BDA Mini Project

By:

Akshat Shah 19162121036

Ashil Shah 19162121037

Dishwa Shah 19162121038

AIM

- The aim of this project is to identify the unique sellers, unique customers, their delivery location and their payment options using the dataset containing information about customers and their location.

Creating tables and uploading data into them

```
hive> CREATE TABLE product_category(product_category_name STRING, product_category_name_english STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.074 seconds
hive> DESCRIBE product_category;
OK
product_category_name   string
product_category_name_english   string
Time taken: 0.087 seconds, Fetched: 2 row(s)
hive> LOAD DATA LOCAL INPATH 'product_category_name_transition.csv' OVERWRITE INTO TABLE product_category;
FAILED: SemanticException Line 1:23 Invalid path ''product_category_name_transition.csv'': No files matching path file:/home/clo
udera/Desktop/product_category_name_transition.csv
hive> LOAD DATA LOCAL INPATH 'product_category_name_translation.csv' OVERWRITE INTO TABLE product_category;
Loading data to table project.product_category
Table project.product_category stats: [numFiles=1, numRows=0, totalSize=2613, rawDataSize=0]
OK
Time taken: 0.355 seconds
hive> SELECT * FROM product_category LIMIT 5;
OK
product_category_name   product_category_name_english
beleza_saude    health_beauty
informatica_acessorios  computers_accessories
automotivo      auto
cama_mesa_banho bed_bath_table
Time taken: 0.098 seconds, Fetched: 5 row(s)
hive>
```

Create a table, describe it to verify, load the data into it and print it. We will do this to other tables as well.

