**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

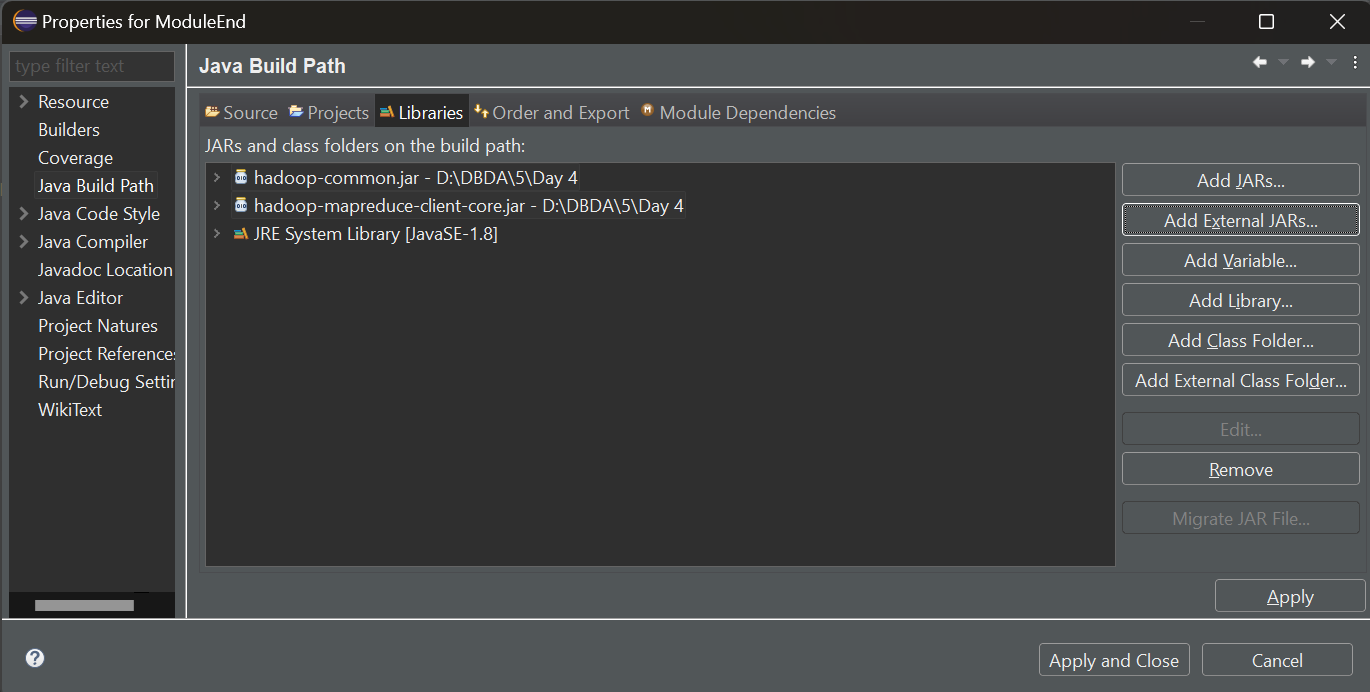
7. Closing price of the stock

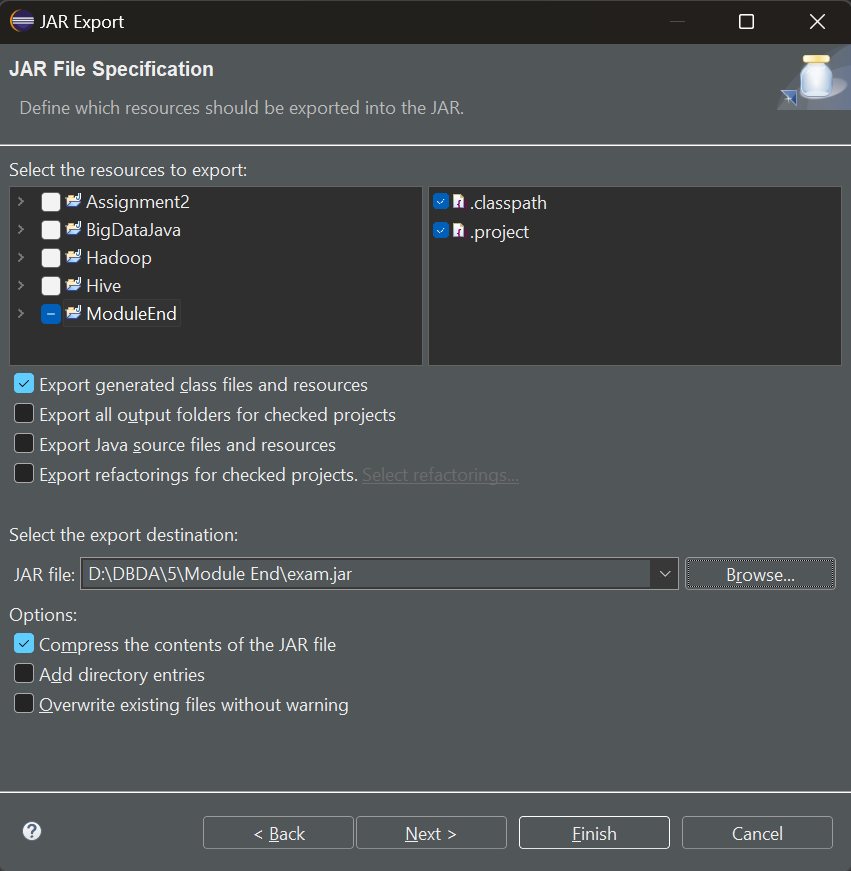
8. Total Volume of the stock on the particular day

