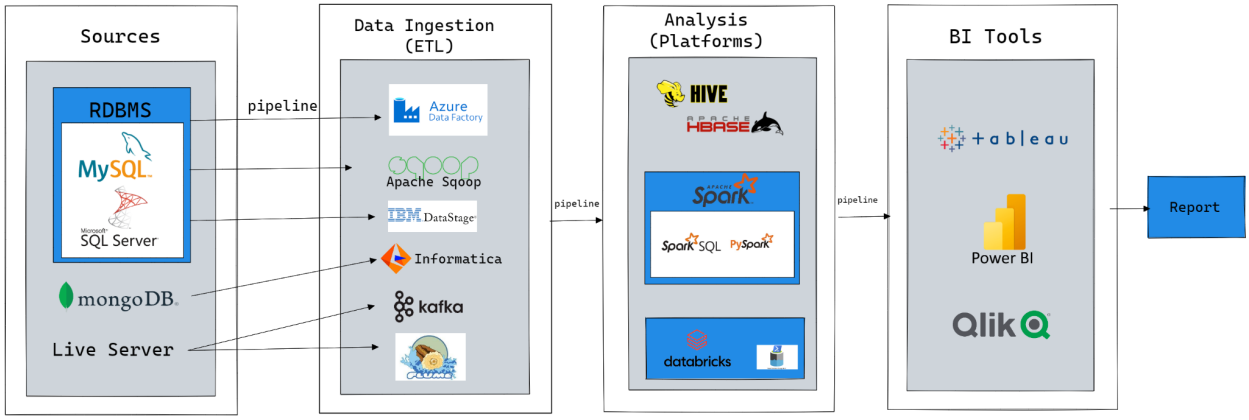
E-Commerce Data Analysis

Faryar Memon - FT641

Big Data Architecture



Sources to Report

Step 01: Data Ingestion

1. **Loading Data from HDFS to Hive**

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir hproject
[cloudera@quickstart ~]$ hdfs dfs -put hproject/ecom.csv hproject/ecom.csv
```

2. **Creating an internal table and loading the data into it**

```
hive> create table ecom(oid string, cid string, quantity int, costprice float,
payment float, timestamp string, rating int, pcat string, pid string, paymenttype
string, ostatus string, pweight_g float, plen_cm float, pheight_cm float, pwidth_cm
float, ccity string, cstate string, sid string, scity string, sstate string,
installments int) row format delimited fields terminated by ','
tblproperties("skip.header.line.count"="1");

hive> load data inpath '/user/cloudera/hproject/ecom.csv' into table ecom;
```

3. **Creating another internal table and storing the processed data**

```
hive> create table cecom(oid string, cid string, quantity int, costprice float,
payment float, timestamp timestamp, rating int, pcat string, pid string,
```

```
paymenttype string, ostatus string, pweight_g float, plen_cm float, pheight_cm
float, pwidth_cm float, ccity string, cstate string, sid string, scity string,
sstate string, installments int) row format delimited fields terminated by ','
tblproperties("skip.header.line.count"="1");

#
hive> insert overwrite table cecom select distinct oid, cid, quantity, costprice,
payment, from_unixtime(unix_timestamp(timestamp, 'dd-MM-yyyy HH:mm')),rating, pcat,
pid, paymenttype, ostatus, pweight_g, plen_cm, pheight_cm, pwidth_cm, ccity,
cstate, sid, scity, sstate, installments from (select *, dense_rank()
over(partition by oid, pid order by quantity desc) as `ranking` from ecom) t where
ranking = 1;
```

Step 02: Data Analysis

01. Customer Segmentation

Question 01: Categorizing customers based on their spendings
4 categories are decided on basis of customer spendings and installments per product

- *Premium Payers*: High Spendings with High Installments
- *Savings Seekers*: High Spendings with Low Installments
- *Budget Balancers*: Low Spendings with High Installments
- *Minimalists*: Low Spendings with Low Installments

```
hive> create external table customer_segmentation(cid string, spendings float,
avg_installments float, cus_category string) row format delimited fields terminated
by ',' location '/user/hive/warehouse/hproject/customer_segmentation';
OK
Time taken: 0.282 seconds

hive>
WITH cte AS (
      SELECT cid, ROUND(SUM(payment),2) AS tpayment,
            ROUND(AVG(installments),2) AS ainstallments,
            avg_spending, avg_installments
      FROM (
            SELECT cid, payment, installments,
                  AVG(payment) OVER () AS avg_spending,
                  AVG(installments) OVER() as `avg_installments`
            FROM cecom) c1
            GROUP BY cid, avg_spending, avg_installments)
INSERT OVERWRITE TABLE customer_segmentation
```

