Created by Yash Patel

Question	Status	Comment
Q1	Fully Working	
Q2	Fully Working	For Reporting neighbors of top 50, Only showing the first 20 neighbors

How to Run:

Inside of this package contains a jar called: managementproject3_2.12-0.1.jar
To run Question 1:

```
spark-submit --class FilterTransactions managementproject3_2.12-0.1.jar
```

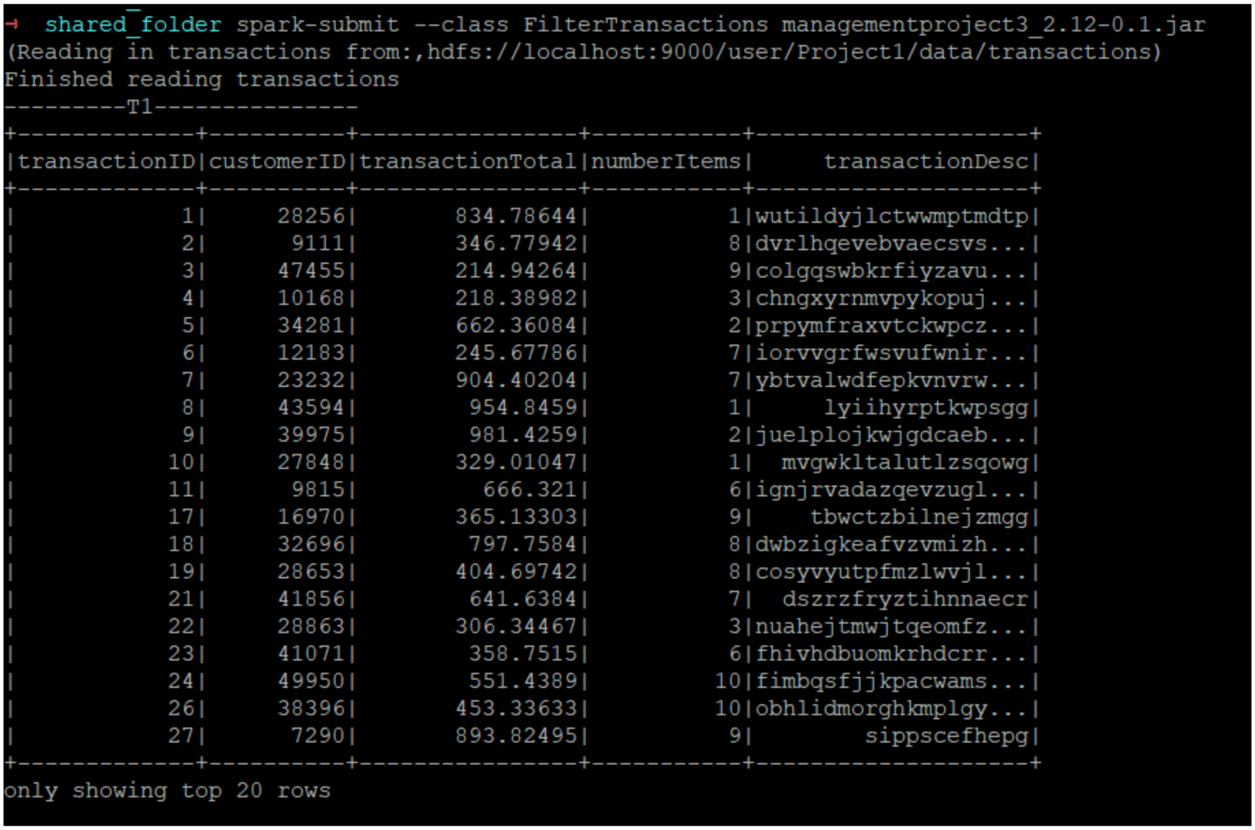
To run Question 2:

```
spark-submit --class GridCells managementproject3_2.12-0.1.jar
```

