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

```
[cloudera@quickstart ~]$ ls
array_data.csv                  kerberos
cloudera-manager                lib
cm_api.py                       map_data.csv
country_wise_latest.csv         Music
covid_19_clean_complete.csv     parcels
day_wise.csv                    Pictures
dept_data.csv                   Public
Desktop                         sales_data.csv
Documents                       sales_order_data.csv
Downloads                       sampledata.csv
eclipse                         temp
employee.csv                    Templates
enterprise-deployment.json      test2.txt
express-deployment.json         usa_county_wise.csv
full_grouped.csv                Videos
hive-hcatalog-core-0.14.0.jar   workspace
json_data.json                  worldometer_data.csv
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hdfs dfs -put '/home/cloudera/sales_order_data.csv' /
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 9 items
drwxrwxrwx   - hdfs     supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - cloudera supergroup          0 2022-09-06 11:10 /data
drwxr-xr-x   - hbase    supergroup          0 2022-09-13 22:34 /hbase
drwxr-xr-x   - cloudera supergroup          0 2022-08-30 03:17 /praj
-rw-r--r--   1 cloudera supergroup     360233 2022-09-13 22:38 /sales_order_data.csv
drwxr-xr-x   - solr     solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs     supergroup          0 2022-08-27 23:22 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /var
```

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

cloudera@quickstart:~

```
hive> create table sales_order_data_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: 6.738 seconds
hive>
```

4. Load data from hdfs path into "sales_order_csv"

```
hive> load data inpath '/sales_order_data.csv' into table sales_order_data_csv;
Loading data to table hive_class_b1.sales_order_data_csv
Table hive_class_b1.sales_order_data_csv stats: [numFiles=1, numRows=0, totalSize=360233, rawDataSize=0]
OK
Time taken: 1.002 seconds
hive>
```

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

cloudera@quickstart:~

```
hive> create table sales_order_data_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: 0.17 seconds
hive>
hive>
```

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

```
cloudera@quickstart:~
hive> from sales_order_data_csv insert overwrite table sales_order_data_orc select *;
Query ID = cloudera_20220914002222_84d0a1a3-ba4d-4734-8339-f2cc42a2f18a
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_1663133620600_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663133620600_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663133620600_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-09-14 00:22:43,009 Stage-1 map = 0%,  reduce = 0%
2022-09-14 00:23:09,402 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.95 sec
MapReduce Total cumulative CPU time: 6 seconds 950 msec
Ended Job = job_1663133620600_0001
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/hive_class_b1.db/sales_order_data_orc/.hive-staging_hive_2022-09-14_00-22-05_876_8887567916657544714-1/-ext-10000
Loading data to table hive_class_b1.sales_order_data_orc
Table hive_class_b1.sales_order_data_orc stats: [numFiles=1, numRows=2823, totalSize=37548, rawDataSize=3153291]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 6.95 sec   HDFS Read: 367446 HDFS Write: 37645 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 950 msec
OK
Time taken: 67.381 seconds
hive>
```

Perform below menioned queries on "sales_order_orc" table :

a. Calculate total sales per year

```
cloudera@quickstart:~
hive> select year_id,sum(sales) Total_Sales
    > from sales_order_data_orc
    > group by year_id;
Query ID = cloudera_20220914003939_084b8ed3-4e95-4929-9181-1f184b8c10ef
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_1663133620600_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663133620600_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663133620600_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-14 00:39:51,042 Stage-1 map = 0%,  reduce = 0%
2022-09-14 00:40:16,982 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.14 sec
2022-09-14 00:40:28,392 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.34 sec
MapReduce Total cumulative CPU time: 5 seconds 340 msec
Ended Job = job_1663133620600_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.34 sec   HDFS Read: 36851 HDFS Write: 70 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 340 msec
OK
2003    3516979.547241211
2004    4724162.593383789
2005    1791486.7086791992
Time taken: 58.704 seconds, Fetched: 3 row(s)
hive>
```

b. Find a product for which maximum orders were placed

cloudera@quickstart:~

```
hive> select productline,productcode,count(*) Total_Orders
    > from sales_order_data_orc s
    > group by productline,productcode
    > having count(*) in (
    > select max(q.cnt) max_orders
    > from
    > (select count(1) cnt
    > from sales_order_data_orc
    > group by productcode)q
    > );
Query ID = cloudera_20220914021919_5c372ce3-8821-43f4-9c9a-8ccc43e77e9e
Total jobs = 5
```

