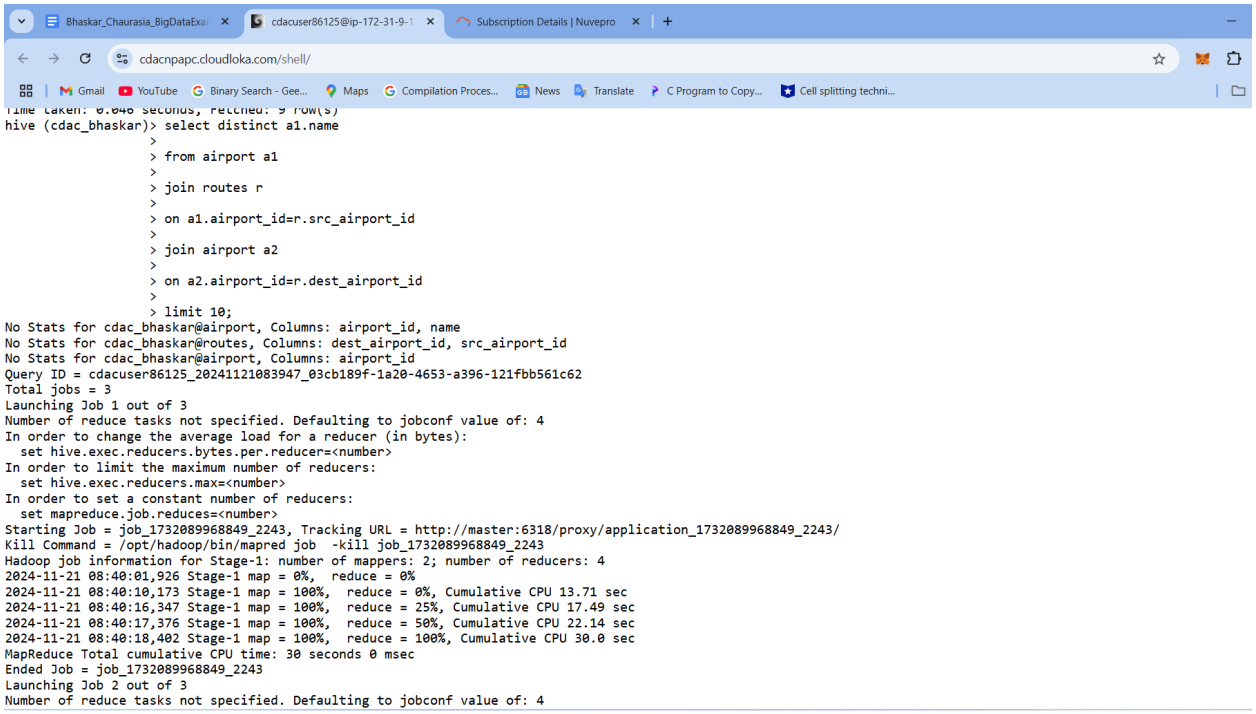
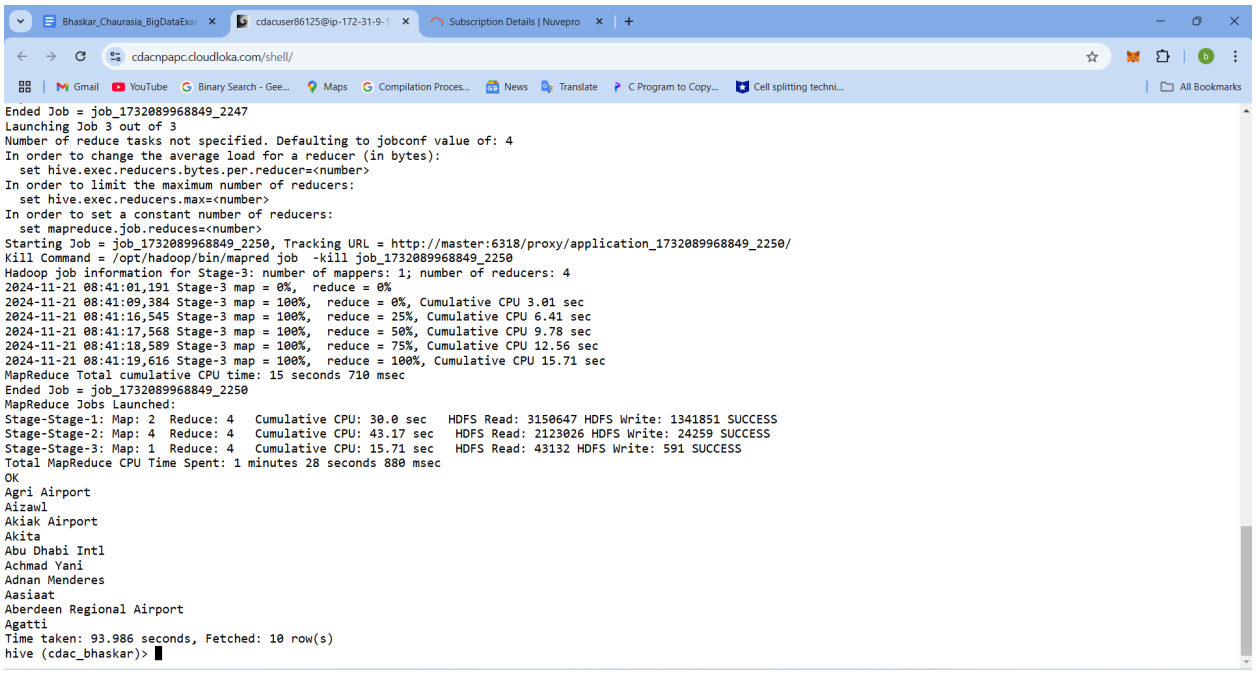
# HIVE