9. Adjustment Closing price of the stock

Field Separator – comma

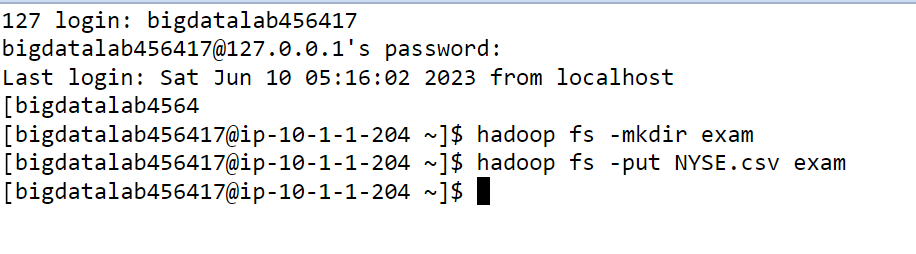
Solution –

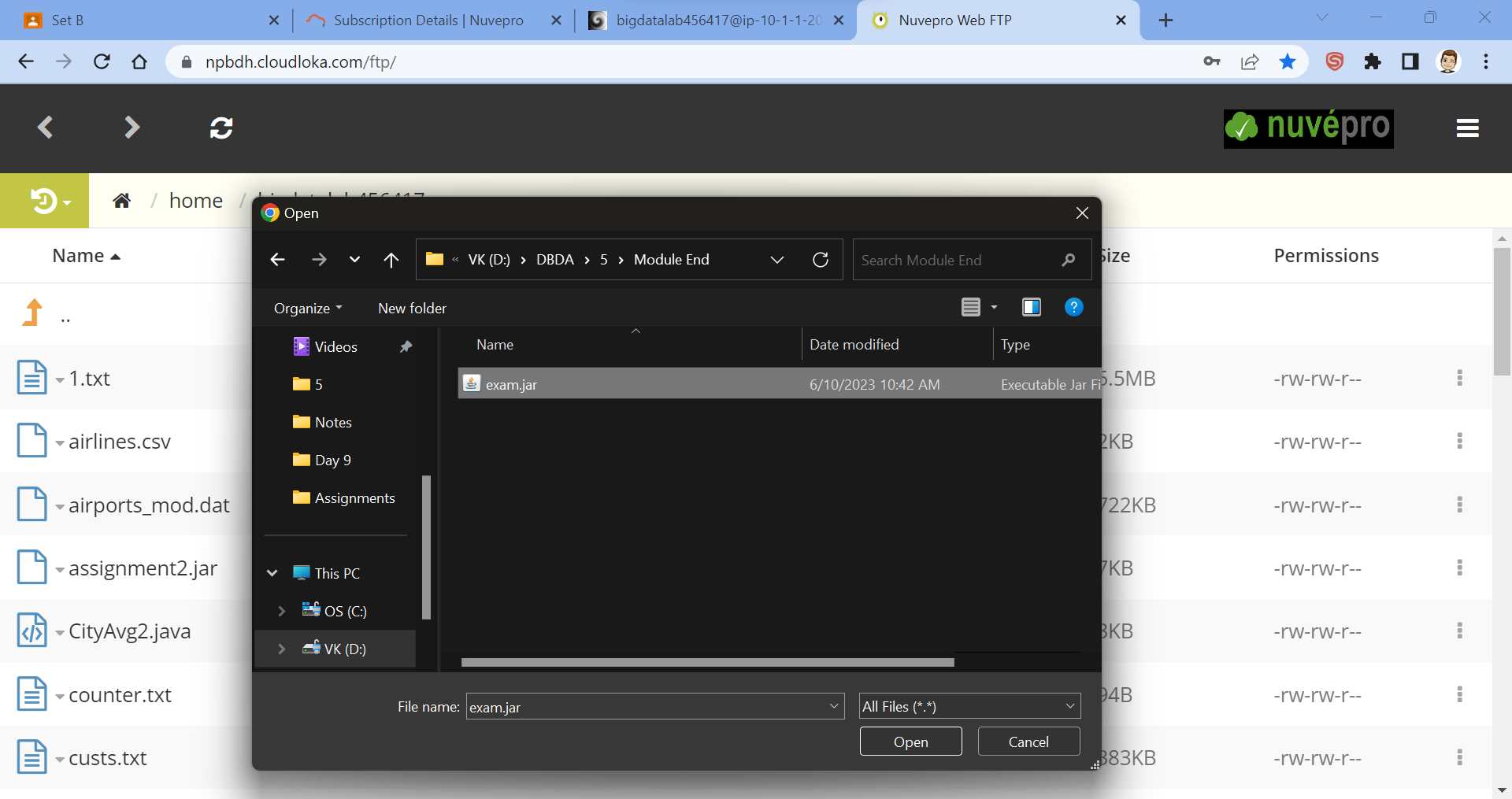




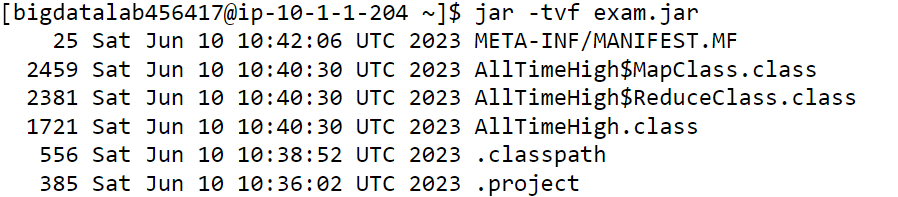
[bigdatalab456417@ip-10-1-1-204 ~]$ hadoop fs -mkdir exam