Creating tables and uploading data into them

```
hive> CREATE TABLE customers(customer_id STRING, customer_unique_id STRING, customer_zip_code_prefix INT, customer_city STRING,
customer_state STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.603 seconds
hive> DESCRIBE customers;
OK
customer_id             string
customer_unique_id      string
customer_zip_code_prefix        int
customer_city           string
customer_state          string
Time taken: 0.248 seconds, Fetched: 5 row(s)
hive> LOAD DATA LOCAL INPATH 'olist_customers_dataset.csv' OVERWRITE INTO TABLE customers;
Loading data to table project.customers
Table project.customers stats: [numFiles=1, numRows=0, totalSize=9033957, rawDataSize=0]
OK
Time taken: 0.875 seconds
hive> SELECT * FROM customers LIMIT 5;
OK
"customer_id"   "customer_unique_id"    NULL    "customer_city" "customer_state"
"06b8999e2fba1a1fbc88172c00ba8bc7"      "861eff4711a542e4b93843c6dd7febb0"      NULL    franca  SP
"18955e83d337fd6b2def6b18a428ac77"      "290c77bc529b7ac935b93aa66c333dc3"      NULL    sao bernardo do campo   SP
"4e7b3e00288586ebd08712fdd0374a03"      "060e732b5b29e8181a18229c7b0b2b5e"      NULL    sao paulo       SP
b2b6027bc5c5109e529d4dc6358b12c3        "259dac757896d24d7702b9acbbff3f3c"      NULL    mogi das cruzes SP
Time taken: 0.397 seconds, Fetched: 5 row(s)
hive>
```

```
hive> CREATE TABLE geolocation(geolocation_zip_code_prefix INT, geolocation_lat FLOAT, geolocation_lng FLOAT, geolocation_city S
TRING, geolocation_state STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.103 seconds
hive> DESCRIBE geolocation;
OK
geolocation_zip_code_prefix     int
geolocation_lat         float
geolocation_lng         float
geolocation_city        string
geolocation_state       string
Time taken: 0.124 seconds, Fetched: 5 row(s)
hive> LOAD DATA LOCAL INPATH 'olist_geolocation_dataset.csv' OVERWRITE INTO TABLE geolocation;
Loading data to table project.geolocation
Table project.geolocation stats: [numFiles=1, numRows=0, totalSize=61273883, rawDataSize=0]
OK
Time taken: 0.968 seconds
hive> SELECT * FROM geolocation LIMIT 5;
OK
NULL    NULL    NULL    "geolocation_city"      "geolocation_state"
NULL    -23.545622      -46.639294      sao paulo       SP
NULL    -23.546082      -46.64482       sao paulo       SP
NULL    -23.54613       -46.642952      sao paulo       SP
NULL    -23.544392      -46.6395        sao paulo       SP
Time taken: 0.098 seconds, Fetched: 5 row(s)
hive>
```

```
hive> CREATE TABLE order_items(order_id STRING, order_item_id INT, product_id STRING, seller_id STRING, shipping_limit_date TIME
STAMP, price FLOAT, freight_value FLOAT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.11 seconds
hive> DESCRIBE order_items;
OK
order_id                string
order_item_id           int
product_id              string
seller_id               string
shipping_limit_date     timestamp
price                   float
freight_value           float
Time taken: 0.106 seconds, Fetched: 7 row(s)
hive> LOAD DATA LOCAL INPATH 'olist_order_items_dataset.csv' OVERWRITE INTO TABLE order_items;
Loading data to table project.order_items
Table project.order_items stats: [numFiles=1, numRows=0, totalSize=15438671, rawDataSize=0]
OK
Time taken: 0.627 seconds
hive> SELECT * FROM order_items LIMIT 5;
OK
"order_id"      NULL    "product_id"    "seller_id"     NULL    NULL    NULL
"00010242fe8c5a6d1ba2dd792cb16214"      1       "4244733e06e7ecb4970a6e2683c13e61"      "48436dade18ac8b2bce089ec2a041202"      2
017-09-19 09:45:35      58.9    13.29
"00018f77f2f0320c557190d7a144bdd3"      1       e5f2d52b802189ee658865ca93d83a8f        dd7ddc04e1b6c2c614352b383efe2d36        2
017-05-03 11:05:13      239.9   19.93
"000229ec398224ef6ca0657da4fc703e"      1       c777355d18b72b67abbeef9df44fd0fd        "5b51032eddd242adc84c38acab88f23d"      2
018-01-18 14:48:30      199.0   17.87
"00024acbcdf0a6daa1e931b038114c75"      1       "7634da152a4610f1595efa32f14722fc"      "9d7a1d34a5052409006425275ba1c2b4"      2
018-08-15 10:10:18      12.99   12.79
Time taken: 0.086 seconds, Fetched: 5 row(s)
hive>
```

```
hive> CREATE TABLE order_payments(order_id STRING, payment_sequential INT, payment_type STRING, payment_installments INT, paymen
t_value FLOAT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.112 seconds
hive> DESCRIBE order_payments;
OK
order_id                string
payment_sequential      int
payment_type            string
payment_installments    int
payment_value           float
Time taken: 0.112 seconds, Fetched: 5 row(s)
hive> LOAD DATA LOCAL INPATH 'olist_order_payments_dataset.csv' OVERWRITE INTO TABLE order_payments;
Loading data to table project.order_payments
Table project.order_payments stats: [numFiles=1, numRows=0, totalSize=5777138, rawDataSize=0]
OK
Time taken: 0.478 seconds
hive> SELECT * FROM order_payments LIMIT 5;
OK
"order_id"      NULL    "payment_type"  NULL    NULL
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
Time taken: 0.077 seconds, Fetched: 5 row(s)
hive>
```

