


HIVE-

Q1.

Setup hive using hive command-



```
OK
airport_id      int
name            string
city            string
country         string
iata            string
icao            string
latitude        double
longitude        double
altitude        int
timezone        double
dst             string
tz             string
Time taken: 0.82 seconds, Fetched: 12 row(s)
hive> desc routes;
OK
airline_iata    string
airline_id      int
src_airport_iata string
src_airport_id  int
dest_airport_iata string
dest_airport_id int
codeshare       string
stops           int
equipment       string
Time taken: 0.045 seconds, Fetched: 9 row(s)
hive> use cdac_aayush;
OK
Time taken: 0.054 seconds
hive> set hive.cli.print.current.db=true;
hive (cdac_aayush)>
```

1. Select * from airports where iata in (select src from routes) and iata in (select dest from routes) limit 10;

```
Subscription Details | Nuvepro x cdacuser90116@ip-172-31-9-1 x 240840325001_Big_data_exam x +
cdacnpac.cloudloka.com/shell/
dest          string
dest_id       int
codeshare     string
stops         int
equip         string
Time taken: 0.043 seconds, Fetched: 9 row(s)
hive (cdac_aayush)> Select * from airports where iata in (select src from routes) and iata in (select dest from routes) limit 10;
Query ID = cdacuser90116_20241121084718_3c06b6d8-303b-4b12-8165-a23d49a634b0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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_2278, Tracking URL = http://master:6318/proxy/application_1732089968849_2278/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2278
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 4
2024-11-21 08:47:33,813 Stage-2 map = 0%, reduce = 0%
2024-11-21 08:47:42,076 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 6.23 sec
2024-11-21 08:47:48,257 Stage-2 map = 100%, reduce = 50%, Cumulative CPU 13.48 sec
2024-11-21 08:47:50,311 Stage-2 map = 100%, reduce = 75%, Cumulative CPU 17.04 sec
2024-11-21 08:47:51,345 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 20.53 sec
MapReduce Total cumulative CPU time: 20 seconds 530 msec
Ended Job = job_1732089968849_2278
Launching Job 2 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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>
```

```
Subscription Details | Nuvepro x cdacuser90116@ip-172-31-9-1 x 240840325001_Big_data_exam x +
cdacnpac.cloudloka.com/shell/
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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_2278, Tracking URL = http://master:6318/proxy/application_1732089968849_2278/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2278
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 4
2024-11-21 08:47:33,813 Stage-2 map = 0%, reduce = 0%
2024-11-21 08:47:42,076 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 6.23 sec
2024-11-21 08:47:48,257 Stage-2 map = 100%, reduce = 50%, Cumulative CPU 13.48 sec
2024-11-21 08:47:50,311 Stage-2 map = 100%, reduce = 75%, Cumulative CPU 17.04 sec
2024-11-21 08:47:51,345 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 20.53 sec
MapReduce Total cumulative CPU time: 20 seconds 530 msec
Ended Job = job_1732089968849_2278
Launching Job 2 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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_2284, Tracking URL = http://master:6318/proxy/application_1732089968849_2284/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2284
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 4
2024-11-21 08:48:04,560 Stage-1 map = 0%, reduce = 0%
2024-11-21 08:48:12,777 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 19.45 sec
2024-11-21 08:48:18,936 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 30.72 sec
2024-11-21 08:48:20,982 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 34.59 sec
```

```
Subscription Details | Nuvepro x cdacuser90116@ip-172-31-9-1 x 240840325001_Big_data_exam x +
cdacnpac.cloudloka.com/shell/

In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1732089968849_2284, Tracking URL = http://master:6318/proxy/application_1732089968849_2284/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2284
Hadoop job information for Stage-1: number of mappers: 3; number of reducers: 4
2024-11-21 08:48:04,560 Stage-1 map = 0%, reduce = 0%
2024-11-21 08:48:12,777 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 19.45 sec
2024-11-21 08:48:18,936 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 30.72 sec
2024-11-21 08:48:20,982 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 34.59 sec
MapReduce Total cumulative CPU time: 34 seconds 590 msec
Ended Job = job_1732089968849_2284
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1 Reduce: 4 Cumulative CPU: 20.53 sec HDFS Read: 2403486 HDFS Write: 71973 SUCCESS
Stage-Stage-1: Map: 3 Reduce: 4 Cumulative CPU: 34.59 sec HDFS Read: 3250280 HDFS Write: 4549 SUCCESS
Total MapReduce CPU Time Spent: 55 seconds 120 msec
OK
220 Annaba Annaba Algeria AAE DABB 36.822224 7.809167 16.0 1.0 N Africa/Algiers
628 Aalborg Aalborg Denmark AAL EKYT 57.09279 9.849164 10.0 1.0 E Europe/Copenhagen
5937 Al Ain International Airport Al Ain United Arab Emirates AAN OMAL 24.261667 55.609165 869.0 4.0 U Asia/
Dubai
7395 Araxa Airport Araxa Brazil AAX SBAX -19.563057 -46.960278 3276.0 -3.0 S America/Sao_Paulo
4355 Lehigh Valley Intl Allentown United States ABE KABE 40.652084 -75.4408 393.0 -5.0 A America/New_Y
ork
7177 Ambler Airport Ambler United States ABL PAFM 67.10639 -157.8575 334.0 -9.0 A America/Anchorage
4019 Albuquerque International Sunport Albuquerque United States ABQ KABQ 35.040222 -106.60919 5355.0 -7.0 A
America/Denver
3333 Albury Albury Australia ABX YMAY -36.06778 146.95805 539.0 10.0 O Australia/Sydney
248 Kotoka Intl Accra Ghana ACC DGAA 5.605186 -0.166786 205.0 0.0 N Africa/Accra
1055 Lanzarote Las Palmas Spain ACE GCRR 28.945463 -13.605225 47.0 0.0 E Atlantic/Canary
Time taken: 65.939 seconds, Fetched: 10 row(s)
hive (cdac_aayush)>
```