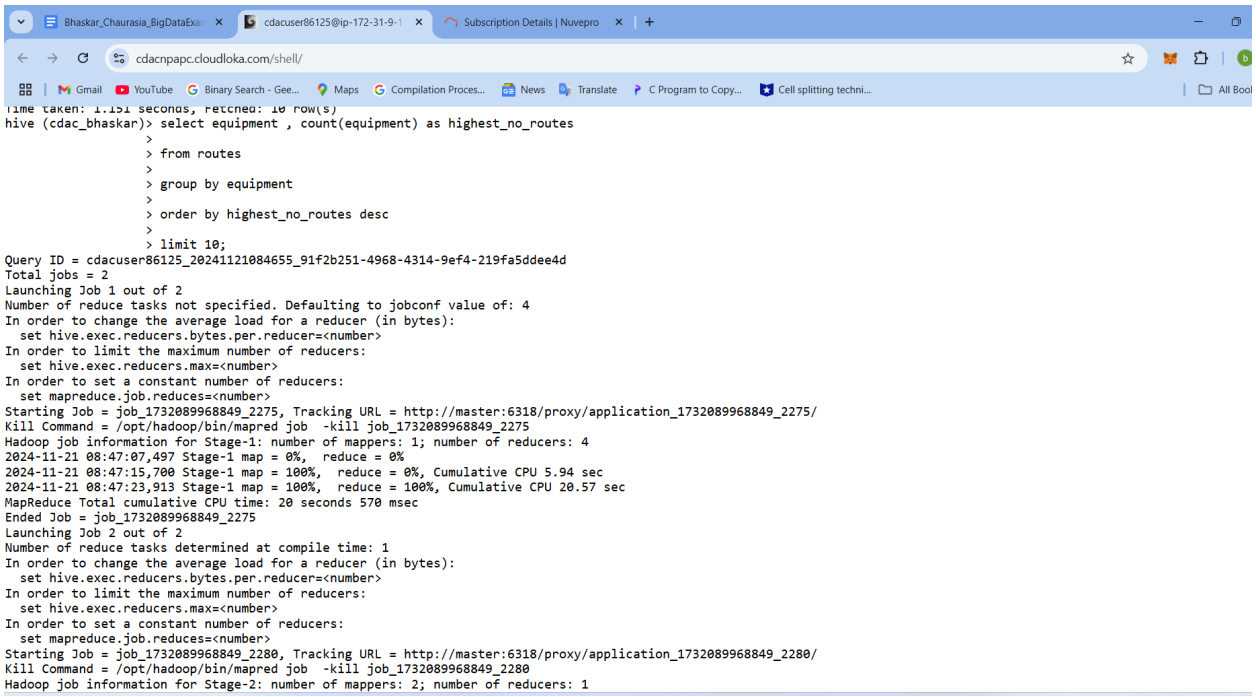
## Question 1:

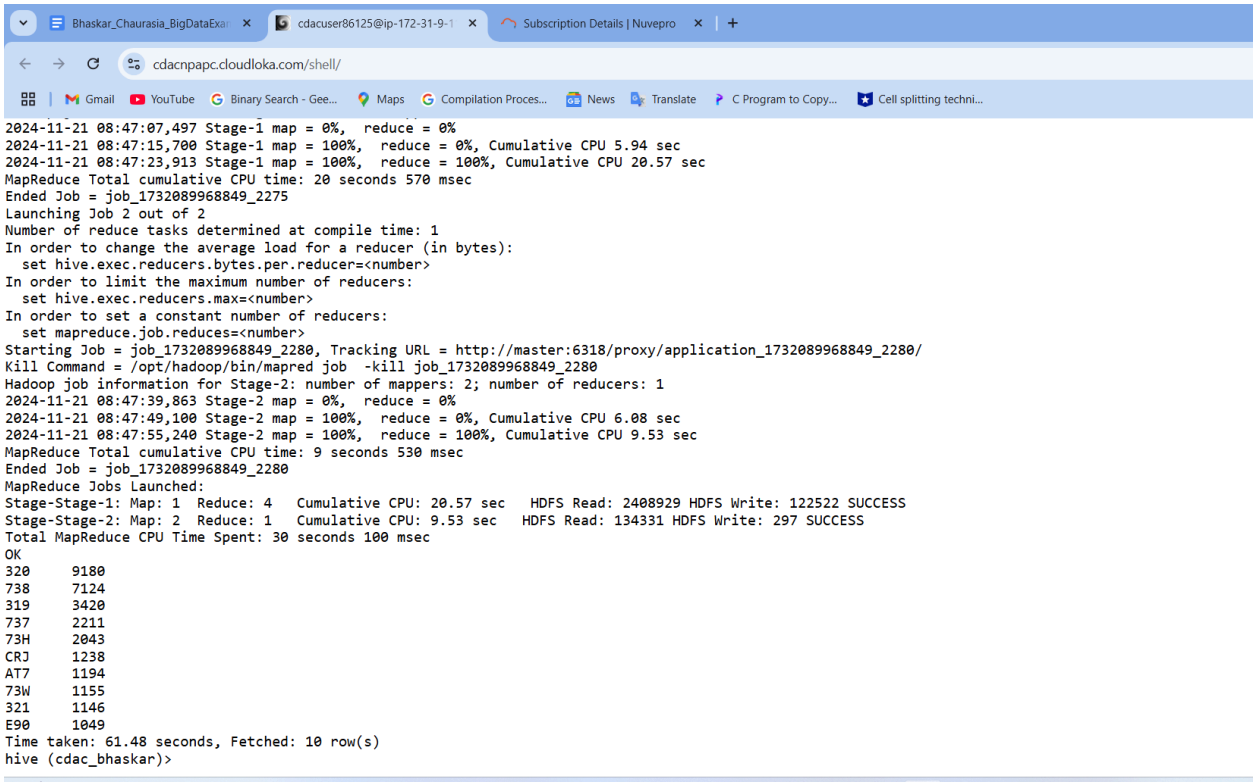
1.



