1. #To download the data
https://github.com/shashank-mishra219/Hive-Class/blob/main/sales_order_data.csv

2. #Once the file is been downloaded, convert the file into CSV and move the file to cloudera with the help of filezilla
#Once the file is been moved then store the file in hdfs location

#to move the file from local machine to hdfs, first create a dir to which the file should me moved

#create a new directory
hadoop fs -mkdir /sales

#to move the file from local to hdfs
hadoop fs -put /'your-file-path'/sales_data.csv /sales

```
[cloudera@quickstart ~]$ cd Downloads
[cloudera@quickstart Downloads]$ ls
csv_table.csv  department_data.csv  f1_drivers.txt  hive-hcatalog-core-0.14.0.jar  json_file.json  sales_data.csv
[cloudera@quickstart Downloads]$ hdfs dfs -put /home/cloudera/Downloads/sales_data.csv /hdfs/sales
put: `/hdfs/sales': No such file or directory
[cloudera@quickstart Downloads]$ hdfs dfs -put /home/cloudera/Downloads/sales_data.csv /sales
[cloudera@quickstart Downloads]$
```

```
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 1 items
-rw-r--r--   1 cloudera cloudera         49 2022-08-28 01:18 testfile2.txt
[cloudera@quickstart ~]$ hdfs dfs -mkdir /sales
[cloudera@quickstart ~]$ hdfd dfs -ls
-bash: hdfd: command not found
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 1 items
-rw-r--r--   1 cloudera cloudera         49 2022-08-28 01:18 testfile2.txt
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 11 items
drwxrwxrwx   - hdfs     supergroup          0 2017-10-23 09:15 /benchmarks
-rw-r--r--   1 cloudera supergroup        328 2022-09-10 03:41 /department_data.csv
drwxrwxrwx   - cloudera supergroup          0 2022-08-28 01:23 /gokul
drwxr-xr-x   - hbase    supergroup          0 2022-09-14 23:21 /hbase
drwxr-xr-x   - cloudera supergroup          0 2022-09-14 23:43 /sales
drwxr-xr-x   - solr     solr                0 2017-10-23 09:18 /solr
-rwxrwxrwx   1 cloudera supergroup         21 2022-08-27 22:31 /test.txt
-rwxrwxrwx   1 cloudera supergroup         49 2022-08-28 01:21 /testfile2.txt
drwxrwxrwt   - hdfs     supergroup          0 2022-08-20 13:11 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2022-09-10 03:44 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$
```

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

```
#create your database
create database sales;

#choose your database
use sales;

#create your table
create table sales_order
(
ordernumber int,
quantityordered int,
priceeach float,
orderlinenumber int,
sales int,
status string,
qtr_id int,
month_id int,
year_id int,
productline string,
msrp string,
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");

```
[cloudera@quickstart Downloads]$ cd ..
[cloudera@quickstart ~]$ clear
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create database sales;
OK
Time taken: 3.755 seconds
hive> show databases;
OK
default
hive_class_b1
sales
test
Time taken: 0.387 seconds, Fetched: 4 row(s)
hive> use sales;
OK
Time taken: 0.121 seconds
```

```
hive> create table sales_order
    > (ordernumber   int,
    > quantityordered   int,
    > priceeach   float,
    > orderlinenumber   int,
    > sales   int,
    > status   string,
    > qtr_id   int,
    > month_id   int,
    > year_id   int,
    > productline   string,
    > msrp   string,
    > 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: 0.477 seconds
```

4. Load data from hadoop hdfs path into "sales_order"
load data inpath '/sales/sales_data.csv' into table sales_order;

```
Time taken: 1.083 seconds, Fetched: 2823 row(s)
hive> load data inpath '/sales/sales_data.csv' into table sales_order;
```

```
hive> show tables;
OK
sales_order
Time taken: 0.074 seconds, Fetched: 1 row(s)
hive>
```

