in Machine 1 – Ubuntu in Local Machine (16 GB RAM)

1. Import joined dataset CSV into HDFS.

```
./hadoop/bin/hdfs dfs -put cleaned6.csv /user/hdfs/inputfolder/
```

```
bklow@LAPTOP-M13O57I5:~$ ./hadoop/bin/hdfs dfs -put cleaned6.csv /user/hdfs/inputfolder/
bklow@LAPTOP-M13O57I5:~$ ./hadoop/bin/hdfs dfs -ls /user/hdfs/inputfolder/
Found 1 items
-rw-r--r--   1 bklow supergroup   54318280 2024-01-20 14:03 /user/hdfs/inputfolder/cleaned6.csv
```

2. Create a table in HBase database.

```
create 'OJD', 'OrderInfo', 'ItemInfo', 'ProductInfo', 'CustomerInfo',
'PaymentInfo', 'ReviewInfo'
```

```
hbase:001:0> create 'OJD', 'OrderInfo', 'ItemInfo', 'ProductInfo', 'CustomerInfo', 'PaymentInfo', 'ReviewInfo'
Created table OJD
Took 0.9616 seconds
=> Hbase::Table - OJD
```

3. Import TSV-formatted data into HBase table.

```
./hbase/bin/hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -
Dimporttsv.separator=, -
Dimporttsv.columns=HBASE_ROW_KEY,OrderInfo:order_id,OrderInfo:customer_id,O
rderInfo:order_status,OrderInfo:order_purchase_timestamp,OrderInfo:order_ap
proved_at,OrderInfo:order_delivered_carrier_date,OrderInfo:order_delivered_
customer_date,OrderInfo:order_estimated_delivery_date,OrderInfo:order_item_
id,ProductInfo:product_id,ProductInfo:seller_id,ProductInfo:shipping_limit_
date,ProductInfo:price,ProductInfo:freight_value,ProductInfo:product_catego
ry_name,ProductInfo:product_name_length,ProductInfo:product_description_len
gth,ProductInfo:product_photos_qty,ProductInfo:product_weight_g,ProductInfo
:product_length_cm,ProductInfo:product_height_cm,ProductInfo:product_width_
cm,PaymentInfo:payment_sequential,PaymentInfo:payment_type,PaymentInfo:paym
ent_installments,PaymentInfo:payment_value,ReviewInfo:review_id,ReviewInfo:
review_score,ReviewInfo:review_comment_title,ReviewInfo:review_comment_mess
age,ReviewInfo:review_creation_date,ReviewInfo:review_answer_timestamp,Cust
omerInfo:customer_unique_id,CustomerInfo:customer_zip_code_prefix,CustomerI
nfo:customer_city,CustomerInfo:customer_state 'OJD'
hdfs://localhost:9000/user/hdfs/inputfolder/cleaned6.csv
```

```
bklow@LAPTOP-M13O57I5:~$ ./hbase/bin/hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=, -Dimporttsv.columns=HBASE_ROW_KEY,OrderInfo:order_id,OrderInfo:customer_id,OrderInfo:order_status,OrderInfo:order_purchase_timestamp,OrderInfo:order_approved_at,OrderInfo:order_delivered_carrier_date,OrderInfo:order_delivered_customer_date,OrderInfo:order_estimated_delivery_date,OrderInfo:order_item_id,ProductInfo:product_id,ProductInfo:seller_id,ProductInfo:shipping_limit_date,ProductInfo:price,ProductInfo:freight_value,ProductInfo:product_category_name,ProductInfo:product_name_length,ProductInfo:product_description_length,ProductInfo:product_photos_qty,ProductInfo:product_weight_g,ProductInfo:product_length_cm,ProductInfo:product_height_cm,ProductInfo:product_width_cm,PaymentInfo:payment_sequential,PaymentInfo:payment_type,PaymentInfo:payment_installments,PaymentInfo:payment_value,ReviewInfo:review_id,ReviewInfo:review_score,ReviewInfo:review_comment_title,ReviewInfo:review_comment_message,ReviewInfo:review_creation_date,ReviewInfo:review_answer_timestamp,CustomerInfo:customer_unique_id,CustomerInfo:customer_zip_code_prefix,CustomerInfo:customer_city,CustomerInfo:customer_state 'OJD' hdfs://localhost:9000/user/hdfs/inputfolder/cleaned6.csv
2024-01-20T14:04:23,326 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-01-20T14:04:23,482 INFO  [main] Configuration.deprecation: hbase.client.pause.cqtbe is deprecated. Instead, use hbase.client.pause.server.overloaded
2024-01-20T14:04:23,535 INFO  [ReadOnlyZKClient-localhost:2181@0x27508c5d] zookeeper.ZooKeeper: Client environment:zookeeper.version=3.8.3-6ad6d364c7c0bcf0de452d54ebefa3058098ab56, built on 2023-10-05 10:34 UTC
2024-01-20T14:04:23,535 INFO  [ReadOnlyZKClient-localhost:2181@0x27508c5d] zookeeper.ZooKeeper: Client environment:host.name=LAPTOP-M13O57I5.
2024-01-20T14:04:23,535 INFO  [ReadOnlyZKClient-localhost:2181@0x27508c5d] zookeeper.ZooKeeper: Client environment:java.version=1.8.0_392
2024-01-20T14:04:23,535 INFO  [ReadOnlyZKClient-localhost:2181@0x27508c5d] zookeeper.ZooKeeper: Client environment:java.vendor=Private Build
2024-01-20T14:04:23,535 INFO  [ReadOnlyZKClient-localhost:2181@0x27508c5d] zookeeper.ZooKeeper: Client environment:java.home=/usr/lib/jvm/java-8-openjdk-amd64/jre
2024-01-20T14:04:23,535 INFO  [ReadOnlyZKClient-localhost:2181@0x27508c5d] zookeeper.ZooKeeper: nce-annotations-0.13.0.jar:/home/bklow/hbase/bin/../lib/client-facing-thirdparty/commons-logging-1.2.jar:/home/bklow/hbase/bin/../lib/client-facing-thirdparty/htrace-core4-4.1.0-incubating.jar:/home/bklow/hbase/bin/../lib/client-facing-thirdparty/jcl-over-slf4j-1.7.33.jar:/home/bklow/hbase/bin/../lib/client-facing-thirdparty/jul-to-slf4j-1.7.33.jar:/home/bklow/hbase/bin/../lib/client-facing-thirdparty/opentelemetry-api-1.15.0.jar:/h
```

```
2024-01-20T14:04:37,162 INFO  [main] mapreduce.Job: Job job_local21319386_0001 completed successfully
2024-01-20T14:04:37,172 INFO  [main] mapreduce.Job: Counters: 24
        File System Counters
                FILE: Number of bytes read=437348
                FILE: Number of bytes written=1024097
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=54318280
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=3
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Map-Reduce Framework
                Map input records=118310
                Map output records=118310
                Input split bytes=121
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=170
                CPU time spent (ms)=5630
                Physical memory (bytes) snapshot=281882624
                Virtual memory (bytes) snapshot=3848052736
                Total committed heap usage (bytes)=125698048
        ImportTsv
                Bad Lines=0
        File Input Format Counters
                Bytes Read=54318280
        File Output Format Counters
                Bytes Written=0
```

4. View all the data in the HBase table.

`scan 'OJD'`

```
hbase:006:0> scan 'OJD'
 fffe41c64501cc87c801fd61db3f6244  column=OrderInfo:order_purchase_timestamp, timestamp=2024-01-20T13:01:28.567, value=9/6/2018 17:10
 fffe41c64501cc87c801fd61db3f6244  column=OrderInfo:order_status, timestamp=2024-01-20T13:01:28.567, value=9/6/2018 17:00
 fffe41c64501cc87c801fd61db3f6244  column=PaymentInfo:payment_installments, timestamp=2024-01-20T13:01:28.567, value=55.79
 fffe41c64501cc87c801fd61db3f6244  column=PaymentInfo:payment_sequential, timestamp=2024-01-20T13:01:28.567, value=credit_card
 fffe41c64501cc87c801fd61db3f6244  column=PaymentInfo:payment_type, timestamp=2024-01-20T13:01:28.567, value=1
 fffe41c64501cc87c801fd61db3f6244  column=PaymentInfo:payment_value, timestamp=2024-01-20T13:01:28.567, value=b2700869a37f1aafc9dda829dc2f9027
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:freight_value, timestamp=2024-01-20T13:01:28.567, value=cama_mesa_banho
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:price, timestamp=2024-01-20T13:01:28.567, value=12.79
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_category_name, timestamp=2024-01-20T13:01:28.567, value=47
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_description_length, timestamp=2024-01-20T13:01:28.567, value=1
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_height_cm, timestamp=2024-01-20T13:01:28.567, value=19
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_id, timestamp=2024-01-20T13:01:28.567, value=f7ccf836d21b2fb1de37564105216cc1
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_length_cm, timestamp=2024-01-20T13:01:28.567, value=3
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_name_length, timestamp=2024-01-20T13:01:28.567, value=511
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_photos_qty, timestamp=2024-01-20T13:01:28.567, value=600
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_weight_g, timestamp=2024-01-20T13:01:28.567, value=30
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:product_width_cm, timestamp=2024-01-20T13:01:28.567, value=1
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:seller_id, timestamp=2024-01-20T13:01:28.567, value=12/6/2018 17:10
 fffe41c64501cc87c801fd61db3f6244  column=ProductInfo:shipping_limit_date, timestamp=2024-01-20T13:01:28.567, value=43
 fffe41c64501cc87c801fd61db3f6244  column=ReviewInfo:review_answer_timestamp, timestamp=2024-01-20T13:01:28.567, value=cd76a00d8e3ca5e6ab9ed9ecb6667ac4
 fffe41c64501cc87c801fd61db3f6244  column=ReviewInfo:review_comment_message, timestamp=2024-01-20T13:01:28.567, value=15/6/2018 0:00
 fffe41c64501cc87c801fd61db3f6244  column=ReviewInfo:review_comment_title, timestamp=2024-01-20T13:01:28.567, value=
 fffe41c64501cc87c801fd61db3f6244  column=ReviewInfo:review_creation_date, timestamp=2024-01-20T13:01:28.567, value=17/6/2018 21:27
 fffe41c64501cc87c801fd61db3f6244  column=ReviewInfo:review_id, timestamp=2024-01-20T13:01:28.567, value=5
 fffe41c64501cc87c801fd61db3f6244  column=ReviewInfo:review_score, timestamp=2024-01-20T13:01:28.567, value=
98666 row(s)
Took 294.5066 seconds
```

