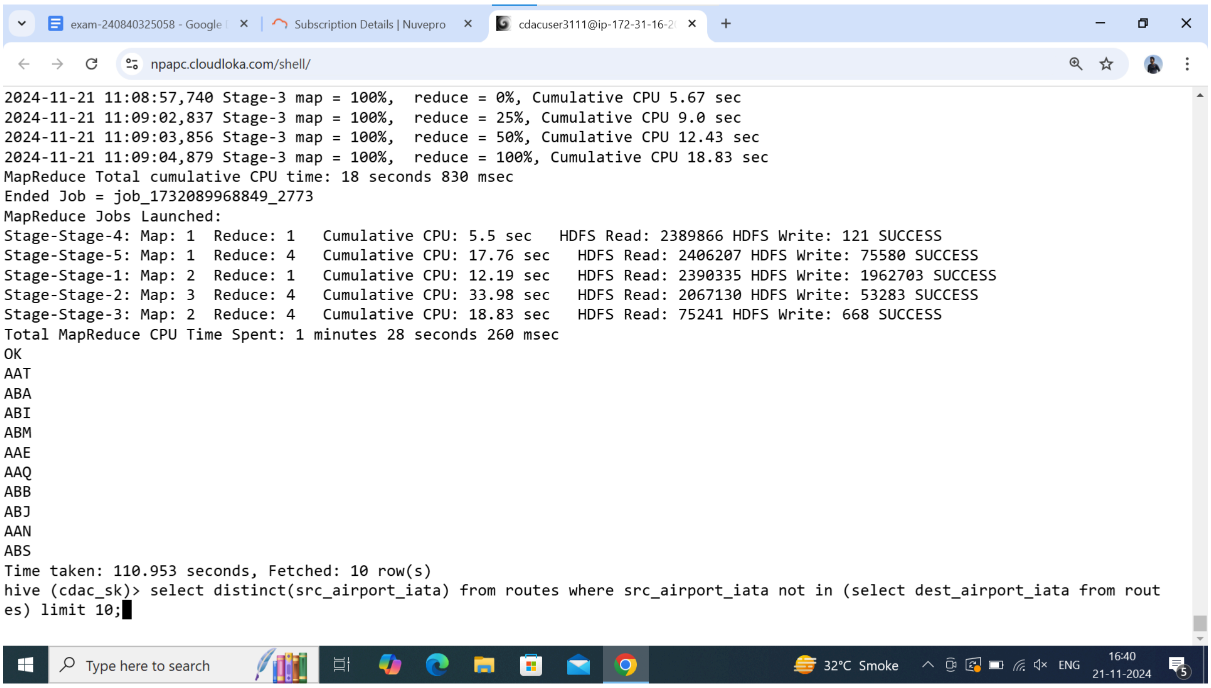
Q1.1.

```
select distinct(src_airport_iata) from routes where
src_airport_iata not in (select dest_airport_iata from rout
es) limit 10;
```