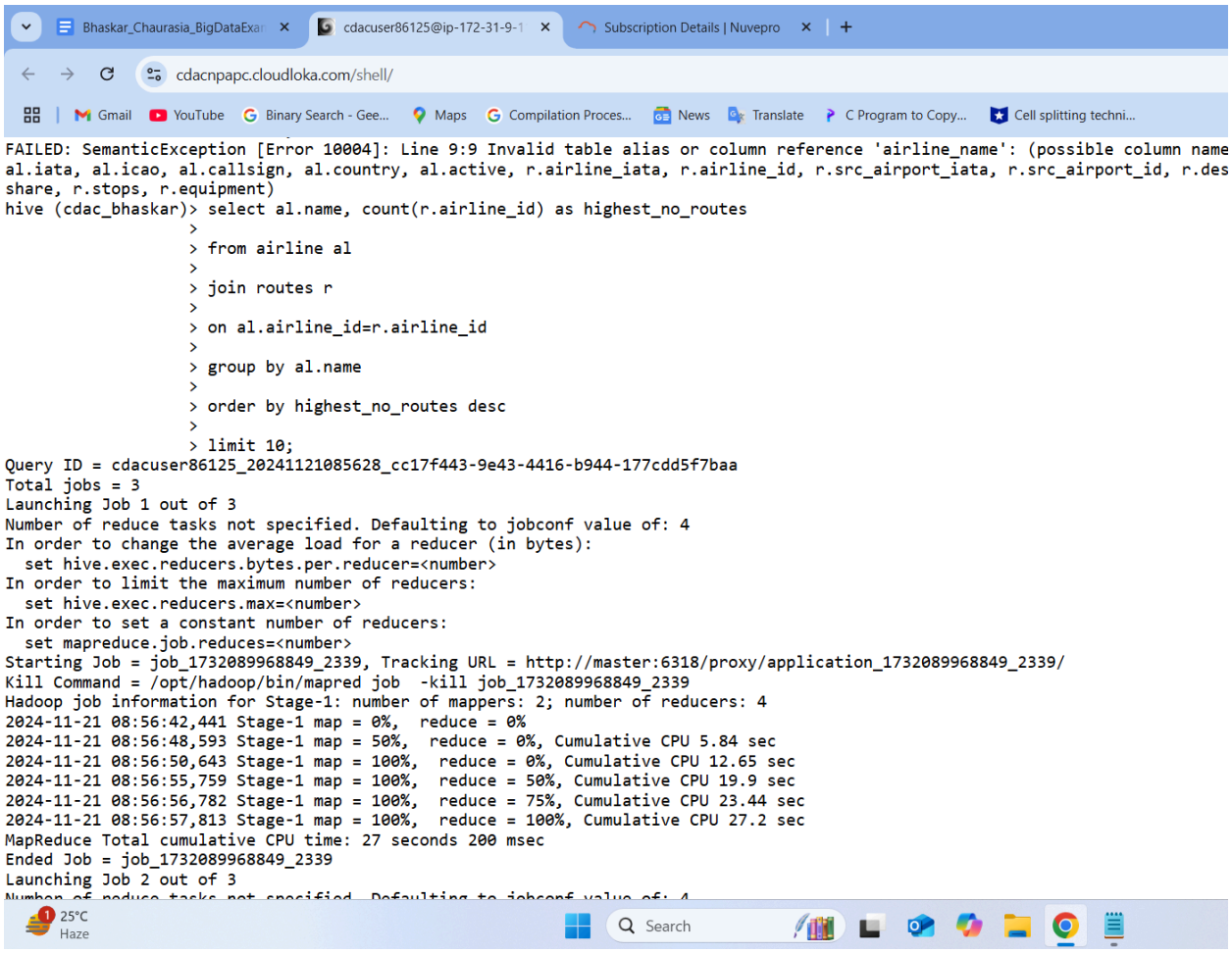

Bhaskar_Chaurasia_BigDataExa
cdacuser86125@ip-172-31-9-1
Subscription Details | Nuvepro
cdacnpapc.cloudloka.com/shell/
Gmail
YouTube
Binary Search - Gee...
Maps
Compilation Proces...
News
Translate
C Program to Copy...
Cell splitting techni...
All Bookmarks
Ended Job = job_1732089968849_2247
Launching Job 3 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_2250, Tracking URL = http://master:6318/proxy/application_1732089968849_2250/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2250
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 4
2024-11-21 08:41:01,191 Stage-3 map = 0%, reduce = 0%
2024-11-21 08:41:09,384 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 3.01 sec
2024-11-21 08:41:16,545 Stage-3 map = 100%, reduce = 25%, Cumulative CPU 6.41 sec
2024-11-21 08:41:17,568 Stage-3 map = 100%, reduce = 50%, Cumulative CPU 9.78 sec
2024-11-21 08:41:18,589 Stage-3 map = 100%, reduce = 75%, Cumulative CPU 12.56 sec
2024-11-21 08:41:19,616 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 15.71 sec
MapReduce Total cumulative CPU time: 15 seconds 710 msec
Ended Job = job_1732089968849_2250
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 4 Cumulative CPU: 30.0 sec HDFS Read: 3150647 HDFS Write: 1341851 SUCCESS
Stage-Stage-2: Map: 4 Reduce: 4 Cumulative CPU: 43.17 sec HDFS Read: 2123026 HDFS Write: 24259 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 4 Cumulative CPU: 15.71 sec HDFS Read: 43132 HDFS Write: 591 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 28 seconds 880 msec
OK
Agri Airport
Aizawl
Akiak Airport
Akita
Abu Dhabi Intl
Achmad Yani
Adnan Menderes
Aasiaat
Aberdeen Regional Airport
Agatti
Time taken: 93.986 seconds, Fetched: 10 row(s)
hive (cdac_bhaskar)>