```
SELECT
	cid, tpayment, ainstallments,
	CASE
		WHEN tpayment > avg_spending AND ainstallments >= avg_installments
THEN 'Premium Payers'
		WHEN tpayment > avg_spending AND ainstallments < avg_installments THEN
'Savings Seekers'
		WHEN tpayment < avg_spending AND ainstallments > avg_installments THEN
'Budget Balancers'
		WHEN tpayment < avg_spending AND ainstallments < avg_installments THEN
'Minimalists'
	END AS`customer_cat`
FROM cte;
```

```
hive> select * from customer_segmentation limit 10;
OK
00012a2ce6f8dcda20d059ce98491703 114.74 8.0    Budget Balancers
000161a058600d5901f007fab4c27140 67.41  5.0    Budget Balancers
0001fd6190edaaf884bcaf3d49edf079 195.42 10.0   Premium Payers
0002414f95344307404f0ace7a26f1d5 179.35 1.0    Savings Seekers
000379cdec625522490c315e70c7a9fb 107.01 1.0    Minimalists
0004164d20a9e969af783496f3408652 71.8   1.0    Minimalists
000419c5494106c306a97b5635748086 49.4   4.0    Budget Balancers
00046a560d407e99b969756e0b10f282 166.59 5.0    Premium Payers
00050bf6e01e69d5c0fd612f1bcfb69c 85.23  8.0    Budget Balancers
000598caf2ef4117407665ac33275130 1255.71       10.0   Premium Payers
Time taken: 0.942 seconds, Fetched: 10 row(s)
```

```
mysql> create table customer_segmentation(cid varchar(100), payment float,
avg_installments float, cust_category varchar(70));
Query OK, 0 rows affected (0.01 sec)
```

```
mysql> select * from customer_segmentation limit 10;
+----------------------------------+---------+------------------+------------------+
| cid                              | payment | avg_installments | cust_category    |
+----------------------------------+---------+------------------+------------------+
| bf40545815ca71487bf80ee2843bce7f |   38.48 |                3 | Budget Balancers |
| bf43e6b74f7c46f38f7b4d40e02b4b4a |   32.68 |                1 | Minimalists      |
| bf4492055d844225c39468984e78f450 |   54.13 |                5 | Budget Balancers |
| bf44c65b0e07eaa3659c0c5c9ecbce69 |   49.42 |                2 | Minimalists      |
| bf4539869e1b193726a27a66da9ba8f4 |   53.56 |                1 | Minimalists      |
| bf45a4524fffba803efa07f225cdfda5 |  187.41 |                1 | Savings Seekers  |
| bf46322ae97e484afe52808f7e357ead |  322.96 |               15 | Premium Payers   |
| bf4672f1d0d89b6a477e91f6ee8ff166 |  136.26 |                1 | Minimalists      |
| bf46f5628006409ce5740517ad6506f1 |   62.42 |                1 | Minimalists      |
| bf4786deca0a046ee35f4ce5c1514c89 |   68.23 |                6 | Budget Balancers |
+----------------------------------+---------+------------------+------------------+
10 rows in set (0.00 sec)
```

02. Monthly Trend Forecasting

Question 01: The monthly trend of sales

```
hive> create external table monthly_sales(years int, months int, sales float) row
format delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/monthly_sales';

hive> insert overwrite table monthly_sales select year(timestamp) as years,
month(timestamp) as months, sum(payment) as sales from cecom where ostatus not in
('canceled', 'unavailable') group by year(timestamp), month(timestamp) order by
months, years;

hive> select * from monthly_sales;
OK
2017   1      140503.1
2018   1      1152889.0
2017   2      287885.0
2018   2      1012354.1
2017   3      439264.97
2018   3      1190650.4
2017   4      411414.06
2018   4      1226668.6
2017   5      603772.6
2018   5      1228077.5
2017   6      527792.56
2018   6      1077228.2
2017   7      619411.44
2018   7      1090424.4
2017   8      697514.0
2018   8      1074313.9
2016   9      272.46
2017   9      746808.56
2018   9      166.46
2016   10     55726.05
2017   10     796992.5
2017   11     1248183.6
2016   12     19.62
2017   12     889477.6
Time taken: 0.095 seconds, Fetched: 24 row(s)
```

```
mysql> create table monthly_sales(years int, months int, sales float);
Query OK, 0 rows affected (0.01 sec)
```

```
[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
monthly_sales --export-dir /user/hive/warehouse/hproject/monthly_sales/000000_0
--input-fields-terminated-by ','
```

