**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 ;

**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)'

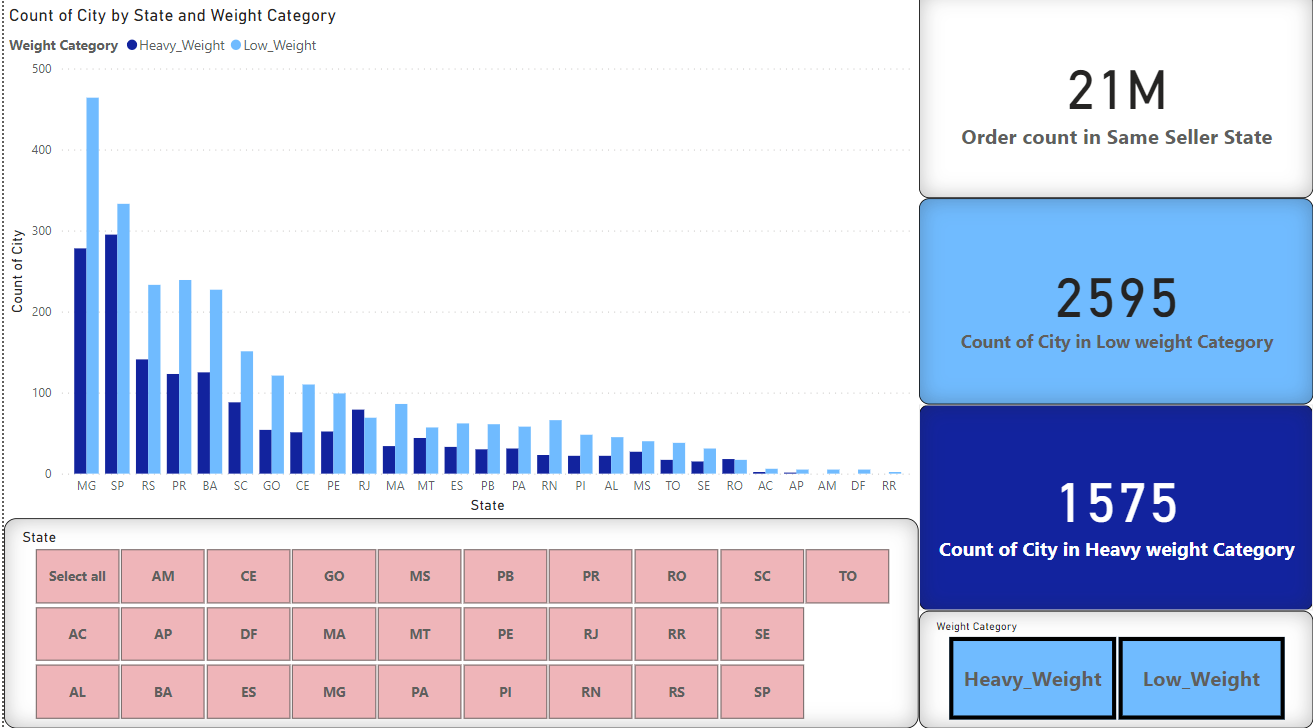
sqoop export --connect jdbc:mysql://127.0.0.1:3306/ecom --username root --password cloudera --table op8\_1 --export-dir '/user/hive/warehouse/ecom\_op/op8\_1'

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'

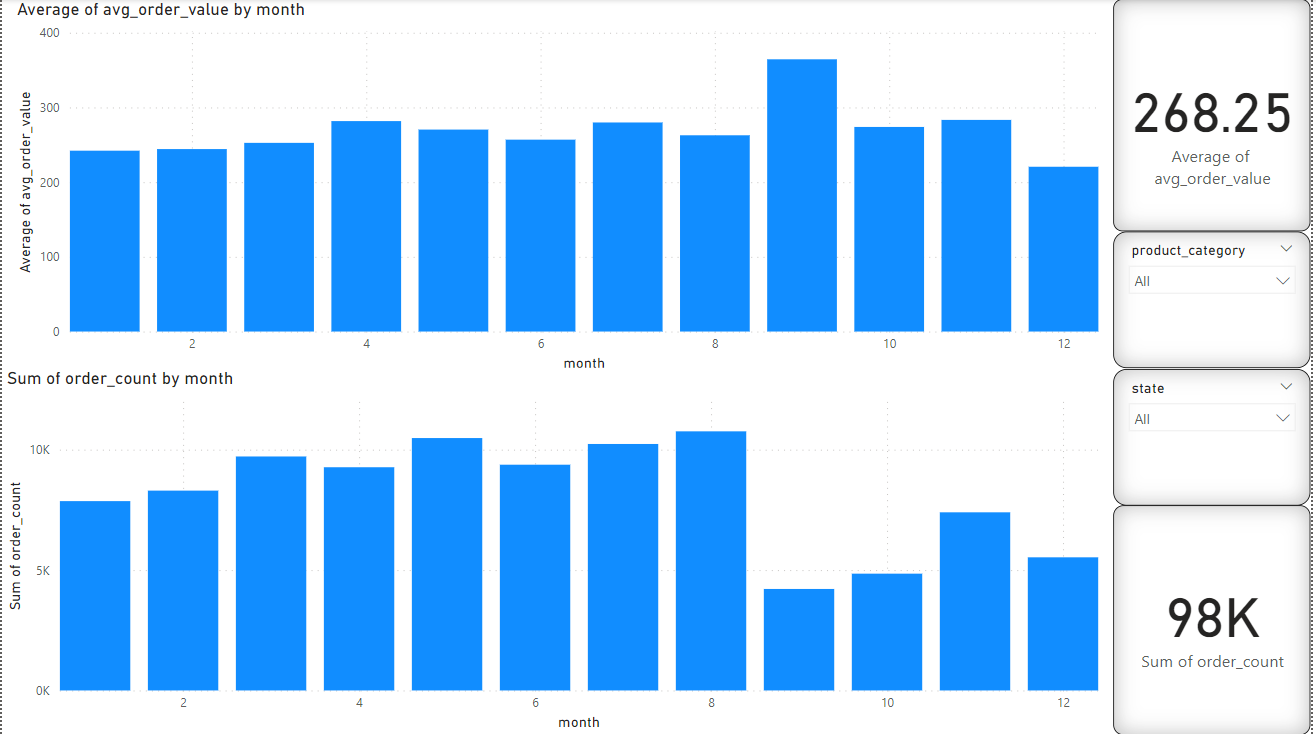
**Visualization:**

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