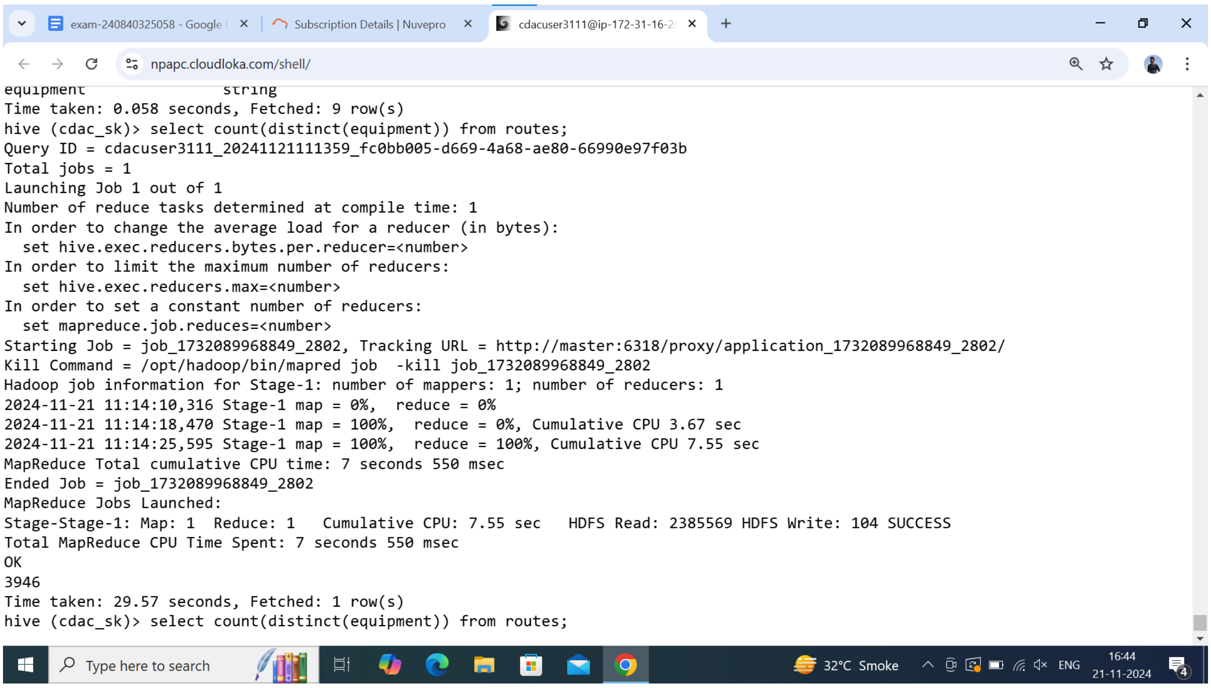

exam-240840325058 - Google
Subscription Details | Nuvepro
cdacuser3111@ip-172-31-16-2
npapc.cloudloka.com/shell/
Type here to search
32°C Smoke
ENG
16:40
21-11-2024

```
2024-11-21 11:08:57,740 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 5.67 sec
2024-11-21 11:09:02,837 Stage-3 map = 100%,  reduce = 25%, Cumulative CPU 9.0 sec
2024-11-21 11:09:03,856 Stage-3 map = 100%,  reduce = 50%, Cumulative CPU 12.43 sec
2024-11-21 11:09:04,879 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 18.83 sec
MapReduce Total cumulative CPU time: 18 seconds 830 msec
Ended Job = job_1732089968849_2773
MapReduce Jobs Launched:
Stage-Stage-4: Map: 1  Reduce: 1   Cumulative CPU: 5.5 sec   HDFS Read: 2389866 HDFS Write: 121 SUCCESS
Stage-Stage-5: Map: 1  Reduce: 4   Cumulative CPU: 17.76 sec   HDFS Read: 2406207 HDFS Write: 75580 SUCCESS
Stage-Stage-1: Map: 2  Reduce: 1   Cumulative CPU: 12.19 sec   HDFS Read: 2390335 HDFS Write: 1962703 SUCCESS
Stage-Stage-2: Map: 3  Reduce: 4   Cumulative CPU: 33.98 sec   HDFS Read: 2067130 HDFS Write: 53283 SUCCESS
Stage-Stage-3: Map: 2  Reduce: 4   Cumulative CPU: 18.83 sec   HDFS Read: 75241 HDFS Write: 668 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 28 seconds 260 msec
OK
AAT
ABA
ABI
ABM
AAE
AAQ
ABB
ABJ
AAN
ABS
Time taken: 110.953 seconds, Fetched: 10 row(s)
hive (cdac_sk)> select distinct(src_airport_iata) from routes where src_airport_iata not in (select dest_airport_iata from rout
es) limit 10;
```

Q1.2.

Q1.3.

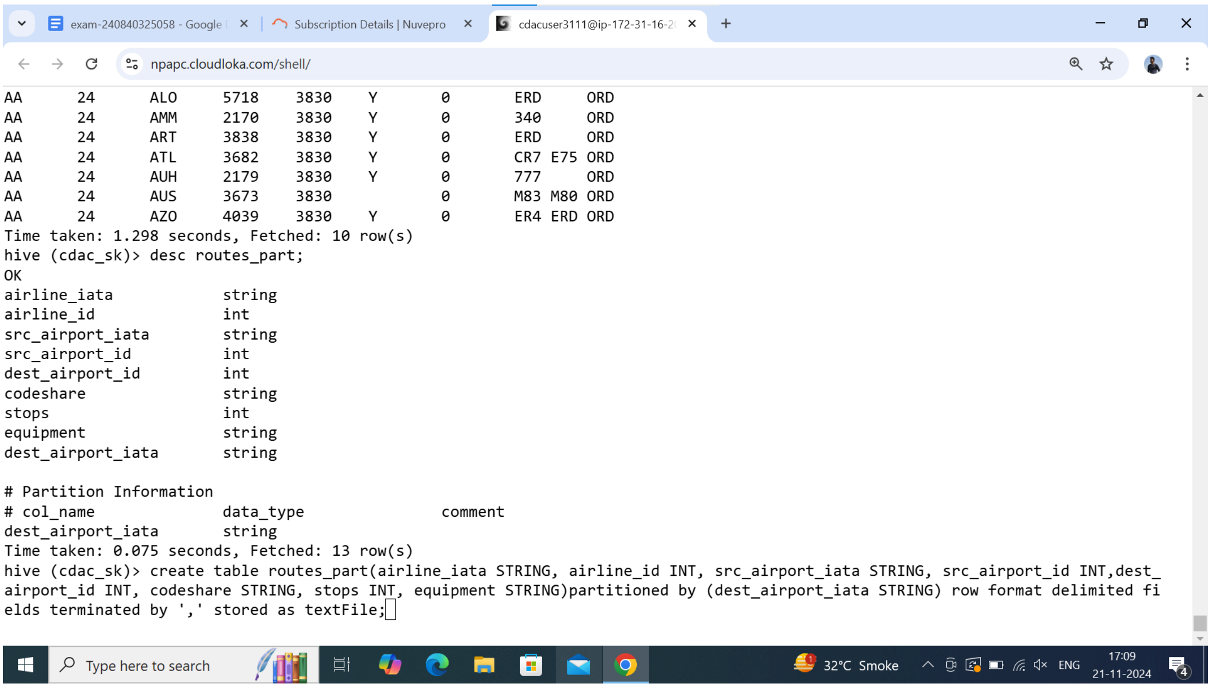
```
select count(distinct(equipment)) from routes;
```


