1. Download vechile sales data -> https://github.com/shashank-mishra219/Hive-Class/blob/main/sales_order_data.csv

2. Store raw data into hdfs location

3. Create an internal hive table "sales_order_csv" which will store csv data sales_order_csv .. make sure to skip header row while creating table

4. Load data from hdfs path into "sales_order_csv"

5. Create an internal hive table which will store data in ORC format "sales_order_orc"

6. Load data from "sales_order_csv" into "sales_order_orc"

```
hive (d1)> select * from sales_order_orc limit 5;
OK
sales_order_orc.ordernumber     sales_order_orc.quantityordered sales_order_orc.priceeach       sales_order_orc.orderlinen
umber   sales_order_orc.sales   sales_order_orc.status  sales_order_orc.qtr_id  sales_order_orc.month_id        sales_orde
r_orc.year_id   sales_order_orc.productline     sales_order_orc.msrp    sales_order_orc.productcode     sales_order_orc.ph
one     sales_order_orc.city    sales_order_orc.state   sales_order_orc.postalcode      sales_order_orc.country sales_orde
r_orc.territory sales_order_orc.contactlastname sales_order_orc.contactfirstname        sales_order_orc.dealsize
10107   30      95.7    2       2871.0  Shipped 1       2       2003    Motorcycles     95      S10_1678        2125557818
NYC     NY      10022   USA     NA      Yu      Kwai    Small
10121   34      81.35   5       2765.9  Shipped 2       5       2003    Motorcycles     95      S10_1678        26.47.1555
Reims           51100   France  EMEA    Henriot Paul    Small
10134   41      94.74   2       3884.34 Shipped 3       7       2003    Motorcycles     95      S10_1678        +33 1 46 6
2 7555  Paris           75508   France  EMEA    Da Cunha        Daniel  Medium
10145   45      83.26   6       3746.7  Shipped 3       8       2003    Motorcycles     95      S10_1678        6265557265
Pasadena        CA      90003   USA     NA      Young   Julie   Medium
10159   49      100.0   14      5205.27 Shipped 4       10      2003    Motorcycles     95      S10_1678        6505551386
San Francisco   CA              USA     NA      Brown   Julie   Medium
Time taken: 0.294 seconds, Fetched: 5 row(s)
hive (d1)>
```

```
cloudera@quickstart:~
hive (d1)> from sales_order_csv insert overwrite table sales_order_orc select *;
Query ID = cloudera_20220918222626_6ce517d3-6aa5-4c20-9fb2-5306149e852b
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1663478033784_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663478033784_0012
/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663478033784_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-09-18 22:26:44,916 Stage-1 map = 0%,  reduce = 0%
2022-09-18 22:27:07,525 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.9 sec
MapReduce Total cumulative CPU time: 3 seconds 900 msec
Ended Job = job_1663478033784_0012
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to: hdfs://quickstart.cloudera:8020/user/hive/warehouse/d1.db/sales_order_orc/.hive-staging_hive_2022-09-18_22
-26-23_397_2706699473746204818-1/-ext-10000
Loading data to table d1.sales_order_orc
Table d1.sales_order_orc stats: [numFiles=1, numRows=2823, totalSize=37548, rawDataSize=3153291]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.9 sec   HDFS Read: 367238 HDFS Write: 37629 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 900 msec