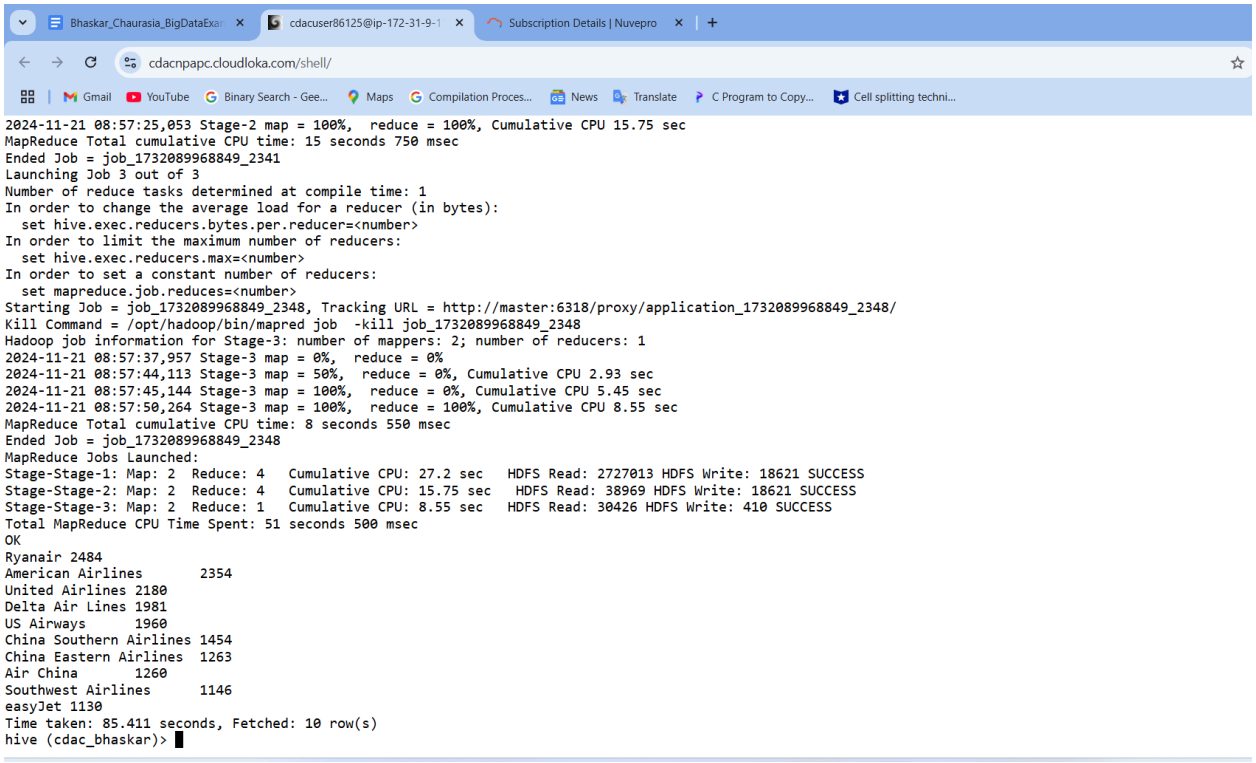
2.


Bhaskar_Chaurasia_BigDataExa
cdacuser86125@ip-172-31-9-1
Subscription Details | Nuvepro
cdacnpapc.cloudloka.com/shell/
Gmail
YouTube
Binary Search - Gee...
Maps
Compilation Proces...
News
Translate
C Program to Copy...
Cell splitting techni...
Time taken: 1.151 seconds, Fetched: 10 row(s)
hive (cdac_bhaskar)> select equipment , count(equipment) as highest_no_routes
> from routes
> group by equipment
> order by highest_no_routes desc
> limit 10;
Query ID = cdacuser86125_20241121084655_91f2b251-4968-4314-9ef4-219fa5ddee4d
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_2275, Tracking URL = http://master:6318/proxy/application_1732089968849_2275/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2275
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 4
2024-11-21 08:47:07,497 Stage-1 map = 0%, reduce = 0%
2024-11-21 08:47:15,700 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.94 sec
2024-11-21 08:47:23,913 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 20.57 sec
MapReduce Total cumulative CPU time: 20 seconds 570 msec
Ended Job = job_1732089968849_2275
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_2280, Tracking URL = http://master:6318/proxy/application_1732089968849_2280/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2280
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1