exam-240840325058 - Google
Subscription Details | Nuvepro
cdacuser3111@ip-172-31-16-2
npapc.cloudloka.com/shell/
Type here to search
32°C Smoke
ENG
16:44
21-11-2024

```
equipment               string
Time taken: 0.058 seconds, Fetched: 9 row(s)
hive (cdac_sk)> select count(distinct(equipment)) from routes;
Query ID = cdacuser3111_20241121111359_fc0bb005-d669-4a68-ae80-66990e97f03b
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_1732089968849_2802, Tracking URL = http://master:6318/proxy/application_1732089968849_2802/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2802
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2024-11-21 11:14:10,316 Stage-1 map = 0%,  reduce = 0%
2024-11-21 11:14:18,470 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.67 sec
2024-11-21 11:14:25,595 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.55 sec
MapReduce Total cumulative CPU time: 7 seconds 550 msec
Ended Job = job_1732089968849_2802
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.55 sec   HDFS Read: 2385569 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 550 msec
OK
3946
Time taken: 29.57 seconds, Fetched: 1 row(s)
hive (cdac_sk)> select count(distinct(equipment)) from routes;
```

Q2.1

```
create table routes_part(airline_iata STRING, airline_id INT,
src_airport_iata STRING, src_airport_id INT,dest_
airport_id INT, codeshare STRING, stops INT, equipment
STRING)partitioned by (dest_airport_iata STRING) row format
delimited fields terminated by ',' stored as textFile;
```