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

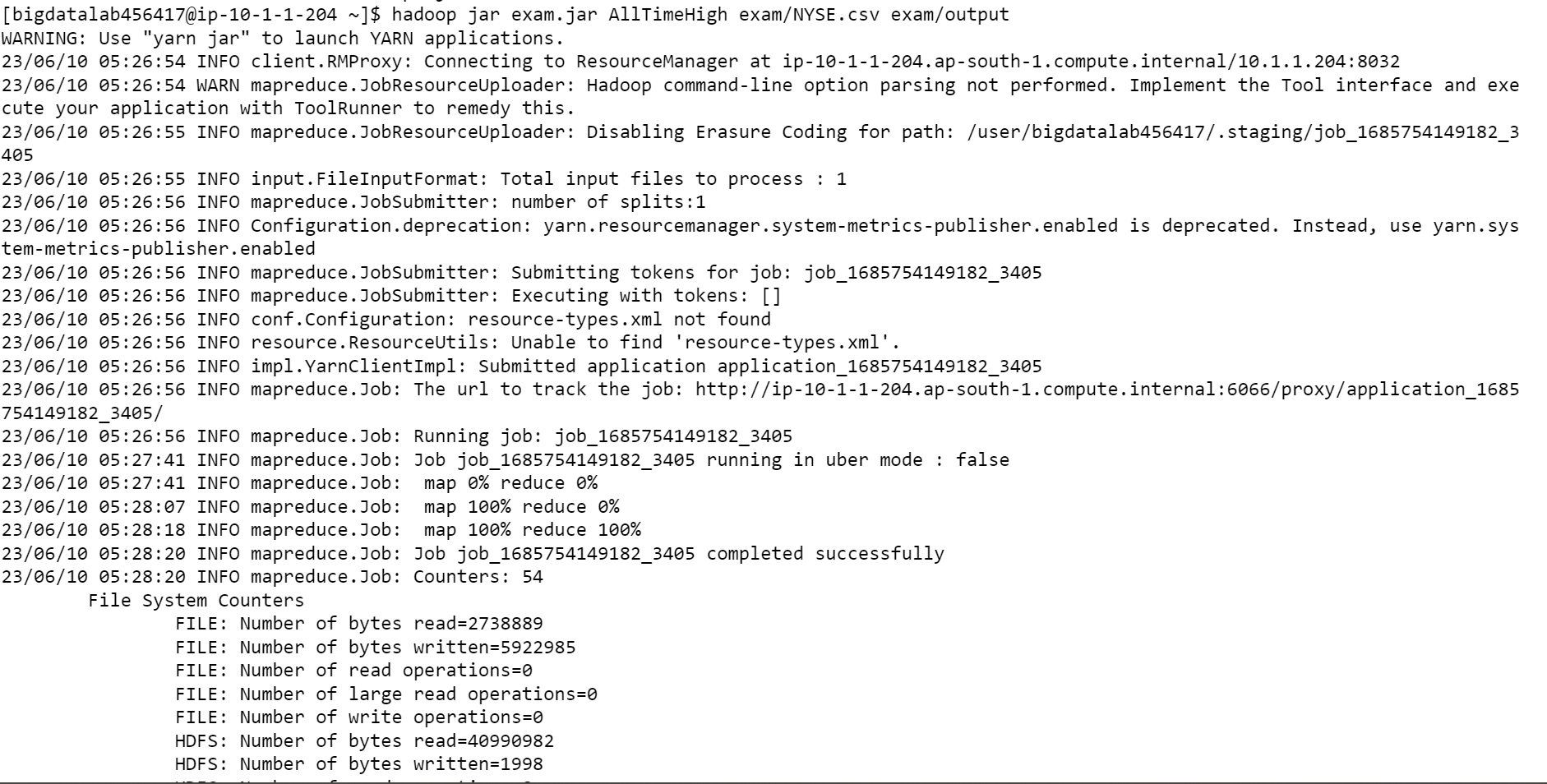