Creating tables and uploading data into them

```
hive> CREATE TABLE order_reviews(review_id STRING, order_id STRING, review_score INT, review_comment_title STRING, review_commen
t_message STRING, review_creation_date TIMESTAMP, review_answer_timestamp TIMESTAMP)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.106 seconds
hive> DESCRIBE order_reviews;
OK
review_id               string
order_id                string
review_score            int
review_comment_title    string
review_comment_message  string
review_creation_date    timestamp
review_answer_timestamp timestamp
Time taken: 0.123 seconds, Fetched: 7 row(s)
hive> LOAD DATA LOCAL INPATH 'olist_order_reviews_dataset.csv' OVERWRITE INTO TABLE order_reviews;
Loading data to table project.order_reviews
Table project.order_reviews stats: [numFiles=1, numRows=0, totalSize=14409007, rawDataSize=0]
OK
Time taken: 0.453 seconds
hive> SELECT * FROM order_reviews LIMIT 5;
OK
"review_id"     "order_id"      NULL    "review_comment_title"  "review_comment_message"        NULL    NULL
"7bc2406110b926393aa56f80a40eba40"      "73fc7af87114b39712e6da79b0a377eb"      4                       2018-01-18 00:00:00     2
018-01-18 21:46:59
"80e641a11e56f04c1ad469d5645fdfde"      a548910a1c6147796b98fdf73dbeba33        5                       2018-03-10 00:00:00     2
018-03-11 03:05:13
"228ce5500dc1d8e020d8d1322874b6f0"      f9e4b658b201a9f2ecdecbb34bed034b        5                       2018-02-17 00:00:00     2
018-02-18 14:36:24
e64fb393e7b32834bb789ff8bb30750e        "658677c97b385a9be170737859d3511b"      5               Recebi bem antes do prazo estipu
lado.   2017-04-21 00:00:00     2017-04-21 22:02:06
Time taken: 0.078 seconds, Fetched: 5 row(s)
hive>
```

```
hive> CREATE TABLE orders(order_id STRING, customer_id STRING, order_status STRING, order_purchase_timestamp TIMESTAMP, order_ap
proved_at TIMESTAMP, order_delivered_carrier_date TIMESTAMP, order_delivered_customer_date TIMESTAMP, order_estimated_delivery_d
ate TIMESTAMP)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.119 seconds
hive> DESCRIBE orders;
OK
order_id                string
customer_id             string
order_status            string
order_purchase_timestamp        timestamp
order_approved_at       timestamp
order_delivered_carrier_date    timestamp
order_delivered_customer_date   timestamp
order_estimated_delivery_date   timestamp
Time taken: 0.114 seconds, Fetched: 8 row(s)
hive> LOAD DATA LOCAL INPATH 'olist_orders_dataset.csv' OVERWRITE INTO TABLE orders;
Loading data to table project.orders
Table project.orders stats: [numFiles=1, numRows=0, totalSize=17654914, rawDataSize=0]
OK
Time taken: 0.507 seconds
hive> SELECT * FROM orders LIMIT 5;
OK
"order_id"      "customer_id"   "order_status"  NULL    NULL    NULL    NULL    NULL
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
"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
Time taken: 0.128 seconds, Fetched: 5 row(s)
hive>
```

```
hive> CREATE TABLE products(product_id STRING, product_category_name STRING, product_name_length INT, product_description_length
 INT, products_photos_qty INT, product_weight_g INT, product_length_cm INT, product_height_cm INT, product_width_cm INT)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.121 seconds
hive> DESCRIBE products;
OK
product_id              string
product_category_name   string
product_name_length     int
product_description_length      int
products_photos_qty     int
product_weight_g        int
product_length_cm       int
product_height_cm       int
product_width_cm        int
Time taken: 0.115 seconds, Fetched: 9 row(s)
hive> LOAD DATA LOCAL INPATH 'olist_products_dataset.csv' OVERWRITE INTO TABLE products;
Loading data to table project.products
Table project.products stats: [numFiles=1, numRows=0, totalSize=2379446, rawDataSize=0]
OK
Time taken: 0.244 seconds
hive> SELECT * FROM products LIMIT 5;
OK
"product_id"    "product_category_name" NULL    NULL    NULL    NULL    NULL    NULL    NULL
"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
Time taken: 0.057 seconds, Fetched: 5 row(s)
hive>
```

```
hive> CREATE TABLE sellers(seller_id STRING, seller_zip_code_prefix INT, seller_city STRING, seller_state STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.081 seconds
hive> DESCRIBE sellers;
OK
seller_id               string
seller_zip_code_prefix  int
seller_city             string
seller_state            string
Time taken: 0.109 seconds, Fetched: 4 row(s)
hive> LOAD DATA LOCAL INPATH 'olist_sellers_dataset.csv' OVERWRITE INTO TABLE sellers;
Loading data to table project.sellers
Table project.sellers stats: [numFiles=1, numRows=0, totalSize=174703, rawDataSize=0]
OK
Time taken: 0.879 seconds
hive> SELECT * FROM sellers LIMIT 5;
OK
"seller_id"     NULL    "seller_city"   "seller_state"
"3442f8959a84dea7ee197c632cb2df15"      NULL    campinas        SP
d1b65fc7debc3361ea86b5f14c68d2e2        NULL    mogi guacu      SP
ce3ad9de960102d0677a81f5d0bb7b2d        NULL    rio de janeiro  RJ
c0f3eea2e14555b6faeea3dd58c1b1c3        NULL    sao paulo       SP
Time taken: 0.125 seconds, Fetched: 5 row(s)
hive>
```

What percentage of users paid for their order by credit cards?