Bhaskar_Chaurasia_BigDataExa
cdacuser86125@ip-172-31-9-1
Subscription Details | Nuvepro
cdacnpapc.cloudloka.com/shell/
Gmail
YouTube
Binary Search - Gee...
Maps
Compilation Proces...
News
Translate
C Program to Copy...
Cell splitting techni...
2024-11-21 08:47:07,497 Stage-1 map = 0%,  reduce = 0%
2024-11-21 08:47:15,700 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.94 sec
2024-11-21 08:47:23,913 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 20.57 sec
MapReduce Total cumulative CPU time: 20 seconds 570 msec
Ended Job = job_1732089968849_2275
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_2280, Tracking URL = http://master:6318/proxy/application_1732089968849_2280/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2280
Hadoop job information for Stage-2: number of mappers: 2; number of reducers: 1
2024-11-21 08:47:39,863 Stage-2 map = 0%,  reduce = 0%
2024-11-21 08:47:49,100 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 6.08 sec
2024-11-21 08:47:55,240 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 9.53 sec
MapReduce Total cumulative CPU time: 9 seconds 530 msec
Ended Job = job_1732089968849_2280
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 4   Cumulative CPU: 20.57 sec   HDFS Read: 2408929 HDFS Write: 122522 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 1   Cumulative CPU: 9.53 sec   HDFS Read: 134331 HDFS Write: 297 SUCCESS
Total MapReduce CPU Time Spent: 30 seconds 100 msec
OK
320	9180
738	7124
319	3420
737	2211
73H	2043
CRJ	1238
AT7	1194
73W	1155
321	1146
E90	1049
Time taken: 61.48 seconds, Fetched: 10 row(s)
hive (cdac_bhaskar)>

3.


Bhaskar_Chaurasia_BigDataExa
cdacuser86125@ip-172-31-9-1
Subscription Details | Nuvepro
cdacnpapc.cloudloka.com/shell/
Gmail
YouTube
Binary Search - Gee...
Maps
Compilation Proces...
News
Translate
C Program to Copy...
Cell splitting techni...
FAILED: SemanticException [Error 10004]: Line 9:9 Invalid table alias or column reference 'airline_name': (possible column name
al.iata, al.icao, al.callsign, al.country, al.active, r.airline_iata, r.airline_id, r.src_airport_iata, r.src_airport_id, r.des
share, r.stops, r.equipment)
hive (cdac_bhaskar)> select al.name, count(r.airline_id) as highest_no_routes
> from airline al
> join routes r
> on al.airline_id=r.airline_id
> group by al.name
> order by highest_no_routes desc
> limit 10;
Query ID = cdacuser86125_20241121085628_cc17f443-9e43-4416-b944-177cdd5f7baa
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_2339, Tracking URL = http://master:6318/proxy/application_1732089968849_2339/
Kill Command = /opt/hadoop/bin/mapred job -kill job_1732089968849_2339
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 4
2024-11-21 08:56:42,441 Stage-1 map = 0%, reduce = 0%
2024-11-21 08:56:48,593 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 5.84 sec
2024-11-21 08:56:50,643 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.65 sec
2024-11-21 08:56:55,759 Stage-1 map = 100%, reduce = 50%, Cumulative CPU 19.9 sec
2024-11-21 08:56:56,782 Stage-1 map = 100%, reduce = 75%, Cumulative CPU 23.44 sec
2024-11-21 08:56:57,813 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 27.2 sec
MapReduce Total cumulative CPU time: 27 seconds 200 msec
Ended Job = job_1732089968849_2339
Launching Job 2 out of 3
25°C
Haze
Search