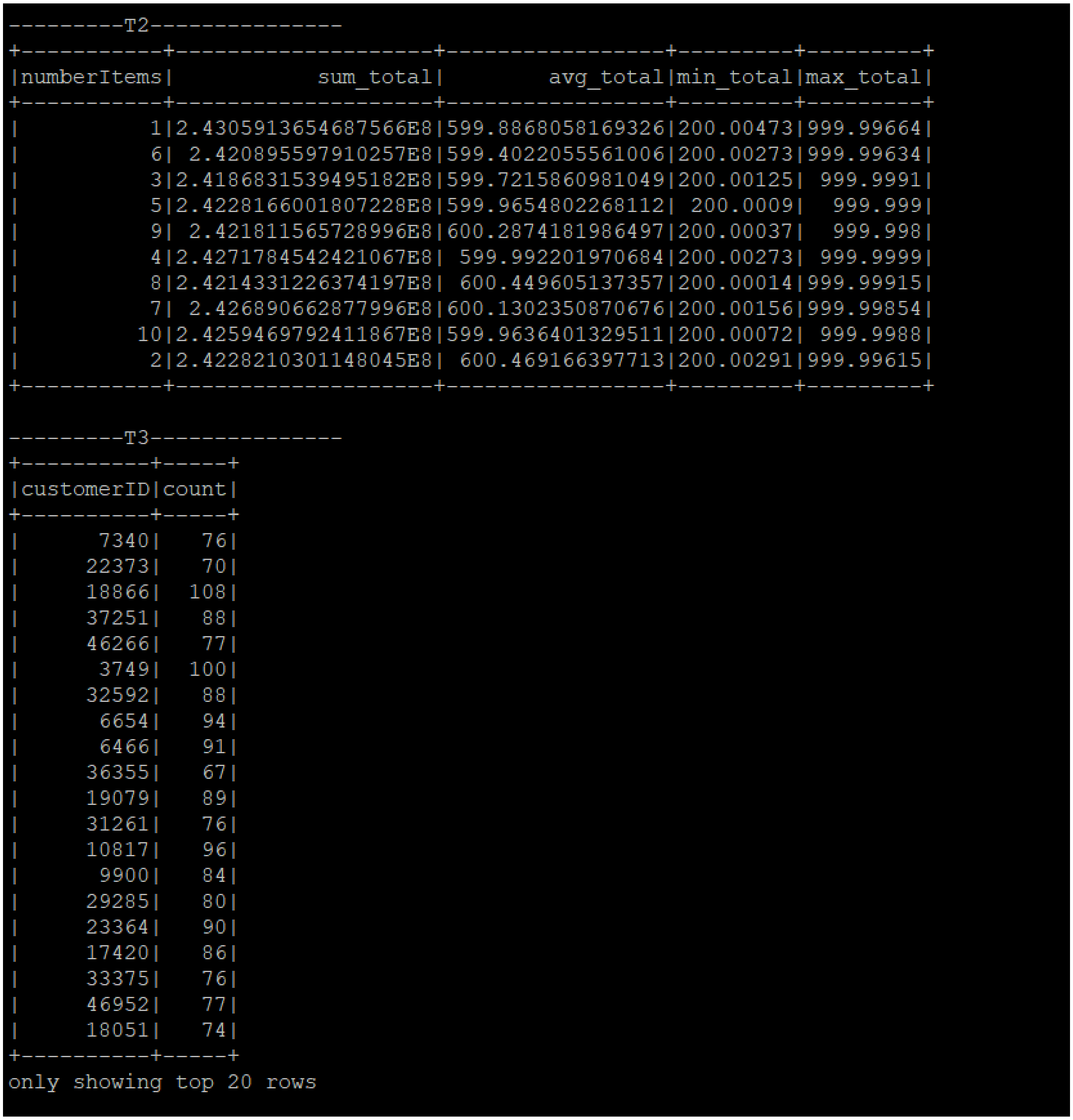
Question 1

I only print out T2 and T6 right now however, T6 doesn't have anything because nothing satisfies the condition for `T5("count") * 5 < T3("count")`