2. Select equip,count(*) from routes group by equip order by count(*) limit 1 ;

```
Subscription Details | Nuvepro x cdacuser90116@ip-172-31-9-1 x cdacuser90116@ip-172-31-9-1 x 240840325001_Big_data_exam x +
cdacnpac.cloudloka.com/shell/

Time taken: 59.125 seconds, Fetched: 10 row(s)
hive (cdac_aayush)> Select equip,count(*) from routes group by equip order by count(*) desc limit 1 ;
Query ID = cdacuser90116_20241121090249_d6effdac-7530-4cff-badd-ef38833a1e45
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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_2391, Tracking URL = http://master:6318/proxy/application_1732089968849_2391/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2391
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 09:03:01,260 Stage-1 map = 0%, reduce = 0%
2024-11-21 09:03:09,455 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.31 sec
2024-11-21 09:03:15,595 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 13.59 sec
2024-11-21 09:03:17,650 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.71 sec
MapReduce Total cumulative CPU time: 20 seconds 710 msec
Ended Job = job_1732089968849_2391
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_1732089968849_2396, Tracking URL = http://master:6318/proxy/application_1732089968849_2396/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2396
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
```

```
Subscription Details | Nuvepro x | cdacuser90116@ip-172-31-9-1 x | cdacuser90116@ip-172-31-9-1 x | 240840325001_sig_data_exam x | +
cdacnpac.cloudloka.com/shell/
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 09:03:01,260 Stage-1 map = 0%, reduce = 0%
2024-11-21 09:03:09,455 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.31 sec
2024-11-21 09:03:15,595 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 13.59 sec
2024-11-21 09:03:17,650 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.71 sec
MapReduce Total cumulative CPU time: 20 seconds 710 msec
Ended Job = job_1732089968849_2391
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_1732089968849_2396, Tracking URL = http://master:6318/proxy/application_1732089968849_2396/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2396
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
2024-11-21 09:03:30,274 Stage-2 map = 0%, reduce = 0%
2024-11-21 09:03:38,471 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 5.99 sec
2024-11-21 09:03:45,624 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 9.18 sec
MapReduce Total cumulative CPU time: 9 seconds 180 msec
Ended Job = job_1732089968849_2396
MapReduce Jobs Launched:
  Stage-Stage-1: Map: 1 Reduce: 4 Cumulative CPU: 20.71 sec HDFS Read: 2407671 HDFS Write: 122522 SUCCESS
  Stage-Stage-2: Map: 2 Reduce: 1 Cumulative CPU: 9.18 sec HDFS Read: 134102 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 29 seconds 890 msec
OK
320 9180
Time taken: 59.161 seconds, Fetched: 1 row(s)
hive (cdac_aayush)>
```