```
mysql> select * from monthly_sales;
+-------+--------+---------+
| years | months | sales   |
+-------+--------+---------+
|  2017 |      1 |  140503 |
|  2018 |      1 | 1152890 |
|  2017 |      2 |  287885 |
|  2018 |      2 | 1012350 |
|  2017 |      3 |  439265 |
|  2018 |      3 | 1190650 |
|  2017 |      4 |  411414 |
|  2018 |      4 | 1226670 |
|  2017 |      5 |  603773 |
|  2018 |      5 | 1228080 |
|  2017 |      6 |  527793 |
|  2018 |      6 | 1077230 |
|  2017 |      7 |  619411 |
|  2018 |      7 | 1090420 |
|  2017 |      8 |  697514 |
|  2018 |      8 | 1074310 |
|  2016 |      9 |  272.46 |
|  2017 |      9 |  746809 |
|  2018 |      9 |  166.46 |
|  2016 |     10 | 55726.1 |
|  2017 |     10 |  796992 |
|  2017 |     11 | 1248180 |
|  2016 |     12 |   19.62 |
|  2017 |     12 |  889478 |
+-------+--------+---------+
24 rows in set (0.00 sec)
```

Observations: November 2017 contributed the highest to the total sales

03. Hourly Sales Analysis

Question 01: Which hour has more no. of sales

```
hive> create external table hourly_most_sales(hour int, sales float) row format
delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/hourly_most_sales';
OK
```

```
Time taken: 0.082 seconds

hive> insert overwrite table hourly_most_sales select hours, sales from (select
hour(timestamp) as hours, sum(quantity) as sales, rank() over(order by
sum(quantity) desc) AS `max_sales` from cecom group by hour(timestamp)) t where
max_sales = 1;

hive> select * from hourly_most_sales;
OK
16     8223.0
Time taken: 0.056 seconds, Fetched: 1 row(s)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
pcat_unit_sold --export-dir /user/hive/warehouse/hproject/pcat_unit_sold/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from hourly_most_sales;
+-------+-------+
| hours | sales |
+-------+-------+
|    16 |  8223 |
+-------+-------+
1 row in set (0.00 sec)
```

04. Product Based Analysis

Question 01: Which category product has sold more?

```
hive> create table pcat_unit_sold(pcat string, tcount int) row format delimited
fields terminated by ',' location '/user/hive/warehouse/hproject/pcat_unit_sold';
OK
Time taken: 1.003 seconds

hive> insert overwrite table pcat_unit_sold select pcat, sum(quantity) as tcount
from cecom where ostatus in ('delivered', 'shipped') group by pcat order by tcount
desc;
Time taken: 211.288 seconds

hive> select * from pcat_unit_sold;
OK
bed_bath_table      13048
health_beauty10130
furniture_decor     9334
sports_leisure      9020
computers_accessories      8314
```

```
housewares   7528
watches_gifts6357
telephony    4761
garden_tools 4724
auto   4454
toys   4351
cool_stuff   4008
perfumery    3627
baby   3263
electronics  2867
stationery   2662
fashion_bags_accessories  2292
pet_shop     2101
office_furniture    1839
luggage_accessories 1183
consoles_games      1169
construction_tools_construction  1047
home_appliances     791
musical_instruments 690
small_appliances    689
home_construction   669
books_general_interest    579
furniture_living_room     544
food   526
home_confort 516
drinks 394
audio  382
market_place 340
construction_tools_lights  334
fashion_shoes302
air_conditioning    301
food_drink   296
kitchen_dining_laundry_garden_furniture 292
industry_commerce_and_business   281
books_technical     274
costruction_tools_garden  272
fixed_telephony     262
home_appliances_2   257
agro_industry_and_commerce 252
computers    216
art    215
signaling_and_security    207
construction_tools_safety 196
christmas_supplies  155
fashion_male_clothing     152
fashion_underwear_beach   139
furniture_bedroom   120
```

```
costruction_tools_tools    105
tablets_printing_image     89
small_appliances_home_oven_and_coffee   78
cine_photo    73
dvds_blu_ray 68
books_imported      67
fashio_female_clothing     47
party_supplies      44
music  41
furniture_mattress_and_upholstery       41
diapers_and_hygiene 38
home_comfort_2      33
flowers       33
fashion_sport31
arts_and_craftmanship      26
la_cuisine    17
cds_dvds_musicals   14
fashion_childrens_clothes 7
security_and_services      2
Time taken: 0.312 seconds, Fetched: 71 row(s)
```

```
mysql> create table pcat_unit_sold(pcat varchar(50), totalcount int);
Query OK, 0 rows affected (0.01 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
pcat_unit_sold --export-dir /user/hive/warehouse/hproject/pcat_unit_sold/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from pcat_unit_sold limit 5;
+----------------------+------------+
| pcat                 | totalcount |
+----------------------+------------+
| bed_bath_table       |      13048 |
| health_beauty        |      10130 |
| furniture_decor      |       9334 |
| sports_leisure       |       9020 |
| computers_accessories |       8314 |
+----------------------+------------+
5 rows in set (0.00 sec)
```

Question 02: Which category product has more rating?