SELECT COUNT(*) FROM order_payments;

```
hive> SELECT COUNT(*) FROM order_payments;
Query ID = cloudera_20211025032828_ad1c956f-436c-4f15-8551-88c7f349691d
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:28:59,778 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:29:10,959 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.8 sec
2021-10-25 03:29:16,276 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.18 sec
MapReduce Total cumulative CPU time: 8 seconds 180 msec
Ended Job = job_1634116013333_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.18 sec   HDFS Read: 5784427 HDFS Write: 7 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 180 msec
OK
103887
Time taken: 33.166 seconds, Fetched: 1 row(s)
```

SELECT COUNT(*) FROM order_payments WHERE payment_type == 'credit_card';

```
hive> SELECT COUNT(*) FROM order_payments WHERE payment_type == "credit_card";
Query ID = cloudera_20211025033030_17395177-2d44-46b5-9e5b-65c03ffaf043
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:30:10,018 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:30:27,032 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 31.47 sec
2021-10-25 03:30:33,367 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 32.8 sec
MapReduce Total cumulative CPU time: 32 seconds 800 msec
Ended Job = job_1634116013333_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 32.8 sec   HDFS Read: 5785297 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 32 seconds 800 msec
OK
76795
Time taken: 29.284 seconds, Fetched: 1 row(s)
```

SELECT ((76795/103887)*100);

```
hive> SELECT ((76795/103887)*100);
OK
73.92166488588563
Time taken: 0.102 seconds, Fetched: 1 row(s)
```

73.921 % of people paid for their order using a **credit card**.

How many orders did people from Rio de Janeiro place?

SELECT COUNT(*) FROM customers WHERE customer_city == 'rio de janeiro';

```
hive> SELECT COUNT(*) FROM customers WHERE customer_city == "rio de janeiro";
Query ID = cloudera_20211025033535_def90cf7-d166-4f8c-a86c-2655c5e58d38
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:35:23,542 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:35:37,527 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 27.86 sec
2021-10-25 03:35:44,906 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 29.48 sec
MapReduce Total cumulative CPU time: 29 seconds 480 msec
Ended Job = job_1634116013333_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 29.48 sec   HDFS Read: 9042012 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 29 seconds 480 msec
OK
6882
Time taken: 27.134 seconds, Fetched: 1 row(s)
```

People from **Rio de Janeiro** placed **6882** orders.

How many orders were placed by a customer with ID 003822434f91204da0a51fe4cf2aba18?

SELECT COUNT(*) FROM orders WHERE order_id == '"003822434f91204da0a51fe4cf2aba18"';

```
hive> SELECT COUNT(*) FROM orders WHERE order_id == '"003822434f91204da0a51fe4cf2aba18"';
Query ID = cloudera_20211025033636_e860f3dd-64bb-486a-9f6b-e0be13657ee4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:36:49,001 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:36:58,558 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 19.21 sec
2021-10-25 03:37:04,858 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 20.53 sec
MapReduce Total cumulative CPU time: 20 seconds 530 msec
Ended Job = job_1634116013333_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 20.53 sec   HDFS Read: 17663572 HDFS Write: 2 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 530 msec
OK
1
Time taken: 22.125 seconds, Fetched: 1 row(s)
```

SELECT customer_id FROM orders WHERE oredr_id == '"003822434f91204da0a51fe4cf2aba18"';

```
hive> SELECT customer_id FROM orders WHERE order_id == '"003822434f91204da0a51fe4cf2aba18"';
OK
"327679cc34d41d4c48ee5e55246aa6d6"
Time taken: 0.092 seconds, Fetched: 1 row(s)
```

One order **(327679cc34d41d4c48ee5e55246aa6d6)** was placed by customer with an ID 003822434f91204da0a51fe4cf2aba18.

How many orders were placed from people in Piracicaba?