3. Select r.code,count(r.code) from routes r inner join airlines ar on r.code=ar.iata group by r.code order by count(*) desc limit 1;

```
Subscri x | cdacuser x | cdacuser x | 240840 x | Home x | cdacuser x | rough_u x | Big dat x | Submis x | Aayush: x | +
cdacnpac.cloudloka.com/shell/
hive (cdac_aayush)> Select r.code,count(r.code) from routes r inner join airlines ar on r.code=ar.iata group by r.code order by count(*) desc limit 1
;
Query ID = cdacuser90116_20241121102623_dc5a9b37-001f-4b0b-b5f2-19f6618f6c26
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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_2676, Tracking URL = http://master:6318/proxy/application_1732089968849_2676/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2676
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 10:26:33,930 Stage-1 map = 0%, reduce = 0%
2024-11-21 10:26:41,096 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 6.68 sec
2024-11-21 10:26:42,119 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.53 sec
2024-11-21 10:26:47,228 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 20.29 sec
2024-11-21 10:26:48,249 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 23.77 sec
2024-11-21 10:26:49,271 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 27.2 sec
MapReduce Total cumulative CPU time: 27 seconds 200 msec
Ended Job = job_1732089968849_2676
Launching Job 2 out of 3
Number of reduce tasks not specified. Defaulting to jobconf value of: 4
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_2677, Tracking URL = http://master:6318/proxy/application_1732089968849_2677/
```

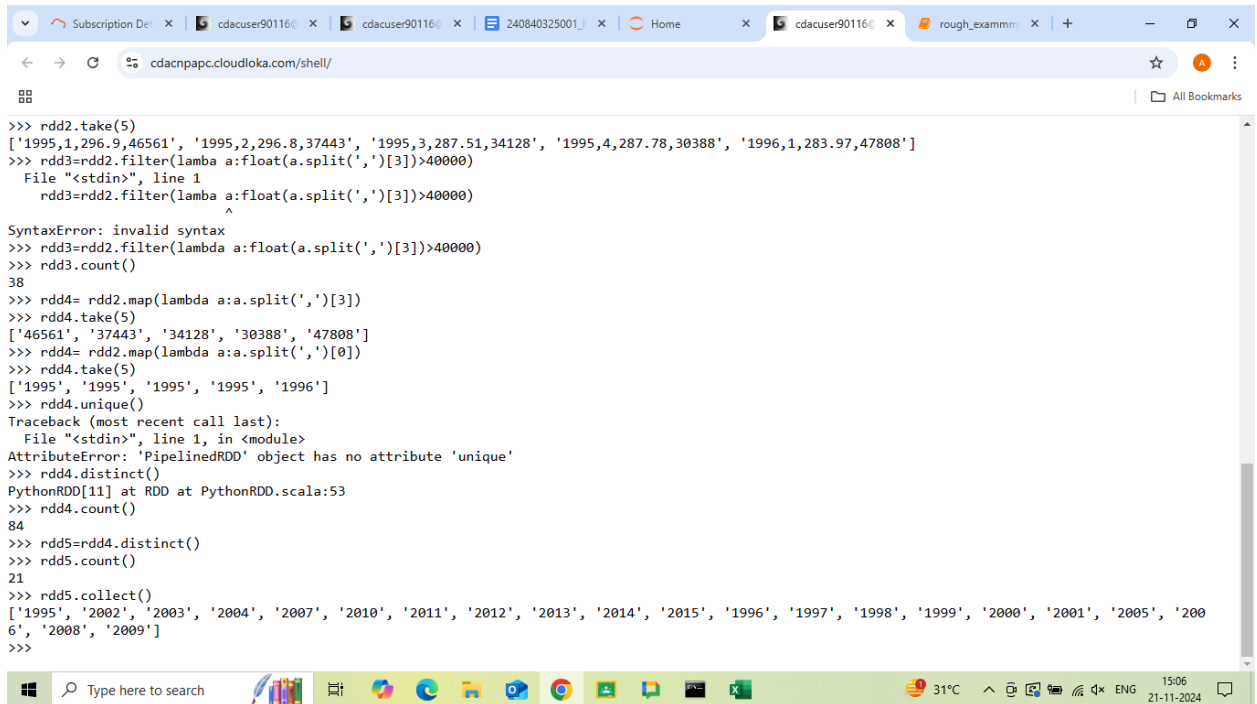
```
cdacnpac.cloudloka.com/shell/

Hadoop job information for Stage-2: number of mappers: 3; number of reducers: 4
2024-11-21 10:24:24,146 Stage-2 map = 0%, reduce = 0%
2024-11-21 10:24:30,295 Stage-2 map = 33%, reduce = 0%, Cumulative CPU 2.77 sec
2024-11-21 10:24:32,342 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 8.55 sec
2024-11-21 10:24:36,431 Stage-2 map = 100%, reduce = 75%, Cumulative CPU 16.55 sec
2024-11-21 10:24:38,475 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 19.19 sec
MapReduce Total cumulative CPU time: 19 seconds 190 msec
Ended Job = job_1732089968849_2673
Launching Job 3 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_1732089968849_2675, Tracking URL = http://master:6318/proxy/application_1732089968849_2675/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2675
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2024-11-21 10:24:49,987 Stage-3 map = 0%, reduce = 0%
2024-11-21 10:24:57,145 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 2.72 sec
2024-11-21 10:25:05,315 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 5.94 sec
MapReduce Total cumulative CPU time: 5 seconds 940 msec
Ended Job = job_1732089968849_2675
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 4 Cumulative CPU: 26.5 sec HDFS Read: 2724024 HDFS Write: 12576 SUCCESS
Stage-Stage-2: Map: 3 Reduce: 4 Cumulative CPU: 19.19 sec HDFS Read: 37530 HDFS Write: 12576 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 5.94 sec HDFS Read: 21386 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 51 seconds 630 msec
OK
FR 2484
Time taken: 79.119 seconds, Fetched: 1 row(s)
```