```
hive> select count(1) from sales_order;
Query ID = cloudera_20220915004040_d648d420-f141-4d56-bbbc-1dcb518570fb
Total jobs = 1
Launching Job 1 out of 1
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_1663220710245_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663220710245_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663220710245_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-15 00:40:45,760 Stage-1 map = 0%,  reduce = 0%
2022-09-15 00:40:56,085 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.42 sec
2022-09-15 00:41:08,083 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.23 sec
MapReduce Total cumulative CPU time: 5 seconds 230 msec
Ended Job = job_1663220710245_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.23 sec   HDFS Read: 367418 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 230 msec
OK
2823
Time taken: 51.77 seconds, Fetched: 1 row(s)
hive>
```

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 string,
productcode string,
phone string,
city string,
state string,
postalcode string,
country string,
territory string,
contactlastname string,
contactfirstname string,
dealsize string
)
stored as orc;

```
Time taken: 31.77 seconds, Fetched: 1 row(s)
hive> 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   string,
    > 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: 0.699 seconds
hive>
```

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

#here we load the data already existing table sales_order into orc table
Insert into table sales_order_orc select * from sales_order;

```
Time taken: 0.899 seconds
hive> insert into table sales_order_orc select * from sales_order;
Query ID = cloudera_20220915004848_4c4753da-a613-41a5-b247-c63e9b7cff68
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_1663220710245_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663220710245_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663220710245_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-09-15 00:48:19,548 Stage-1 map = 0%,  reduce = 0%
2022-09-15 00:48:31,269 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.95 sec
MapReduce Total cumulative CPU time: 3 seconds 950 msec
Ended Job = job_1663220710245_0002
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/sales.db/sales_order_orc/.hive-staging_hive_2022-09-15_00-48-05_809_3860123
/-ext-10000
Loading data to table sales.sales_order_orc
Table sales.sales_order_orc stats: [numFiles=1, numRows=2823, totalSize=33441, rawDataSize=3384777]
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.95 sec   HDFS Read: 364504 HDFS Write: 33525 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 950 msec
OK
Time taken: 30.256 seconds
hive>
```

Perform below menioned queries on "sales_order_orc" table :

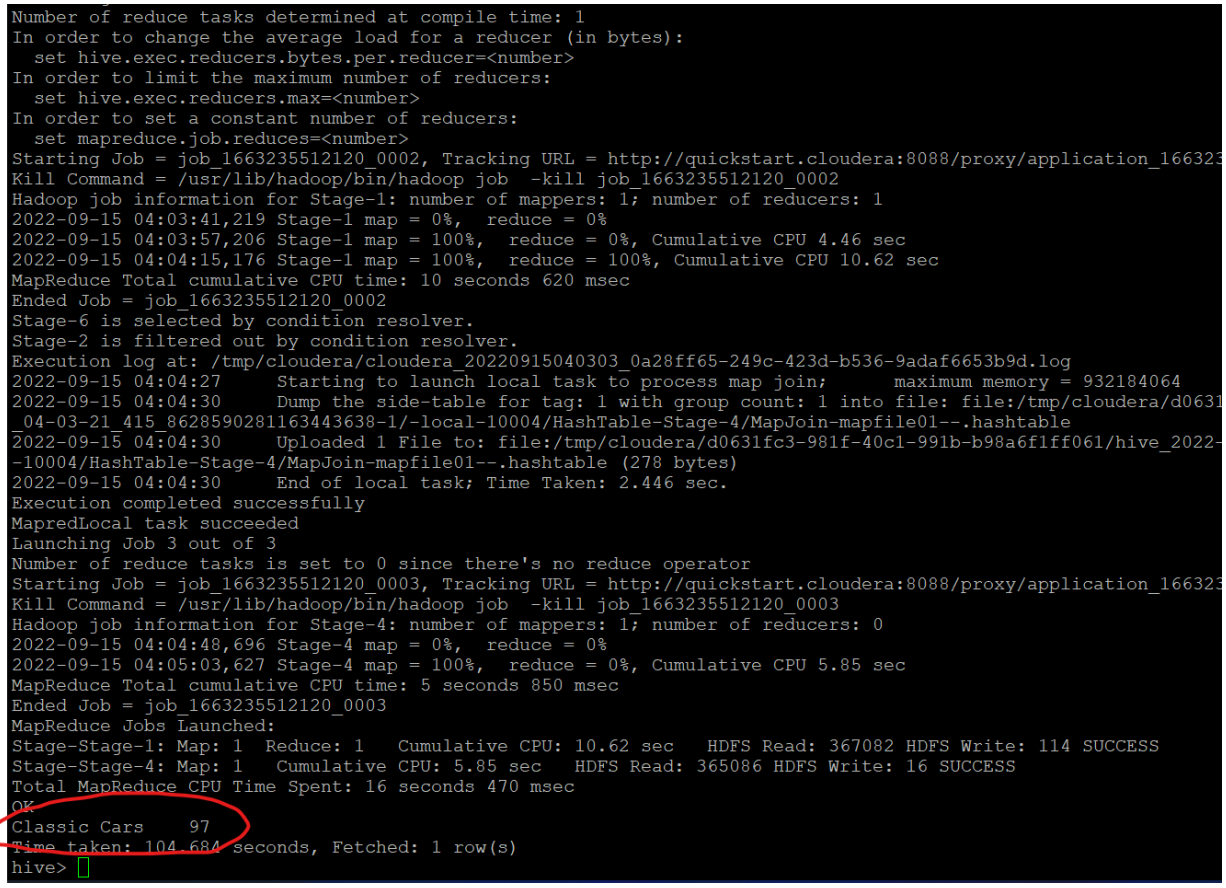
a. Calculatye total sales per year

select year_id, sum(sales) t_sales from sales_order_orc
group by year_id;

