Group 1
Abhishek, Anurag, Himanshu, Kaviya, Nikita, Shreyash

Data Pre-processing

File saved as csv but actually it is an excel file so convert it to csv file.

```
hdfs dfs -put ecom_data.csv ecom
```

```
create database ecom;

use ecom;

CREATE TABLE ecom_data (order_id STRING, customer_id STRING, quantity INT, price_MRP FLOAT, payment FLOAT, timestamp STRING, rating INT, product_category STRING, product_id STRING, payment_type STRING, order_status STRING, product_weight INT, product_length INT, product_height INT, product_width INT,customer_city STRING, customer_state STRING, seller_id STRING,seller_city STRING,payment_installments INT)  row format delimited fields terminated by ',' tblproperties('skip.header.line.count'='1');

load data inpath 'ecom/ecom_data.csv' into table ecom_data;

CREATE TABLE ecom_data_orc (order_id STRING, customer_id STRING, quantity INT, price_MRP FLOAT, payment FLOAT, timestamp STRING, rating INT, product_category STRING, product_id STRING, payment_type STRING, order_status STRING, product_weight INT, product_length INT, product_height INT, product_width INT, customer_city STRING, customer_state STRING, seller_id STRING, seller_city STRING, payment_installments INT) stored as orc;

insert overwrite table ecom_data_orc select order_id, customer_id , max(quantity), price_MRP , payment , timestamp , rating , product_category , product_id , payment_type , order_status , product_weight , product_length , product_height , product_width , customer_city , customer_state, seller_id ,   seller_city , payment_installments
from ecom_data
group by order_id, customer_id , price_MRP , payment , timestamp , rating , product_category , product_id , payment_type , order_status , product_weight , product_length , product_height , product_width , customer_city , customer_state , seller_id , seller_city , payment_installments ;
```

. . → HDFS CLI Commands
. . → Hive queries
. . → Problem Statements

Problem Statement 1

Categorizing customers based on their spendings

```
create external table op1 (customer_id string, avg_spending double, spend_category string) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op1';

with cte as (select customer_id, avg(payment* quantity) as avg_spending from ecom_data_orc group by customer_id)
insert overwrite table op1 select customer_id, avg_spending, concat(floor(avg_spending/1000)*1000,'-',floor(avg_spending/1000)*1000+1000) as spend_category from cte;
```

```
sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op1 (customer_id varchar(100), avg_spending float, spend_category varchar(100));'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op1 --export-dir '/user/hive/warehouse/ecom_op/op1'
```

Problem Statement 2

the monthly trend of sales

```
create external table op2 (month int, product_categoty string, customer_state string, order_count int, avg_of_order double) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op2';

insert overwrite table op2 select substr(timestamp, 4,2) as month, product_category, customer_state, count(distinct order_id), round(avg(quantity*payment),2) from ecom_data group by substr(timestamp,4,2), product_category, customer_state;
```

```
sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op2 (month int, product_categoty varchar(100) , customer_state varchar(100) , order_count int, avg_of_order float);'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op2 --export-dir '/user/hive/warehouse/ecom_op/op2'
```

Problem statement 3

Hourly Sales Analysis

```
create external table op3 (hour int, product_category string, customer_state string, order_count int) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op3'

insert overwrite table op3 select substr(timestamp, 12,2) as hour, product_category, customer_state, count(distinct order_id) from ecom_data group by substr(timestamp,12,2), product_category, customer_state;
```

```
sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op3 (hour int, product_categoty varchar(100) , customer_state varchar(100) , order_count int);'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op3 --export-dir '/user/hive/warehouse/ecom_op/op3      '
```

Problem Statement 4

Product Based Analysis

Which category product has sold more?
Which category product has more rating?
Which product has sold more?
Top 10 highest & least product rating?
Order Count for each rating

```
create table part_cate (product_id string, quantity int, rating int) partitioned by (product_category string) clustered by (product_id) into 3 buckets;

insert overwrite table part_cate partition(product_category) select product_id, quantity, rating, product_category from ecom_data_orc;

create external table op4_1 (product_category string,count_of_products int,avg_rating float) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op4_1';

insert overwrite table op4_1 select product_category, sum(quantity) as count_of_products, round(avg(rating),2) as avg_rating from part_cate group by product_category;

create external table op4_3 (product_id string,count_of_products_sold int, avg_rating float) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op4_3';

insert overwrite table op4_3 select product_id, sum(quantity), round(avg(rating),2) from ecom_data group by product_id;

create external table op4_5 (rating int,count_of_orders int) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op4_5';

insert overwrite table op4_5 select rating, count(distinct order_id) from ecom_data group by rating;

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op4_1 (product_category varchar(100), count_of_products int, avg_rating float) '

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op4_3 (product_id varchar(100) , count_of_products_sold int, avg_rating float)  '

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op4_5 (rating int,count_of_orders int)'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op4_1 --export-dir '/user/hive/warehouse/ecom_op/op4_1'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op4_3 --export-dir '/user/hive/warehouse/ecom_op/op4_3'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op4_5 --export-dir '/user/hive/warehouse/ecom_op/op4_5'
```