SPARK-
Q1.

2.

```
rdd4= rdd2.map(lambda a:a.split(',')[0])
rdd4.take(5)
rdd5=rdd4.distinct()
rdd5.count()
rdd5.collect()
```



The screenshot shows a web browser window with multiple tabs. The active tab is titled 'cdacnppc.cloudloka.com/shell/'. The browser's address bar shows the URL. Below the browser window, a terminal window displays the following code and output:

```
>>> rdd2.take(5)
['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']
>>> rdd3=rdd2.filter(lambda a:float(a.split(',')[3])>40000)
File "<stdin>", line 1
    rdd3=rdd2.filter(lambda a:float(a.split(',')[3])>40000)
                        ^
SyntaxError: invalid syntax
>>> rdd3=rdd2.filter(lambda a:float(a.split(',')[3])>40000)
>>> rdd3.count()
38
>>> rdd4= rdd2.map(lambda a:a.split(',')[3])
>>> rdd4.take(5)
['46561', '37443', '34128', '30388', '47808']
>>> rdd4= rdd2.map(lambda a:a.split(',')[0])
>>> rdd4.take(5)
['1995', '1995', '1995', '1995', '1996']
>>> rdd4.unique()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'unique'
>>> rdd4.distinct()
PythonRDD[11] at RDD at PythonRDD.scala:53
>>> rdd4.count()
84
>>> rdd5=rdd4.distinct()
>>> rdd5.count()
21
>>> rdd5.collect()
['1995', '2002', '2003', '2004', '2007', '2010', '2011', '2012', '2013', '2014', '2015', '1996', '1997', '1998', '1999', '2000', '2001', '2005', '2006', '2008', '2009']
>>>
```

Q2.

1.