5. View the output of HBase table structure.

```
describe 'OJD'
```

```
hbase:004:0> describe 'OJD'
Table OJD is ENABLED
OJD, {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}
COLUMN FAMILIES DESCRIPTION
{NAME => 'CustomerInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE
', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE'
, BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

{NAME => 'ItemInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', T
TL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BL
OCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

{NAME => 'OrderInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE',
TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', B
LOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

{NAME => 'PaymentInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE'
, TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE',
 BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

{NAME => 'ProductInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE'
, TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE',
 BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

{NAME => 'ReviewInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE',
 TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE',
BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

6 row(s)
Quota is disabled
Took 0.1184 seconds
```

6. View the first 10 rows of data in the HBase table.

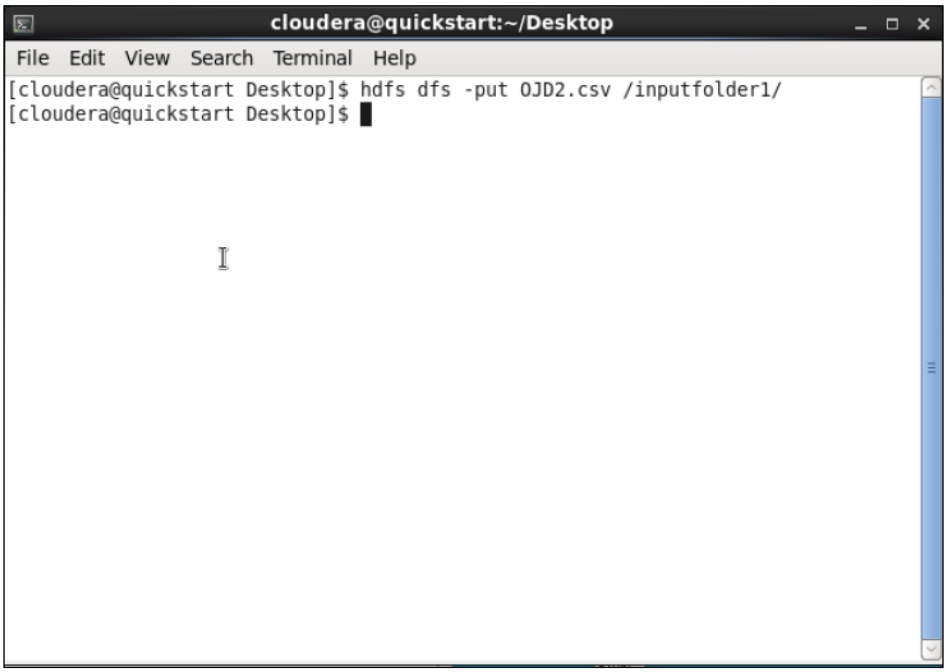
```
Scan 'OJD', {LIMIT=>10}
```

```
hbase:005:0> scan 'OJD', {LIMIT=>10}
ROW                                   COLUMN+CELL
 00010242fe8c5a6d1ba2dd792cb16214     column=CustomerInfo:customer_city, timestamp=2024-01-20T13:01:28.567, value=RJ
 00010242fe8c5a6d1ba2dd792cb16214     column=CustomerInfo:customer_unique_id, timestamp=2024-01-20T13:01:28.567, value=28013
 00010242fe8c5a6d1ba2dd792cb16214     column=CustomerInfo:customer_zip_code_prefix, timestamp=2024-01-20T13:01:28.567, value=campos dos
                                      goytacazes
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:customer_id, timestamp=2024-01-20T13:01:28.567, value=delivered
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:order_approved_at, timestamp=2024-01-20T13:01:28.567, value=19/9/2017 18:34
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:order_delivered_carrier_date, timestamp=2024-01-20T13:01:28.567, value=20/9/2017
                                      23:43
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:order_delivered_customer_date, timestamp=2024-01-20T13:01:28.567, value=29/9/2017
                                       0:00
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:order_estimated_delivery_date, timestamp=2024-01-20T13:01:28.567, value=1
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:order_id, timestamp=2024-01-20T13:01:28.567, value=3ce436f183e68e07877b285a838db1
                                      1a
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:order_item_id, timestamp=2024-01-20T13:01:28.567, value=4244733e06e7ecb4970a6e268
                                      3c13e61
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:order_purchase_timestamp, timestamp=2024-01-20T13:01:28.567, value=13/9/2017 9:45
 00010242fe8c5a6d1ba2dd792cb16214     column=OrderInfo:order_status, timestamp=2024-01-20T13:01:28.567, value=13/9/2017 8:59
 00010242fe8c5a6d1ba2dd792cb16214     column=PaymentInfo:payment_installments, timestamp=2024-01-20T13:01:28.567, value=72.19
 00010242fe8c5a6d1ba2dd792cb16214     column=PaymentInfo:payment_sequential, timestamp=2024-01-20T13:01:28.567, value=credit_card
 00010242fe8c5a6d1ba2dd792cb16214     column=PaymentInfo:payment_type, timestamp=2024-01-20T13:01:28.567, value=2
 00010242fe8c5a6d1ba2dd792cb16214     column=PaymentInfo:payment_value, timestamp=2024-01-20T13:01:28.567, value=97ca439bc427b48bc1cd717
                                      7abe71365
 00010242fe8c5a6d1ba2dd792cb16214     column=ProductInfo:freight_value, timestamp=2024-01-20T13:01:28.567, value=cool_stuff
 00010242fe8c5a6d1ba2dd792cb16214     column=ProductInfo:price, timestamp=2024-01-20T13:01:28.567, value=13.29
 00010242fe8c5a6d1ba2dd792cb16214     column=ProductInfo:product_category_name, timestamp=2024-01-20T13:01:28.567, value=58
 00010242fe8c5a6d1ba2dd792cb16214     column=ProductInfo:product_description_length, timestamp=2024-01-20T13:01:28.567, value=4
 00010242fe8c5a6d1ba2dd792cb16214     column=ProductInfo:product_height_cm, timestamp=2024-01-20T13:01:28.567, value=14
 00010242fe8c5a6d1ba2dd792cb16214     column=ProductInfo:product_id, timestamp=2024-01-20T13:01:28.567, value=48436dade18ac8b2bce089ec2a
                                      041202
 00010242fe8c5a6d1ba2dd792cb16214     column=ProductInfo:product_length_cm, timestamp=2024-01-20T13:01:28.567, value=9
 0005f50442cb953dcd1d21e1fb923495     column=PaymentInfo:payment_type, timestamp=2024-01-20T13:01:28.567, value=1
 0005f50442cb953dcd1d21e1fb923495     column=PaymentInfo:payment_value, timestamp=2024-01-20T13:01:28.567, value=5c0b7e34ed85ec659bb0649
                                      02d878e7a
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:freight_value, timestamp=2024-01-20T13:01:28.567, value=livros_tecnicos
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:price, timestamp=2024-01-20T13:01:28.567, value=11.4
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_category_name, timestamp=2024-01-20T13:01:28.567, value=52
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_description_length, timestamp=2024-01-20T13:01:28.567, value=1
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_height_cm, timestamp=2024-01-20T13:01:28.567, value=21
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_id, timestamp=2024-01-20T13:01:28.567, value=ba143b05f0110f0dc71ad71b44
                                      66ce92
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_length_cm, timestamp=2024-01-20T13:01:28.567, value=3
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_name_length, timestamp=2024-01-20T13:01:28.567, value=1192
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_photos_qty, timestamp=2024-01-20T13:01:28.567, value=850
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_weight_g, timestamp=2024-01-20T13:01:28.567, value=29
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:product_width_cm, timestamp=2024-01-20T13:01:28.567, value=1
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:seller_id, timestamp=2024-01-20T13:01:28.567, value=6/7/2018 14:10
 0005f50442cb953dcd1d21e1fb923495     column=ProductInfo:shipping_limit_date, timestamp=2024-01-20T13:01:28.567, value=53.99
 0005f50442cb953dcd1d21e1fb923495     column=ReviewInfo:review_answer_timestamp, timestamp=2024-01-20T13:01:28.567, value=0782c41380992a
                                      5a533489063df0eef6
 0005f50442cb953dcd1d21e1fb923495     column=ReviewInfo:review_comment_message, timestamp=2024-01-20T13:01:28.567, value=5/7/2018 0:00
 0005f50442cb953dcd1d21e1fb923495     column=ReviewInfo:review_comment_title, timestamp=2024-01-20T13:01:28.567, value=
 0005f50442cb953dcd1d21e1fb923495     column=ReviewInfo:review_creation_date, timestamp=2024-01-20T13:01:28.567, value=5/7/2018 23:17
 0005f50442cb953dcd1d21e1fb923495     column=ReviewInfo:review_id, timestamp=2024-01-20T13:01:28.567, value=4
 0005f50442cb953dcd1d21e1fb923495     column=ReviewInfo:review_score, timestamp=2024-01-20T13:01:28.567, value=
10 row(s)
Took 0.0479 seconds
```

