Name – Vaibhav Khandekar

Enrollment No. - 230340325073

Set - B

Q.1) Find out the average High price for each stock.

Here, we have chosen the stock market dataset on which we have performed map-reduce operations. Following is the structure of the data. Kindly Find the solutions to the questions below.

Data Structure

1. Exchange Name

2 Stock symbol

3. Transaction date

4. Opening price of the stock

5. Intra day high price of the stock

6. Intra day low price of the stock

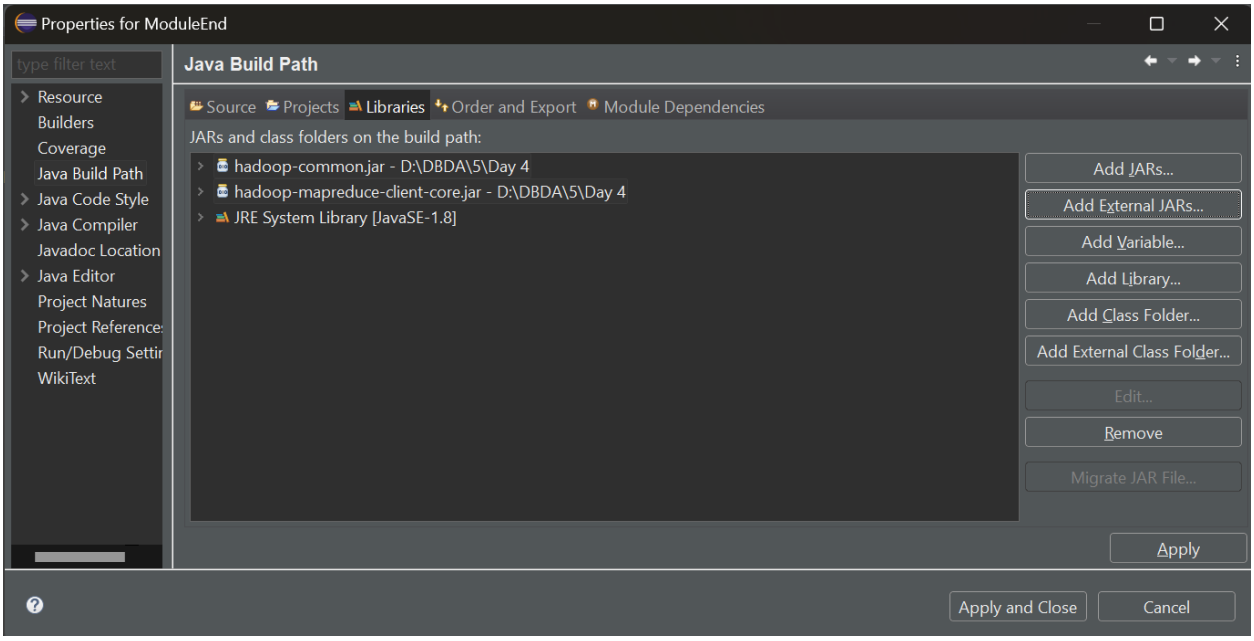
7. Closing price of the stock

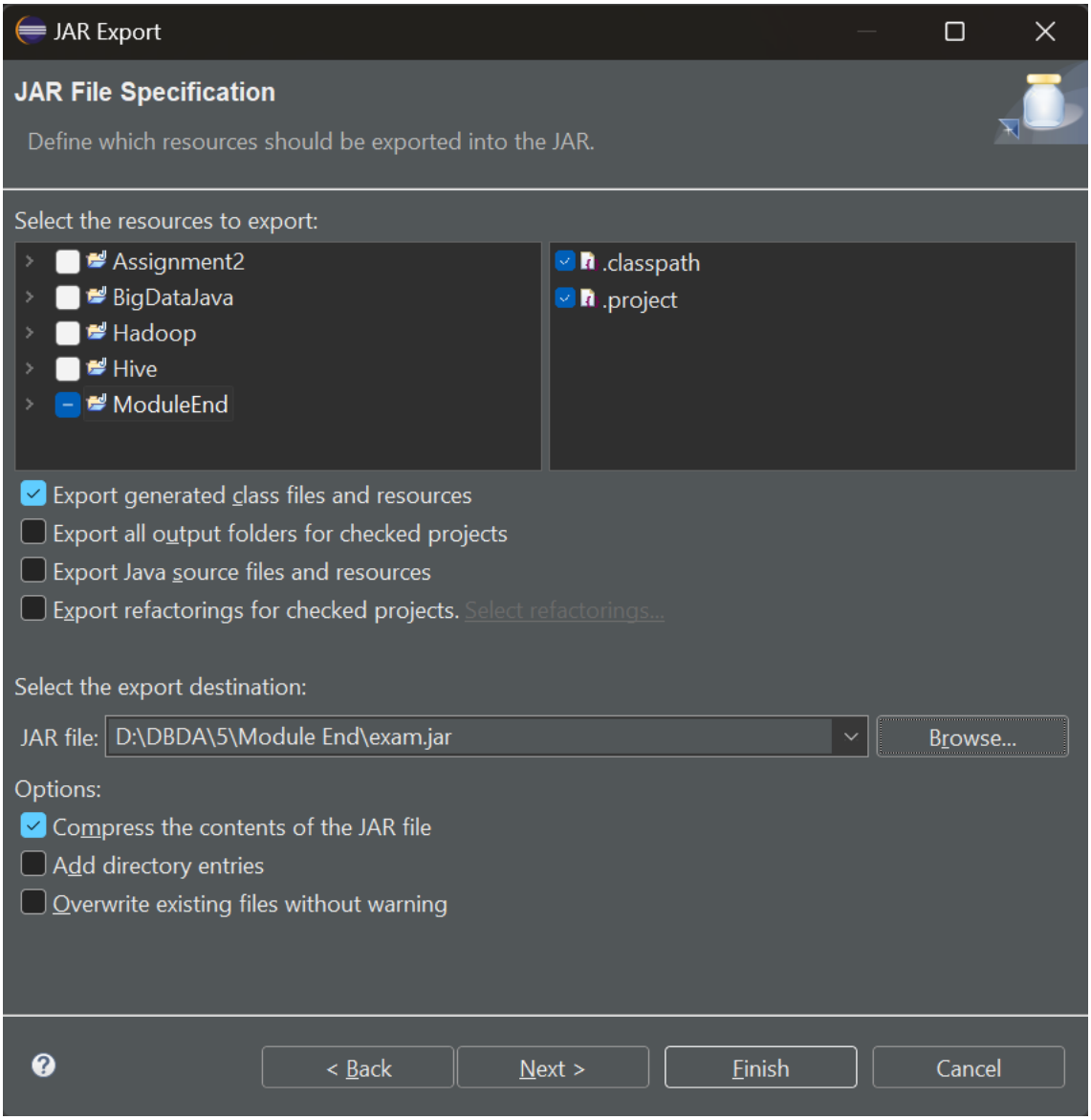
8. Total Volume of the stock on the particular day