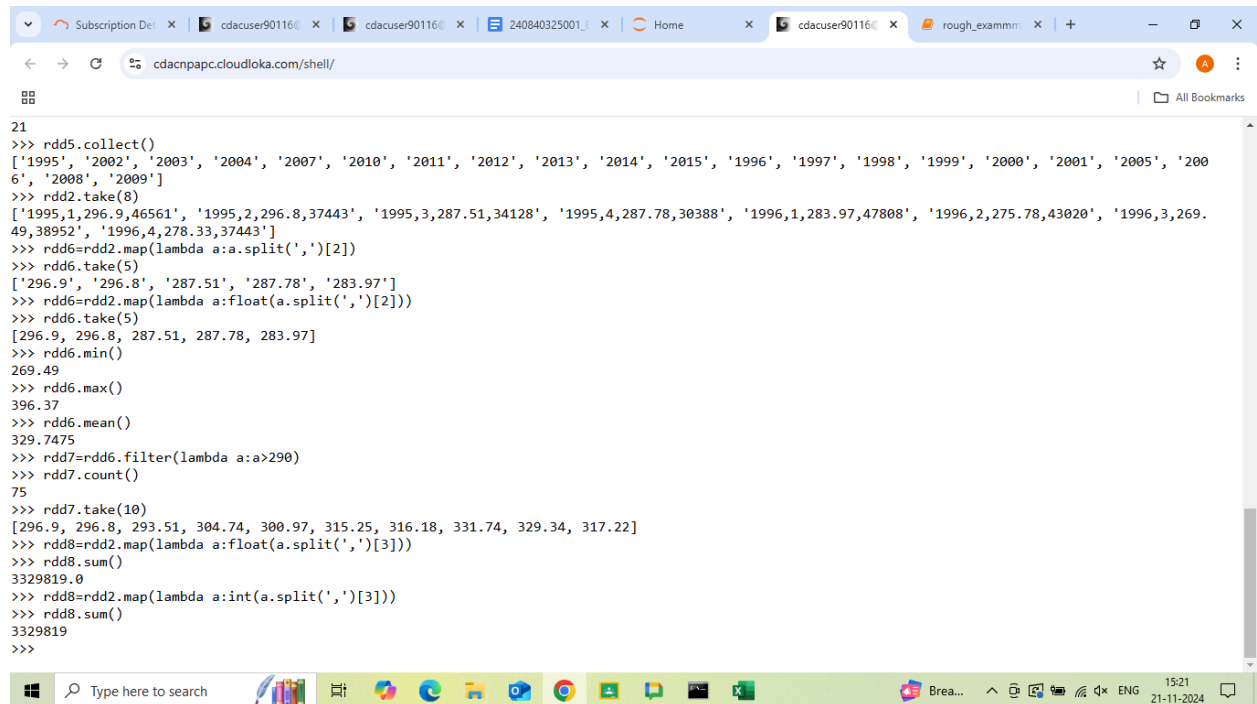
```
rdd6=rdd2.map(lambda a:float(a.split(',')[2]))
rdd6.take(5)
rdd6.min()
rdd6.max()
rdd6.mean()
```

```
Subscription De x cdacuser90116 x cdacuser90116 x 240840325001 x Home x cdacuser90116 x rough_exammm x + - x
cdacnpapc.cloudloka.com/shell/
[ '1995', '1995', '1995', '1995', '1996' ]
>>> rdd4.unique()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'PipelinedRDD' object has no attribute 'unique'
>>> rdd4.distinct()
PythonRDD[11] at RDD at PythonRDD.scala:53
>>> rdd4.count()
84
>>> rdd5=rdd4.distinct()
>>> rdd5.count()
21
>>> rdd5.collect()
[ '1995', '2002', '2003', '2004', '2007', '2010', '2011', '2012', '2013', '2014', '2015', '1996', '1997', '1998', '1999', '2000', '2001', '2005', '2006', '2008', '2009' ]
>>> rdd2.take(8)
[ '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', '1996,3,269.49,38952', '1996,4,278.33,37443' ]
>>> rdd6=rdd2.map(lambda a:a.split(',')[2])
>>> rdd6.take(5)
[ '296.9', '296.8', '287.51', '287.78', '283.97' ]
>>> rdd6=rdd2.map(lambda a:float(a.split(',')[2]))
>>> rdd6.take(5)
[296.9, 296.8, 287.51, 287.78, 283.97]
>>> rdd6.min()
269.49
>>> rdd6.max()
396.37
>>> rdd6.mean()
329.7475
>>>
```

```
2.
rdd6=rdd2.map(lambda a:float(a.split(',')[2]))
rdd7=rdd6.filter(lambda a:a>290)
rdd7.count()
```

```
Subscription De x cdacuser90116 x cdacuser90116 x 240840325001 x Home x cdacuser90116 x rough_exammm x + - x
cdacnpapc.cloudloka.com/shell/
>>> rdd4.distinct()
PythonRDD[11] at RDD at PythonRDD.scala:53
>>> rdd4.count()
84
>>> rdd5=rdd4.distinct()
>>> rdd5.count()
21
>>> rdd5.collect()
[ '1995', '2002', '2003', '2004', '2007', '2010', '2011', '2012', '2013', '2014', '2015', '1996', '1997', '1998', '1999', '2000', '2001', '2005', '2006', '2008', '2009' ]
>>> rdd2.take(8)
[ '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', '1996,3,269.49,38952', '1996,4,278.33,37443' ]
>>> rdd6=rdd2.map(lambda a:a.split(',')[2])
>>> rdd6.take(5)
[ '296.9', '296.8', '287.51', '287.78', '283.97' ]
>>> rdd6=rdd2.map(lambda a:float(a.split(',')[2]))
>>> rdd6.take(5)
[296.9, 296.8, 287.51, 287.78, 283.97]
>>> rdd6.min()
269.49
>>> rdd6.max()
396.37
>>> rdd6.mean()
329.7475
>>> rdd7=rdd6.filter(lambda a:a>290)
>>> rdd7.count()
75
>>> rdd7.take(10)
[296.9, 296.8, 293.51, 304.74, 300.97, 315.25, 316.18, 331.74, 329.34, 317.22]
>>>
```