Running HBase in Machine 2 – Virtual Machine, Cloudera (4GB RAM)

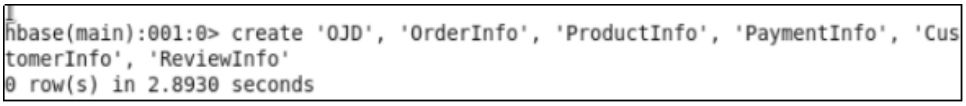
1. Import joined dataset CSV into HDFS.

```
hdfs dfs -put OJD2.csv /inputfolder1/
```



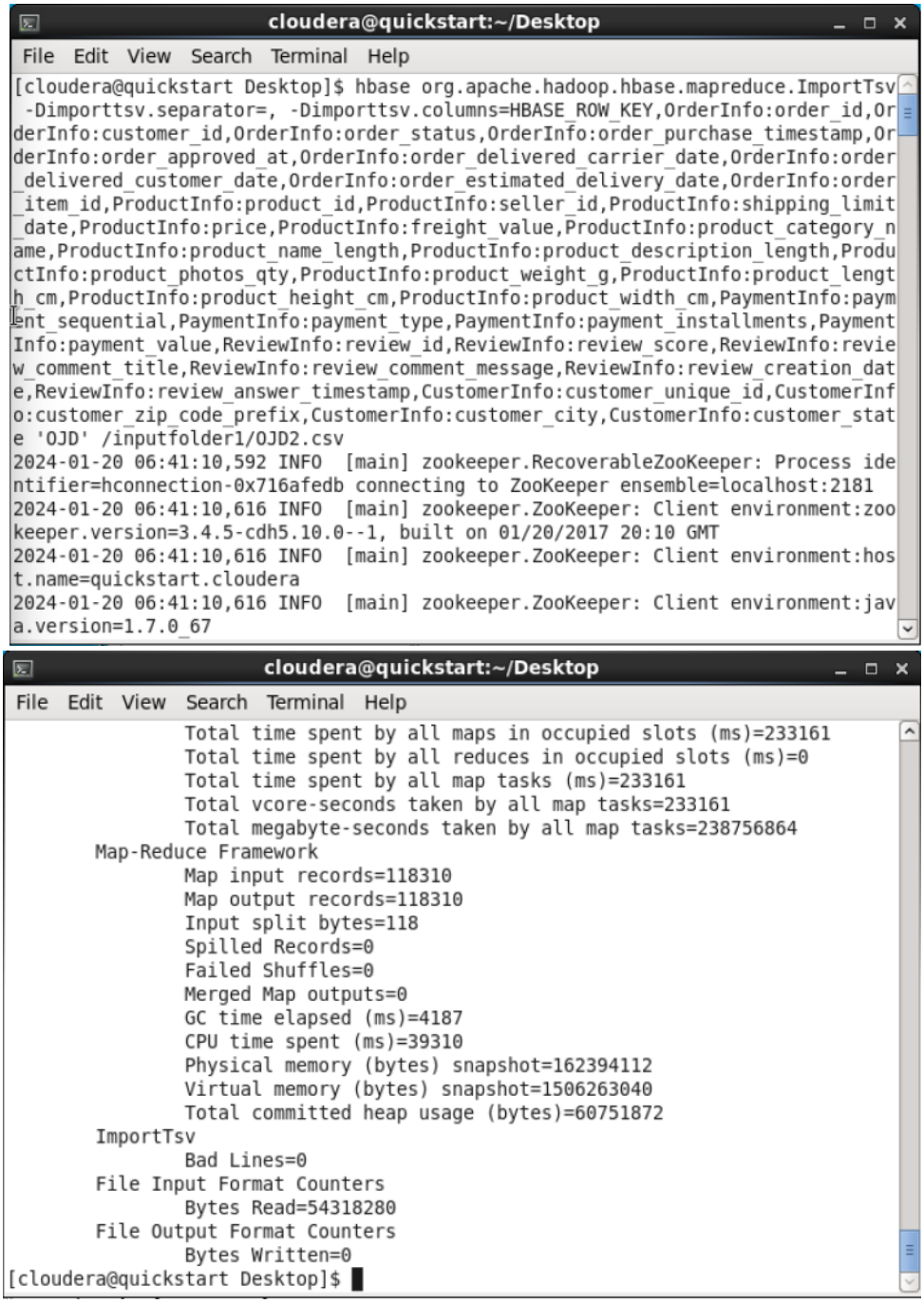
2. Create a table in HBase database.

```
create 'OJD', 'OrderInfo', 'ItemInfo', 'ProductInfo', 'CustomerInfo', 'PaymentInfo', 'ReviewInfo'
```



3. Import TSV-formatted data into HBase table.

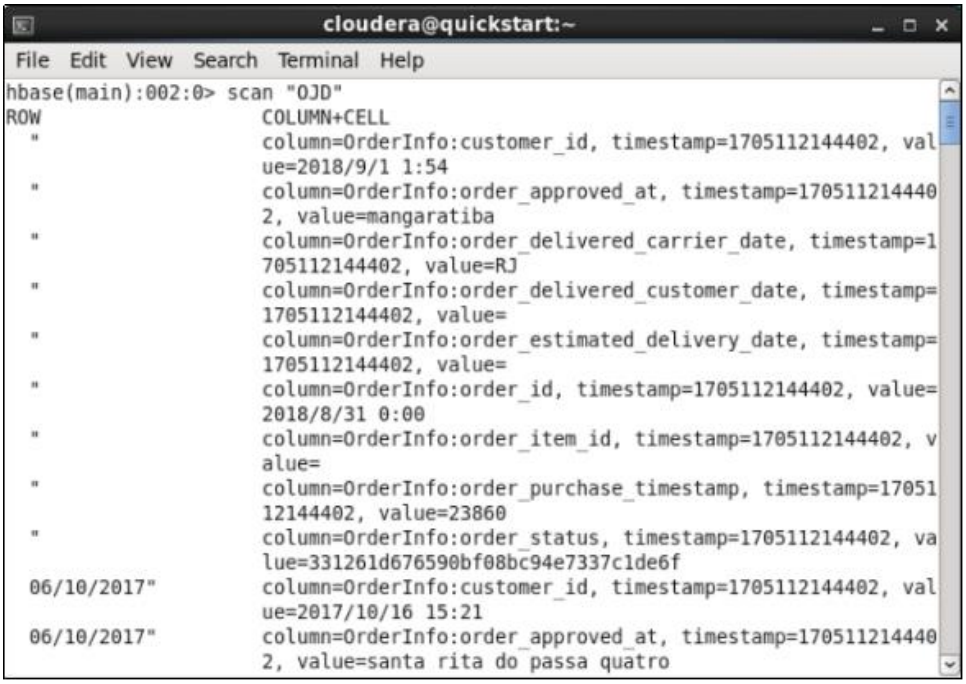
```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator=, -Dimporttsv.columns=HBASE_ROW_KEY,OrderInfo:order_id,OrderInfo:customer_id,OrderInfo:order_status,OrderInfo:order_purchase_timestamp,OrderInfo:order_approved_at,OrderInfo:order_delivered_carrier_date,OrderInfo:order_delivered_customer_date,OrderInfo:order_estimated_delivery_date,OrderInfo:order_item_id,ProductInfo:product_id,ProductInfo:seller_id,ProductInfo:shipping_limit_date,ProductInfo:price,ProductInfo:freight_value,ProductInfo:product_category_name,ProductInfo:product_name_length,ProductInfo:product_description_length,ProductInfo:product_photos_qty,ProductInfo:product_weight_g,ProductInfo:product_length_cm,ProductInfo:product_height_cm,ProductInfo:product_width_cm,PaymentInfo:payment_sequential,PaymentInfo:payment_type,PaymentInfo:payment_installments,PaymentInfo:payment_value,ReviewInfo:review_id,ReviewInfo:review_score,ReviewInfo:review_comment_title,ReviewInfo:review_comment_message,ReviewInfo:review_creation_date,ReviewInfo:review_answer_timestamp,CustomerInfo:customer_unique_id,CustomerInfo:customer_zip_code_prefix,CustomerInfo:customer_city,CustomerInfo:customer_state 'OJD' /inputfolder1/OJD2.csv
```