Problem Statement 5

Payment Preference

What are the most commonly used payment types?
Count of Orders With each No. of Payment Instalments

```
create external table op5_1 (payment_type string,count_of_orders int) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op5_1';

insert overwrite table op5_1 select payment_type, count(distinct order_id) from ecom_data_orc group by payment_type;

create external table op5_2 (payment_installment int,count_of_orders int) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op5_2';

insert overwrite table op5_2 select coalesce(payment_installments, 'NO'), count(distinct order_id) from ecom_data_orc group by payment_installments;

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op5_1 (payment_type varchar(100) ,count_of_orders int) '

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op5_2 (payment_installment int,count_of_orders int)  '

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op5_1 --export-dir '/user/hive/warehouse/ecom_op/op5_1'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op5_2 --export-dir '/user/hive/warehouse/ecom_op/op5_2' --input-null-string '\\N' --input-null-non-string '\\N'
```

Problem Statement 6

Where do most customers come from?

```
create external table op6 (customer_state string, customer_city string, customer_count int) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op6';

insert overwrite table op6 select customer_state, customer_city, count(distinct customer_id) from ecom_data_orc group by customer_state, customer_city;
```

```
sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op6 (customer_state varchar(100), customer_city varchar(100), customer_count int)'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op6 --export-dir '/user/hive/warehouse/ecom_op/op6'
```

Problem Statement 7

Which seller sold more?
Which seller got more rating?

```
create external table op7_1 (seller_id string, products_sold int) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op7_1';

insert overwrite table op7_1 select seller_id, sum(quantity) from ecom_data_orc group by seller_id;

create external table op7_2 (seller_id string, average_rating float, order_count int) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op7_2';

insert overwrite table op7_2 select seller_id, round(avg(rating),2) as arating, count(order_id) as orders from ecom_data_orc group by seller_id;

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op7_1 (seller_id varchar(100), products_sold int)'

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op7_2 (seller_id varchar(100), average_rating float, order_count int) '

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op7_1 --export-dir '/user/hive/warehouse/ecom_op/op7_1'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op7_2 --export-dir '/user/hive/warehouse/ecom_op/op7_2'
```

Problem Statement 8

Which city buys heavy weight products and low weight products?
How much products sold within seller state?

select avg(product_weight) from ecom_data_orc; ➔ 2018

create external table op8_1(city string, state string, weight_category string) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op8_1';

insert overwrite table op8_1 select customer_city,customer_state, if (avg(product_weight) >2018 , 'Heavy_Weight', 'Low_Weight') from ecom_data_orc group by customer_city, customer_state;

create external table op8_2(state string, order_count int) row format delimited fields terminated by ',' location '/user/hive/warehouse/ecom_op/op8_2';

insert overwrite table op8_2 select seller_state ,count(distinct order_id) from ecom_data where seller_state = customer_state group by seller_state;

sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op8_1(city varchar(100), state varchar(100), weight_category varchar(100)) '