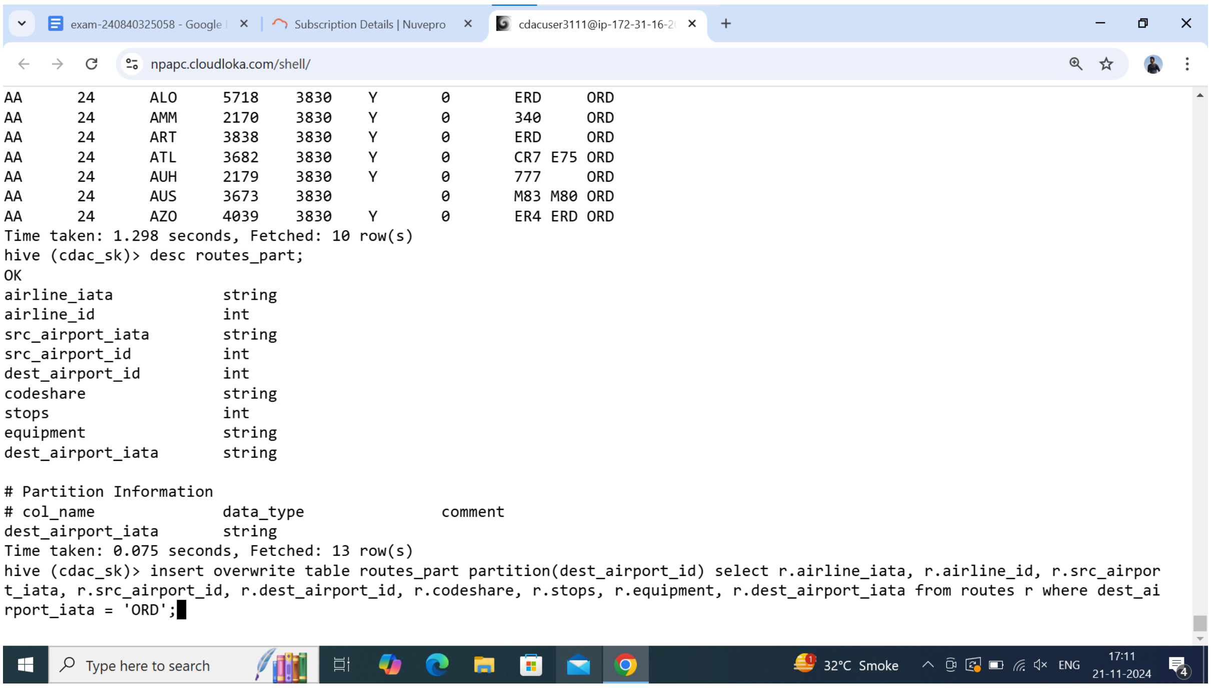

exam-240840325058 - Google
Subscription Details | Nuvepro
cdacuser3111@ip-172-31-16-2
npapc.cloudloka.com/shell/
Type here to search
32°C Smoke
ENG
17:09
21-11-2024

```
AA      24      ALO     5718    3830    Y       0       ERD     ORD
AA      24      AMM     2170    3830    Y       0       340     ORD
AA      24      ART     3838    3830    Y       0       ERD     ORD
AA      24      ATL     3682    3830    Y       0       CR7 E75 ORD
AA      24      AUH     2179    3830    Y       0       777     ORD
AA      24      AUS     3673    3830            0       M83 M80 ORD
AA      24      AZO     4039    3830    Y       0       ER4 ERD ORD
Time taken: 1.298 seconds, Fetched: 10 row(s)
hive (cdac_sk)> desc routes_part;
OK
airline_iata            string
airline_id              int
src_airport_iata        string
src_airport_id          int
dest_airport_id         int
codeshare               string
stops                   int
equipment               string
dest_airport_iata       string

# Partition Information
# col_name              data_type               comment
dest_airport_iata       string
Time taken: 0.075 seconds, Fetched: 13 row(s)
hive (cdac_sk)> create table routes_part(airline_iata STRING, airline_id INT, src_airport_iata STRING, src_airport_id INT,dest_
airport_id INT, codeshare STRING, stops INT, equipment STRING)partitioned by (dest_airport_iata STRING) row format delimited fi
elds terminated by ',' stored as textFile;
```

Q2.2

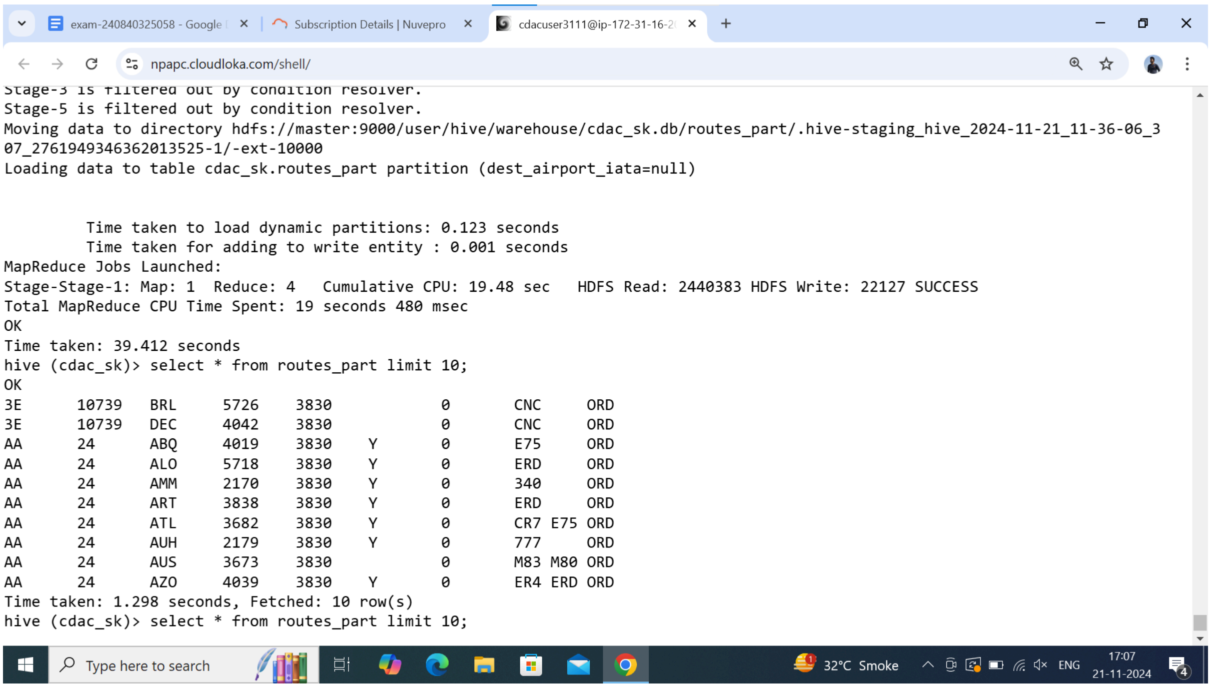
```
insert overwrite table routes_part partition(dest_airport_id)
select r.airline_iata, r.airline_id, r.src_airpor
t_iata, r.src_airport_id, r.dest_airport_id, r.codeshare, r.stops,
r.equipment, r.dest_airport_iata from routes r where dest_ai
rport_iata = 'ORD';
```