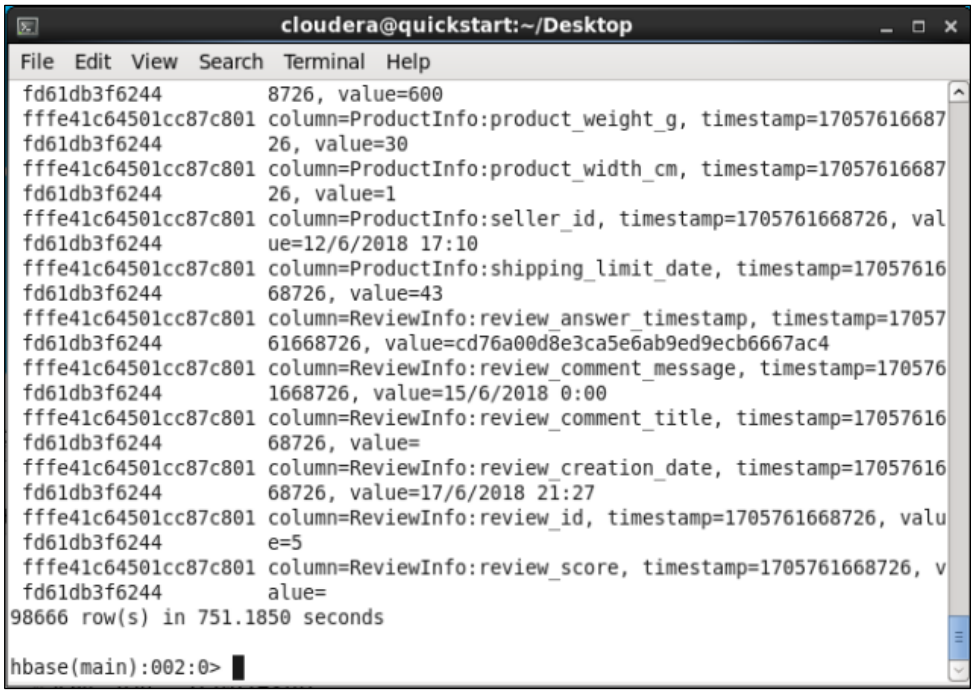

cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ hbase org.apache.hadoop.hbase.mapreduce.ImportTsv
-Dimporttsv.separator=, -Dimporttsv.columns=HBASE_ROW_KEY,OrderInfo:order_id,Or
derInfo:customer_id,OrderInfo:order_status,OrderInfo:order_purchase_timestamp,Or
derInfo:order_approved_at,OrderInfo:order_delivered_carrier_date,OrderInfo:order
_delivered_customer_date,OrderInfo:order_estimated_delivery_date,OrderInfo:order
_item_id,ProductInfo:product_id,ProductInfo:seller_id,ProductInfo:shipping_limit
_date,ProductInfo:price,ProductInfo:freight_value,ProductInfo:product_category_n
ame,ProductInfo:product_name_length,ProductInfo:product_description_length,Produ
ctInfo:product_photos_qty,ProductInfo:product_weight_g,ProductInfo:product_lengt
h_cm,ProductInfo:product_height_cm,ProductInfo:product_width_cm,PaymentInfo:paym
ent_sequential,PaymentInfo:payment_type,PaymentInfo:payment_installments,Payment
Info:payment_value,ReviewInfo:review_id,ReviewInfo:review_score,ReviewInfo:revie
w_comment_title,ReviewInfo:review_comment_message,ReviewInfo:review_creation_dat
e,ReviewInfo:review_answer_timestamp,CustomerInfo:customer_unique_id,CustomerInf
o:customer_zip_code_prefix,CustomerInfo:customer_city,CustomerInfo:customer_stat
e 'OJD' /inputfolder1/OJD2.csv
2024-01-20 06:41:10,592 INFO [main] zookeeper.RecoverableZooKeeper: Process ide
ntifier=hconnection-0x716afedb connecting to ZooKeeper ensemble=localhost:2181
2024-01-20 06:41:10,616 INFO [main] zookeeper.ZooKeeper: Client environment:zoo
keeper.version=3.4.5-cdh5.10.0--1, built on 01/20/2017 20:10 GMT
2024-01-20 06:41:10,616 INFO [main] zookeeper.ZooKeeper: Client environment:hos
t.name=quickstart.cloudera
2024-01-20 06:41:10,616 INFO [main] zookeeper.ZooKeeper: Client environment:jav
a.version=1.7.0_67

cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
Total time spent by all maps in occupied slots (ms)=233161
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=233161
Total vcore-seconds taken by all map tasks=233161
Total megabyte-seconds taken by all map tasks=238756864
Map-Reduce Framework
Map input records=118310
Map output records=118310
Input split bytes=118
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=4187
CPU time spent (ms)=39310
Physical memory (bytes) snapshot=162394112
Virtual memory (bytes) snapshot=1506263040
Total committed heap usage (bytes)=60751872
ImportTsv
Bad Lines=0
File Input Format Counters
Bytes Read=54318280
File Output Format Counters
Bytes Written=0
[cloudera@quickstart Desktop]$

4. View all the data in the HBase table.

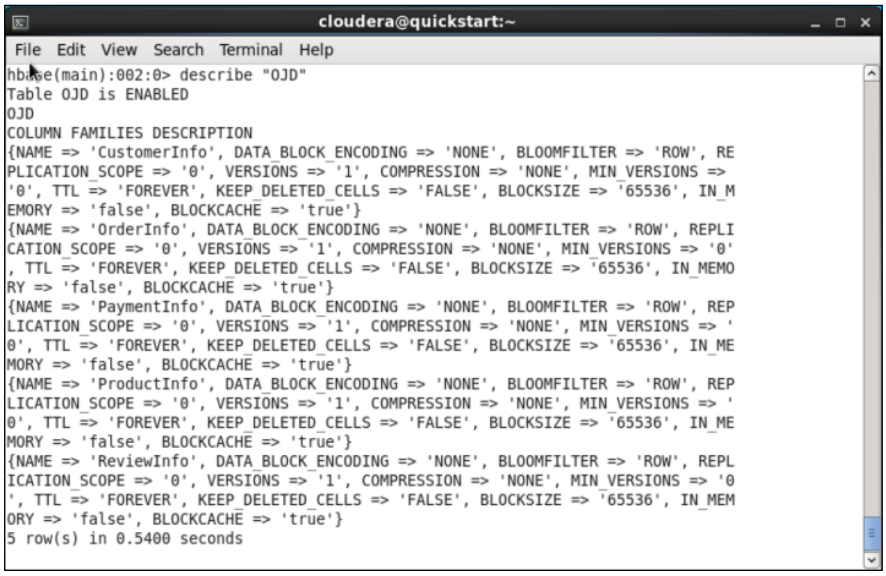
```
Scan "OJD"
```





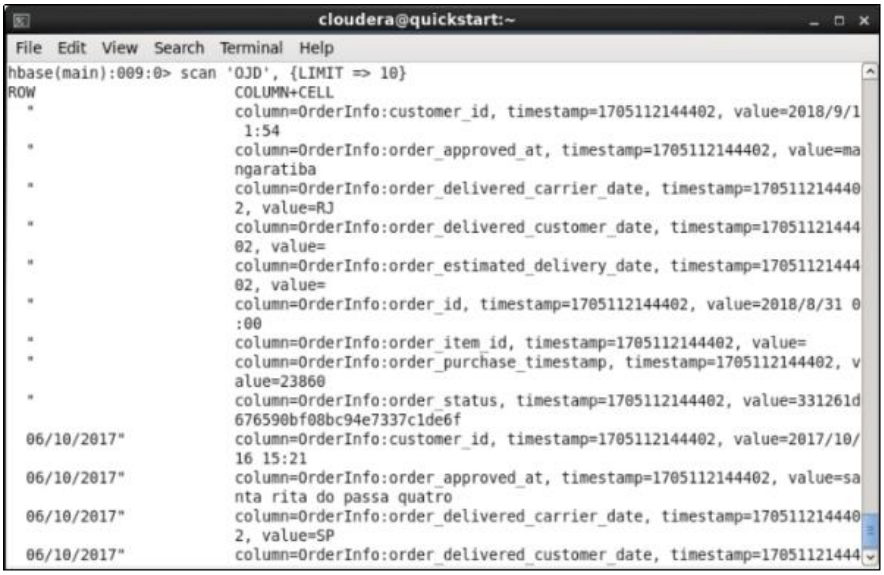
5. View the output of HBase table structure.

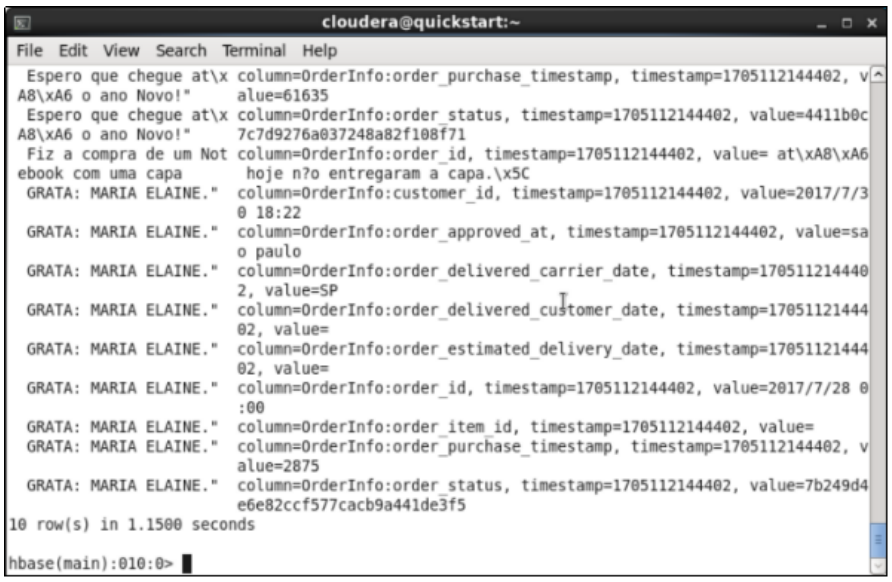
```
Describe "OJD"
```



6. View the first 10 rows of data in the HBase table.

`Scan 'OJD', {LIMIT=>10}`





6.4 Data Access – Hive

1. CSV from local is uploaded to HDFS.
hdfs dfs -put /mnt/hgfs/cleaned8.csv /user/hdfs/data_2.csv

2. Hive is executed using:
Hive

3. Create table.

```
CREATE TABLE IF NOT EXISTS olist_order2 (
   order_id STRING,
   customer_id STRING,
   order_status STRING,
   order_purchase_STRING STRING,
   order_approved_at STRING,
   order_delivered_carrier_date STRING,
   order_delivered_customer_date STRING,
   order_estimated_delivery_date STRING,
   order_item_id INT,
   product_id STRING,
   seller_id STRING,
   shipping_limit_date STRING,
   price FLOAT,
   freight_value FLOAT,
   product_category_name STRING,
   product_name_length FLOAT,
   product_description_length FLOAT,
   product_photos_qty FLOAT,
   product_weight_g FLOAT,
   product_length_cm FLOAT,
   product_height_cm FLOAT,
   product_width_cm FLOAT,
   payment_sequential INT,
   payment_type STRING,
   payment_installments INT,
   payment_value FLOAT,
   review_id STRING,
   review_score INT,
   review_comment_title STRING,
   review_comment_message STRING,
   review_creation_date STRING,
   review_answer_STRING STRING,
   customer_unique_id STRING,
   customer_zip_code_prefix STRING,
   customer_city STRING,
   customer_state STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
TBLPROPERTIES("skip.header.line.count"="1");
```

```
    >     payment_sequential INT,
    >     payment_type STRING,
    >     payment_installments INT,
    >     payment_value FLOAT,
    >     review_id STRING,
    >     review_score INT,
    >     review_comment_title STRING,
    >     review_comment_message STRING,
    >     review_creation_date STRING,
    >     review_answer_STRING STRING,
    >     customer_unique_id STRING,
    >     customer_zip_code_prefix STRING,
    >     customer_city STRING,
    >     customer_state STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > TBLPROPERTIES("skip.header.line.count"="1");
OK
Time taken: 0.073 seconds
```

```
    > TBLPROPERTIES("skip.header.line.count"="1");
OK
Time taken: 0.624 seconds
hive>
```