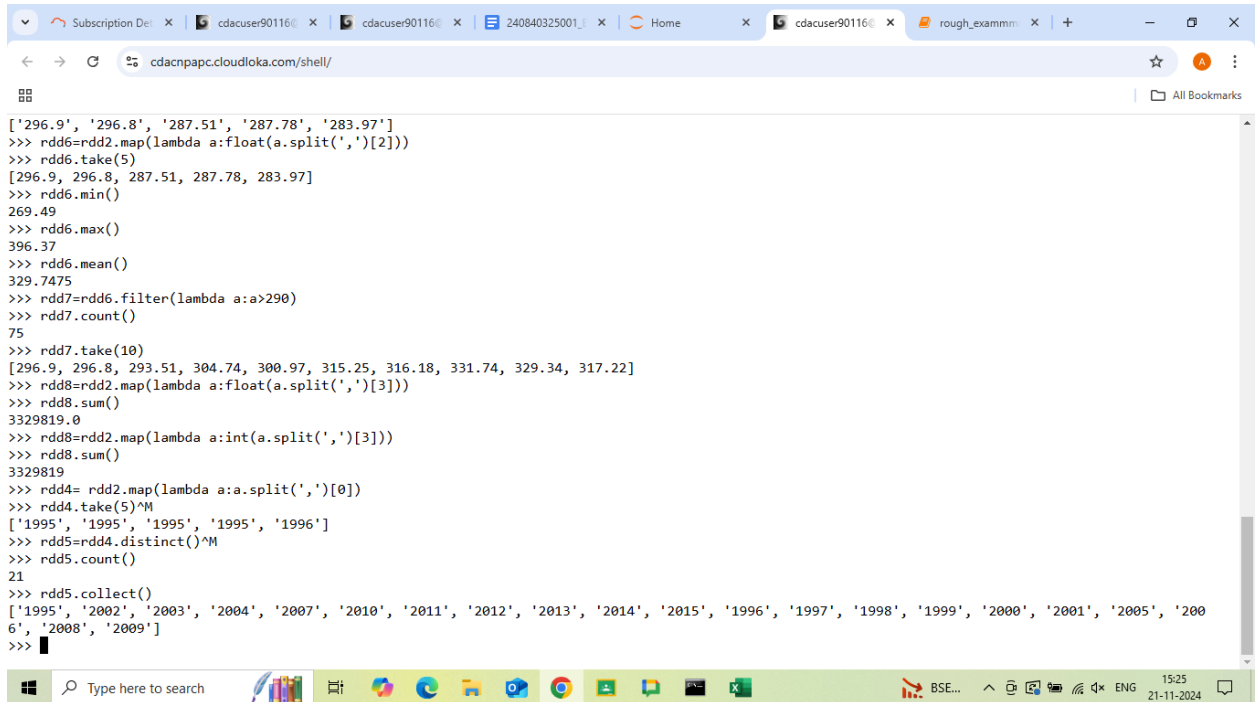

3. `rdd8=rdd2.map(lambda a:int(a.split(',')[3]))`
`rdd8.sum()`



The screenshot shows a web browser window with multiple tabs. The active tab is titled 'cdacuser90116@' and shows a Jupyter Notebook interface. The notebook contains Scala code for RDD operations. The output of the code is displayed below the code cells. The code defines RDDs, performs map operations with lambda functions, and calculates sum, min, max, mean, and count for various RDDs. The final output shows the sum of RDD8 as 3329819.

```
21
>>> rdd5.collect()
['1995', '2002', '2003', '2004', '2007', '2010', '2011', '2012', '2013', '2014', '2015', '1996', '1997', '1998', '1999', '2000', '2001', '2005', '2006', '2008', '2009']
>>> rdd2.take(8)
['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', '1996,3,269.49,38952', '1996,4,278.33,37443']
>>> rdd6=rdd2.map(lambda a:a.split(',')[2])
>>> rdd6.take(5)
['296.9', '296.8', '287.51', '287.78', '283.97']
>>> rdd6=rdd2.map(lambda a:float(a.split(',')[2]))
>>> rdd6.take(5)
[296.9, 296.8, 287.51, 287.78, 283.97]
>>> rdd6.min()
269.49
>>> rdd6.max()
396.37
>>> rdd6.mean()
329.7475
>>> rdd7=rdd6.filter(lambda a:a>290)
>>> rdd7.count()
75
>>> rdd7.take(10)
[296.9, 296.8, 293.51, 304.74, 300.97, 315.25, 316.18, 331.74, 329.34, 317.22]
>>> rdd8=rdd2.map(lambda a:float(a.split(',')[3]))
>>> rdd8.sum()
3329819.0
>>> rdd8=rdd2.map(lambda a:int(a.split(',')[3]))
>>> rdd8.sum()
3329819
>>>
```

4. `rdd4= rdd2.map(lambda a:a.split(',')[0])`