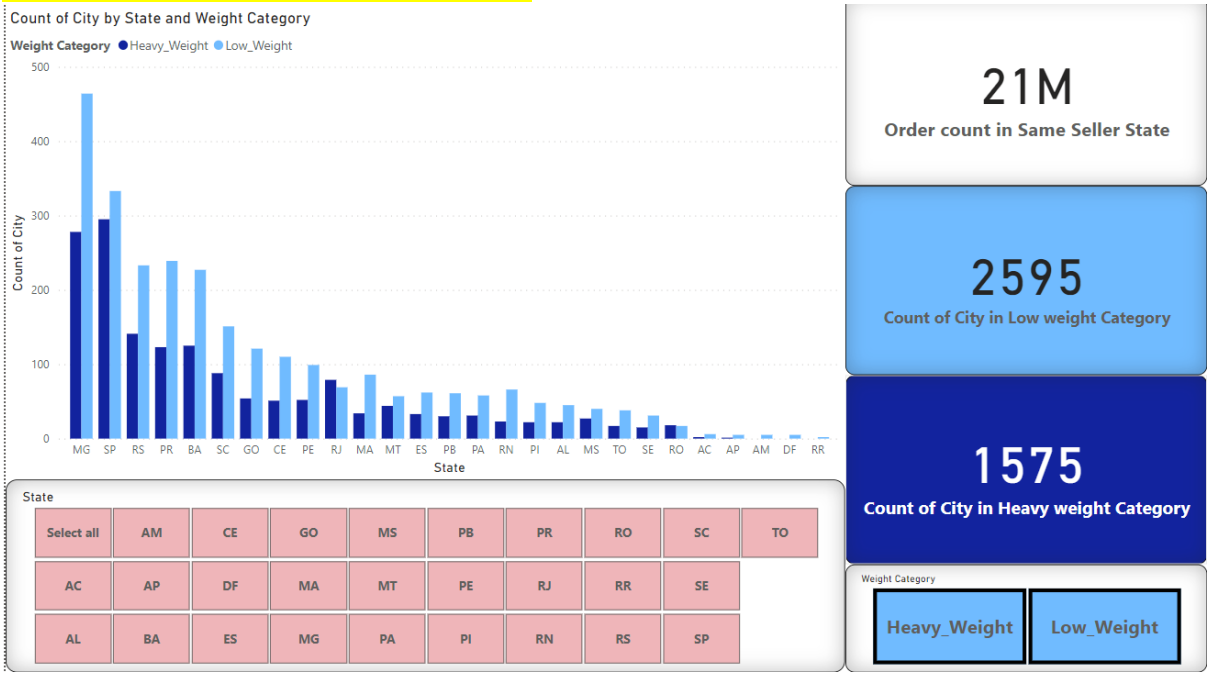
sqoop eval --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --query 'create table op8_2(state varchar(100), order_count int)'

sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op8_1 --export-dir '/user/hive/warehouse/ecom_op/op8_1'

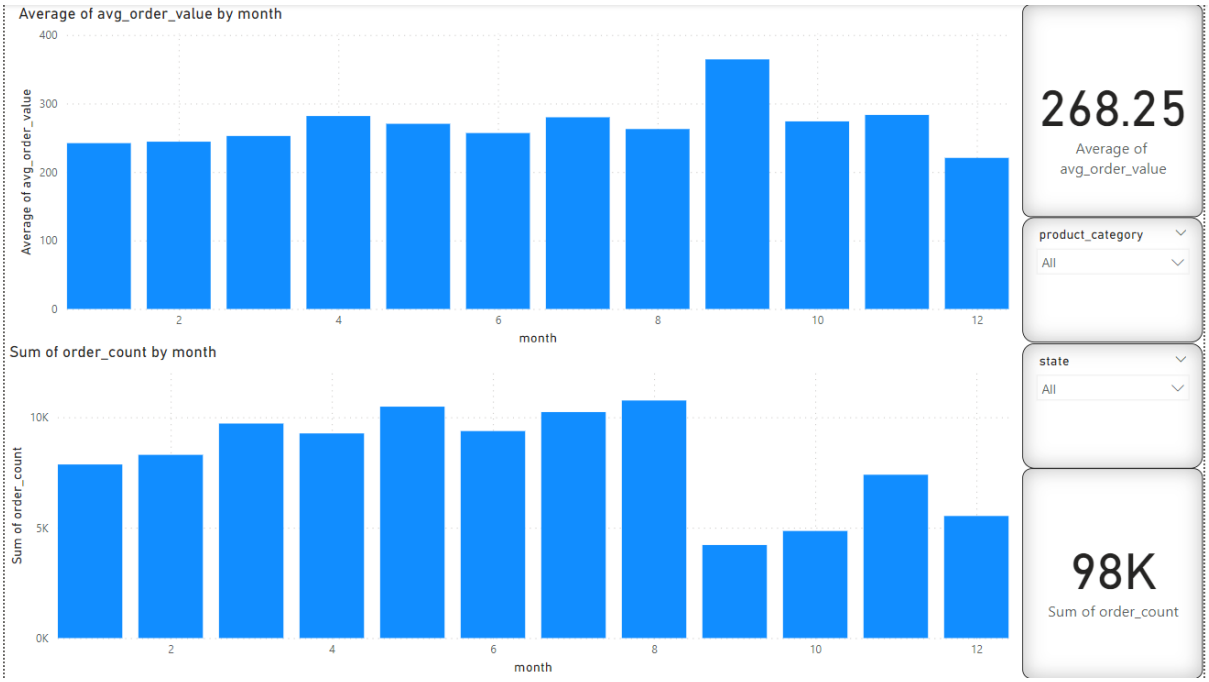
sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op8_2 --export-dir '/user/hive/warehouse/ecom_op/op8_2'

Visualization:

Which city buys heavy weight products and low weight products?
How much products sold within seller state?



the monthly trend of sales



What are the most commonly used payment types?

Count of Orders With each No. of Payment Installments