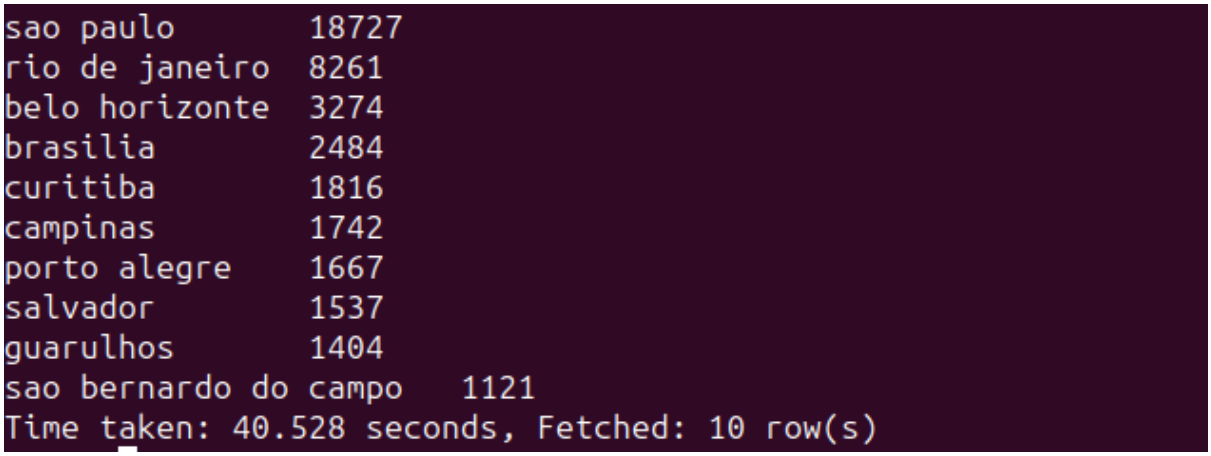
4. CSV is loaded into Hive.
LOAD DATA INPATH '/user/hdfs/data2.csv' INTO TABLE olist_order2;

```
hive> LOAD DATA INPATH '/user/hdfs/data_2.csv' INTO TABLE olist_order2;
Loading data to table default.olist_order2
OK
Time taken: 0.765 seconds
```

```
hive> LOAD DATA INPATH '/user/hdfs/data2.csv' INTO TABLE olist_order2;
Loading data to table default.olist_order2
OK
Time taken: 0.796 seconds
```

5. Which cities are generating the highest orders?
SELECT customer_city, SUM(payment_value) as total_revenue
FROM olist_order2
GROUP BY customer_city
ORDER BY total_revenue DESC
LIMIT 10;

```
sao paulo       18727
rio de janeiro  8261
belo horizonte  3274
brasilia        2484
curitiba        1816
campinas        1742
porto alegre    1667
salvador        1537
guarulhos       1404
sao bernardo do campo   1121
Time taken: 37.329 seconds, Fetched: 10 row(s)
```



6. Which cities are generating the most revenue?

```
SELECT customer_city, SUM(payment_value) as total_revenue
FROM olist_order2
GROUP BY customer_city
ORDER BY total_revenue DESC
LIMIT 10;
```

```
sao paulo       2868846.6671375316
rio de janeiro  1574182.8406030685
belo horizonte  504234.4404745791
brasilia        434082.1994677782
curitiba        331052.0200969875
porto alegre    317384.380638659
salvador        290277.6100103259
campinas        268914.90963315964
goiania 213737.6708741188
guarulhos       205642.69019404054
Time taken: 37.527 seconds, Fetched: 10 row(s)
```

```
sao paulo       2868846.6671375316
rio de janeiro  1574182.8406030685
belo horizonte  504234.4404745791
brasilia        434082.1994677782
curitiba        331052.0200969875
porto alegre    317384.380638659
salvador        290277.6100103259
campinas        268914.90963315964
goiania 213737.6708741188
guarulhos       205642.69019404054
Time taken: 36.615 seconds, Fetched: 10 row(s)
```

7. What are the peak order timings?

```
SELECT
  hour(from_unixtime(unix_timestamp(order_purchase_STRING, 'd/M/yyyy H:mm'))) as order_hour,
  COUNT(*) as total_orders
FROM olist_order2
GROUP BY hour(from_unixtime(unix_timestamp(order_purchase_STRING, 'd/M/yyyy H:mm')));
```

```
0       2917
1       1347
2       611
3       325
4       255
5       225
6       573
7       1425
8       3525
9       5673
10      7377
11      7809
12      7228
13      7735
14      7951
15      7654
16      8022
17      7301
18      6902
19      7014
20      7270
21      7290
22      6976
23      4905
Time taken: 35.118 seconds, Fetched: 24 row(s)
```

```
0       2917
1       1347
2       611
3       325
4       255
5       225
6       573
7       1425
8       3525
9       5673
10      7377
11      7809
12      7228
13      7735
14      7951
15      7654
16      8022
17      7301
18      6902
19      7014
20      7270
21      7290
22      6976
23      4905
Time taken: 16.126 seconds, Fetched: 24 row(s)
```

8. What are the peak order days?

```
SELECT
 date_format(TO_DATE(FROM_UNIXTIME(UNIX_TIMESTAMP(order_purchase_STRING,
'dd/MM/yyyy HH:mm'))), 'EEEE') as order_day_of_week,
  COUNT(*) as total_orders
FROM
  olist_order2
GROUP BY
  date_format(TO_DATE(FROM_UNIXTIME(UNIX_TIMESTAMP(order_purchase_STRING
, 'dd/MM/yyyy HH:mm'))), 'EEEE');
```

```
Friday  16881
Monday  19234
Saturday        12807
Sunday  14022
Thursday        17695
Tuesday 19182
Wednesday       18489
Time taken: 23.336 seconds, Fetched: 7 row(s)
```


```
Friday  16881
Monday  19234
Saturday        12807
Sunday  14022
Thursday        17695
Tuesday 19182
Wednesday       18489
Time taken: 15.998 seconds, Fetched: 7 row(s)
```

9. What are the best performing product categories in terms of number of orders?

```
SELECT product_category_name, COUNT(*) AS order_count
FROM olist_order2
GROUP BY product_category_name
ORDER BY order_count DESC
LIMIT 10;
```

```
bed_bath_table  11988
health_beauty   10032
sports_leisure  9004
furniture_decor 8832
computers_accessories   8150
housewares      7380
watches_gifts   6213
telephony       4726
garden_tools    4590
auto    4400
Time taken: 40.356 seconds, Fetched: 10 row(s)
```

```
bed_bath_table  11988
health_beauty   10032
sports_leisure  9004
furniture_decor 8832
computers_accessories   8150
housewares      7380
watches_gifts   6213
telephony       4726
garden_tools    4590
auto    4400
Time taken: 34.303 seconds, Fetched: 10 row(s)
```

10. What are the worst performing product categories in terms of number of orders?

```
SELECT product_category_name, COUNT(*) AS order_count
FROM olist_order2
GROUP BY product_category_name
ORDER BY order_count ASC
LIMIT 10;
```

```
security_and_services   2
fashion_childrens_clothes       8
cds_dvds_musicals       14
la_cuisine      16
arts_and_craftmanship   24
fashion_sport   31
home_comfort_2  31
flowers 33
diapers_and_hygiene     39
music   40
Time taken: 35.252 seconds, Fetched: 10 row(s)
```

```
security_and_services   2
fashion_childrens_clothes       8
cds_dvds_musicals       14
la_cuisine      16
arts_and_craftmanship   24
fashion_sport   31
home_comfort_2  31
flowers 33
diapers_and_hygiene     39
music   40
Time taken: 35.182 seconds, Fetched: 10 row(s)
```

11. What are the best performing product categories in terms of total revenue generated?

```
SELECT product_category_name, SUM(payment_value) as total_revenue
FROM olist_order2
GROUP BY product_category_name
ORDER BY total_revenue DESC
LIMIT 10;
```

```
bed_bath_table  1743998.8011998404
health_beauty   1662963.589410035
computers_accessories   1599481.0562016796
furniture_decor 1443963.6117437426
watches_gifts   1430553.4785659462
sports_leisure  1400223.0691631027
housewares      1097900.090309674
auto    855095.6811037064
garden_tools    840721.5907554775
cool_stuff      781933.9700460881
Time taken: 41.348 seconds, Fetched: 10 row(s)
```

```
bed_bath_table  1743998.8011998404
health_beauty   1662963.589410035
computers_accessories   1599481.0562016796
furniture_decor 1443963.6117437426
watches_gifts   1430553.4785659462
sports_leisure  1400223.0691631027
housewares      1097900.090309674
auto    855095.6811037064
garden_tools    840721.5907554775
cool_stuff      781933.9700460881
Time taken: 34.844 seconds, Fetched: 10 row(s)
```