`rdd4.take(5)`
`rdd5=rdd4.distinct()`
`rdd5.count()`
`rdd5.collect()`



The screenshot shows a web browser window with multiple tabs. The active tab is titled 'cdacuser90116' and the address bar shows 'cdacnpapc.cloudloka.com/shell/'. The browser window displays a terminal session with the following commands and output:

```
[ '296.9', '296.8', '287.51', '287.78', '283.97' ]
>>> rdd6=rdd2.map(lambda a:float(a.split(',')[2]))
>>> rdd6.take(5)
[296.9, 296.8, 287.51, 287.78, 283.97]
>>> rdd6.min()
269.49
>>> rdd6.max()
396.37
>>> rdd6.mean()
329.7475
>>> rdd7=rdd6.filter(lambda a:a>290)
>>> rdd7.count()
75
>>> rdd7.take(10)
[296.9, 296.8, 293.51, 304.74, 300.97, 315.25, 316.18, 331.74, 329.34, 317.22]
>>> rdd8=rdd2.map(lambda a:float(a.split(',')[3]))
>>> rdd8.sum()
3329819.0
>>> rdd8=rdd2.map(lambda a:int(a.split(',')[3]))
>>> rdd8.sum()
3329819
>>> rdd4= rdd2.map(lambda a:a.split(',')[0])
>>> rdd4.take(5)^M
['1995', '1995', '1995', '1995', '1996']
>>> rdd5=rdd4.distinct()^M
>>> rdd5.count()
21
>>> rdd5.collect()
['1995', '2002', '2003', '2004', '2007', '2010', '2011', '2012', '2013', '2014', '2015', '1996', '1997', '1998', '1999', '2000', '2001', '2005', '2006', '2008', '2009']
>>>
```

The Windows taskbar is visible at the bottom of the browser window, showing the search bar and various application icons.

5.

```
rdd9=rdd2.map(lambda a:(a.split(',')[0],float(a.split(',')[2])*float(a.split(',')[3])))
rdd9.take(5)
rdd10=rdd9.reduceByKey(lambda a,b:a+b)
rdd10.take(5)
rdd10.collect()
```