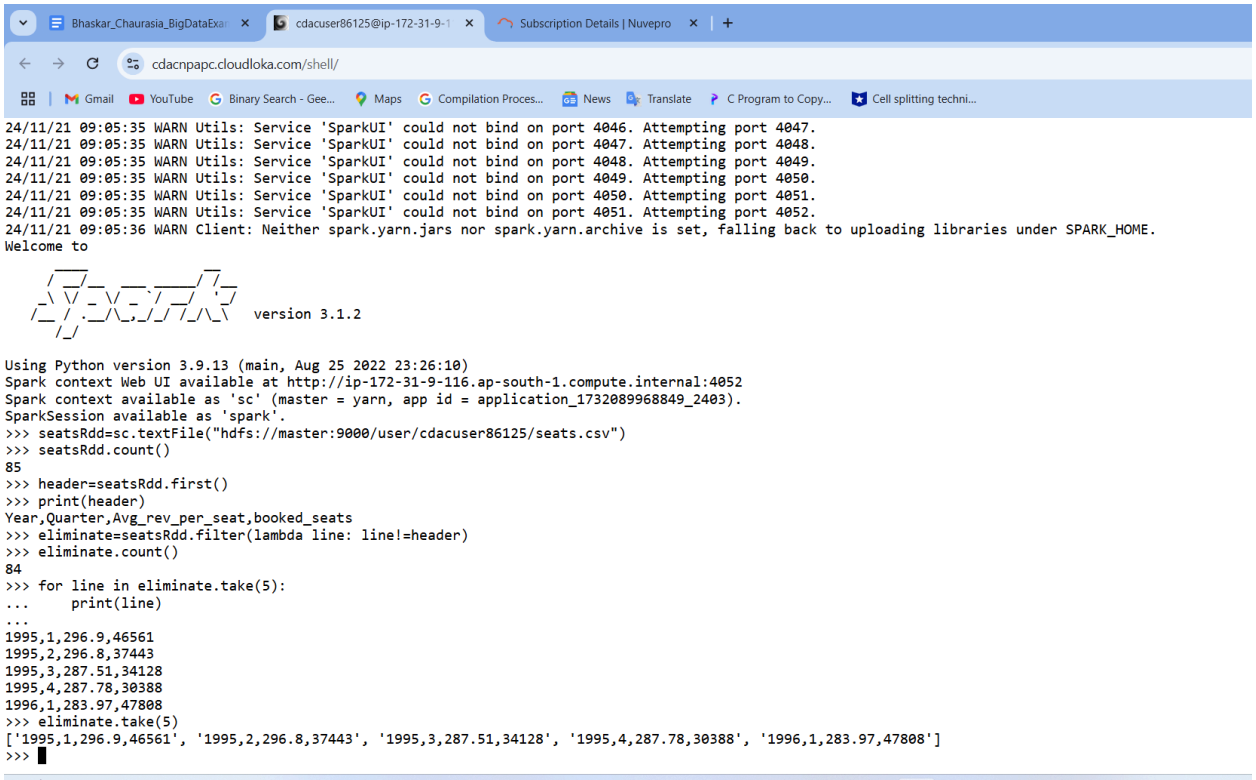

Bhaskar_Chaurasia_BigDataExa
cdacuser86125@ip-172-31-9-1
Subscription Details | Nuvepro
cdacnpapc.cloudloka.com/shell/
Gmail
YouTube
Binary Search - Gee...
Maps
Compilation Proces...
News
Translate
C Program to Copy...
Cell splitting techni...
2024-11-21 08:57:25,053 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 15.75 sec
MapReduce Total cumulative CPU time: 15 seconds 750 msec
Ended Job = job_1732089968849_2341
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_2348, Tracking URL = http://master:6318/proxy/application_1732089968849_2348/
Kill Command = /opt/hadoop/bin/mapred job  -kill job_1732089968849_2348
Hadoop job information for Stage-3: number of mappers: 2; number of reducers: 1
2024-11-21 08:57:37,957 Stage-3 map = 0%,  reduce = 0%
2024-11-21 08:57:44,113 Stage-3 map = 50%,  reduce = 0%, Cumulative CPU 2.93 sec
2024-11-21 08:57:45,144 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 5.45 sec
2024-11-21 08:57:50,264 Stage-3 map = 100%,  reduce = 100%, Cumulative CPU 8.55 sec
MapReduce Total cumulative CPU time: 8 seconds 550 msec
Ended Job = job_1732089968849_2348
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2  Reduce: 4   Cumulative CPU: 27.2 sec   HDFS Read: 2727013 HDFS Write: 18621 SUCCESS
Stage-Stage-2: Map: 2  Reduce: 4   Cumulative CPU: 15.75 sec   HDFS Read: 38969 HDFS Write: 18621 SUCCESS
Stage-Stage-3: Map: 2  Reduce: 1   Cumulative CPU: 8.55 sec   HDFS Read: 30426 HDFS Write: 410 SUCCESS
Total MapReduce CPU Time Spent: 51 seconds 500 msec
OK
Ryanair 2484
American Airlines       2354
United Airlines 2180
Delta Air Lines 1981
US Airways      1960
China Southern Airlines 1454
China Eastern Airlines  1263
Air China       1260
Southwest Airlines      1146
easyJet 1130
Time taken: 85.411 seconds, Fetched: 10 row(s)
hive (cdac_bhaskar)>

