HIVE – ASSIGNMENT 1

1. Download vehicle sales data ->

https://github.com/shashank-mishra219/Hive-Class/blob/main/sales_order_data.csv

2. Store raw data into hdfs location

hadoop fs -copyFromLocal /home/cloudera/Desktop/sales_order_data.csv /

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

Use hive_class_b1;

Create table sales_order_csv(ordernumber int, quantityordered int, priceeach float, orderlinenumber int, sales int, status string, qtr_id int, month_id int, year_id int, productline string, msrp int, productcode int, phone int, city string, state string, postalcode int, country string, territory string, contactlastname string, contactfirstname string, dealsize string) row format delimited fields terminated by ',' tblproperties("skip.header.line.count"="1");

4. Load data from hdfs path into "sales_order_csv"

Load data local inpath 'file://user/hive/sales_order_csv' into table sales_order_csv;

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

Create table sales_order_orc(ordernumber int, quantityordered int, priceeach float, orderlinenumber int, sales int, status string, qtr_id int, month_id int, year_id int, productline string, msrp int, productcode int, phone int, city string, state string, postalcode int, country string, territory string, contactlastname string, contactfirstname string, dealsize string) row format delimited fields terminated by ',' stored as ORC;

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

Insert into table sales_order_orc select * from sales_order_csv;

Perform below mentioned queries on "sales_order_orc" table :

a. Calculate total sales per year

Select year_id, sum(sales) total_sales from sales_order_orc group by year_id;

```
hive> show tables;
OK
sales_order_csv
sales_order_orc
Time taken: 0.031 seconds, Fetched: 2 row(s)
hive> Select year_id, sum(sales) total_sales from sales_order_csv;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'year_id'
hive> select year_id, sum(sales) total_sales from sales_order_csv;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'year_id'
hive> select year_id, sum(sales) total_sales from sales_order_orc;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'year_id'
hive> select year_id, sum(sales) total_sales from sales_order_orc group by year_id;
Query ID = cloudera_20221103010202_0c452a45-7cbf-4439-b892-c1334aeed2e0
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_1667453642337_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/applica
tion_1667453642337_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-03 01:02:40,028 Stage-1 map = 0%,  reduce = 0%
2022-11-03 01:02:52,645 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.22 sec
2022-11-03 01:03:04,708 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.67 sec
MapReduce Total cumulative CPU time: 6 seconds 670 msec
Ended Job = job_1667453642337_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.67 sec   HDFS Read: 32481 HDFS Write: 39 SUCCE
SS
Total MapReduce CPU Time Spent: 6 seconds 670 msec
OK
2003    3516514
2004    4723531
2005    1791264
Time taken: 44.96 seconds, Fetched: 3 row(s)
hive>
```

b. Find a product for which maximum orders were placed

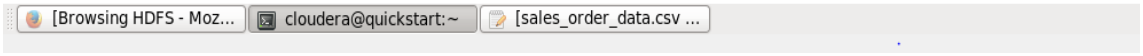
Select a.productline, a.quantityordered from sales_order_orc a left semi join (select max(quantityordered) max_order from sales_order_orc) b on (a.quantityordered=b.max_order);

```
hive> Select a.productline, a.quantityordered from sales_order_orc a left semi join (select max(quantityordered) max_order from sales_order_orc) b on (a.quantityordered=b.max_order);
Query ID = cloudera_20221103010505_bd828099-06b1-415a-a20a-e477fbeac91a
Total jobs = 3
Launching Job 1 out of 3
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_1667453642337_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1667453642337_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-03 01:05:37,557 Stage-1 map = 0%,  reduce = 0%
2022-11-03 01:05:48,482 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.05 sec
2022-11-03 01:05:59,283 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.91 sec
MapReduce Total cumulative CPU time: 6 seconds 910 msec
Ended Job = job_1667453642337_0005
Stage-6 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Execution log at: /tmp/cloudera/cloudera_20221103010505_bd828099-06b1-415a-a20a-e477fbeac91a.log
2022-11-03 01:06:08     Starting to launch local task to process map join;      maximum memory = 932
184064
2022-11-03 01:06:10     Dump the side-table for tag: 1 with group count: 1 into file: file:/tmp/clou
dera/ef192242-d9a9-471d-a1fc-667b657812b0/hive_2022-11-03_01-05-25_119_9124603777831644843-1/-local-
10004/HashTable-Stage-4/MapJoin-mapfile01--.hashtable
2022-11-03 01:06:10     Uploaded 1 File to: file:/tmp/cloudera/ef192242-d9a9-471d-a1fc-667b657812b0/
hive_2022-11-03_01-05-25_119_9124603777831644843-1/-local-10004/HashTable-Stage-4/MapJoin-mapfile01-
-.hashtable (278 bytes)
2022-11-03 01:06:10     End of local task; Time Taken: 2.171 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1667453642337_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/applicat
ion_1667453642337_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0006
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 0
2022-11-03 01:06:23,604 Stage-4 map = 0%,  reduce = 0%
2022-11-03 01:06:34,426 Stage-4 map = 100%,  reduce = 0%, Cumulative CPU 4.66 sec
MapReduce Total cumulative CPU time: 4 seconds 660 msec
Ended Job = job_1667453642337_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.91 sec   HDFS Read: 28424 HDFS Write: 114 SUCCE
SS
Stage-Stage-4: Map: 1   Cumulative CPU: 4.66 sec   HDFS Read: 26686 HDFS Write: 16 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 570 msec
OK
Classic Cars    97
Time taken: 70.486 seconds, Fetched: 1 row(s)
```