```
hive> select YEAR_ID, sum(SALES) TOTAL_SALES from sales_order_orc
    > group by YEAR_ID;
Query ID = cloudera_20220915010202_54eb81bb-e457-423b-964c-c7843098e8d6
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_1663220710245_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663220710245_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663220710245_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-15 01:02:26,302 Stage-1 map = 0%,  reduce = 0%
2022-09-15 01:02:36,537 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.04 sec
2022-09-15 01:02:48,479 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.67 sec
MapReduce Total cumulative CPU time: 6 seconds 670 msec
Ended Job = job_1663220710245_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.67 sec   HDFS Read: 32443 HDFS Write: 39 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 670 msec
OK
2003    3516514
2004    4723531
2005    1791264
Time taken: 37.208 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 join
(select max(quantityordered) max_o from sales_order_orc) b
on
(a.quantityordered=b.max_o);
```



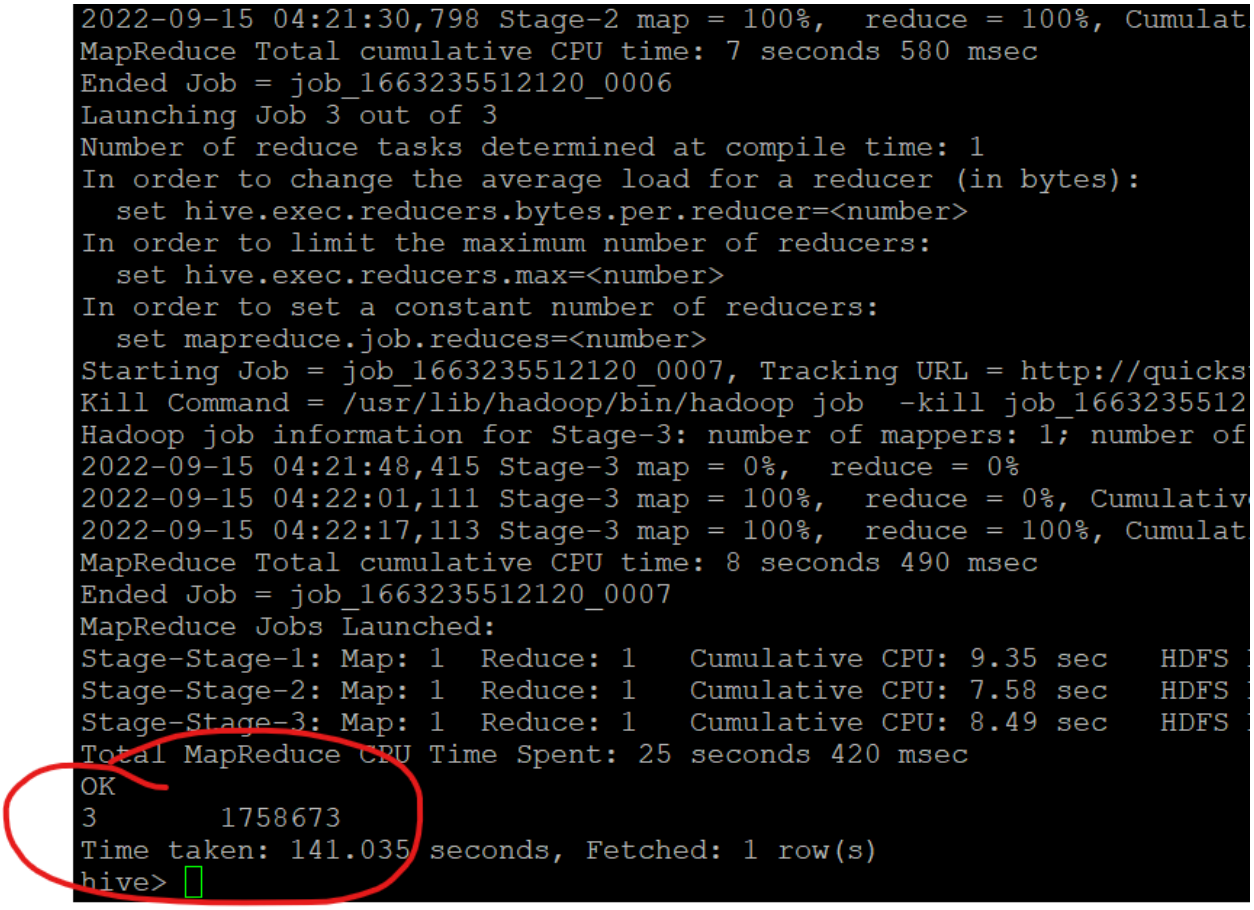
c. Calculate the total sales for each quarter