```
hive> create external table pcat_rating(pcat string, arating float) row format
delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/pcat_rating';
```

```
OK
Time taken: 0.119 seconds

hive> insert overwrite table pcat_rating select pcat, avg(rating) as arating from
cecom where ostatus in ('delivered') group by pcat order by arating desc;

hive> select * from pcat_rating;
OK
fashion_childrens_clothes 5.0
cds_dvds_musicals   4.6666665
books_imported      4.5172415
books_general_interest    4.498099
food_drink    4.470339
small_appliances_home_oven_and_coffee   4.4533334
books_technical     4.391635
fashion_sport4.3846154
costruction_tools_tools    4.3434343
luggage_accessories 4.3427787
food   4.3048244
cine_photo    4.292308
flowers       4.275862
stationery    4.2701917
dvds_blu_ray 4.2622952
pet_shop      4.235974
drinks 4.233108
small_appliances    4.232595
perfumery     4.215871
agro_industry_and_commerce4.2156863
fashion_shoes4.209738
toys   4.2096896
health_beauty4.205295
musical_instruments 4.202194
home_appliances     4.2015505
music  4.2
industry_commerce_and_business   4.19917
sports_leisure      4.198964
cool_stuff    4.1971154
home_appliances_2   4.193548
computers     4.191489
fashion_bags_accessories   4.171875
housewares    4.161332
costruction_tools_garden   4.1549296
garden_tools 4.1530337
la_cuisine    4.133333
consoles_games      4.132964
arts_and_craftmanship      4.125
construction_tools_lights 4.1190476
```

```
auto   4.1147256
party_supplies      4.1025643
construction_tools_construction  4.102402
electronics  4.1013436
kitchen_dining_laundry_garden_furniture4.1
market_place 4.081081
watches_gifts4.0809565
baby   4.076586
tablets_printing_image    4.072289
air_conditioning    4.071713
christmas_supplies  4.0697675
art    4.0682926
furniture_bedroom   4.0612245
computers_accessories     4.045507
signaling_and_security    4.041958
telephony    4.0254745
furniture_decor     4.016332
furniture_living_room     4.0135746
fashion_underwear_beach   3.984127
construction_tools_safety 3.9818182
home_construction   3.9769673
diapers_and_hygiene 3.96
home_confort 3.943439
fixed_telephony     3.935484
bed_bath_table      3.9122791
fashio_female_clothing    3.891892
furniture_mattress_and_upholstery      3.875
audio  3.8264463
home_comfort_2      3.72
fashion_male_clothing     3.6269841
office_furniture    3.6231344
security_and_services     2.5
Time taken: 0.045 seconds, Fetched: 71 row(s)

mysql> create table pcat_rating(pcat varchar(50), avgRating float);
Query OK, 0 rows affected (0.00 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
pcat_rating --export-dir /user/hive/warehouse/hproject/pcat_rating/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from pcat_rating limit 5;
+--------------------------+-----------+
| pcat                     | avgRating |
+--------------------------+-----------+
| furniture_decor          |   4.01633 |
| furniture_living_room    |   4.01357 |
| fashion_underwear_beach  |   3.98413 |
| construction_tools_safety |   3.98182 |
| home_construction        |   3.97697 |
+--------------------------+-----------+
5 rows in set (0.00 sec)
```

Question 03: Which product has sold more?

```
hive> create external table products_sold(pid string, tsold int) row format
delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/products_sold';

hive> insert overwrite table products_sold select pid, sum(quantity) as tquantity
from cecom where ostatus not in ('canceled', 'unavailable') group by pid order by
tquantity desc;

hive> select * from products_sold limit 5;
OK
99a4788cb24856965c36a24e339b6058 561
aca2eb7d00ea1a7b8ebd4e68314663af 536
422879e10f46682990de24d770e7f83d 534
389d119b48cf3043d311335e499d9c6b 434
53759a2ecddad2bb87a079a1f1519f73 418
Time taken: 0.042 seconds, Fetched: 5 row(s)

mysql> create table products_sold(pid varchar(100), totalSold int);
Query OK, 0 rows affected (0.01 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
products_sold --export-dir /user/hive/warehouse/hproject/products_sold/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from products_sold order by totalSold desc limit 5;
+----------------------------------+-----------+
| pid                              | totalSold |
+----------------------------------+-----------+
| 99a4788cb24856965c36a24e339b6058 |       561 |
| aca2eb7d00ea1a7b8ebd4e68314663af |       536 |
| 422879e10f46682990de24d770e7f83d |       534 |
| 389d119b48cf3043d311335e499d9c6b |       434 |
| 53759a2ecddad2bb87a079a1f1519f73 |       418 |
+----------------------------------+-----------+
5 rows in set (0.04 sec)
```

Question 04: Top 10 highest & lowest product rating?

A. Top 10 products with high ratings