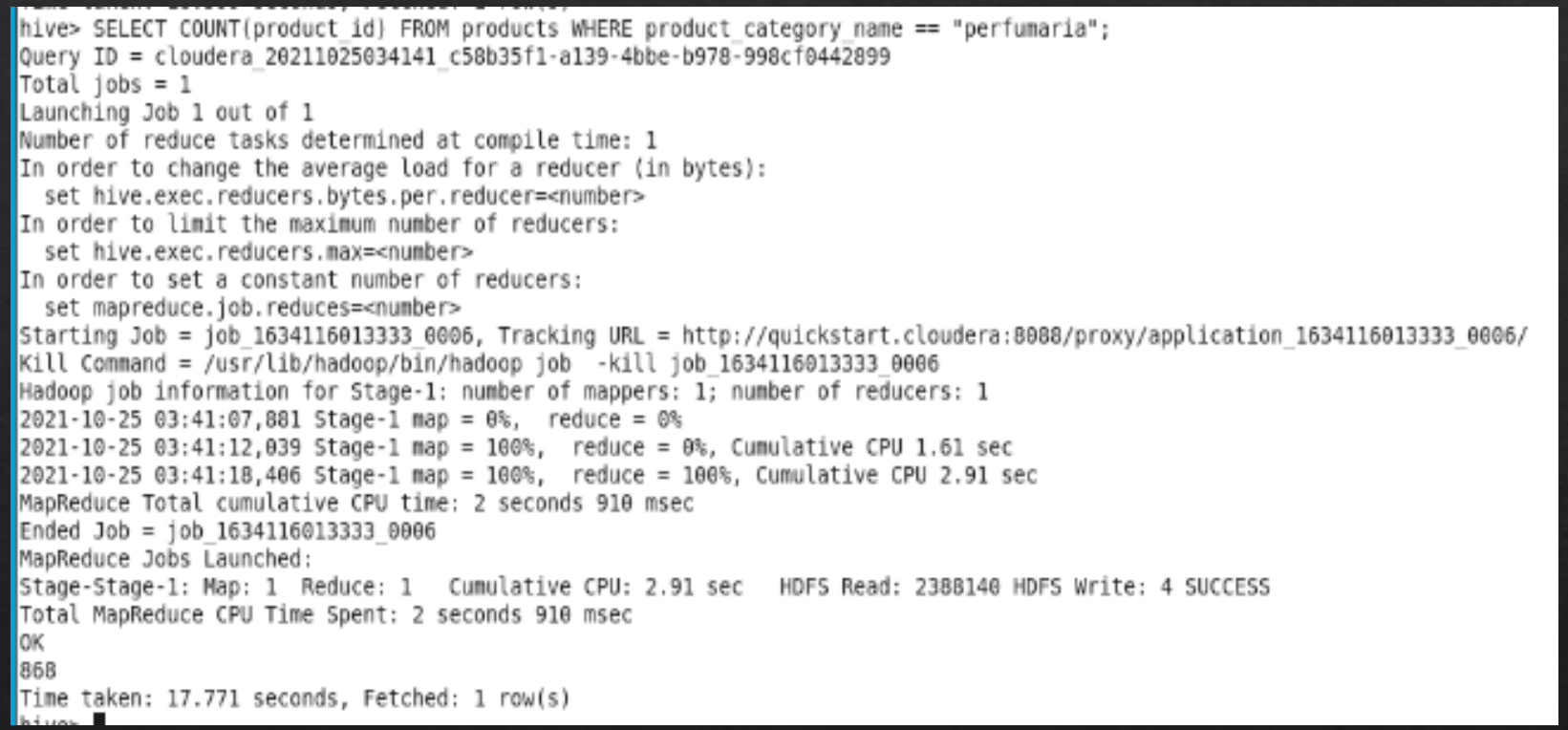
SELECT COUNT(*) FROM customers WHERE customer_city == "piracicaba";

```
hive> SELECT COUNT(*) FROM customers WHERE customer_city == "piracicaba";
Query ID = cloudera_20211025033939_140393fa-32b9-4b6f-a300-fc84197f857b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:39:58,252 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:40:05,774 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.04 sec
2021-10-25 03:40:11,028 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.3 sec
MapReduce Total cumulative CPU time: 8 seconds 300 msec
Ended Job = job_1634116013333_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.3 sec   HDFS Read: 9042087 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 300 msec
OK
369
Time taken: 18.599 seconds, Fetched: 1 row(s)
```

369 orders were placed from the people in **Piracicaba.**

How many orders of the category “perfumaria” were placed?

SELECT COUNT(product_id) FROM products WHERE product_category_name == "perfumaria";



```
hive> SELECT COUNT(product_id) FROM products WHERE product_category_name == "perfumaria";
Query ID = cloudera_20211025034141_c58b35f1-a139-4bbe-b978-998cf0442899
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:41:07,881 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:41:12,039 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.61 sec
2021-10-25 03:41:18,406 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.91 sec
MapReduce Total cumulative CPU time: 2 seconds 910 msec
Ended Job = job_1634116013333_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.91 sec   HDFS Read: 2388140 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 910 msec
OK
868
Time taken: 17.771 seconds, Fetched: 1 row(s)
hive>
```

868 orders were placed in the category of **perfumaria**.