12. What are the worst performing product categories in terms of total revenue generated?

```
SELECT product_category_name, SUM(payment_value) as total_revenue
FROM olist_order2
GROUP BY product_category_name
ORDER BY total_revenue ASC
LIMIT 10;
```

```
security_and_services   324.50999450683594
fashion_childrens_clothes       785.6700019836426
cds_dvds_musicals       1199.4300155639648
home_comfort_2  1710.540005683899
flowers 2213.009984970093
arts_and_craftmanship   2326.1699829101562
la_cuisine      2913.5300216674805
fashion_sport   3685.0099754333496
diapers_and_hygiene     4221.249980926514
fashio_female_clothing  5220.069961547852
Time taken: 40.834 seconds, Fetched: 10 row(s)
```


```
security_and_services   324.50999450683594
fashion_childrens_clothes       785.6700019836426
cds_dvds_musicals       1199.4300155639648
home_comfort_2  1710.540005683899
flowers 2213.009984970093
arts_and_craftmanship   2326.1699829101562
la_cuisine      2913.5300216674805
fashion_sport   3685.0099754333496
diapers_and_hygiene     4221.249980926514
fashio_female_clothing  5220.069961547852
Time taken: 35.529 seconds, Fetched: 10 row(s)
```


13. What are the best performing product categories in terms of review scores?

```
SELECT product_category_name, SUM(payment_value) as total_revenue
FROM olist_order2
GROUP BY product_category_name
ORDER BY total_revenue DESC
LIMIT 10;
```

```
cds_dvds_musicals       4.642857142857143
fashion_childrens_clothes       4.5
books_general_interest  4.438502673796791
flowers 4.419354838709677
books_imported  4.419354838709677
costruction_tools_tools 4.415841584158416
books_technical 4.37546468401487
food_drink      4.324137931034483
small_appliances_home_oven_and_coffee   4.32051282051282
luggage_accessories     4.295944779982744
Time taken: 39.625 seconds, Fetched: 10 row(s)
```

```
cds_dvds_musicals       4.642857142857143
fashion_childrens_clothes       4.5
books_general_interest  4.438502673796791
flowers 4.419354838709677
books_imported  4.419354838709677
costruction_tools_tools 4.415841584158416
books_technical 4.37546468401487
food_drink      4.324137931034483
small_appliances_home_oven_and_coffee   4.32051282051282
luggage_accessories     4.295944779982744
Time taken: 36.935 seconds, Fetched: 10 row(s)
```

14. What are the worst performing product categories in terms of review scores?

```
SELECT product_category_name, AVG(review_score) AS average_review_score
FROM olist_order2
GROUP BY product_category_name
ORDER BY average_review_score ASC
LIMIT 10;
```

```
security_and_services   2.5
diapers_and_hygiene     3.2564102564102564
office_furniture        3.526790750141004
fashion_male_clothing   3.548611111111111
home_comfort_2  3.642857142857143
fixed_telephony 3.6728624535315983
fashio_female_clothing  3.78
furniture_mattress_and_upholstery       3.8048780487804876
        3.833720930232558
audio   3.8408488063660475
Time taken: 38.024 seconds, Fetched: 10 row(s)
```


```
security_and_services   2.5
diapers_and_hygiene     3.2564102564102564
office_furniture        3.526790750141004
fashion_male_clothing   3.548611111111111
home_comfort_2  3.642857142857143
fixed_telephony 3.6728624535315983
fashio_female_clothing  3.78
furniture_mattress_and_upholstery       3.8048780487804876
        3.833720930232558
audio   3.8408488063660475
Time taken: 34.889 seconds, Fetched: 10 row(s)
```


15. What are the most popular payment methods?

```
SELECT payment_type, COUNT(*) AS order_count
FROM olist_order2
GROUP BY payment_type
ORDER BY order_count DESC;
```

```
credit_card     87258
boleto  23018
voucher 6332
debit_card      1699
        3
Time taken: 37.27 seconds, Fetched: 5 row(s)
```

```
credit_card     87258
boleto  23018
voucher 6332
debit_card      1699
        3
Time taken: 34.224 seconds, Fetched: 5 row(s)
```

16. How long does it take for the products to be delivered?

```
SELECT
  AVG(DATEDIFF(
    FROM_UNIXTIME(UNIX_TIMESTAMP(order_delivered_customer_date, 'd/M/yyyy H:mm')),
    FROM_UNIXTIME(UNIX_TIMESTAMP(order_purchase_STRING, 'd/M/yyyy H:mm'))
  )) as average_delivery_time
FROM olist_order2;
```

```
12.427299908401169
Time taken: 17.844 seconds, Fetched: 1 row(s)
```

```
OK
12.427299908401169
Time taken: 16.433 seconds, Fetched: 1 row(s)
```

17. Average days between order date and delivered date are calculated for review score:

```
SELECT
  review_score,
  AVG(datediff(
    from_unixtime(unix_timestamp(order_delivered_customer_date, 'd/M/yyyy H:mm')),
    from_unixtime(unix_timestamp(order_purchase_STRING, 'd/M/yyyy H:mm'))
  )) as avg_days_diff
FROM
  olist_order2
GROUP BY
  review_score;
```

```
OK
NULL    18.05574912891986
1       19.498818687600032
2       15.782764811490125
3       13.958860103626943
4       12.179954853273138
5       10.610014990687603
Time taken: 18.837 seconds, Fetched: 6 row(s)
```

```
NULL    18.05574912891986
1       19.498818687600032
2       15.782764811490125
3       13.958860103626943
4       12.179954853273138
5       10.610014990687603
Time taken: 16.799 seconds, Fetched: 6 row(s)
```

6.5 Data Access - Pig

1. CSV from local is updated to HDFS:

hadoop/bin/hdfs dfs -put cleaned_latest.csv /user/hdfs/cleaned10.csv

2. Pig is executed using “Pig”.

3. CSV is loaded into Pig:

data4 = load '/user/hdfs/cleaned10.csv' using PigStorage(',');

4. Column “city” is grouped and the number of occurrences is count:

grunt> city_row = GROUP data4 by $34;
grunt> count = FOREACH city_row GENERATE group AS city, COUNT(data4) AS count;
grunt> city_order = ORDER count by count DESC;
grunt> lmt = LIMIT city_order 10;
grunt> DUMP lmt;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 22:38:00     2024-01-17 22:39:53     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0052  1       1       23      23      23      23      3       3       3       3       city_row,count,
data4   GROUP_BY,COMBINER
job_1705415922749_0053  1       1       2       2       2       2       3       3       3       3       city_order    S
AMPLER
job_1705415922749_0054  1       1       3       3       3       3       3       3       3       3       city_order    O
RDER_BY,COMBINER
job_1705415922749_0055  1       1       2       2       2       2       2       2       2       2       city_order    h
dfs://localhost:9000/tmp/temp-200880814/tmp18525151,

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 20:38:39     2024-01-19 20:40:01        GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0010  1       1       4       4       4 42       2       2       2       city_row,count,data4    GROU
P_BY,COMBINER
job_1704497996174_0011  1       1       1       1       1 11       1       1       1       city_order      SAMPLER
job_1704497996174_0012  1       1       1       1       1 11       1       1       1       city_order      ORDER_BY,COM
BINER
job_1704497996174_0013  1       1       1       1       1 11       1       1       1       city_order              hdfs
://localhost:9000/tmp/temp847527422/tmp740462216,

Input(s):
```

```
(sao paulo,18727)
(rio de janeiro,8261)
(belo horizonte,3274)
(brasilia,2484)
(curitiba,1816)
(campinas,1742)
(porto alegre,1667)
(salvador,1537)
(guarulhos,1404)
(sao bernardo do campo,1121)
```

5. City is grouped and total of revenue is calculated using SUM():

grunt> city_revenue_0 = FOREACH data4 GENERATE $34 AS city, $25 AS payment;
grunt> grouped = GROUP city_revenue_0 by city;
grunt> city_revenue_1 = FOREACH grouped GENERATE group as city2,
SUM(city_revenue_0.payment) as revenue;
grunt> city_revenue = ORDER city_revenue_1 by revenue DESC;
grunt> lmt = LIMIT city_revenue 10;
grunt> DUMP lmt;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 22:43:44     2024-01-17 22:45:21     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0056  1       1       5       5       5       3       3       3       city_revenue_0,
city_revenue_1,data4,grouped    GROUP_BY,COMBINER
job_1705415922749_0057  1       1       3       3       3       3       3       3       3       3       city_revenue  S
AMPLER
job_1705415922749_0058  1       1       3       3       3       3       3       3       3       3       city_revenue  O
RDER_BY,COMBINER
job_1705415922749_0059  1       1       3       3       3       3       3       3       3       3       city_revenue  h
dfs://localhost:9000/tmp/temp-200880814/tmp-1278056006,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 20:42:21     2024-01-19 20:43:36        GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0014  1       1       3       3       3 31       1       1       1       city_revenue_0,city_revenue_
1,data4,grouped         GROUP_BY,COMBINER
job_1704497996174_0015  1       1       1       1       1 11       1       1       1       city_revenue    SAMPLER
job_1704497996174_0016  1       1       1       1       1 11       1       1       1       city_revenue    ORDER_BY,COM
BINER
job_1704497996174_0017  1       1       1       1       1 11       1       1       1       city_revenue            hdfs
://localhost:9000/tmp/temp847527422/tmp-951449670,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(sao paulo,2868846.670000025)
(rio de janeiro,1574182.8399999968)
(belo horizonte,504234.4400000007)
(brasilia,434082.20000000054)
(curitiba,331052.0200000012)
(porto alegre,317384.3800000012)
(salvador,290277.61000000034)
(campinas,268914.9099999998)
(goiania,213737.66999999998)
(guarulhos,205642.68999999983)
```