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.13 sec   HDFS Read: 27723 HDFS Write: 4386 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 3.87 sec   HDFS Read: 27317 HDFS Write: 114 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 10.07 sec   HDFS Read: 4297 HDFS Write: 114 SUCCESS
Stage-Stage-5: Map: 1   Cumulative CPU: 1.78 sec   HDFS Read: 9157 HDFS Write: 25 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 850 msec
OK
productline     productcode     total_orders
Classic Cars    S18_3232        52
Time taken: 111.564 seconds, Fetched: 1 row(s)
hive>
```

c. Calculate the total sales for each quarter

select qtr_id,sum(Sales) Total_Sales
from sales_order_data_orc
group by qtr_id;

cloudera@quickstart:~

```
hive> select qtr_id,sum(Sales) Total_Sales
    > from sales_order_data_orc
    > group by qtr_id;
Query ID = cloudera_20220914031919_78b25a7d-ab6c-46ce-9469-10e5e05cb11b
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_1663133620600_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663133620600_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663133620600_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-14 03:19:22,164 Stage-1 map = 0%,  reduce = 0%
2022-09-14 03:19:30,665 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.9 sec
2022-09-14 03:19:38,989 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.22 sec
MapReduce Total cumulative CPU time: 3 seconds 220 msec
Ended Job = job_1663133620600_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.22 sec   HDFS Read: 37113 HDFS Write: 81 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 220 msec
OK
qtr_id  total_sales
1       2350817.726501465
2       2048120.3029174805
3       1758910.808959961
4       3874780.010925293
Time taken: 30.008 seconds, Fetched: 4 row(s)
hive>
```

d. In which quarter sales was minimum

```
select qtr_id,sum(sales) total_sales
from sales_order_data_orc
group by qtr_id
having sum(sales) in(
select max(Total_Sales) from(
select sum(Sales) Total_Sales
from sales_order_data_orc
group by qtr_id)q
);
```

cloudera@quickstart:~

```
MapReduce Total cumulative CPU time: 2 seconds 160 msec
Ended Job = job_1663133620600_0024
Launching Job 3 out of 5
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_1663133620600_0025, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663133620600_0025/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663133620600_0025
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2022-09-14 03:31:37,140 Stage-4 map = 0%,  reduce = 0%
2022-09-14 03:31:44,687 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 6.97 sec
2022-09-14 03:31:54,038 Stage-4 map = 100%,  reduce = 100%, Cumulative CPU 8.63 sec
MapReduce Total cumulative CPU time: 8 seconds 630 msec
Ended Job = job_1663133620600_0025
Stage-7 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20220914033030_5ee8846b-6055-4a3e-a788-cf4c865cbb4d.log
2022-09-14 03:31:58     Starting to launch local task to process map join;      maximum memory = 932184064
2022-09-14 03:31:59     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/cloudera/c83d6c25-f0e5-492f-8211-2a
5-1/-local-10006/HashTable-Stage-5/MapJoin-mapfile31--.hashtable
2022-09-14 03:31:59     Uploaded 1 File to: file:/tmp/cloudera/c83d6c25-f0e5-492f-8211-2a2fed390100/hive_2022-09-14_03-30-51_369_7
-mapfile31--.hashtable (285 bytes)
2022-09-14 03:31:59     End of local task; Time Taken: 0.96 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 5 out of 5
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1663133620600_0026, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663133620600_0026/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663133620600_0026
Hadoop job information for Stage-5: number of mappers: 1; number of reducers: 0
2022-09-14 03:32:06,847 Stage-5 map = 0%,  reduce = 0%
2022-09-14 03:32:11,986 Stage-5 map = 100%,  reduce = 0%, Cumulative CPU 1.07 sec
MapReduce Total cumulative CPU time: 1 seconds 70 msec
Ended Job = job_1663133620600_0026
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.63 sec   HDFS Read: 36529 HDFS Write: 200 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 2.16 sec   HDFS Read: 36612 HDFS Write: 121 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 8.63 sec   HDFS Read: 4317 HDFS Write: 121 SUCCESS
Stage-Stage-5: Map: 1   Cumulative CPU: 1.07 sec   HDFS Read: 4821 HDFS Write: 20 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 490 msec
OK
qtr_id  total_sales
4       3874780.010925293
Time taken: 82.735 seconds, Fetched: 1 row(s)
hive>
```

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