c. Calculate the total sales for each quarter

Select qtr_id, sum(sales) total_sales from sales_order_orc group by qtr_id;

```
hive> Select qtr_id, sum(sales) total_sales from sales_order_orc group by qtr_id;
Query ID = cloudera_20221103011717_1ea97cc6-0015-4f20-a039-d935e3cfe18f
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_1667453642337_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1667453642337_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-03 01:17:37,711 Stage-1 map = 0%,  reduce = 0%
2022-11-03 01:17:48,780 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.78 sec
2022-11-03 01:18:00,588 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.24 sec
MapReduce Total cumulative CPU time: 9 seconds 240 msec
Ended Job = job_1667453642337_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.24 sec   HDFS Read: 32743 HDFS Write: 40 SUCCES
S
Total MapReduce CPU Time Spent: 9 seconds 240 msec
OK
1       2350510
2       2047855
3       1758673
4       3874271
Time taken: 36.167 seconds, Fetched: 4 row(s)
hive>
```



d. In which quarter sales was minimum

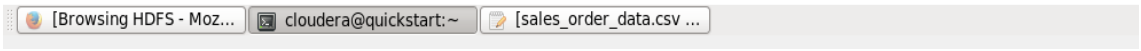
Select qtr_id, sum(sales) total_sales from sales_order_orc group by qtr_id order by total_sales asc limit 1;

```
hive> Select qtr_id, sum(sales) total_sales from sales_order_orc group by qtr_id order by total_sales asc limit 1;
Query ID = cloudera_20221103011313_8858b341-8c7a-48f1-9301-254403ef0a50
Total jobs = 2
Launching Job 1 out of 2
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_1667453642337_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1667453642337_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-03 01:14:05,607 Stage-1 map = 0%,  reduce = 0%
2022-11-03 01:14:17,432 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.72 sec
2022-11-03 01:14:28,213 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.8 sec
MapReduce Total cumulative CPU time: 8 seconds 800 msec
Ended Job = job_1667453642337_0008
Launching Job 2 out of 2
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_1667453642337_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/applicat
ion_1667453642337_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0009
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-11-03 01:14:39,972 Stage-2 map = 0%,  reduce = 0%
2022-11-03 01:14:49,663 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.31 sec
2022-11-03 01:15:00,424 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.72 sec
MapReduce Total cumulative CPU time: 5 seconds 720 msec
Ended Job = job_1667453642337_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.8 sec   HDFS Read: 31813 HDFS Write: 184 SUCCES
S
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.72 sec   HDFS Read: 5223 HDFS Write: 10 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 520 msec
OK
3       1758673
Time taken: 67.941 seconds, Fetched: 1 row(s)
hive>
```

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

Select country, max(sales) from sales_order_orc group by country;

```
hive> Select country, max(sales) from sales_order_orc group by country;
Query ID = cloudera_20221103012121_6d8970e4-b379-4c29-b74d-d3b11b604f10
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_1667453642337_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1667453642337_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-03 01:21:30,953 Stage-1 map = 0%,  reduce = 0%
2022-11-03 01:21:41,789 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.31 sec
2022-11-03 01:21:53,638 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.74 sec
MapReduce Total cumulative CPU time: 9 seconds 740 msec
Ended Job = job_1667453642337_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.74 sec   HDFS Read: 33594 HDFS Write: 248 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 740 msec
OK
Australia       9774
Austria 9240
Belgium 6804
Canada  9064
Denmark 10468
Finland 10606
France  11739
Germany 8940
Ireland 8258
Italy   9160
Japan   10758
Norway  8844
Philippines     7483
Singapore       10993
Spain   12001
Sweden  7209
Switzerland     6761
UK      11886
USA     14082
Time taken: 37.982 seconds, Fetched: 19 row(s)
hive>
```



Select country, min(sales) from sales_order_orc group by country;

```
hive> Select country, min(sales) from sales_order_orc group by country;
Query ID = cloudera_20221103012424_177fe40e-1230-4134-ba95-f988f5c059ee
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_1667453642337_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1667453642337_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-03 01:24:14,638 Stage-1 map = 0%,  reduce = 0%
2022-11-03 01:24:26,567 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.88 sec
2022-11-03 01:24:38,435 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.46 sec
MapReduce Total cumulative CPU time: 8 seconds 460 msec
Ended Job = job_1667453642337_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.46 sec   HDFS Read: 33594 HDFS Write: 228 SUCCE
SS
Total MapReduce CPU Time Spent: 8 seconds 460 msec
OK
Australia       652
Austria 640
Belgium 881
Canada  1119
Denmark 1146
Finland 891
France  482
Germany 948
Ireland 1056
Italy   577
Japan   553
Norway  1129
Philippines     1173
Singapore       785
Spain   683
Sweden  1467
Switzerland     1205
UK      710
USA     541
Time taken: 39.565 seconds, Fetched: 19 row(s)
hive>
```