9. Adjustment Closing price of the stock

Field Separator – comma

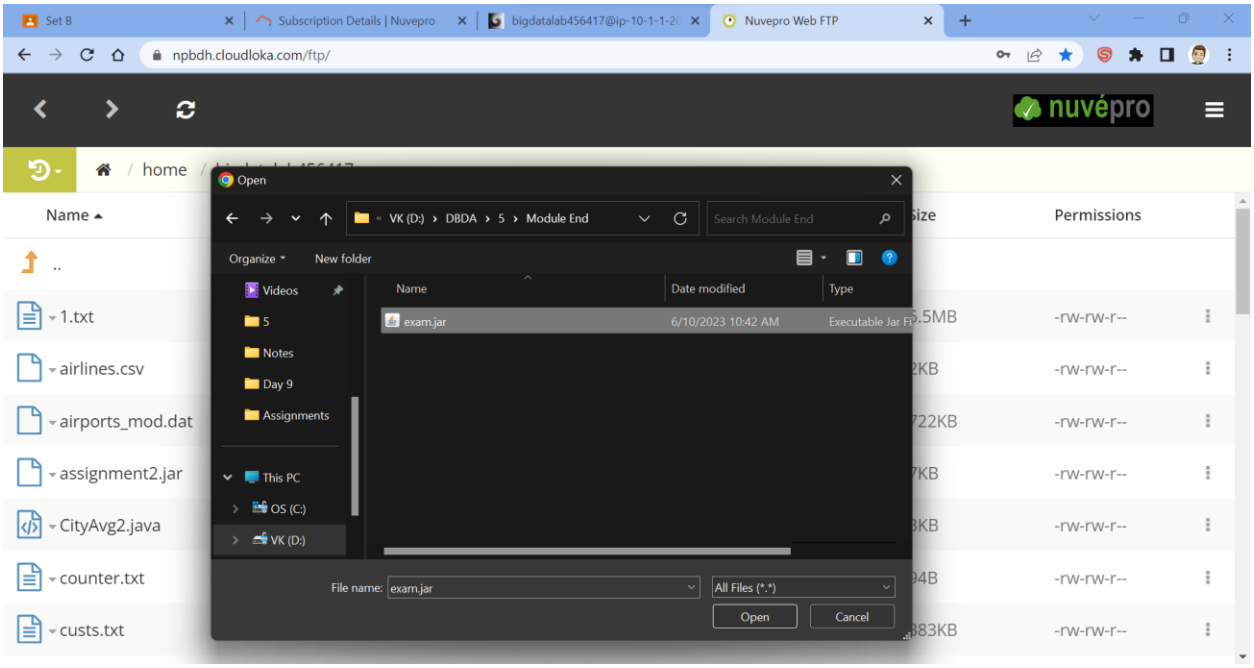
Solution –


Properties for ModuleEnd
type filter text
Resource
Builders
Coverage
Java Build Path
Java Code Style
Java Compiler
Javadoc Location
Java Editor
Project Natures
Project Reference:
Run/Debug Settir
WikiText
Java Build Path
Source
Projects
Libraries
Order and Export
Module Dependencies
JARs and class folders on the build path:
hadoop-common.jar - D:\DBDA\5\Day 4
hadoop-mapreduce-client-core.jar - D:\DBDA\5\Day 4
JRE System Library [JavaSE-1.8]
Add JARs...
Add External JARs...
Add Variable...
Add Library...
Add Class Folder...
Add External Class Folder...
Edit...
Remove
Migrate JAR File...
Apply
Apply and Close
Cancel




```
[bigdatalab456417@ip-10-1-1-204 ~]$ hadoop fs -mkdir exam
[bigdatalab456417@ip-10-1-1-204 ~]$ hadoop fs -put NYSE.csv
exam
```

```
127 login: bigdatalab456417
bigdatalab456417@127.0.0.1's password:
Last login: Sat Jun 10 05:16:02 2023 from localhost
[bigdatalab4564
[bigdatalab456417@ip-10-1-1-204 ~]$ hadoop fs -mkdir exam
[bigdatalab456417@ip-10-1-1-204 ~]$ hadoop fs -put NYSE.csv exam
[bigdatalab456417@ip-10-1-1-204 ~]$
```




```
[bigdatalab456417@ip-10-1-1-204 ~]$ jar -tvf exam.jar
```

```
[bigdatalab456417@ip-10-1-1-204 ~]$ jar -tvf exam.jar
    25 Sat Jun 10 10:42:06 UTC 2023 META-INF/MANIFEST.MF
  2459 Sat Jun 10 10:40:30 UTC 2023 AllTimeHigh$MapClass.class
  2381 Sat Jun 10 10:40:30 UTC 2023 AllTimeHigh$ReduceClass.class
  1721 Sat Jun 10 10:40:30 UTC 2023 AllTimeHigh.class
   556 Sat Jun 10 10:38:52 UTC 2023 .classpath
   385 Sat Jun 10 10:36:02 UTC 2023 .project
```