```
# using the below query retrieves 13755 rows
hive> create external table high_rating_products(pid string, arating float) row
format delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/high_rating_products';

hive> insert overwrite table high_rating_products select pid, round(arating, 3)
from (select pid, avg(rating) as arating, dense_rank() over(order by avg(rating)
desc) AS `ranking` from cecom group by pid) t where ranking <= 10;

hive> select * from high_rating_products limit 10;
OK
495dd714eb32a76c84c675fc016206f6 5.0
495f9290868e1bde785d3c372608ad54 5.0
eae99ba0d11a9e56ddc878519bdb33b1 5.0
6dc8936547770265403 69b32aebda760 5.0
6dc1e760dde478aea03467d92737a0a6 5.0
49659dedf501090b249d0b09ca7faaf7 5.0
6dbcbca84288705e65660c9b4f369134 5.0
0a9bba9c02d484c391416587002dae47 5.0
6db72cc3d861dbea370e6959aa850b8c 5.0
6db02fce12f88341876dbb10e2c16d8a 5.0
Time taken: 0.041 seconds, Fetched: 10 row(s)

mysql> create table high_rating_products (pid varchar(100), avgRating float);
Query OK, 0 rows affected (0.00 sec)

sqoop export --connect jdbc:mysql://localhost:3306/hproject --username root
--password cloudera --table high_rating_products --export-dir
/user/hive/warehouse/hproject/high_rating_products/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from high_rating_products limit 5;
+----------------------------------+-----------+
| pid                              | avgRating |
+----------------------------------+-----------+
| 37ff4de377ca7a4bc299086c73de26b1 |         5 |
| 922ee69fced8c701d47470e02811d22a |         5 |
| 922ec0e0e2ca6416b4671198410cbf50 |         5 |
| 92298b2c1c8b487f7029ce5ea3b87018 |         5 |
| 37ffcf68086496e428431de29b0b394d |         5 |
+----------------------------------+-----------+
5 rows in set (0.00 sec)
```

B. Top 10 products with low ratings

```
hive> create external table low_rating_products(pid string, arating float) row
format delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/low_rating_products';

hive> insert overwrite table low_rating_products select pid, round(arating, 3) from
(select pid, avg(rating) as arating, dense_rank() over(order by avg(rating)) AS
`ranking` from cecom group by pid) t where ranking <= 10;

hive> select * from low_rating_products limit 10;
OK
b70adcd90b3dc72e1b0243fbffd2b625 1.0
b7082b40a807413582afbfac88facc4b 1.0
9393358ebcb4f182f4e42d0c3b863cea 1.0
b6f334a1ae0731e790e39cab5da09600 1.0
b6eb2752e40bb6f33a3b608e78f5ffec 1.0
b6e406be1aa00db8be5dd3ec6b524d4d 1.0
1583253d78b23f7808214770298e0118 1.0
93a5fdb44a53e09f4a7d659420947201 1.0
93a99b40878c0d888fa4f44459fcac05 1.0
93cbda98bb66aaaae2ab0a7789dfc87b 1.0
Time taken: 0.054 seconds, Fetched: 10 row(s)

mysql> create table low_rating_products like high_rating_products;
Query OK, 0 rows affected (0.00 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
low_rating_products --export-dir
/user/hive/warehouse/hproject/low_rating_products/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from low_rating_products limit 5;
+----------------------------------+-----------+
| pid                              | avgRating |
+----------------------------------+-----------+
| 72f027cff2922e7c85a28976ca21a1bb |         1 |
| c686d4fd1e845efd1b98db8220f0af3b |         1 |
| 72f7a39ebe43db1530241755c8df490e |         1 |
| 735a36bd5c680c3b10f1c48c99c0559b |         1 |
| 1003992d2a8d1e7f870643148854ddc7 |         1 |
+----------------------------------+-----------+
5 rows in set (0.00 sec)
```

Question 04: Order Count for each rating?

```
hive> create external table rating_count(rating int, tcount int) row format
delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/rating_count';
OK
Time taken: 0.084 seconds

hive> insert overwrite table rating_count @;

hive> select * from rating_count;
OK
1       15110
2       4133
3       9842
4       22017
5       65478
Time taken: 0.048 seconds, Fetched: 5 row(s)

mysql> create table rating_count (rating int, totalCount int);
Query OK, 0 rows affected (0.03 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
rating_count --export-dir /user/hive/warehouse/hproject/rating_count/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from rating_count;
+--------+------------+
| rating | totalCount |
+--------+------------+
|      5 |      65478 |
|      1 |      15110 |
|      2 |       4133 |
|      3 |       9842 |
|      4 |      22017 |
+--------+------------+
5 rows in set (0.00 sec)
```

05. Payment Preference

Question 01: What are the most commonly used payment types?