```
Select country,sum(sales) max_sales
from sales_order_data_orc
group by country
having sum(sales) in
(Select max(total_sales) from
  (select sum(sales) total_sales
  from sales_order_data_orc
  group by country
  )q
);
```

```
hive> Select country,sum(sales) max_sales
    > from sales_order_data_orc
    > group by country
    > having sum(sales) in
    > (Select max(total_sales) from
    >   (select sum(sales) total_sales
    >   from sales_order_data_orc
    >   group by country
    >   )q
    > );
Query ID = cloudera_20220914033838_a538844c-f22d-47ec-9a01-66ba1b7f37b7
Total jobs = 5
```

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.76 sec   HDFS Read: 37417 HDFS Write: 716 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 12.65 sec   HDFS Read: 37478 HDFS Write: 121 SUCCESS
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 6.99 sec   HDFS Read: 4317 HDFS Write: 121 SUCCESS
Stage-Stage-5: Map: 1   Cumulative CPU: 0.97 sec   HDFS Read: 5353 HDFS Write: 22 SUCCESS
Total MapReduce CPU Time Spent: 23 seconds 370 msec
OK
country max_sales
USA     3627982.825744629
Time taken: 78.759 seconds, Fetched: 1 row(s)
hive>
```

```
Select country,sum(sales) min_sales
from sales_order_data_orc
group by country
having sum(sales) in
(Select min(total_sales) from
  (select sum(sales) total_sales
  from sales_order_data_orc
  group by country
  )q
);
```

```
hive>
hive> Select country,sum(sales) min_sales
    > from sales_order_data_orc
    > group by country
    > having sum(sales) in
    > (Select min(total_sales) from
    >   (select sum(sales) total_sales
    >   from sales_order_data_orc
    >   group by country
    >   )q
    > );
Query ID = cloudera_20220914034747_5a44eafa-210c-47b7-b4ad-3469e048c110
Total jobs = 5
Launching Job 1 out of 5
```

```
country min_sales
Ireland 57756.43029785156
Time taken: 76.693 seconds, Fetched: 1 row(s)
```

f. Calculate quarterly sales for each city

Select city,qtr_id,sum(sales) quarterly_sales

from sales_order_data_orc

group by city,qtr_id;

```
hive> Select city,qtr_id,sum(sales) quarterly_sales
    > from sales_order_data_orc
    > group by city,qtr_id;
Query ID = cloudera_20220914034444_8ce30043-21f4-4c67-b138-8a3b475e5688
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
```

```
city     qtr_id  quarterly_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    1       31606.72021484375
Boras    3       53941.68981933594
Boras    4       48710.92053222656
Boston   2       74994.240234375
```

h. Find a month for each year in which maximum number of quantities were sold

```
with cte as(
Select year_id,month_id,sum(QuantityOrdered) QuantityOrdered,
rank() over(partition by year_id order by sum(QuantityOrdered) desc) ranking
from sales_order_data_orc
group by year_id,month_id
)
Select year_id,month_id from cte where ranking=1;
```

```
hive> with cte as(
    > Select year_id,month_id,sum(QuantityOrdered) QuantityOrdered,
    > rank() over(partition by year_id order by sum(QuantityOrdered) desc) ranking
    > from sales_order_data_orc
    > group by year_id,month_id
    > )
    > Select year_id,month_id from cte where ranking=1;
Query ID = cloudera_20220914044444_8ed95c72-988d-4a03-8f29-64f7186c1ab1
Total jobs = 2
```

```
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU:
Total MapReduce CPU Time Spent: 9 seconds 990 msec
OK
year_id month_id
2003    11
2004    11
2005    5
```