[bigdatalab456417@ip-10-1-1-204 ~]$ hadoop jar exam.jar AllTimeHigh exam/NYSE.csv exam/output

```
[bigdatalab456417@ip-10-1-1-204 ~]$ hadoop jar exam.jar AllTimeHigh exam/NYSE.csv exam/output
WARNING: Use "yarn jar" to launch YARN applications.
23/06/10 05:26:54 INFO client.RMProxy: Connecting to ResourceManager at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/06/10 05:26:54 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and exe
cute your application with ToolRunner to remedy this.
23/06/10 05:26:55 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/bigdatalab456417/.staging/job_1685754149182_3
405
23/06/10 05:26:55 INFO input.FileInputFormat: Total input files to process : 1
23/06/10 05:26:56 INFO mapreduce.JobSubmitter: number of splits:1
23/06/10 05:26:56 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.sys
tem-metrics-publisher.enabled
23/06/10 05:26:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1685754149182_3405
23/06/10 05:26:56 INFO mapreduce.JobSubmitter: Executing with tokens: []
23/06/10 05:26:56 INFO conf.Configuration: resource-types.xml not found
23/06/10 05:26:56 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
23/06/10 05:26:56 INFO impl.YarnClientImpl: Submitted application application_1685754149182_3405
23/06/10 05:26:56 INFO mapreduce.Job: The url to track the job: http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/application_1685
754149182_3405/
23/06/10 05:26:56 INFO mapreduce.Job: Running job: job_1685754149182_3405
23/06/10 05:27:41 INFO mapreduce.Job: Job job_1685754149182_3405 running in uber mode : false
23/06/10 05:27:41 INFO mapreduce.Job:  map 0% reduce 0%
23/06/10 05:28:07 INFO mapreduce.Job:  map 100% reduce 0%
23/06/10 05:28:18 INFO mapreduce.Job:  map 100% reduce 100%
23/06/10 05:28:20 INFO mapreduce.Job: Job job_1685754149182_3405 completed successfully
23/06/10 05:28:20 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=2738889
                FILE: Number of bytes written=5922985
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=40990982
                HDFS: Number of bytes written=1998
```

```
Job Counters
        Launched map tasks=1
        Launched reduce tasks=1
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=23348
        Total time spent by all reduces in occupied slots (ms)=7576
        Total time spent by all map tasks (ms)=23348
        Total time spent by all reduce tasks (ms)=7576
        Total vcore-milliseconds taken by all map tasks=23348
        Total vcore-milliseconds taken by all reduce tasks=7576
        Total megabyte-milliseconds taken by all map tasks=23908352
        Total megabyte-milliseconds taken by all reduce tasks=7757824
Map-Reduce Framework
        Map input records=735026
        Map output records=735026
        Map output bytes=8781587
        Map output materialized bytes=2738885
        Input split bytes=120
        Combine input records=0
        Combine output records=0
        Reduce input groups=203
        Reduce shuffle bytes=2738885
        Reduce input records=735026
        Reduce output records=203
        Spilled Records=1470052
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=530
        CPU time spent (ms)=7560
        Physical memory (bytes) snapshot=927473664
        Virtual memory (bytes) snapshot=5182926848
        Total committed heap usage (bytes)=1104150528
```

Home / user / bigdatalab456417 / **exam**

Trash

Name	Size	User	Group	Permissions	Date
↑		bigdatalab456417	bigdatalab456417	drwxr-xr-x	June 09, 2023 10:20 PM
.		bigdatalab456417	bigdatalab456417	drwxr-xr-x	June 09, 2023 10:27 PM
NYSE.csv	39.1 MB	bigdatalab456417	bigdatalab456417	-rw-r--r--	June 09, 2023 10:20 PM
output		bigdatalab456417	bigdatalab456417	drwxr-xr-x	June 09, 2023 10:28 PM

Show 45 of 2 items

Page 1 of 1

Home / user / bigdatalab456417 / exam / **output** Trash

Name	Size	User	Group	Permissions	Date
↑		bigdatalab456417	bigdatalab456417	drwxr-xr-x	June 09, 2023 10:27 PM
.		bigdatalab456417	bigdatalab456417	drwxr-xr-x	June 09, 2023 10:28 PM
_SUCCESS	0 bytes	bigdatalab456417	bigdatalab456417	-rw-r--r--	June 09, 2023 10:28 PM
part-r-00000	2.0 KB	bigdatalab456417	bigdatalab456417	-rw-r--r--	June 09, 2023 10:28 PM