exam-240840325058 - Google
Subscription Details | Nuvepro
cdacuser3111@ip-172-31-16-2
npapc.cloudloka.com/shell/
Type here to search
32°C Smoke
ENG
17:11
21-11-2024

```
AA      24      ALO     5718    3830    Y       0       ERD     ORD
AA      24      AMM     2170    3830    Y       0       340     ORD
AA      24      ART     3838    3830    Y       0       ERD     ORD
AA      24      ATL     3682    3830    Y       0       CR7 E75 ORD
AA      24      AUH     2179    3830    Y       0       777     ORD
AA      24      AUS     3673    3830            0       M83 M80 ORD
AA      24      AZO     4039    3830    Y       0       ER4 ERD ORD
Time taken: 1.298 seconds, Fetched: 10 row(s)
hive (cdac_sk)> desc routes_part;
OK
airline_iata            string
airline_id              int
src_airport_iata        string
src_airport_id          int
dest_airport_id         int
codeshare               string
stops                   int
equipment               string
dest_airport_iata       string

# Partition Information
# col_name              data_type               comment
dest_airport_iata       string
Time taken: 0.075 seconds, Fetched: 13 row(s)
hive (cdac_sk)> insert overwrite table routes_part partition(dest_airport_id) select r.airline_iata, r.airline_id, r.src_airpor
t_iata, r.src_airport_id, r.dest_airport_id, r.codeshare, r.stops, r.equipment, r.dest_airport_iata from routes r where dest_ai
rport_iata = 'ORD';
```

Q2.3
Select * from routes_part limit 10;