```
hive> create external table common_paymenttype(paymenttype string, tcount int) row
format delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/common_paymenttype';
OK
Time taken: 0.085 seconds

hive> insert overwrite table common_paymenttype select paymenttype, count(oid) as
tcount from cecom group by paymenttype order by tcount desc;

hive> select * from common_paymenttype;
OK
credit_card  78279
boleto 20057
voucher      5197
debit_card   1555
Time taken: 0.044 seconds, Fetched: 4 row(s)

mysql> create table common_paymenttype(paymenttype varchar(50), totalCount int);
Query OK, 0 rows affected (0.00 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
common_paymenttype --export-dir
/user/hive/warehouse/hproject/common_paymenttype/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from common_paymenttype;
+-------------+------------+
| paymenttype | totalCount |
+-------------+------------+
| credit_card |      78279 |
| boleto      |      20057 |
| voucher     |       5197 |
| debit_card  |       1555 |
+-------------+------------+
4 rows in set (0.00 sec)
```

Question 02: Count of Orders With each No. of Payment Installments

```
hive> create external table installment_count(installments int, tcount int) row
format delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/installment_count';
OK
Time taken: 0.067 seconds

hive> insert overwrite table installment_count  select installments, count(oid) as
tcount from cecom group by installments;


hive> select * from installment_count;
OK
NULL   3
0      2
1      52297
2      12481
3      10638
4      7273
5      5384
6      4050
7      1667
8      4512
9      665
10     5740
11     24
12     150
13     17
14     15
15     80
16     6
17     7
```

```
18      31
20      19
21      3
22      1
23      1
24      22
Time taken: 0.057 seconds, Fetched: 25 row(s)

mysql> create table installment_count(installments int, totalCount int);
Query OK, 0 rows affected (0.01 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
installment_count --export-dir
/user/hive/warehouse/hproject/installment_count/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from installment_count;
+--------------+------------+
| installments | totalCount |
+--------------+------------+
|           17 |          7 |
|           18 |         31 |
|           20 |         19 |
|           21 |          3 |
|           22 |          1 |
|           23 |          1 |
|           24 |         22 |
|            0 |          2 |
|            1 |      52297 |
|            2 |      12481 |
|            3 |      10638 |
|            4 |       7273 |
|            5 |       5384 |
|            6 |       4050 |
|            7 |       1667 |
|            8 |       4512 |
|            9 |        665 |
|           10 |       5740 |
|           11 |         24 |
|           12 |        150 |
|           13 |         17 |
|           14 |         15 |
|           15 |         80 |
|           16 |          6 |
+--------------+------------+
24 rows in set (0.00 sec)
```

06. Potential Customer's Location

Question 01: Where do most customers come from?

```
hive> create external table potential_customer_location(cstate string, tcount int)
row format delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/potential_customer_location';
OK
Time taken: 0.066 seconds

hive> insert overwrite table potential_customer_location select cstate,
```

```
count(distinct cid) as tcount from cecom group by cstate order by tcount desc;

hive> select * from potential_customer_location limit 10;
OK
SP     40800
RJ     12568
MG     11375
RS     5350
PR     4928
SC     3554
BA     3314
DF     2094
ES     2008
GO     1959
Time taken: 0.056 seconds, Fetched: 10 row(s)

mysql> create table potential_customer_location (cstate varchar(20), totalCount
int);
Query OK, 0 rows affected (0.01 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
potential_customer_location --export-dir
/user/hive/warehouse/hproject/potential_customer_location/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from potential_customer_location order by totalCount desc limit 5;
+--------+------------+
| cstate | totalCount |
+--------+------------+
| SP     |      40800 |
| RJ     |      12568 |
| MG     |      11375 |
| RS     |       5350 |
| PR     |       4928 |
+--------+------------+
5 rows in set (0.00 sec)
```

07. Seller Rating

Question 01: Which seller sold more?

```
hive> create external table seller_sold(sid string, tcount int) row format
delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/seller_sold';
OK
```

```
hive> insert overwrite table seller_sold select sid, sum(quantity) as tcount from
cecom group by sid order by tcount desc;

hive> select * from seller_sold limit 10;
OK
4a3ca9315b744ce9f8e9374361493884 2298
6560211a19b47992c3666cc44a7e94c0 2285
1f50f920176fa81dab994f9023523100 2169
cc419e0650a3c5ba77189a1882b7556a 1870
da8622b14eb17ae2831f4ac5b9dab84a 1830
1025f0e2d44d7041d6cf58b6550e0bfa 1552
955fee9216a65b617aa5c0531780ce60 1535
7c67e1448b00f6e969d365cea6b010ab 1524
ea8482cd71df3c1969d7b9473ff13abc 1278
3d871de0142ce09b7081e2b9d1733cb1 1258
Time taken: 0.037 seconds, Fetched: 10 row(s)

mysql> create table seller_sold(sid varchar(100), totalCount int);
Query OK, 0 rows affected (0.01 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
seller_sold --export-dir /user/hive/warehouse/hproject/seller_sold/000000_0
--input-fields-terminated-by ','
```

```
mysql> select * from seller_sold order by totalCount desc limit 5;
+----------------------------------+------------+
| sid                              | totalCount |
+----------------------------------+------------+
| 4a3ca9315b744ce9f8e9374361493884 |       2298 |
| 6560211a19b47992c3666cc44a7e94c0 |       2285 |
| 1f50f920176fa81dab994f9023523100 |       2169 |
| cc419e0650a3c5ba77189a1882b7556a |       1870 |
| da8622b14eb17ae2831f4ac5b9dab84a |       1830 |
+----------------------------------+------------+
5 rows in set (0.01 sec)
```

Question 02: Which seller got more rating?