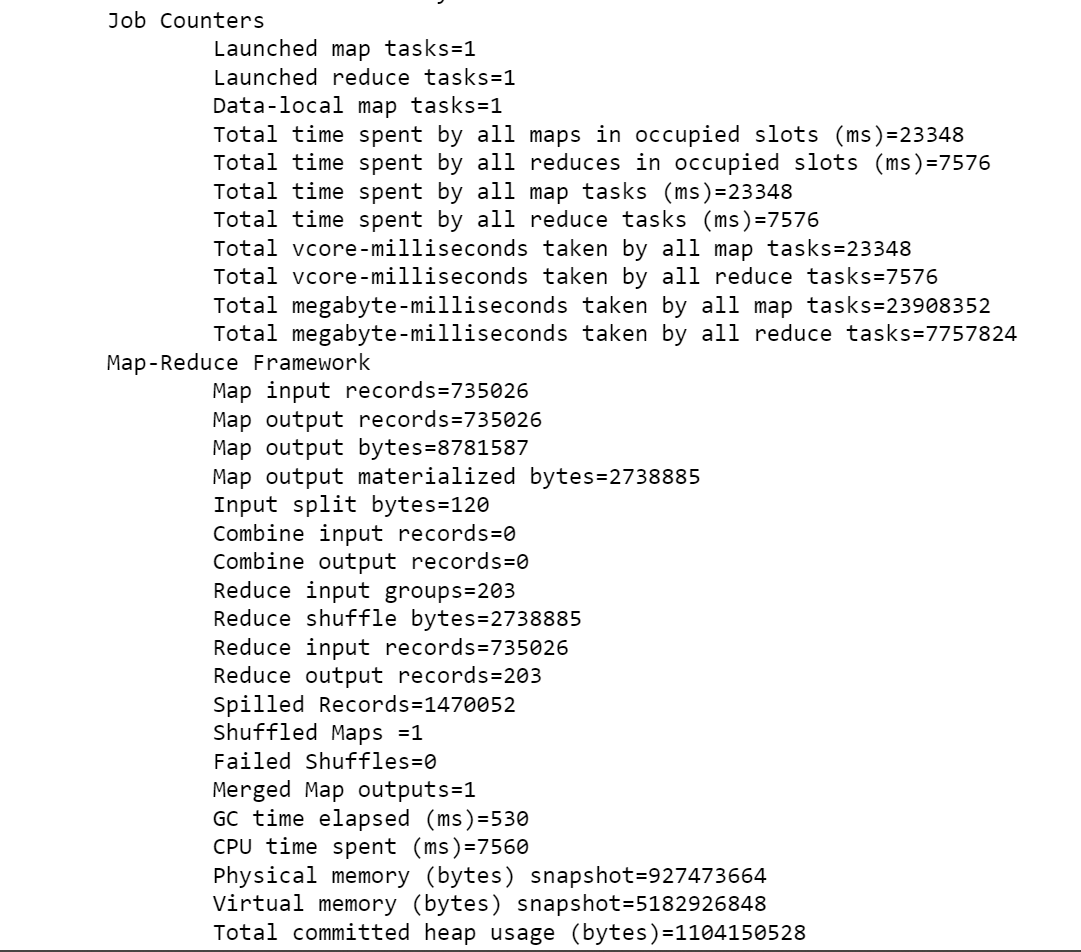


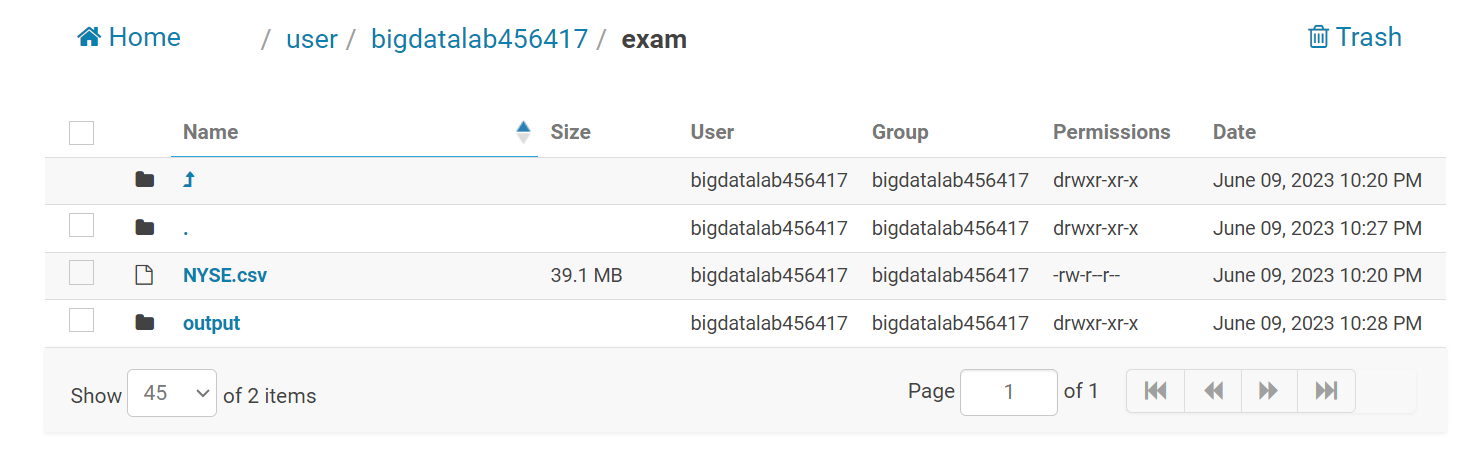
[bigdatalab456417@ip-10-1-1-204 ~]$ jar -tvf exam.jar