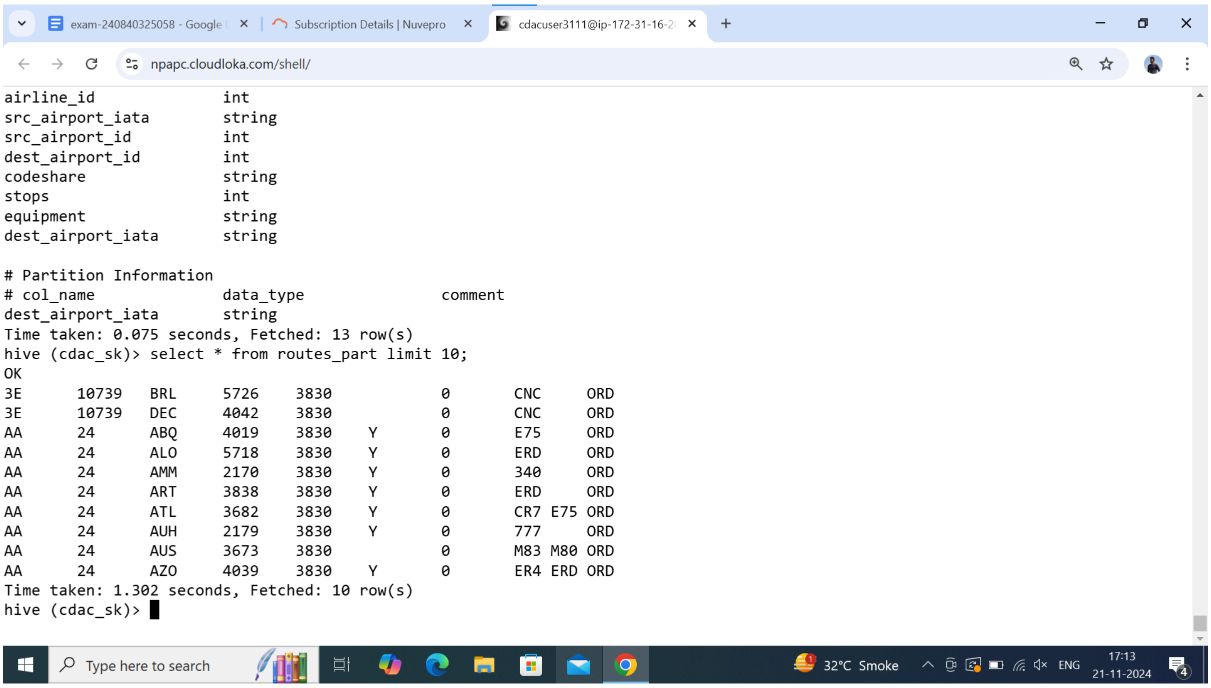

exam-240840325058 - Google
Subscription Details | Nuvepro
cdacuser3111@ip-172-31-16-2
npapc.cloudloka.com/shell/
Type here to search
32°C Smoke
ENG
17:07
21-11-2024

```
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://master:9000/user/hive/warehouse/cdac_sk.db/routes_part/.hive-staging_hive_2024-11-21_11-36-06_3
07_2761949346362013525-1/-ext-10000
Loading data to table cdac_sk.routes_part partition (dest_airport_iata=null)


         Time taken to load dynamic partitions: 0.123 seconds
         Time taken for adding to write entity : 0.001 seconds
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 19.48 sec   HDFS Read: 2440383 HDFS Write: 22127 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 480 msec
OK
Time taken: 39.412 seconds
hive (cdac_sk)> select * from routes_part limit 10;
OK
3E      10739   BRL     5726    3830            0       CNC     ORD
3E      10739   DEC     4042    3830            0       CNC     ORD
AA      24      ABQ     4019    3830    Y       0       E75     ORD
AA      24      ALO     5718    3830    Y       0       ERD     ORD
AA      24      AMM     2170    3830    Y       0       340     ORD
AA      24      ART     3838    3830    Y       0       ERD     ORD
AA      24      ATL     3682    3830    Y       0       CR7 E75 ORD
AA      24      AUH     2179    3830    Y       0       777     ORD
AA      24      AUS     3673    3830            0       M83 M80 ORD
AA      24      AZO     4039    3830    Y       0       ER4 ERD ORD