OK
sales_order_csv.ordernumber     sales_order_csv.quantityordered sales_order_csv.priceeach       sales_order_csv.orderlinen
umber   sales_order_csv.sales   sales_order_csv.status  sales_order_csv.qtr_id  sales_order_csv.month_id        sales_orde
r_csv.year_id   sales_order_csv.productline     sales_order_csv.msrp    sales_order_csv.productcode     sales_order_csv.ph
one     sales_order_csv.city    sales_order_csv.state   sales_order_csv.postalcode      sales_order_csv.country sales_orde
r_csv.territory sales_order_csv.contactlastname sales_order_csv.contactfirstname        sales_order_csv.dealsize
Time taken: 49.323 seconds
hive (d1)>
```

```
hive (d1)> create table sales_order_orc
         > (
         > ORDERNUMBER int,
         > QUANTITYORDERED int,
         > PRICEEACH float,
         > ORDERLINENUMBER int,
         > SALES float,
         > STATUS string,
         > QTR_ID int,
         > MONTH_ID int,
         > YEAR_ID int,
         > PRODUCTLINE string,
         > MSRP int,
         > PRODUCTCODE string,
         > PHONE string,
         > CITY string,
         > STATE string,
         > POSTALCODE string,
         > COUNTRY string,
         > TERRITORY string,
         > CONTACTLASTNAME string,
         > CONTACTFIRSTNAME string,
         > DEALSIZE string
         > )
         > stored as orc
         > ;
OK
Time taken: 0.249 seconds
hive (d1)>
```

```
Time taken: 0.411 seconds
hive> show tables
    > ;
OK
emp_avro
emp_ext
employee
employee_array
employee_map
employee_orc
employee_parq
sales_dynamic
sales_order_csv
sales_order_data
sales_static
Time taken: 0.054 seconds, Fetched: 11 row(s)
hive> load data inpath '/tmp/data/' into table slaes_order_csv;
FAILED: SemanticException [Error 10001]: Line 1:41 Table not found 'slaes_order_csv'
hive> load data inpath '/tmp/data/' into table sales_order_csv;
Loading data to table d1.sales_order_csv
Table d1.sales_order_csv stats: [numFiles=1, totalSize=360233]
OK
Time taken: 0.953 seconds
hive> select * from sales_order_csv limit 5;
OK
10107   30      95.7    2       2871.0  Shipped 1       2       2003    Motorcycles     95      S10_1678        2125557818      NYC     NY      10022   USA     NA
u       Kwai    Small
10121   34      81.35   5       2765.9  Shipped 2       5       2003    Motorcycles     95      S10_1678        26.47.1555      Reims           51100   France  EMEA
enriot  Paul    Small
10134   41      94.74   2       3884.34 Shipped 3       7       2003    Motorcycles     95      S10_1678        +33 1 46 62 7555        Paris           75508   France
```

```
Time taken: 0.44 seconds, Fetched: 10 row(s)
hive> create table sales_order_csv
    > (
    > ORDERNUMBER int,
    > QUANTITYORDERED int,
    > PRICEEACH float,
    > ORDERLINENUMBER int,
    > SALES float,
    > STATUS string,
    > QTR_ID int,
    > MONTH_ID int,
    > YEAR_ID int,
    > PRODUCTLINE string,
    > MSRP int,
    > PRODUCTCODE string,
    > PHONE string,
    > CITY string,
    > STATE string,
    > POSTALCODE string,
    > COUNTRY string,
    > TERRITORY string,
    > CONTACTLASTNAME string,
    > CONTACTFIRSTNAME string,
    > DEALSIZE string
    > )
    > row format delimited
    > fields terminated by ','
    > tblproperties("skip.header.line.count"="1")
    > ;
OK
Time taken: 0.411 seconds
hive> show tables
```