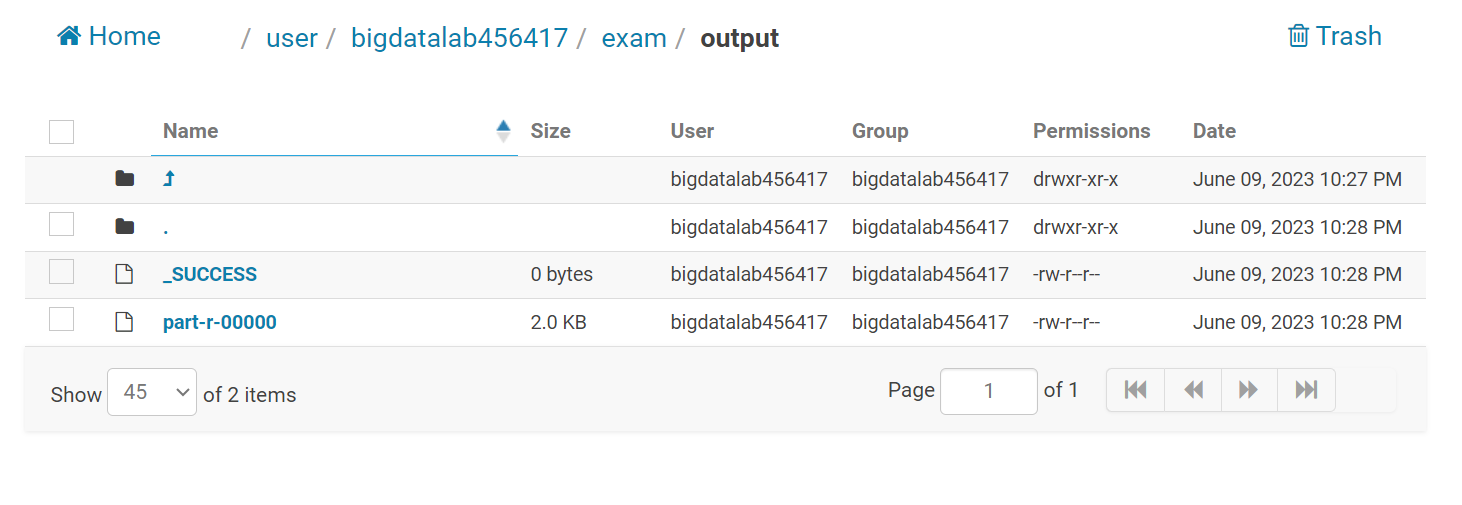


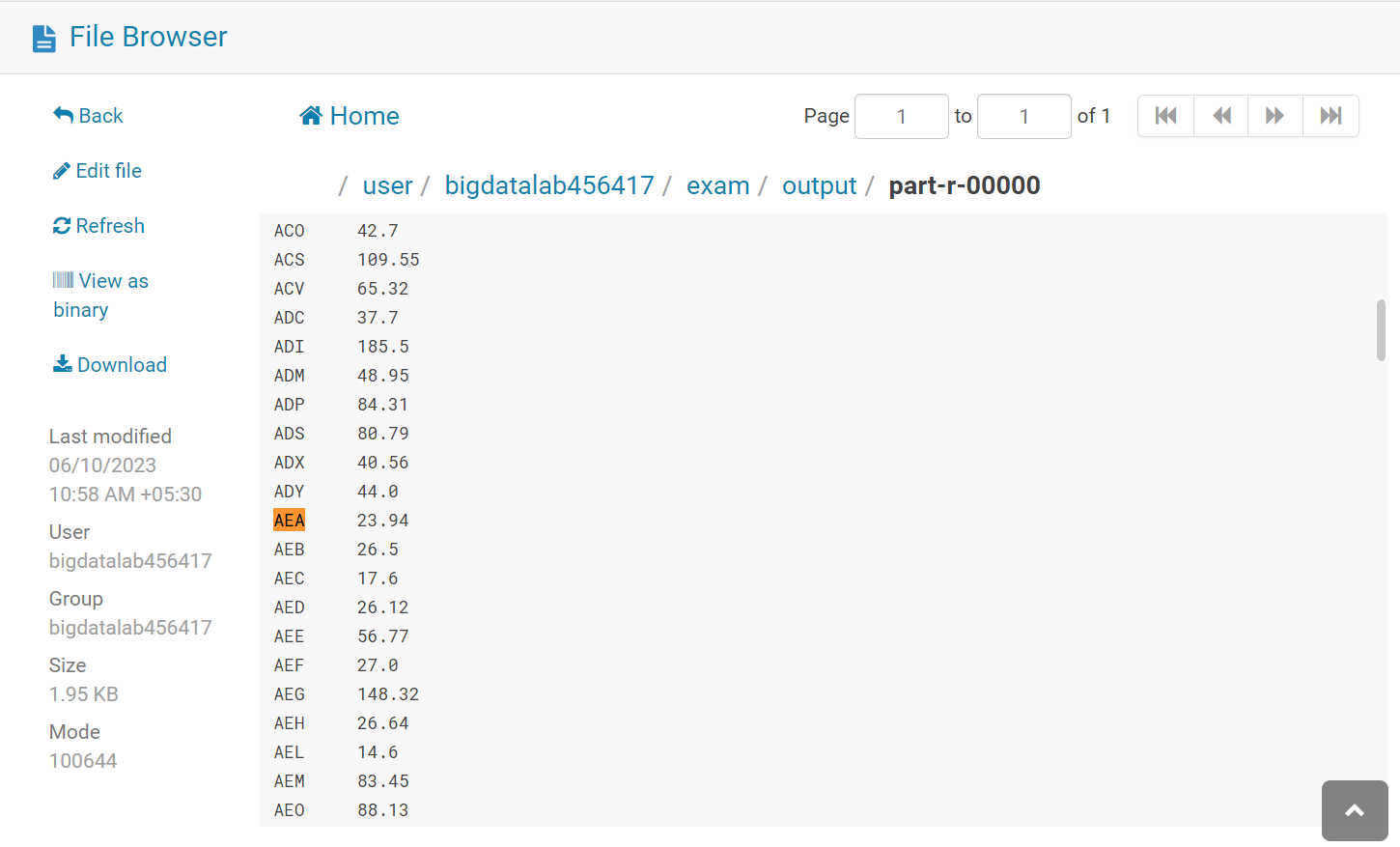
[bigdatalab456417@ip-10-1-1-204 ~]$ hadoop jar exam.jar AllTimeHigh exam/NYSE.csv exam/output