Time taken: 1.298 seconds, Fetched: 10 row(s)
hive (cdac_sk)> select * from routes_part limit 10;
```

Q2.4

```
select * from routes_part limit 10;
```


exam-240840325058 - Google
Subscription Details | Nuvepro
cdacuser3111@ip-172-31-16-2
npapc.cloudloka.com/shell/
Type here to search
32°C Smoke
ENG
17:13
21-11-2024

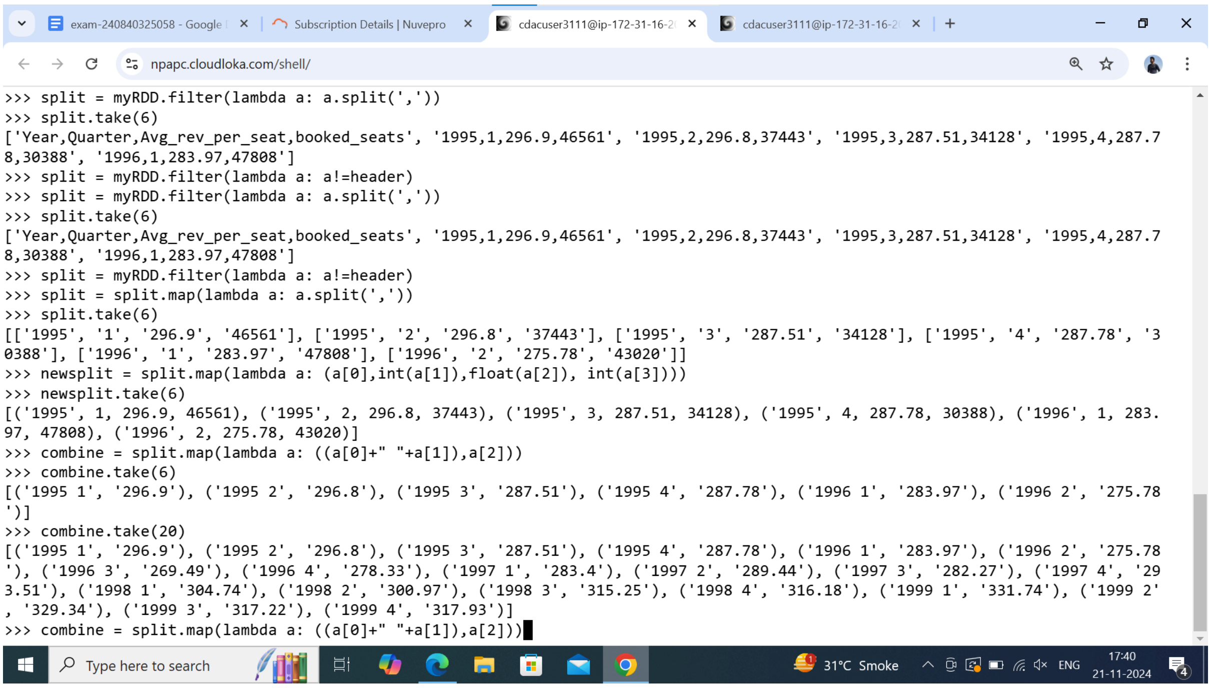
```
airline_id              int
src_airport_iata        string
src_airport_id          int
dest_airport_id         int
codeshare               string
stops                   int
equipment               string
dest_airport_iata       string

# Partition Information
# col_name              data_type               comment
dest_airport_iata       string
Time taken: 0.075 seconds, Fetched: 13 row(s)
hive (cdac_sk)> select * from routes_part limit 10;
OK
3E      10739   BRL     5726    3830            0       CNC     ORD
3E      10739   DEC     4042    3830            0       CNC     ORD
AA      24      ABQ     4019    3830    Y       0       E75     ORD
AA      24      ALO     5718    3830    Y       0       ERD     ORD
AA      24      AMM     2170    3830    Y       0       340     ORD
AA      24      ART     3838    3830    Y       0       ERD     ORD
AA      24      ATL     3682    3830    Y       0       CR7 E75 ORD
AA      24      AUH     2179    3830    Y       0       777     ORD
AA      24      AUS     3673    3830            0       M83 M80 ORD
AA      24      AZO     4039    3830    Y       0       ER4 ERD ORD
Time taken: 1.302 seconds, Fetched: 10 row(s)
hive (cdac_sk)> █
```

