#####Ashish Mani Acharya####

2

#####AXA190076###########

3

#########Assignment 3 ###############

4

############04/27/2022###############

##########################Question 1 ##########################

####################loading into table ####################################

flights = spark.read.option("header","true").option("inferSchema", "true").csv("/FileStore/tables/T_T100D_SEGMENT_US_CARRIER_ONLY-1.csv")

(2) Spark Jobs

Command took 1.04 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:03:30 PM on assignment 3 may 29

##############displaying flights ####################

display(flights)

(1) Spark Jobs

Table

ORIGIN

ORIGIN_CITY_NAME

DEST

DEST_CITY_NAME

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

06A

Kizhuyak, AK

A43

Kodiak Island, AK

09A

Homer, AK

ADQ

Kodiak, AK

1G4

Peach Springs, AZ

BLD

Boulder City, NV

1G4

Peach Springs, AZ

BLD

Boulder City, NV

1NY

Penn Yan, NY

HPN

White Plains, NY

1TX

Dumas, TX

PHX

Phoenix, AZ

2AK

Deer Park, AK

PTD

Port Alexander, AK

2NC

Asheboro, NC

MXE

Maxton, NC

2TX

Brenham, TX

T4X

Llano, TX

7AK

Akun, AK

DUT

Unalaska, AK

7AK

Akun, AK

ANC

Anchorage, AK

7AK

Akun, AK

DUT

Unalaska, AK

7AK

Akun, AK

STG

St. George Island, AK

7AK

Akun, AK

CDB

Cold Bay, AK

A20

Alpine, AK

ANC

Anchorage, AK

A20

Alpine, AK

SCC

Deadhorse, AK

A27

Pogo Mines, AK

FAI

Fairbanks, AK

1,000 rows

|

Truncated data

|

0.29 seconds runtime

Refreshed 353 days ago

Command took 0.29 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:03:33 PM on assignment 3 may 29

################finding distinct airports giving us total number of airports ##############################

```
airports = flights.select("ORIGIN").toDF("id").distinct()
```

Command took 0.06 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:04:35 PM on assignment 3 may 29

```
airportEdges = flights.select("ORIGIN", "DEST").toDF("src","dst")
```

Command took 0.06 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:04:43 PM on assignment 3 may 29

################### importing graphframes ####################

```
from graphframes import GraphFrame
```

Command took 0.03 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:04:47 PM on assignment 3 may 29

```
airportGraph = GraphFrame(airports, airportEdges)
```

```
airportGraph.cache()
```

Out[24]: GraphFrame(v:[id: string], e:[src: string, dst: string])

Command took 0.05 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:04:51 PM on assignment 3 may 29

#printing total number of stations

```
print("Total Number of Stations: " + str(airportGraph.vertices.count()))
```

(2) Spark Jobs

Total Number of Stations: 866

Command took 2.86 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:05:59 PM on assignment 3 may 29

#printing total number of trips in graph#

print("Total Number of Trips in Graph: " + str(airportGraph.edges.count()))

(2) Spark Jobs

Total Number of Trips in Graph: 33673

Command took 0.18 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:06:16 PM on assignment 3 may 29

#printing total number of trips in original data #

print("Total Number of Trips in Original Data: " + str(airportEdges.count()))

(2) Spark Jobs

Total Number of Trips in Original Data: 33673

Command took 0.13 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:06:26 PM on assignment 3 may 29

 # importing sql functions #

from pyspark.sql.functions import *

Command took 0.04 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:06:33 PM on assignment 3 may 29

 # answering q 1 a that says "a. Find the top 5 nodes with the highest outdegree and find the count of the number of outgoingedges in each"

#Top 5 OutDegrees

outDeg = airportGraph.outDegrees

outDeg.orderBy(desc("outDegree")).show(5, truncate = False)

(2) Spark Jobs

+---+---------+

|id |outDegree|

+---+---------+

|ORD|1075 |

|DEN|983 |

|ATL|821 |

|LAS|688 |

|LAX|681 |

+---+---------+

only showing top 5 rows

Command took 1.00 second -- by mongodbalex31@gmail.com at 4/29/2022, 3:06:36 PM on assignment 3 may 29

 # answering question 1 b that says "Find the top 5 nodes with the highest indegree and find the count of the number of incoming edges in each" #

#finding indegress in the ascending order

inDeg = airportGraph.inDegrees

inDeg.orderBy(desc("inDegree")).show(5, truncate = False)

(2) Spark Jobs

+---+--------+

|id |inDegree|

+---+--------+

|ORD|1096 |

|DEN|997 |

|ATL|875 |

|LAS|686 |

|DFW|662 |

+---+--------+

only showing top 5 rows

Command took 0.41 seconds -- by mongodbalex31@gmail.com at 4/29/2022, 3:06:42 PM on assignment 3 may 29

 # answering question 1c that says " . Calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values. You are free to define any suitable parameters."

#PageRank command

ranks = airportGraph.pageRank(resetProbability=0.15, maxIter=10)

ranks.vertices.orderBy(desc("pagerank")).select("id", "pagerank").show(5)

(18) Spark Jobs

+---+------------------+

| id| pagerank|

+---+------------------+

|ORD| 19.0770741959917|

|DEN|18.423589881631706|

|ANC|14.902854843696757|

|ATL| 14.55629426824534|

|DFW|12.584833097708746|

+---+------------------+

only showing top 5 rows

Command took 10.39 minutes -- by mongodbalex31@gmail.com at 4/29/2022, 3:06:47 PM on assignment 3 may 29

 # answering question 1 d that says " Run the strongly connected components algorithm on it and find the top 5 components with thelargest number of nodes. "

#finding Strongly Connected Components

result = airportGraph.stronglyConnectedComponents(maxIter=10)

result.select("id", "component").orderBy(desc("component")).show(5)

(14) Spark Jobs

Command complete

1

 # answering question 1 d that says "Run the triangle counts algorithm on each of the vertices and output the top 5 vertices with the largest triangle count."

2

#running the Triangle Count#

3

results = airportGraph.triangleCount()

4

results.select("id", "count").orderBy(desc("count")).show(5)

(4) Spark Jobs

+---+-----+

| id|count|

+---+-----+

|ATL| 3473|

|DFW| 3358|

|ORD| 3311|

|DEN| 3164|

|MEM| 3137|

+---+-----+

only showing top 5 rows

Command took 36.53 seconds -- by mongodbalex31@gmail.com at 4/27/2022, 7:59:03 AM on vv

Shift+Enter to run

Shift+Ctrl+Enter to run selected text