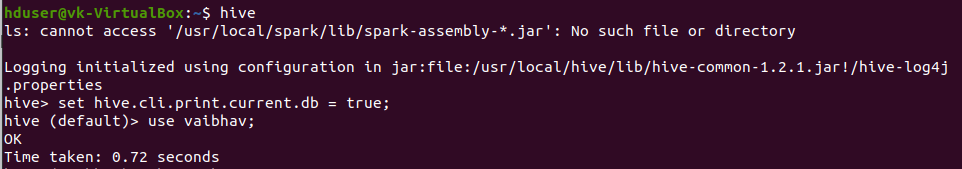


**Hive**

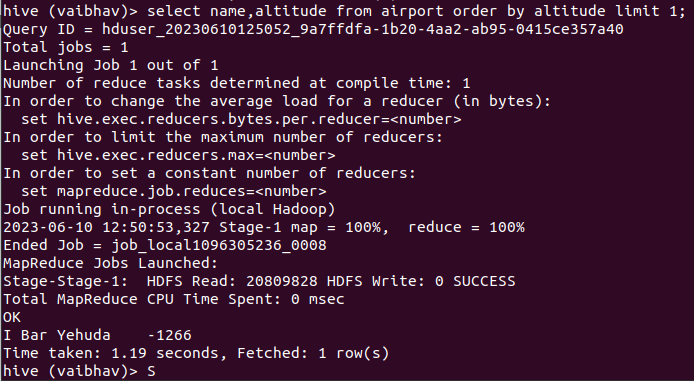
hive

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

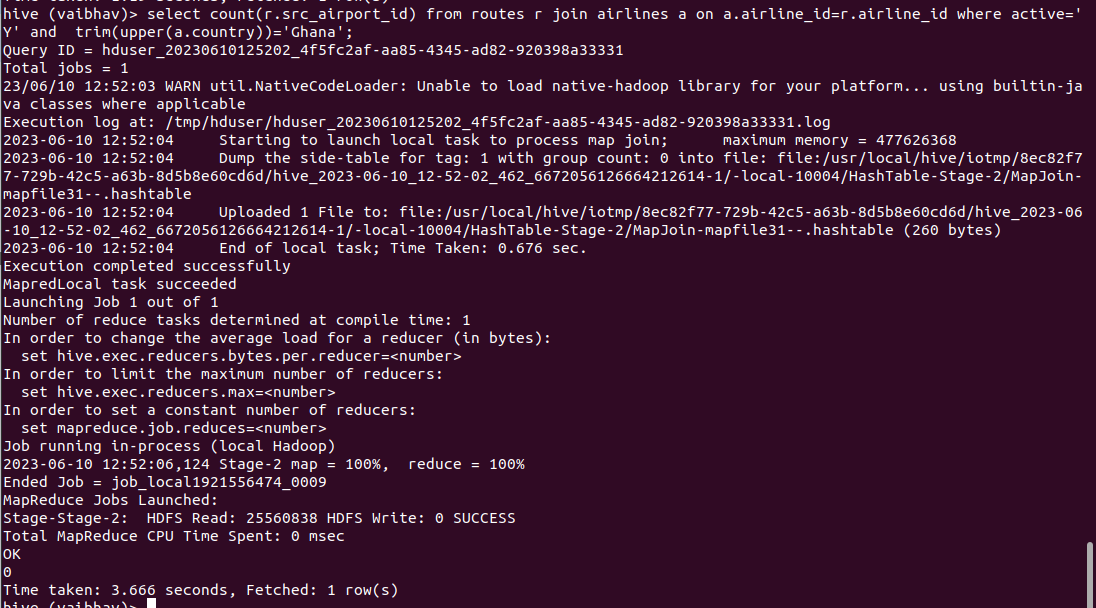
hive (default)> use vaibhav\_training;



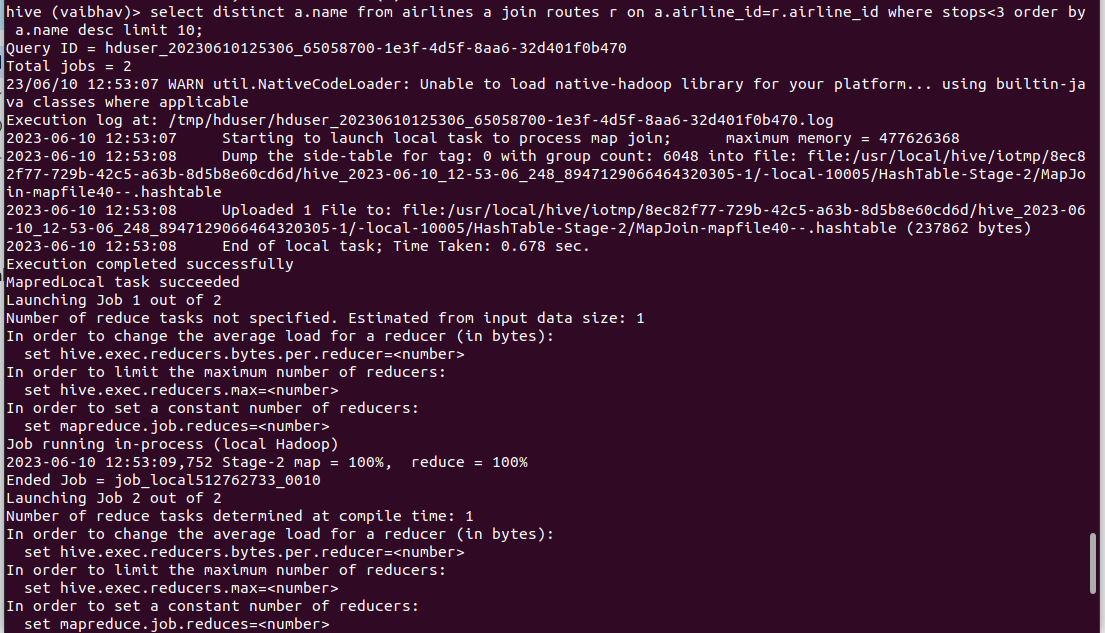
1. Which airports have the lowest altitude?