```
select qtr_id, min(sales) t_sales from sales_order_orc
group  by qtr_id;
```

```
hive> select qtr_id, sum(sales) total_sales from sales_order
    > group by qtr_id;
Query ID = cloudera_20220915041717_c9dabfc5-f722-4727-b730-a848435854fb
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_1663235512120_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1663235512120_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1663235512120_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-15 04:17:30,970 Stage-1 map = 0%,  reduce = 0%
2022-09-15 04:17:48,150 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.41 sec
2022-09-15 04:18:02,912 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.27 sec
MapReduce Total cumulative CPU time: 9 seconds 270 msec
Ended Job = job_1663235512120_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.27 sec   HDFS Read: 367781 HDFS Write: 40 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 270 msec
OK
1       2350510
2       2047855
3       1758673
4       3874271
Time taken: 53.62 seconds, Fetched: 4 row(s)
hive>
```

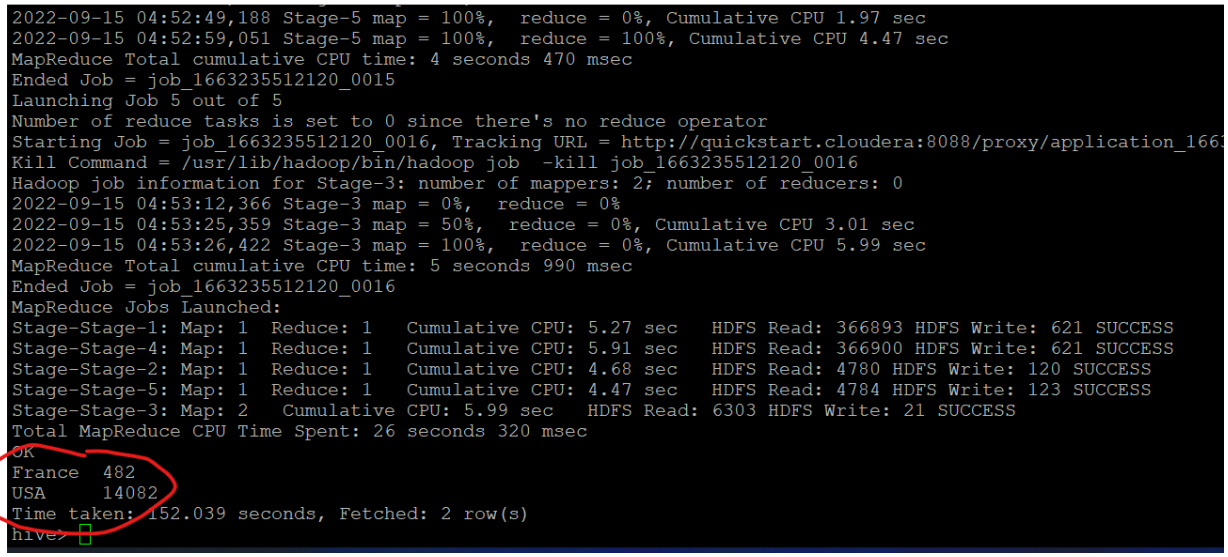
d. In which quarter sales was minimum