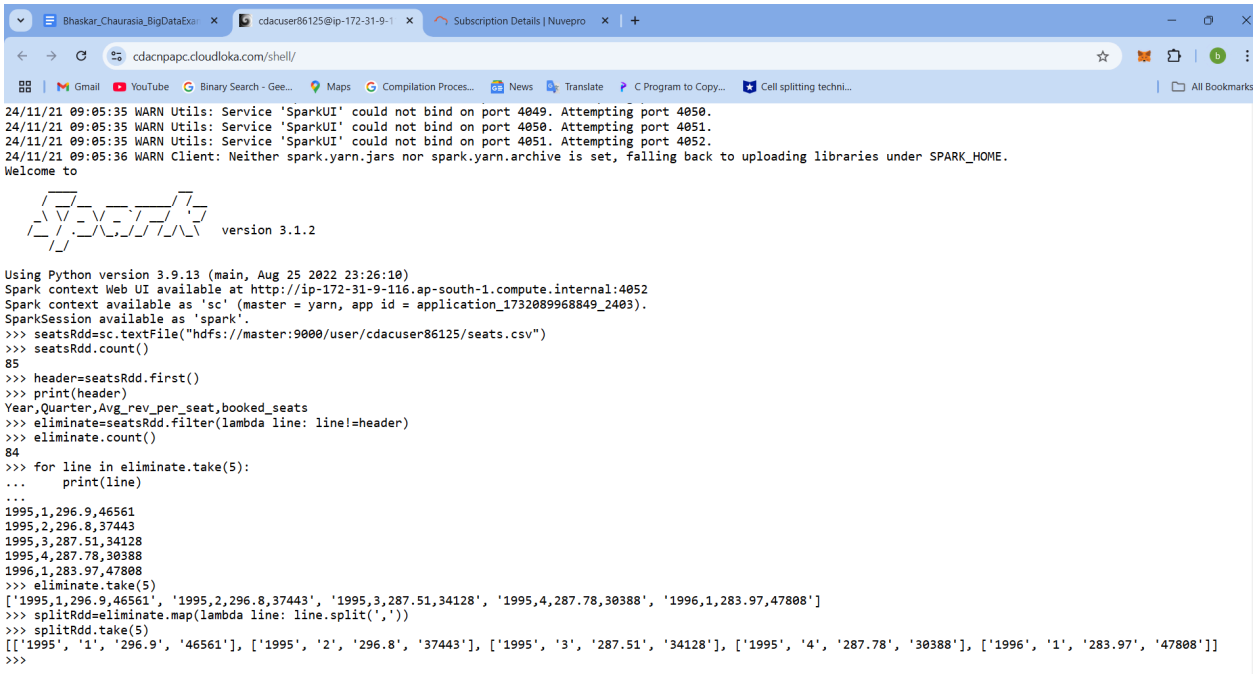
# SPARK

Home / user / cdacuser86125 / **seats.csv**

Page 1 to 1 of 1

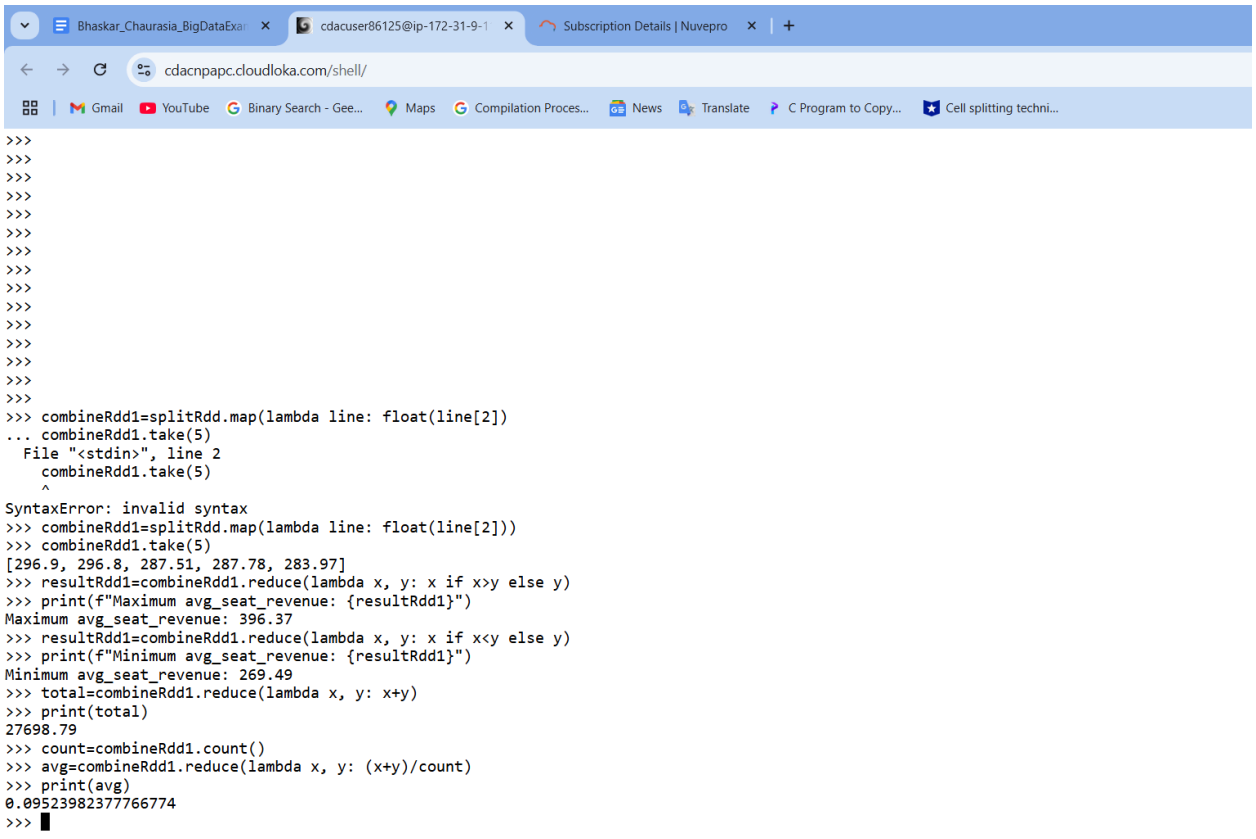
```
Year,Quarter,Avg_rev_per_seat,booked_seats
1995,1,296.9,46561
1995,2,296.8,37443
1995,3,287.51,34128
1995,4,287.78,30388
1996,1,283.97,47808
1996,2,275.78,43020
1996,3,269.49,38952
1996,4,278.33,37443
1997,1,283.4,35067
1997,2,289.44,46565
1997,3,282.27,38886
1997,4,293.51,37454
1998,1,304.74,31315
1998,2,300.97,30852
1998,3,315.25,38118
1998,4,316.18,35393
1999,1,331.74,47453
1999,2,329.34,38243
1999,3,317.22,33048
1999,4,317.93,31256
2000,1,340.23,48159
2000,2,339.16,38329
2000,3,336.66,37785
2000,4,340.08,30103
```


Bhaskar_Chaurasia_BigDataExa
cdacuser86125@ip-172-31-9-1
Subscription Details | Nuvepro
cdacnpapc.cloudloka.com/shell/
Gmail
YouTube
Binary Search - Gee...
Maps
Compilation Proces...
News
Translate
C Program to Copy...
Cell splitting techni...
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4046. Attempting port 4047.
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4047. Attempting port 4048.
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4048. Attempting port 4049.
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4049. Attempting port 4050.
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4050. Attempting port 4051.
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4051. Attempting port 4052.
24/11/21 09:05:36 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
version 3.1.2
Using Python version 3.9.13 (main, Aug 25 2022 23:26:10)
Spark context Web UI available at http://ip-172-31-9-116.ap-south-1.compute.internal:4052
Spark context available as 'sc' (master = yarn, app id = application_1732089968849_2403).
SparkSession available as 'spark'.
>>> seatsRdd=sc.textFile("hdfs://master:9000/user/cdacuser86125/seats.csv")
>>> seatsRdd.count()
85
>>> header=seatsRdd.first()
>>> print(header)
Year,Quarter,Avg_rev_per_seat,booked_seats
>>> eliminate=seatsRdd.filter(lambda line: line!=header)
>>> eliminate.count()
84
>>> for line in eliminate.take(5):
...     print(line)
...
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
>>> eliminate.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']
>>>