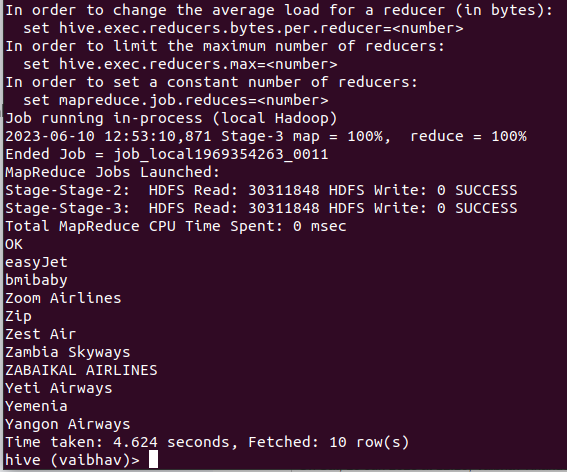


1. How many routes are operated by active airlines from Ghana?

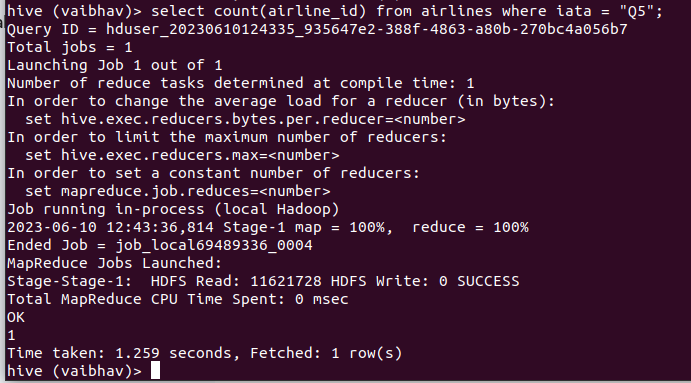


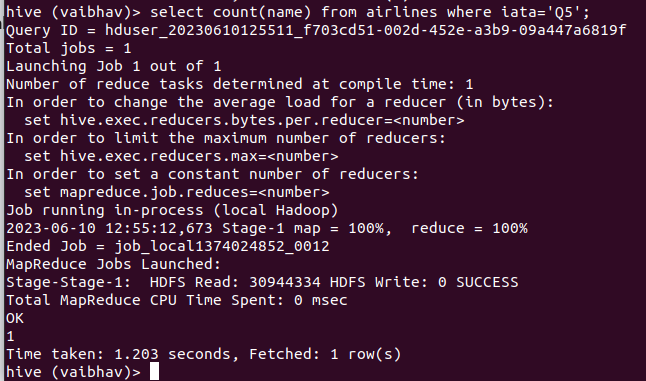
1. Which airlines operate routes that have less than 3 stops number of stop bottom 10 alphabetically?





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



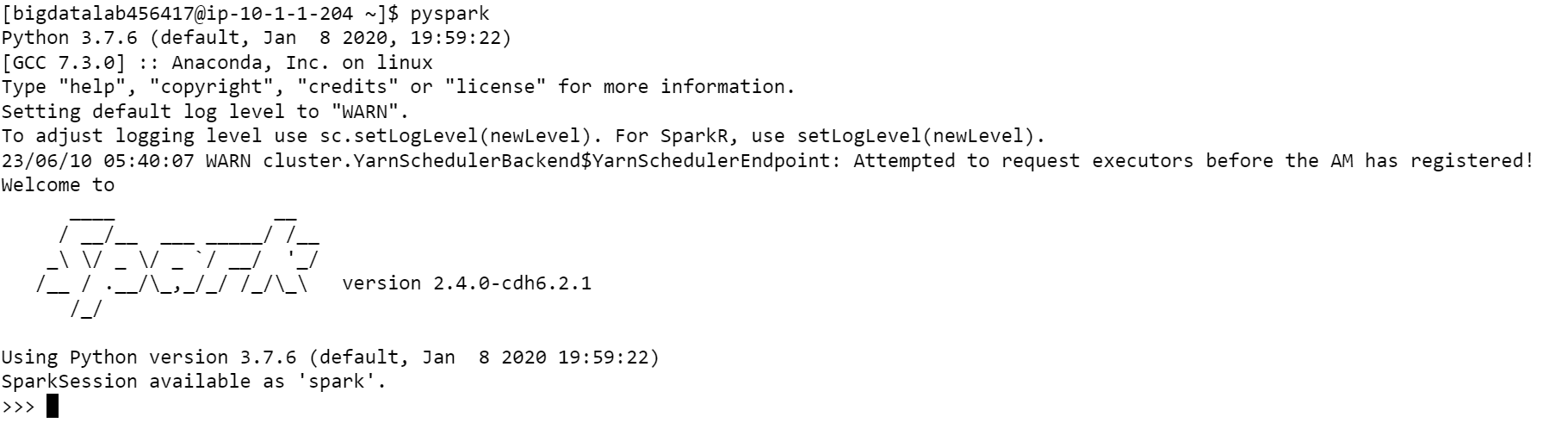


1. Find the airlines that operate routes with a specific equipment as ‘A81’ or codeshare enabled.

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



>>> 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",Integer

Type(),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()