What percentage of sellers are from Curitiba?

SELECT COUNT(*) FROM sellers;

```
hive> SELECT COUNT(*) FROM sellers;
Query ID = cloudera_20211025034242_aea6a223-b36e-4a16-bbb2-9e3309ca7579
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:42:27,578 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:42:31,749 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.94 sec
2021-10-25 03:42:38,147 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.2 sec
MapReduce Total cumulative CPU time: 2 seconds 200 msec
Ended Job = job_1634116013333_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.2 sec   HDFS Read: 181800 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 200 msec
OK
3096
Time taken: 16.393 seconds, Fetched: 1 row(s)
```

SELECT COUNT(*) FROM sellers WHERE seller_city == "curitiba";

```
hive> SELECT COUNT(*) FROM sellers WHERE seller_city == "curitiba";
Query ID = cloudera_20211025034343_a7ee1a4f-fe3c-40c6-aa1b-1b7a01f9ab16
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:43:33,608 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:43:38,859 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.32 sec
2021-10-25 03:43:44,180 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.67 sec
MapReduce Total cumulative CPU time: 2 seconds 670 msec
Ended Job = job_1634116013333_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.67 sec   HDFS Read: 182662 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 670 msec
OK
127
Time taken: 17.428 seconds, Fetched: 1 row(s)
```

SELECT ((127/3096)*100);

```
hive> SELECT ((127/3096)*100);
OK
4.102067183462532
Time taken: 0.11 seconds, Fetched: 1 row(s)
```

4.102% sellers are from **Curitiba**.

How many unique sellers are present on the platform?

SELECT COUNT(DISTINCT(seller_id)) FROM sellers;

```
hive> SELECT COUNT(DISTINCT(seller_id)) FROM sellers;
Query ID = cloudera_20211025034545_8cc5c040-3c63-49b0-8464-822318ec2c9b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:45:15,172 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:45:20,428 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.3 sec
2021-10-25 03:45:25,688 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.86 sec
MapReduce Total cumulative CPU time: 2 seconds 860 msec
Ended Job = job_1634116013333_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.86 sec   HDFS Read: 182196 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 860 msec
OK
3096
Time taken: 16.225 seconds, Fetched: 1 row(s)
```

3096 unique sellers are present on the platform.

How many unique orders were placed in the data provided to you?

SELECT COUNT(DISTINCT(order_id)) FROM orders;

```
hive> SELECT COUNT(DISTINCT(order_id)) FROM orders;
Query ID = cloudera_20211025034646_bfd089bd-cca4-496c-9956-7d43fc40ec75
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:46:17,956 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:46:23,222 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
2021-10-25 03:46:29,514 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.93 sec
MapReduce Total cumulative CPU time: 4 seconds 930 msec
Ended Job = job_1634116013333_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.93 sec   HDFS Read: 17663143 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 930 msec
OK
99442
Time taken: 18.733 seconds, Fetched: 1 row(s)
hive>
```

99442 unique orders were placed.

How many products with the category “moveis_decoracao” have a product height less than 10cm?

SELECT COUNT(*) FROM products WHERE product_category_name == "moveis_decoracao" AND product_height_cm < 10;

```
hive> SELECT COUNT(*) FROM products WHERE product_category_name == "moveis_decoracao" AND product_height_cm < 10;
Query ID = cloudera_20211025034747_da69ae7e-3bc4-4422-9ba8-0ef32485254c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1634116013333_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1634116013333_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1634116013333_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2021-10-25 03:47:24,633 Stage-1 map = 0%,  reduce = 0%
2021-10-25 03:47:38,296 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 28.51 sec
2021-10-25 03:47:43,573 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 29.85 sec
MapReduce Total cumulative CPU time: 29 seconds 850 msec
Ended Job = job_1634116013333_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 29.85 sec   HDFS Read: 2388515 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 29 seconds 850 msec
OK
899
Time taken: 25.232 seconds, Fetched: 1 row(s)
```

899 products with the category “**moveis_decoracao**” have a product **height less than 10cm**.

Conclusion

- In this mini project, we identified the unique sellers, unique customers, their delivery location, their product categories and their payment options using the dataset containing information about sellers, customers and their location.

THANK YOU!