```
select qtr_id, t_sales
from (
  select qtr_id, sum(sales) t_sales from sales_order_orc
  group  by qtr_id
) min
sort by t_sales asc
 limit 1;
```



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

```
select country,max(sales) sales from sales_order
group by country
order by sales desc
limit 1
union all
select country,min(sales) sales from sales_order
group by country
order by sales asc
limit 1;
```



f. Calculate quartelry sales for each city

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

```
Aaarhus 4       100583
Allentown       2       6166
Allentown       3       71924
Allentown       4       44038
Barcelona       2       4219
Barcelona       4       74182
Bergamo 1       56172
Bergamo 4       81762
Bergen  3       16361
Bergen  4       95266
Boras   1       31603
Boras   3       53933
Boras   4       48704
Boston  2       74982
Boston  3       15342
Boston  4       63724
Brickhaven      1       31470
Brickhaven      2       7276
Brickhaven      3       114957
Brickhaven      4       11527
Bridgewater     2       75771
Bridgewater     4       26113
Brisbane        1       16116
Brisbane        3       34094
Bruxelles       1       18798
Bruxelles       2       8410
Bruxelles       3       47753
Burbank 1       37847
Burbank 4       8233
Burlingame      1       13529
Burlingame      3       42027
Burlingame      4       65213
Cambridge       1       21780
Cambridge       2       14379
Cambridge       3       48822
Cambridge       4       54246
Charleroi       1       16626
Charleroi       2       1711
Charleroi       3       1637
Charleroi       4       13462
Chatswood       2       43965
Chatswood       3       69688
Chatswood       4       37897
```

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

```
select year_id, month_id, quantityordered from (
select year_id, month_id, quantityordered ,rank() over (partition by year_id, month_id order by
cast(quantityordered as int) desc) y from
(
SELECT year_id, month_id, sum(quantityordered)  quantityordered from  sales_order
group by  year_id,month_id   )  x) x
WHERE x.y = 1;
```

```
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-09-15 05:02:48,007 Stage-2 map = 0%,  reduce = 0%
2022-09-15 05:02:56,827 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
2022-09-15 05:03:07,736 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.72
MapReduce Total cumulative CPU time: 5 seconds 720 msec
Ended Job = job_1663235512120_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.47 sec   HDFS Read: 367159
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.72 sec   HDFS Read: 9334 HI
Total MapReduce CPU Time Spent: 11 seconds 190 msec
OK
2003    1       1357
2003    2       1449
2003    3       1755
2003    4       1993
2003    5       2017
2003    6       1649
2003    7       1725
2003    8       1974
2003    9       2510
2003    10      5515
2003    11      10179
2003    12      2489
2004    1       3245
2004    2       3061
2004    3       1978
2004    4       2077
2004    5       2618
2004    6       2971
2004    7       3174
2004    8       4564
2004    9       3171
2004    10      5483
2004    11      10678
2004    12      3804
2005    1       3395
2005    2       3393
2005    3       3852
2005    4       2634
2005    5       4357
Time taken: 62.032 seconds, Fetched: 29 row(s)
hive>
```