6. Order Date is changed to Datetime using ToDate() and changed to hours using GetHour():

```
grunt>  order_date_format = FOREACH data4 GENERATE ToDate($3,'d/M/yyyy H:mm') AS date;
grunt> hour = FOREACH order_date_format GENERATE GetHour(date) As hr;
grunt> hour_order = GROUP hour as hr;
grunt> hour_order = FOREACH hour_order GENERATE group as hours, COUNT(hour.hr) as count;
grunt> DUMP hour_order;
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-16 23:28:43     2024-01-16 23:29:14     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0015  1       1       13      13      13      13      3       3       3       3       data4,hour,hour
_order,order_date_format        GROUP_BY,COMBINER       hdfs://localhost:9000/tmp/temp1610563132/tmp571687836,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 20:45:57     2024-01-19 20:46:18        GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0018  1       1       3       3       3 31       1       1       1       data4,hour,hour_order,order_
date_format     GROUP_BY,COMBINER       hdfs://localhost:9000/tmp/temp847527422/tmp349484501,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"

Output(s):
Successfully stored 24 records (214 bytes) in: "hdfs://localhost:9000/tmp/temp847527422/tmp349484501"
```

```
(0,2917)
(1,1347)
(2,611)
(3,325)
(4,255)
(5,225)
(6,573)
(7,1425)
(8,3525)
(9,5673)
(10,7377)
(11,7809)
(12,7228)
(13,7735)
(14,7951)
(15,7654)
(16,8022)
(17,7301)
(18,6902)
(19,7014)
(20,7270)
(21,7290)
(22,6976)
(23,4905)
```

7. Weekdays are obtained using ToString(date, ‘EEE’):

grunt> weekdays = FOREACH order_date_format GENERATE ToString(date, 'EEE') As days;
grunt> day_order = GROUP weekdays by days;
grunt> day_order = FOREACH day_order GENERATE group as days, COUNT(weekdays.days) as count;
grunt> DUMP day_order;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-16 23:35:10     2024-01-16 23:35:51     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0016  1       1       24      24      24      24      2       2       2       2       data4,day_order
,order_date_format,weekdays     GROUP_BY,COMBINER       hdfs://localhost:9000/tmp/temp1610563132/tmp723298405,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 20:47:33     2024-01-19 20:47:50        GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0019  1       1       3       3       3 31       1       1       1       data4,day_order,order_date_f
ormat,weekdays GROUP_BY,COMBINER       hdfs://localhost:9000/tmp/temp847527422/tmp-1938942218,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(Fri,16881)
(Mon,19234)
(Sat,12807)
(Sun,14022)
(Thu,17695)
(Tue,19182)
(Wed,18489)
```

8. Categories are grouped and count and arrange in descending order:

grunt> category_row = GROUP data4 by $14;
grunt> count = FOREACH category_row GENERATE group AS categories, COUNT(data4) AS count;
grunt> category_order = ORDER count by count DESC;
grunt> lmt = LIMIT category_order 10;
grunt> DUMP lmt;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 22:52:00     2024-01-17 22:53:35     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0060  1       1       4       4       4       4       2       2       2       2       category_row,co
unt,data4       GROUP_BY,COMBINER
job_1705415922749_0061  1       1       3       3       3       3       3       3       3       3       category_orderS
AMPLER
job_1705415922749_0062  1       1       3       3       3       3       3       3       3       3       category_orderO
RDER_BY,COMBINER
job_1705415922749_0063  1       1       2       2       2       2       3       3       3       3       category_orderh
dfs://localhost:9000/tmp/temp1557855082/tmp-42855354,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  krs     2024-01-19 20:49:22     2024-01-19 20:50:41     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1704497996174_0020  1       1       2       2       2       2       1       1       1       1       category_row,co
unt,data4       GROUP_BY,COMBINER
job_1704497996174_0021  1       1       1       1       1       1       4       4       4       4       category_orderS
AMPLER
job_1704497996174_0022  1       1       1       1       1       1       1       1       1       1       category_orderO
RDER_BY,COMBINER
job_1704497996174_0023  1       1       1       1       1       1       1       1       1       1       category_orderh
dfs://localhost:9000/tmp/temp847527422/tmp1150362118,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(bed_bath_table,11988)
(health_beauty,10032)
(sports_leisure,9004)
(furniture_decor,8832)
(computers_accessories,8150)
(housewares,7380)
(watches_gifts,6213)
(telephony,4726)
(garden_tools,4590)
(auto,4400)
```

9. Categories are grouped and count and arrange in ascending order:

grunt> category_row = GROUP data4 by $14;
grunt> count = FOREACH category_row GENERATE group AS categories, COUNT(data4) AS count;
grunt> category_order = ORDER count by count ASC;
grunt> lmt = LIMIT category_order 10;
grunt> DUMP lmt;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 22:59:16     2024-01-17 23:00:49     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0064  1       1       4       4       4       4       3       3       3       3       category_row,co
unt,data4       GROUP_BY,COMBINER
job_1705415922749_0065  1       1       2       2       2       2       2       2       2       2       category_orderS
AMPLER
job_1705415922749_0066  1       1       3       3       3       3       3       3       3       3       category_orderO
RDER_BY,COMBINER
job_1705415922749_0067  1       1       3       3       3       3       3       3       3       3       category_orderh
dfs://localhost:9000/tmp/temp1557855082/tmp-1742513872,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 20:51:55     2024-01-19 20:53:10        GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0024  1       1       2       2       2 21       1       1       1       category_row,count,data4  GR
OUP_BY,COMBINER
job_1704497996174_0025  1       1       1       1       1 12       2       2       2       category_order  SAMPLER
job_1704497996174_0026  1       1       1       1       1 11       1       1       1       category_order  ORDER_BY,COM
BINER
job_1704497996174_0027  1       1       1       1       1 11       1       1       1       category_order          hdfs
://localhost:9000/tmp/temp847527422/tmp649811064,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(security_and_services,2)
(fashion_childrens_clothes,8)
(cds_dvds_musicals,14)
(la_cuisine,16)
(arts_and_craftmanship,24)
(fashion_sport,31)
(home_comfort_2,31)
(flowers,33)
(diapers_and_hygiene,39)
(music,40)
```

10. Categories are grouped and revenue is calculated, arrange in descending order:

grunt> category_revenue = FOREACH data4 GENERATE $14 AS categories, $25 AS payment;
grunt> category_revenue_grouped = GROUP category_revenue by categories;
grunt> category_revenue_list = FOREACH category_revenue_grouped GENERATE group as categories, SUM(category_revenue.payment) as revenue;
grunt> category_revenue = ORDER category_revenue_list BY revenue DESC;
grunt> lmt = LIMIT category_revenue 10;
grunt> DUMP lmt;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 23:06:40     2024-01-17 23:08:38     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0068  1       1       25      25      25      25      3       3       3       3       category_revenu
e,category_revenue_grouped,category_revenue_list,data4  GROUP_BY,COMBINER
job_1705415922749_0069  1       1       4       4       4       4       3       3       3       3       category_revenu
e       SAMPLER
job_1705415922749_0070  1       1       2       2       2       2       3       3       3       3       category_revenu
e       ORDER_BY,COMBINER
job_1705415922749_0071  1       1       3       3       3       3       3       3       3       3       category_revenu
e               hdfs://localhost:9000/tmp/temp1557855082/tmp-474731114,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 20:54:52     2024-01-19 20:56:06        GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0028  1       1       2       2       2 22       2       2       2       category_revenue,category_re
venue_grouped,category_revenue_list,data4       GROUP_BY,COMBINER
job_1704497996174_0029  1       1       1       1       1 11       1       1       1       category_revenue        SAMP
LER
job_1704497996174_0030  1       1       1       1       1 11       1       1       1       category_revenue        ORDE
R_BY,COMBINER
job_1704497996174_0031  1       1       1       1       1 11       1       1       1       category_revenue          hd
fs://localhost:9000/tmp/temp847527422/tmp55439086,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(bed_bath_table,1743998.8000000021)
(health_beauty,1662963.5900000155)
(computers_accessories,1599481.0599999991)
(furniture_decor,1443963.6100000043)
(watches_gifts,1430553.4800000046)
(sports_leisure,1400223.069999989)
(housewares,1097900.0899999985)
(auto,855095.6800000023)
(garden_tools,840721.589999999)
(cool_stuff,781933.9700000016)
```