```
hive> create external table seller_rating(sid string, arating int) row format
delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/seller_rating';
OK
Time taken: 0.091 seconds
```

```
hive> insert overwrite table seller_rating select sid, avg(rating) as arating from
cecom group by sid order by arating desc;

hive> select * from seller_rating limit 10;
OK
1d0646a72178a6fb37ee8082140e06ec 5
1de62b6f2fd96227629786db492433db 5
333c4210e76a1aa2ab817b99437e3ff1 5
98dddbc4601dd4443ca174359b237166 5
95cca791657aabeff15a07eb152d7841 5
98115075dd26cb8835946fc6086f5d30 5
97e50a621f8e801f4baf69e08687c192 5
979e9f8b5b39dd243a2550c8b05aecf0 5
f5b84683a9bf9e1df748cf40f601b39c 5
1d953075c2f0dd990bacf27b83b330f1 5
Time taken: 0.034 seconds, Fetched: 10 row(s)

mysql> create table seller_rating(sid varchar(100), avgRating int);
Query OK, 0 rows affected (0.01 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
seller_rating --export-dir /user/hive/warehouse/hproject/seller_rating/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from seller_rating order by avgRating desc limit 5;
+----------------------------------+-----------+
| sid                              | avgRating |
+----------------------------------+-----------+
| 333c4210e76a1aa2ab817b99437e3ff1 |         5 |
| 98dddbc4601dd4443ca174359b237166 |         5 |
| 1de62b6f2fd96227629786db492433db |         5 |
| 95cca791657aabeff15a07eb152d7841 |         5 |
| 1d0646a72178a6fb37ee8082140e06ec |         5 |
+----------------------------------+-----------+
5 rows in set (0.00 sec)
```

08. Logistics based Optimization Insights

Question 01: Which city buys heavy weight products and low weight products?

A. Heavy weight products

```
hive> create external table ccity_heavyproducts(ccity string, aweight float) row
format delimited fields terminated by ',' location
```

```
'/user/hive/warehouse/hproject/ccity_heavyproducts';
OK
Time taken: 0.064 seconds

hive> insert overwrite table ccity_heavyproducts select ccity, sum(quantity) as
`total` from (select ccity, quantity,  pweight_g, avg(pweight_g) over() as
`avgweight` from cecom) t where pweight_g >= avgweight group by ccity order by
total desc;


hive> select * from ccity_heavyproducts limit 10;
OK
sao paulo    3587
rio de janeiro      2003
belo horizonte      699
brasilia     461
porto alegre 386
campinas     379
curitiba     365
salvador     348
guarulhos    343
niteroi      238

mysql> create table ccity_heavyproducts(ccity varchar(80),total int);
Query OK, 0 rows affected (0.01 sec)
```

```
mysql> select * from ccity_heavyproducts limit 5;
+----------------+-------+
| ccity          | total |
+----------------+-------+
| sao paulo      |  3587 |
| rio de janeiro |  2003 |
| belo horizonte |   699 |
| brasilia       |   461 |
| porto alegre   |   386 |
+----------------+-------+
5 rows in set (0.00 sec)
```

B. Low weight products

```
## heavy weight products
hive> create external table ccity_heavyproducts(ccity string, total int) row format
delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/ccity_heavyproducts';
OK
```

```
Time taken: 0.064 seconds

hive> insert overwrite table ccity_heavyproducts select ccity, sum(quantity) as
`total` from (select ccity, quantity,  pweight_g, avg(pweight_g) over() as
`avgweight` from cecom) t where pweight_g >= avgweight group by ccity order by
total desc;


hive> select * from ccity_heavyproducts limit 10;
OK
sao paulo    3587
rio de janeiro      2003
belo horizonte      699
brasilia     461
porto alegre 386
campinas     379
curitiba     365
salvador     348
guarulhos    343
niteroi      238


mysql> create table ccity_heavyproducts(ccity varchar(80), total int);
Query OK, 0 rows affected (0.01 sec)