Bhaskar_Chaurasia_BigDataExa
cdacuser86125@ip-172-31-9-1
Subscription Details | Nuvepro
cdacnpapc.cloudloka.com/shell/
Gmail
YouTube
Binary Search - Gee...
Maps
Compilation Proces...
News
Translate
C Program to Copy...
Cell splitting techni...
All Bookmarks
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4049. Attempting port 4050.
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4050. Attempting port 4051.
24/11/21 09:05:35 WARN Utils: Service 'SparkUI' could not bind on port 4051. Attempting port 4052.
24/11/21 09:05:36 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
version 3.1.2
Using Python version 3.9.13 (main, Aug 25 2022 23:26:10)
Spark context Web UI available at http://ip-172-31-9-116.ap-south-1.compute.internal:4052
Spark context available as 'sc' (master = yarn, app id = application_1732089968849_2403).
SparkSession available as 'spark'.
>>> seatsRdd=sc.textFile("hdfs://master:9000/user/cdacuser86125/seats.csv")
>>> seatsRdd.count()
85
>>> header=seatsRdd.first()
>>> print(header)
Year,Quarter,Avg_rev_per_seat,booked_seats
>>> eliminate=seatsRdd.filter(lambda line: line!=header)
>>> eliminate.count()
84
>>> for line in eliminate.take(5):
...     print(line)
...
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
>>> eliminate.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']
>>> splitRdd=eliminate.map(lambda line: line.split(','))
>>> splitRdd.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']]
>>>

# Question 2:

1.


Bhaskar_Chaurasia_BigDataExar
cdacuser86125@ip-172-31-9-1
Subscription Details | Nuvepro
cdacnpapc.cloudloka.com/shell/
Gmail
YouTube
Binary Search - Gee...
Maps
Compilation Proces...
News
Translate
C Program to Copy...
Cell splitting techni...
>>> combineRdd1=splitRdd.map(lambda line: float(line[2])
... combineRdd1.take(5)
File "<stdin>", line 2
combineRdd1.take(5)
^
SyntaxError: invalid syntax
>>> combineRdd1=splitRdd.map(lambda line: float(line[2]))
>>> combineRdd1.take(5)
[296.9, 296.8, 287.51, 287.78, 283.97]
>>> resultRdd1=combineRdd1.reduce(lambda x, y: x if x>y else y)
>>> print(f"Maximum avg_seat_revenue: {resultRdd1}")
Maximum avg_seat_revenue: 396.37
>>> resultRdd1=combineRdd1.reduce(lambda x, y: x if x<y else y)
>>> print(f"Minimum avg_seat_revenue: {resultRdd1}")
Minimum avg_seat_revenue: 269.49
>>> total=combineRdd1.reduce(lambda x, y: x+y)
>>> print(total)
27698.79
>>> count=combineRdd1.count()
>>> avg=combineRdd1.reduce(lambda x, y: (x+y)/count)
>>> print(avg)
0.09523982377766774
>>>

2.

```
[True, True, False, False, False]
>>> combineRdd1=splitRdd.map(lambda line: float(line[2]))
>>> combineRdd1.take(5)
[296.9, 296.8, 287.51, 287.78, 283.97]
```

```
resultRdd1=combineRdd1.map(lambda x: x[2] if x[2]>290.0)
resultRdd1.count()
```

3.

```
>>> combineRdd=splitRdd.map(lambda line: (line[1], int(line[3])))
>>> combineRdd.take(5)
[('1', 46561), ('2', 37443), ('3', 34128), ('4', 30388), ('1', 47808)]
>>>
```

4.

```
>>> combineRdd=splitRdd.map(lambda line: int(line[0]))
>>> combineRdd.take(5)
[1995, 1995, 1995, 1995, 1996]
>>> combineRdd.collect()
[1995, 1995, 1995, 1995, 1996, 1996, 1996, 1996, 1997, 1997, 1997, 1997, 1998, 1998, 1998, 1998, 1999, 1999, 1999, 1999, 2000, 2000, 2000, 2000, 2001, 2001, 2001, 2001,
 2002, 2002, 2002, 2002, 2003, 2003, 2003, 2003, 2004, 2004, 2004, 2004, 2005, 2005, 2005, 2005, 2006, 2006, 2006, 2006, 2007, 2007, 2007, 2007, 2008, 2008, 2008, 2008,
 2009, 2009, 2009, 2009, 2010, 2010, 2010, 2010, 2011, 2011, 2011, 2011, 2012, 2012, 2012, 2012, 2013, 2013, 2013, 2013, 2014, 2014, 2014, 2014, 2015, 2015, 2015, 2015]
>>>
```

5.

```
>>> combineRdd=splitRdd.map(lambda line: float(line[2]))
>>> combineRdd.take(5)
[296.9, 296.8, 287.51, 287.78, 283.97]
>>> resultRdd=combineRdd.reduce(lambda x, y: x+y)
>>> print(f"Total Revenue: {resultRdd})
  File "<stdin>", line 1
    print(f"Total Revenue: {resultRdd})
                                       ^
SyntaxError: EOL while scanning string literal
>>> print(f"Total Revenue: {resultRdd}")
Total Revenue: 27698.79
>>>
```