11. Categories are grouped and revenue is calculated, arrange in ascending order:

grunt> category_revenue = FOREACH data4 GENERATE $14 AS categories, $25 AS payment;
grunt> category_revenue_grouped = GROUP category_revenue by categories;
grunt> category_revenue_list = FOREACH category_revenue_grouped GENERATE group as categories, SUM(category_revenue.payment) as revenue;
grunt> category_revenue = ORDER category_revenue_list BY revenue ASC;
grunt> lmt = LIMIT category_revenue 10;
grunt> DUMP lmt;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 23:23:17     2024-01-17 23:24:54     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0072  1       1       11      11      11      11      2       2       2       2       category_revenu
e,category_revenue_grouped,category_revenue_list,data4  GROUP_BY,COMBINER
job_1705415922749_0073  1       1       2       2       2       2       2       2       2       2       category_revenu
e       SAMPLER
job_1705415922749_0074  1       1       2       2       2       2       3       3       3       3       category_revenu
e       ORDER_BY,COMBINER
job_1705415922749_0075  1       1       2       2       2       2       2       2       2       2       category_revenu
e               hdfs://localhost:9000/tmp/temp1557855082/tmp-1832316302,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 20:58:05     2024-01-19 20:59:16        GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0032  1       1       2       2       2 21       1       1       1       category_revenue,category_re
venue_grouped,category_revenue_list,data4       GROUP_BY,COMBINER
job_1704497996174_0033  1       1       1       1       1 11       1       1       1       category_revenue        SAMP
LER
job_1704497996174_0034  1       1       1       1       1 11       1       1       1       category_revenue        ORDE
R_BY,COMBINER
job_1704497996174_0035  1       1       1       1       1 11       1       1       1       category_revenue          hd
fs://localhost:9000/tmp/temp847527422/tmp-1588564054,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(security_and_services,324.51)
(fashion_childrens_clothes,785.6700000000001)
(cds_dvds_musicals,1199.43)
(home_comfort_2,1710.5400000000002)
(flowers,2213.0099999999998)
(arts_and_craftmanship,2326.1700000000005)
(la_cuisine,2913.53)
(fashion_sport,3685.0099999999998)
(diapers_and_hygiene,4221.25)
(fashio_female_clothing,5220.069999999999)
```

12. Categories are grouped and their review score average is calculated, arrange in descending order:

grunt> category_review = FOREACH data4 GENERATE $14 AS categories, $27 AS review;
grunt> category_review_grouped = GROUP category_review BY categories;
grunt> category_review_list = FOREACH category_review_grouped GENERATE group as categories, AVG(category_review.review) as review;
grunt> category_review = ORDER category_review_list BY review DESC;
grunt>lmt = LIMIT category_review 10;
grunt> DUMP lmt;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 23:34:56     2024-01-17 23:36:23     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0082  1       1       4       4       4       4       2       2       2       2       category_review
,category_review_grouped,category_review_list,data4     GROUP_BY,COMBINER
job_1705415922749_0083  1       1       2       2       2       2       3       3       3       3       category_review
SAMPLER
job_1705415922749_0084  1       1       2       2       2       2       2       2       2       2       category_review
ORDER_BY,COMBINER
job_1705415922749_0085  1       1       2       2       2       2       2       2       2       2       category_review
hdfs://localhost:9000/tmp/temp1557855082/tmp258520321,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 21:00:47     2024-01-19 21:02:02        GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0036  1       1       2       2       2 23       3       3       3       category_review,category_rev
iew_grouped,category_review_list,data4 GROUP_BY,COMBINER
job_1704497996174_0037  1       1       1       1       1 12       2       2       2       category_review SAMPLER
job_1704497996174_0038  1       1       1       1       1 11       1       1       1       category_review ORDER_BY,COM
BINER
job_1704497996174_0039  1       1       1       1       1 11       1       1       1       category_review         hdfs
://localhost:9000/tmp/temp847527422/tmp740767054,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(cds_dvds_musicals,4.642857142857143)
(fashion_childrens_clothes,4.5)
(books_general_interest,4.438502673796791)
(books_imported,4.419354838709677)
(flowers,4.419354838709677)
(costruction_tools_tools,4.415841584158416)
(books_technical,4.37546468401487)
(food_drink,4.324137931034483)
(small_appliances_home_oven_and_coffee,4.32051282051282)
(luggage_accessories,4.295944779982744)
```

13. Categories are grouped and their review score average is calculated, arrange in ascending order:

grunt> category_review = FOREACH data4 GENERATE $14 AS categories, $27 AS review;
grunt> category_review_grouped = GROUP category_review BY categories;
grunt> category_review_list = FOREACH category_review_grouped GENERATE group as categories, AVG(category_review.review) as review;
grunt> category_review = ORDER category_review_list BY review DESC;
grunt>lmt = LIMIT category_review 10;
grunt> DUMP lmt;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 23:47:13     2024-01-17 23:49:06     GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0086  1       1       23      23      23      23      2       2       2       2       category_review
,category_review_grouped,category_review_list,data4     GROUP_BY,COMBINER
job_1705415922749_0087  1       1       2       2       2       2       3       3       3       3       category_review
SAMPLER
job_1705415922749_0088  1       1       3       3       3       3       3       3       3       3       category_review
ORDER_BY,COMBINER
job_1705415922749_0089  1       1       2       2       2       2       3       3       3       3       category_review
hdfs://localhost:9000/tmp/temp1557855082/tmp-1450480437,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 21:05:07     2024-01-19 21:06:21        GROUP_BY,ORDER_BY,LIMIT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0044  1       1       2       2       2 21       1       1       1       category_review,category_rev
iew_grouped,category_review_list,data4 GROUP_BY,COMBINER
job_1704497996174_0045  1       1       1       1       1 11       1       1       1       category_review SAMPLER
job_1704497996174_0046  1       1       1       1       1 11       1       1       1       category_review ORDER_BY,COM
BINER
job_1704497996174_0047  1       1       1       1       1 11       1       1       1       category_review         hdfs
://localhost:9000/tmp/temp847527422/tmp-1156489979,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(security_and_services,2.5)
(diapers_and_hygiene,3.2564102564102564)
(office_furniture,3.526790750141004)
(fashion_male_clothing,3.548611111111111)
(home_comfort_2,3.642857142857143)
(fixed_telephony,3.6728624535315983)
(fashio_female_clothing,3.78)
(furniture_mattress_and_upholstery,3.8048780487804876)
```

14. The payment methods are grouped and find their frequency:

grunt> payment_method_row = GROUP data4 by $23;
grunt> count = FOREACH payment_method_row GENERATE group AS payment_method, COUNT(data4) AS count;
grunt> payment_method = ORDER count by count ASC;
grunt> DUMP payment_method;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 00:37:47     2024-01-17 00:39:10     GROUP_BY,ORDER_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0032  1       1       18      18      18      18      3       3       3       3       count,data4,pay
ment_method_row GROUP_BY,COMBINER
job_1705415922749_0033  1       1       2       2       2       2       3       3       3       3       payment_methodS
AMPLER
job_1705415922749_0034  1       1       3       3       3       3       3       3       3       3       payment_methodO
RDER_BY hdfs://localhost:9000/tmp/temp2109082168/tmp2120782852,

Input(s):
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 21:07:25     2024-01-19 21:08:18        GROUP_BY,ORDER_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0048  1       1       2       2       2 21       1       1       1       count,data4,payment_method_r
ow      GROUP_BY,COMBINER
job_1704497996174_0049  1       1       1       1       1 11       1       1       1       payment_method  SAMPLER
job_1704497996174_0050  1       1       1       1       1 11       1       1       1       payment_method  ORDER_BY  hd
fs://localhost:9000/tmp/temp847527422/tmp-1927977317,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(credit_card,87258)
(boleto,23018)
(voucher,6332)
(debit_card,1699)
(,3)
```

15. DaysBetween() is used to find out days between order date and delivered date:

grunt> delivery_days = FOREACH data4 GENERATE DaysBetween(ToDate($6,'d/M/yyyy H:mm'),ToDate($3,'d/M/yyyy H:mm')) as daydiff;
grunt> mean_delivery_grouped = GROUP delivery_days all;
grunt> avg_delivery = FOREACH mean_delivery_grouped GENERATE AVG(delivery_days);
DUMP avg_delivery;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 01:04:54     2024-01-17 01:05:37     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0035  1       1       23      23      23      23      3       3       3       3       avg_delivery,da
ta4,delivery_days,mean_delivery_grouped GROUP_BY,COMBINER       hdfs://localhost:9000/tmp/temp-132876509/tmp-821872792,

Input(s):
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 21:09:26     2024-01-19 21:09:47        GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0051  1       1       3       3       3 31       1       1       1       avg_delivery,data4,delivery_
days,mean_delivery_grouped      GROUP_BY,COMBINER       hdfs://localhost:9000/tmp/temp847527422/tmp1813167489,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
2024-01-17 01:05:37,116 [mai
to process : 1
(12.023003404711291)
grunt> S
```

16. Average days between order date and delivered date are calculated for review score:

grunt> review_delivery= FOREACH data4 GENERATE DaysBetween(ToDate($6,'d/M/yyyy H:mm'),ToDate($3,'d/M/yyyy H:mm')) as daydiff, $27 as review;
grunt> group_corr = GROUP review_delivery by review;
grunt> corr = FOREACH group_corr GENERATE group as score,
AVG(review_delivery.daydiff) as daysdiff;
grunt> DUMP corr;

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
2.10.2  0.16.0  yk      2024-01-17 01:24:36     2024-01-17 01:25:02     GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime A
vgReduceTime    MedianReducetime        Alias   Feature Outputs
job_1705415922749_0038  1       1       5       5       5       5       2       2       2       2       corr,data4,grou
p_corr,review_delivery  GROUP_BY,COMBINER       hdfs://localhost:9000/tmp/temp-132876509/tmp-1193870475,
```

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt Features
2.10.2  0.16.0  krs     2024-01-19 21:18:07     2024-01-19 21:18:23        GROUP_BY

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime MedianMapTime   MaxReduceTime   MinReduceTime   AvgR
educeTime       MedianReducetime        Alias   Feature Outputs
job_1704497996174_0052  1       1       2       2       2 21       1       1       1       corr,data4,group_corr,review
_delivery       GROUP_BY,COMBINER       hdfs://localhost:9000/tmp/temp847527422/tmp1100968551,

Input(s):
Successfully read 118310 records (54105127 bytes) from: "/user/hdfs/data.csv"
```

```
(1,19.100525874552243)
(2,15.382405745062837)
(3,13.553056994818652)
(4,11.778735891647855)
(5,10.203600793446496)
(,17.642276422764226)
```