sqoop export --connect jdbc:mysql://localhost:3306/hproject --username root
--password cloudera --table ccity_lowproducts --export-dir
/user/hive/warehouse/hproject/ccity_lowproducts/000000_0
--input-fields-terminated-by ',';
```

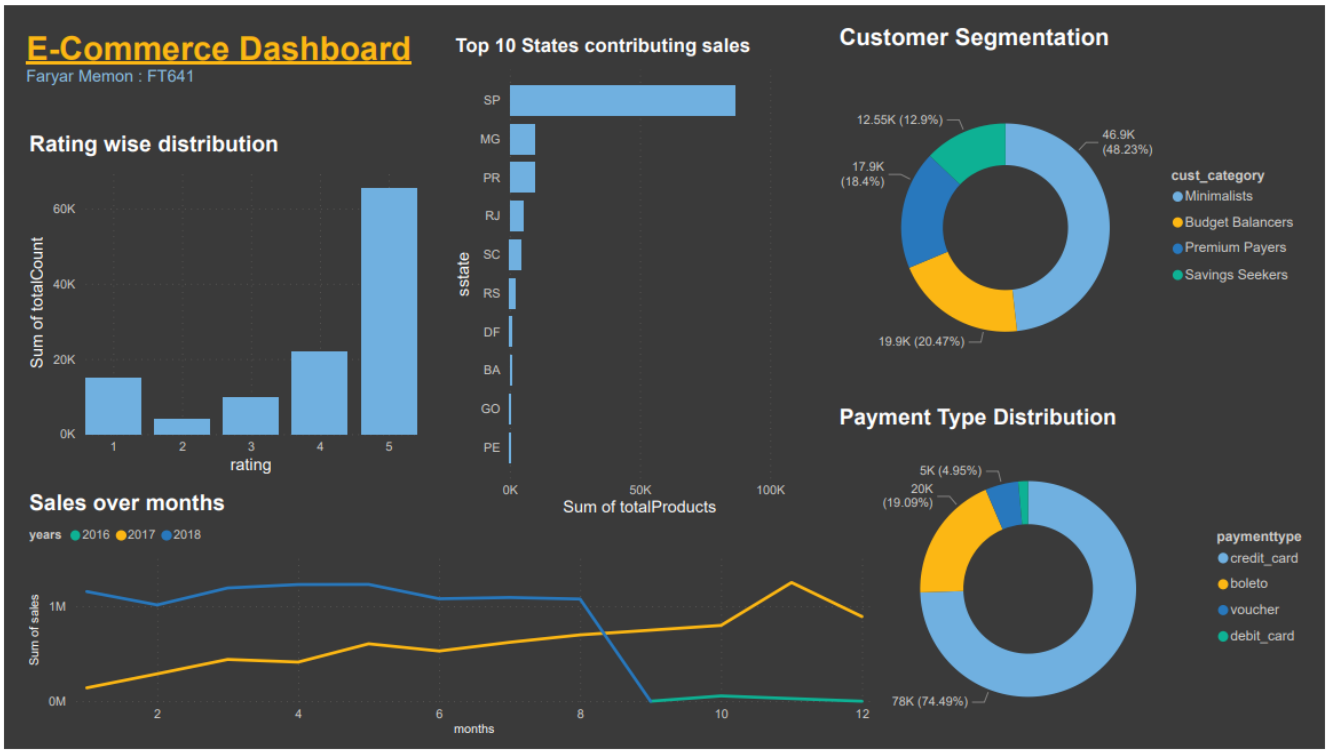
```
mysql> select * from ccity_lowproducts limit 5;
+---------------+-------+
| ccity         | total |
+---------------+-------+
| jardim        |     1 |
| independencia |     1 |
| indiaroba     |     1 |
| inga          |     1 |
| japaratuba    |     1 |
+---------------+-------+
5 rows in set (0.00 sec)
```

Question 02: How many products are sold within the seller state?

```
hive> create external table seller_state(sstate string, tproducts int) row format
delimited fields terminated by ',' location
'/user/hive/warehouse/hproject/seller_state';
OK
Time taken: 0.118 seconds

hive> insert overwrite table seller_state select sstate, sum(quantity) as tquantity
from cecom group by sstate order by tquantity desc;

hive> select * from seller_state limit 10;
OK
SP      86471
MG      9401
PR      9377
RJ      5022
SC      4401
RS      2285
DF      937
BA      683
GO      567
PE      468

mysql> create table seller_state (sstate varchar(30), totalProducts int);
Query OK, 0 rows affected (0.01 sec)

[cloudera@quickstart ~]$ sqoop export --connect
jdbc:mysql://localhost:3306/hproject --username root --password cloudera --table
seller_rating --export-dir /user/hive/warehouse/hproject/seller_rating/000000_0
--input-fields-terminated-by ',';
```

```
mysql> select * from seller_state order by totalProducts desc limit 5;
+--------+---------------+
| sstate | totalProducts |
+--------+---------------+
| SP     |         86471 |
| MG     |          9401 |
| PR     |          9377 |
| RJ     |          5022 |
| SC     |          4401 |
+--------+---------------+
5 rows in set (0.00 sec)
```

Step 03: Data Visualization (Power BI)



Observations:

- *Rating 5 is the most common rating in the overall dataset with **65478** products*
- ***São Paulo (SP)** is the state that contributes the most to the total sales*
- *Frequently used payment type by the customers is **credit card,** followed by **boleto***
- *The **highest sales** were recorded in **Nov 2017***
- *There are more **Minimalists** (low installments), followed by **Budget Balancers** (high installments), i.e., most people spend less as compared to Premium Payers (high installments) and Savings Seekers (low installments) i.e., people who spend more*