Show 45 of 2 items Page 1 of 1

File Browser

Back

Edit file

Refresh

View as binary

Download

Last modified
06/10/2023
10:58 AM +05:30

User
bigdatalab456417

Group
bigdatalab456417

Size
1.95 KB

Mode
100644

Home Page 1 to 1 of 1

/ user / bigdatalab456417 / exam / output / **part-r-00000**

```
ACO     42.7
ACS     109.55
ACV     65.32
ADC     37.7
ADI     185.5
ADM     48.95
ADP     84.31
ADS     80.79
ADX     40.56
ADY     44.0
AEA     23.94
AEB     26.5
AEC     17.6
AED     26.12
AEE     56.77
AEF     27.0
AEG     148.32
AEH     26.64
AEL     14.6
AEM     83.45
AEO     88.13
```

Hive

```
hive

hive> set hive.cli.print.current.db = true;

hive (default)> use vaibhav_training;
```

```
hduser@vk-VirtualBox:~$ hive
ls: cannot access '/usr/local/spark/lib/spark-assembly-*.jar': No such file or directory

Logging initialized using configuration in jar:file:/usr/local/hive/lib/hive-common-1.2.1.jar!/hive-log4j
.properties
hive> set hive.cli.print.current.db = true;
hive (default)> use vaibhav;
OK
Time taken: 0.72 seconds
```

1. Which airports have the lowest altitude?

```
hive (vaibhav)> select name,altitude from airport order by altitude limit 1;
Query ID = hduser_20230610125052_9a7ffdfa-1b20-4aa2-ab95-0415ce357a40
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>
Job running in-process (local Hadoop)
2023-06-10 12:50:53,327 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1096305236_0008
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 20809828 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
I Bar Yehuda    -1266
Time taken: 1.19 seconds, Fetched: 1 row(s)
hive (vaibhav)> S
```

2. How many routes are operated by active airlines from Ghana?

```
hive (vaibhav)> select count(r.src_airport_id) from routes r join airlines a on a.airline_id=r.airline_id where active='
Y' and  trim(upper(a.country))='Ghana';
Query ID = hduser_20230610125202_4f5fc2af-aa85-4345-ad82-920398a33331
Total jobs = 1
23/06/10 12:52:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
Execution log at: /tmp/hduser/hduser_20230610125202_4f5fc2af-aa85-4345-ad82-920398a33331.log
2023-06-10 12:52:04     Starting to launch local task to process map join;      maximum memory = 477626368
2023-06-10 12:52:04     Dump the side-table for tag: 1 with group count: 0 into file: file:/usr/local/hive/iotmp/8ec82f7
7-729b-42c5-a63b-8d5b8e60cd6d/hive_2023-06-10_12-52-02_462_6672056126664212614-1/-local-10004/HashTable-Stage-2/MapJoin-
mapfile31--.hashtable
2023-06-10 12:52:04     Uploaded 1 File to: file:/usr/local/hive/iotmp/8ec82f77-729b-42c5-a63b-8d5b8e60cd6d/hive_2023-06
-10_12-52-02_462_6672056126664212614-1/-local-10004/HashTable-Stage-2/MapJoin-mapfile31--.hashtable (260 bytes)
2023-06-10 12:52:04     End of local task; Time Taken: 0.676 sec.
Execution completed successfully
MapredLocal task succeeded
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>
Job running in-process (local Hadoop)
2023-06-10 12:52:06,124 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local1921556474_0009
MapReduce Jobs Launched:
Stage-Stage-2:  HDFS Read: 25560838 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
0
Time taken: 3.666 seconds, Fetched: 1 row(s)
```