Spark:-

Q.1.1

```
myRDD = sc.textFile("/user/cdacuser3111/airlinesnew.csv")
split = split.map(lambda a: a.split(','))
newsplit = split.map(lambda a: (a[0],int(a[1]),float(a[2]),
int(a[3])))
combine = split.map(lambda a: ((a[0]+" "+a[1]),a[2]))
combine.take(20)
```



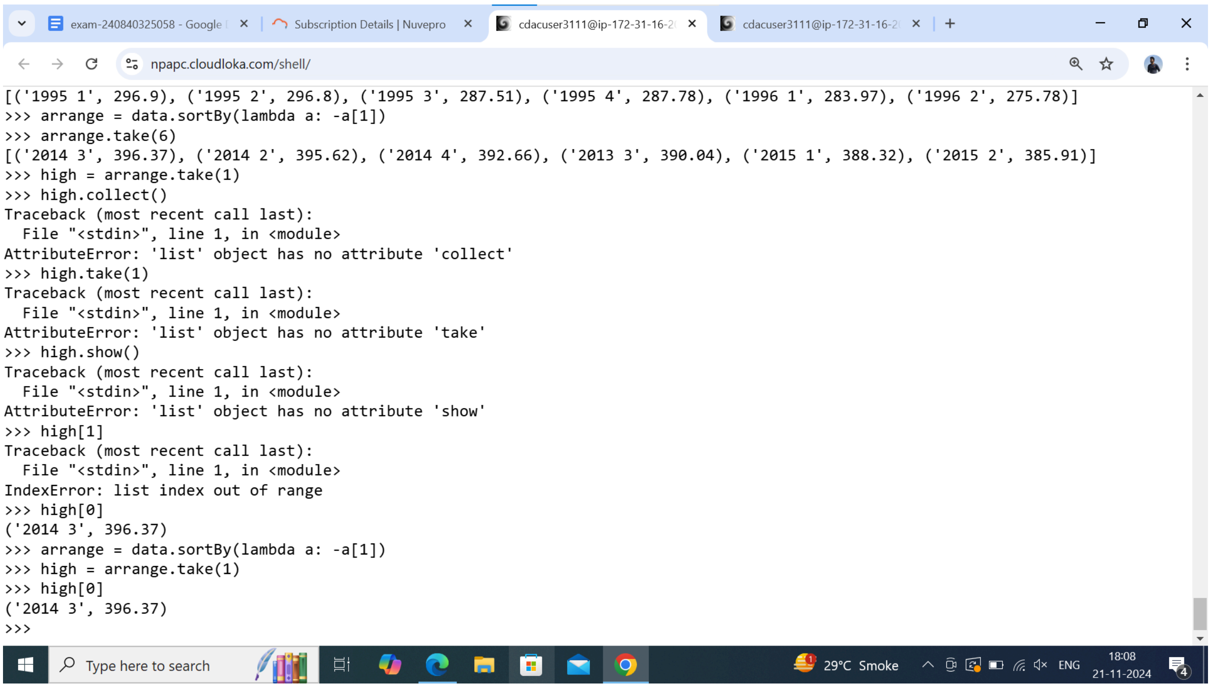
exam-240840325058 - Google
Subscription Details | Nuvepro
cdacuser3111@ip-172-31-16-2
cdacuser3111@ip-172-31-16-2
npapc.cloudloka.com/shell/
Type here to search
31°C Smoke
ENG
17:40
21-11-2024

```
>>> split = myRDD.filter(lambda a: a.split(','))
>>> split.take(6)
['Year,Quarter,Avg_rev_per_seat,booked_seats', '1995,1,296.9,46561', '1995,2,296.8,37443', '1995,3,287.51,34128', '1995,4,287.7
8,30388', '1996,1,283.97,47808']
>>> split = myRDD.filter(lambda a: a!=header)
>>> split = myRDD.filter(lambda a: a.split(','))
>>> split.take(6)
['Year,Quarter,Avg_rev_per_seat,booked_seats', '1995,1,296.9,46561', '1995,2,296.8,37443', '1995,3,287.51,34128', '1995,4,287.7
8,30388', '1996,1,283.97,47808']
>>> split = myRDD.filter(lambda a: a!=header)
>>> split = split.map(lambda a: a.split(','))
>>> split.take(6)
[['1995', '1', '296.9', '46561'], ['1995', '2', '296.8', '37443'], ['1995', '3', '287.51', '34128'], ['1995', '4', '287.78', '3
0388'], ['1996', '1', '283.97', '47808'], ['1996', '2', '275.78', '43020']]
>>> newsplit = split.map(lambda a: (a[0],int(a[1]),float(a[2]), int(a[3])))
>>> newsplit.take(6)
[('1995', 1, 296.9, 46561), ('1995', 2, 296.8, 37443), ('1995', 3, 287.51, 34128), ('1995', 4, 287.78, 30388), ('1996', 1, 283.
97, 47808), ('1996', 2, 275.78, 43020)]
>>> combine = split.map(lambda a: ((a[0]+" "+a[1]),a[2]))
>>> combine.take(6)
[('1995 1', '296.9'), ('1995 2', '296.8'), ('1995 3', '287.51'), ('1995 4', '287.78'), ('1996 1', '283.97'), ('1996 2', '275.78
')]
>>> combine.take(20)
[('1995 1', '296.9'), ('1995 2', '296.8'), ('1995 3', '287.51'), ('1995 4', '287.78'), ('1996 1', '283.97'), ('1996 2', '275.78
'), ('1996 3', '269.49'), ('1996 4', '278.33'), ('1997 1', '283.4'), ('1997 2', '289.44'), ('1997 3', '282.27'), ('1997 4', '29
3.51'), ('1998 1', '304.74'), ('1998 2', '300.97'), ('1998 3', '315.25'), ('1998 4', '316.18'), ('1999 1', '331.74'), ('1999 2'
, '329.34'), ('1999 3', '317.22'), ('1999 4', '317.93')]
>>> combine = split.map(lambda a: ((a[0]+" "+a[1]),a[2]))
```

Q2.1

Q2.5

```
data = combine.map(lambda a: (a[0], float(a[1])))
arrange = data.sortBy(lambda a: -a[1])
```



exam-240840325058 - Google
Subscription Details | Nuvepro
cdacuser3111@ip-172-31-16-2
cdacuser3111@ip-172-31-16-2
npapc.cloudloka.com/shell/
Type here to search
29°C Smoke
ENG
18:08
21-11-2024

```
[('1995 1', 296.9), ('1995 2', 296.8), ('1995 3', 287.51), ('1995 4', 287.78), ('1996 1', 283.97), ('1996 2', 275.78)]
>>> arrange = data.sortBy(lambda a: -a[1])
>>> arrange.take(6)
[('2014 3', 396.37), ('2014 2', 395.62), ('2014 4', 392.66), ('2013 3', 390.04), ('2015 1', 388.32), ('2015 2', 385.91)]
>>> high = arrange.take(1)
>>> high.collect()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'list' object has no attribute 'collect'
>>> high.take(1)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'list' object has no attribute 'take'
>>> high.show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'list' object has no attribute 'show'
>>> high[1]
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
IndexError: list index out of range
>>> high[0]
('2014 3', 396.37)
>>> arrange = data.sortBy(lambda a: -a[1])
>>> high = arrange.take(1)
>>> high[0]
('2014 3', 396.37)
>>>
```