f. Calculate quarterly sales for each city

Select qtr_id, city, sum(sales) from sales_order_orc group by city, qtr_id;

```
hive> Select qtr_id, city, sum(sales) from sales_order_orc group by city, qtr_id;
Query ID = cloudera_20221103012727_1294d0dd-4e9e-437d-abea-11243becc998
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_1667453642337_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1667453642337_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-11-03 01:28:08,554 Stage-1 map = 0%,  reduce = 0%
2022-11-03 01:28:20,785 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.65 sec
2022-11-03 01:28:32,626 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.34 sec
MapReduce Total cumulative CPU time: 10 seconds 340 msec
Ended Job = job_1667453642337_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.34 sec   HDFS Read: 34980 HDFS Write: 3095 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 340 msec
OK
4       Aaarhus 100583
2       Allentown       6166
3       Allentown       71924
4       Allentown       44038
2       Barcelona       4219
4       Barcelona       74182
1       Bergamo 56172
4       Bergamo 81762
3       Bergen  16361
4       Bergen  95266
1       Boras   31603
3       Boras   53933
4       Boras   48704
2       Boston  74982
3       Boston  15342
4       Boston  63724
1       Brickhaven      31470
2       Brickhaven      7276
3       Brickhaven      114957
4       Brickhaven      11527
2       Bridgewater     75771
4       Bridgewater     26113
1       Brisbane        16116
3       Brisbane        34094
1       Bruxelles       18798
2       Bruxelles       8410
3       Bruxelles       47753
1       Burbank 37847
```

```
1       Paris   71486
2       Paris   80206
3       Paris   27794
4       Paris   89428
1       Pasadena        44267
3       Pasadena        55767
4       Pasadena        4512
1       Philadelphia    27396
2       Philadelphia    7286
4       Philadelphia    116486
2       Reggio Emilia   41503
3       Reggio Emilia   56415
4       Reggio Emilia   44665
1       Reims   52020
2       Reims   18970
3       Reims   15145
4       Reims   48889
2       Salzburg        98091
3       Salzburg        6692
4       Salzburg        44997
1       San Diego       87478
1       San Francisco   72890
4       San Francisco   151442
2       San Jose        159994
1       San Rafael      267279
2       San Rafael      7261
3       San Rafael      216271
4       San Rafael      163961
4       Sevilla 54718
1       Singapore       28391
2       Singapore       92018
3       Singapore       90240
4       Singapore       77802
1       South Brisbane  21728
3       South Brisbane  10639
4       South Brisbane  27097
1       Stavern 54694
4       Stavern 61890
2       Strasbourg      80430
3       Torino  94107
1       Toulouse        15137
3       Toulouse        17249
4       Toulouse        38092
2       Tsawassen       31296
3       Tsawassen       43326
4       Vancouver       75227
1       Versailles      5759
4       Versailles      59068
4       White Plains    85545
Time taken: 37.331 seconds, Fetched: 182 row(s)
hive>
```

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

Select distinct a.month_id, a.year_id, a.quantityordered from sales_order_orc a inner join (select month_id, year_id, max(quantityordered) as max_quantity from sales_order_orc group by month_id, year_id)b on (a.month_id = b.month_id and a.year_id=b.year_id and a.quantityordered = b.max_quantity);

```
hive> Select distinct a.month_id, a.year_id, a.quantityordered from sales_order_orc a inner join (select month_id, year_id, max(quantityordered) as max_quantity from sales_order_orc group by month_id, year_id)b on (a.month_id = b.month_id and a.year_id=b.year_id and a.quantityordered = b.max_quantity);
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_1667453642337_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1667453642337_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1667453642337_0015
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2022-11-03 01:32:06,126 Stage-3 map = 0%,  reduce = 0%
2022-11-03 01:32:17,023 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.62 sec
2022-11-03 01:32:29,993 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 6.12 sec
MapReduce Total cumulative CPU time: 6 seconds 120 msec
Ended Job = job_1667453642337_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.67 sec   HDFS Read: 29366 HDFS Write: 734 SUCCESS
Stage-Stage-3: Map: 1  Reduce: 1   Cumulative CPU: 6.12 sec   HDFS Read: 11121 HDFS Write: 296 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 790 msec
OK
1       2003    50
1       2004    50
1       2005    50
2       2003    50
2       2004    50
2       2005    50
3       2003    50
3       2004    50
3       2005    50
4       2003    50
4       2004    49
4       2005    97
5       2003    50
5       2004    50
5       2005    70
6       2003    50
6       2004    50
7       2003    49
7       2004    50
8       2003    49
8       2004    50
9       2003    50
9       2004    50
10      2003    50
10      2004    50
11      2003    50
11      2004    55
12      2003    49
12      2004    50
Time taken: 90.745 seconds, Fetched: 29 row(s)
hive>
```

[Browsing HDFS - Moz...] [cloudera@quickstart:~] [sales_order_data.csv ...]