Perform below menioned queries on "sales_order_orc" table :

a. Calculatye total sales per year

```
hive (d1)> select year_id,sum(sales) as Total_sales from sales_order_orc
         > group by year_id;
```

```
hive (d1)> select year_id as Year, sum(sales) as Total_Sales from sales_order_orc
         > group by year_id;
Query ID = cloudera_20220918222929_881306ba-4226-4826-a2c0-c6b2e99dcde3
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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_1663478033784_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_166347803378
/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663478033784_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-18 22:30:10,591 Stage-1 map = 0%,  reduce = 0%
2022-09-18 22:30:27,692 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 10.65 sec
2022-09-18 22:30:39,404 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.79 sec
MapReduce Total cumulative CPU time: 12 seconds 790 msec
Ended Job = job_1663478033784_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.79 sec   HDFS Read: 36783 HDFS Write: 70 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 790 msec
OK
year    total_sales
2003    3516979.547241211
2004    4724162.593383789
2005    1791486.7086791992
Time taken: 48.723 seconds, Fetched: 3 row(s)
hive (d1)>
```

b. Find a product for which maximum orders were placed

```
hive (d1)> select productline,sum(quantityordered) as Total from sales_order_orc
         > group by productline
         > order by Total desc
         > limit 1;
```

```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.57 sec   HDFS
Total MapReduce CPU Time Spent: 21 seconds 990 msec
OK
productline     total
Classic Cars    33992
Time taken: 80.119 seconds, Fetched: 1 row(s)
hive (d1)>
```

c. Calculate the total sales for each quarter

```
hive (d1)> select qtr_id,sum(sales) as Total_sales from sales_order_orc
         > group by qtr_id;
```

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.91 sec   HDFS
Total MapReduce CPU Time Spent: 11 seconds 510 msec
OK
qtr_id  total_sales
1       2350817.726501465
2       2048120.3029174805
3       1758910.808959961
4       3874780.010925293
Time taken: 35.247 seconds, Fetched: 4 row(s)
hive (d1)>
```

d. In which quarter sales was minimum

```
hive (d1)> select qtr_id,sum(sales) as Total_sales from sales_order_orc
         > group by qtr_id
         > order by Total_sales asc
         > limit 1;
```

```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.87 sec   HDFS
Total MapReduce CPU Time Spent: 17 seconds 870 msec
OK
qtr_id  total_sales
3       1758910.808959961
Time taken: 73.076 seconds, Fetched: 1 row(s)
hive (d1)>
```

e. In which country sales was maximum and in which country sales was minimum

```
 hive (d1)> select country, sum(sales) as total_sales from sales_order_csv
         > group by country
         > order by total_sales
         > limit 1;
hive (d1)> select country, sum(sales) as total_sales from sales_order_csv
         > group by country
         > order by total_sales desc
         > limit 1;
```

```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.33 sec   HDFS Read: 5781 HDFS Write
Total MapReduce CPU Time Spent: 11 seconds 750 msec
OK
country total_sales
Ireland 57756.43029785156
Time taken: 78.0 seconds, Fetched: 1 row(s)
```

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.22 sec   HDFS Read: 369918 HDFS Write: 716 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.16 sec   HDFS Read: 5781 HDFS Write: 22 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 380 msec
OK
country total_sales
USA     3627982.825744629
Time taken: 79.216 seconds, Fetched: 1 row(s)
hive (d1)>
```

f. Calculate quarterly sales for each city

```
hive (d1)> select city,qtr_id, sum(sales) as total_sales from sales_order_orc
         > group by city,qtr_id
         > order by city;
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 18.61 sec   HDFS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.62 sec   HDFS R
Total MapReduce CPU Time Spent: 24 seconds 230 msec
OK
city    qtr_id  total_sales
Aaarhus 4       100595.5498046875
Allentown       2       6166.7998046875
Allentown       3       71930.61041259766
Allentown       4       44040.729736328125
Barcelona       2       4219.2001953125
Barcelona       4       74192.66003417969
Bergamo 1       56181.320068359375
Bergamo 4       81774.40008544922
Bergen  3       16363.099975585938
Bergen  4       95277.17993164062
Boras   3       53941.68981933594
Boras   1       31606.72021484375
Boras   4       48710.92053222656
Boston  2       74994.240234375
Boston  3       15344.640014648438
Boston  4       63730.7802734375
Brickhaven      1       31474.7802734375
Brickhaven      2       7277.35009765625
Brickhaven      3       114974.53967285156
Brickhaven      4       11528.52978515625
Bridgewater     2       75778.99060058594
Bridgewater     4       26115.800537109375
Brisbane        1       16118.479858398438
Brisbane        3       34100.030029296875
```