3. Which airlines operate routes that have less than 3 stops number of stop bottom 10 alphabetically?

```
hive (vaibhav)> select distinct a.name from airlines a join routes r on a.airline_id=r.airline_id where stops<3 order by
 a.name desc limit 10;
Query ID = hduser_20230610125306_65058700-1e3f-4d5f-8aa6-32d401f0b470
Total jobs = 2
23/06/10 12:53:07 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-ja
va classes where applicable
Execution log at: /tmp/hduser/hduser_20230610125306_65058700-1e3f-4d5f-8aa6-32d401f0b470.log
2023-06-10 12:53:07     Starting to launch local task to process map join;      maximum memory = 477626368
2023-06-10 12:53:08     Dump the side-table for tag: 0 with group count: 6048 into file: file:/usr/local/hive/iotmp/8ec8
2f77-729b-42c5-a63b-8d5b8e60cd6d/hive_2023-06-10_12-53-06_248_8947129066464320305-1/-local-10005/HashTable-Stage-2/MapJo
in-mapfile40--.hashtable
2023-06-10 12:53:08     Uploaded 1 File to: file:/usr/local/hive/iotmp/8ec82f77-729b-42c5-a63b-8d5b8e60cd6d/hive_2023-06
-10_12-53-06_248_8947129066464320305-1/-local-10005/HashTable-Stage-2/MapJoin-mapfile40--.hashtable (237862 bytes)
2023-06-10 12:53:08     End of local task; Time Taken: 0.678 sec.
Execution completed successfully
MapredLocal task succeeded
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>
Job running in-process (local Hadoop)
2023-06-10 12:53:09,752 Stage-2 map = 100%,  reduce = 100%
Ended Job = job_local512762733_0010
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>
```

```
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>
Job running in-process (local Hadoop)
2023-06-10 12:53:10,871 Stage-3 map = 100%,  reduce = 100%
Ended Job = job_local1969354263_0011
MapReduce Jobs Launched:
Stage-Stage-2:  HDFS Read: 30311848 HDFS Write: 0 SUCCESS
Stage-Stage-3:  HDFS Read: 30311848 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
easyJet
bmibaby
Zoom Airlines
Zip
Zest Air
Zambia Skyways
ZABAIKAL AIRLINES
Yeti Airways
Yemenia
Yangon Airways
Time taken: 4.624 seconds, Fetched: 10 row(s)
hive (vaibhav)>
```

4. How many airlines have a specific IATA code ‘Q5’?

```
hive (vaibhav)> select count(airline_id) from airlines where iata = "Q5";
Query ID = hduser_20230610124335_935647e2-388f-4863-a80b-270bc4a056b7
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>
Job running in-process (local Hadoop)
2023-06-10 12:43:36,814 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local69489336_0004
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 11621728 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1
Time taken: 1.259 seconds, Fetched: 1 row(s)
hive (vaibhav)>
```

```
hive (vaibhav)> select count(name) from airlines where iata='Q5';
Query ID = hduser_20230610125511_f703cd51-002d-452e-a3b9-09a447a6819f
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>
Job running in-process (local Hadoop)
2023-06-10 12:55:12,673 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1374024852_0012
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 30944334 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1
Time taken: 1.203 seconds, Fetched: 1 row(s)
hive (vaibhav)>
```

5. Find the airlines that operate routes with a specific equipment as ‘A81’ or codeshare enabled.

```
select a.name from airlines a join routes r on
a.airline_id=r.airline_id where equipment='A81';
```

Pyspark

```
[bigdatalab456417@ip-10-1-1-204 ~]$ pyspark
```

```
[bigdatalab456417@ip-10-1-1-204 ~]$ pyspark
Python 3.7.6 (default, Jan  8 2020, 19:59:22)
[GCC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/06/10 05:40:07 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.0-cdh6.2.1
      /_/

Using Python version 3.7.6 (default, Jan  8 2020 19:59:22)
SparkSession available as 'spark'.
>>> 
```

```
>>> from pyspark.sql.types import StructType, StringType, IntegerType, DoubleType, LongType
```

```
>>> schema9 = StructType().add("year",IntegerType(),True).add("quarter",IntegerType(),True).add("rev",DoubleType(),True).add("seats",IntegerType(),True)
```

df =spark.read.format("csv").option("header","False").schema(schema9).load("hdfs://nameservice1/user/bigdatalab456417/training/airlines.csv")

df_air.registerTempTable("airlines")

1.What is the total revenue generated in each year?

df=spark.sql(select year,sum(avrs) from airlines group by year)

df.show()

2. Which year had the highest average revenue per seat??

df=spark.sql("SELECT year, quarter, avg(rev) AS avg_arps from airlines group by year, quarter Order by avg_arps desc limit 1")

df.show()

3. What is the total number of booked seats for each quarter in a given year?

```
df=spark.sql("SELECT year, quarter, SUM(seats) AS total_tickets from airlines group by year, quarter”)

df.show()
```