Here are all of the tables

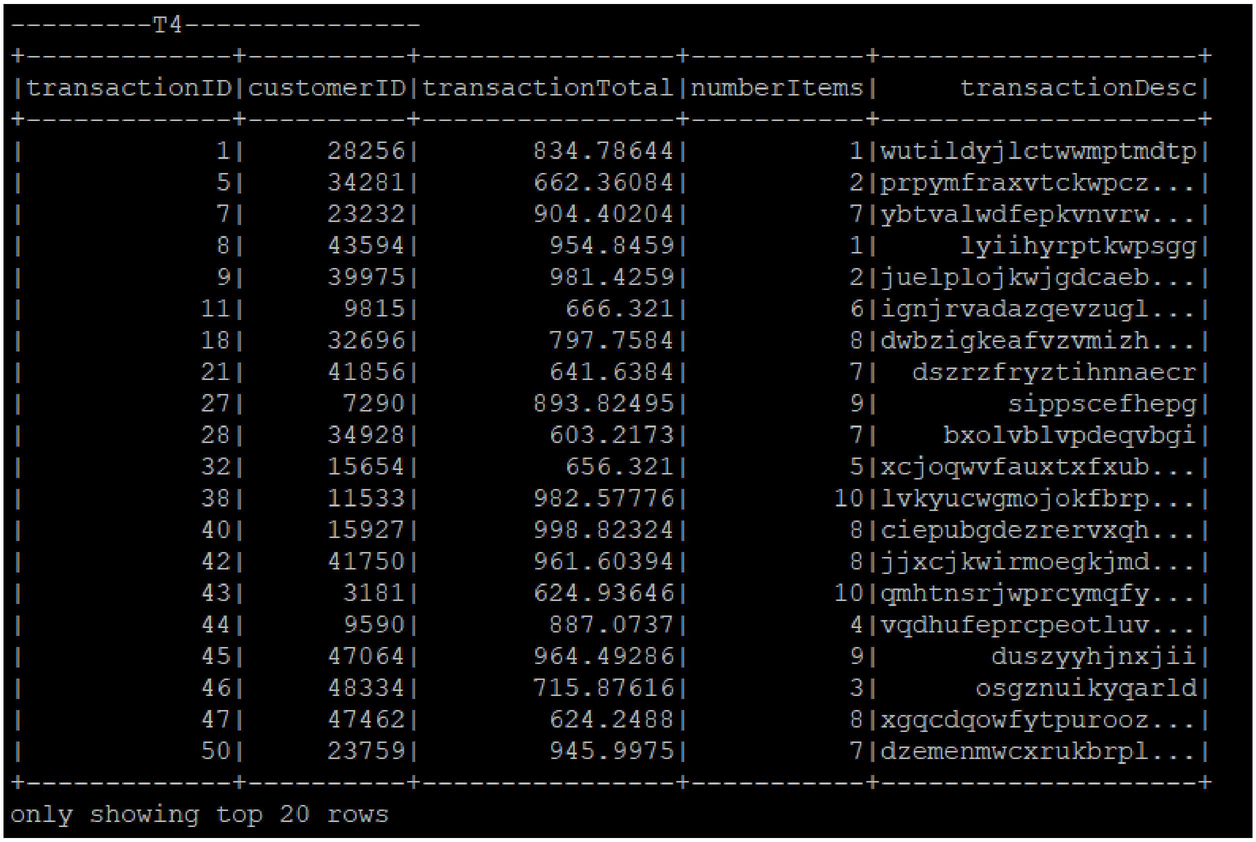


```
➜  shared_folder spark-submit --class FilterTransactions managementproject3_2.12-0.1.jar
(Reading in transactions from:,hdfs://localhost:9000/user/Project1/data/transactions)
Finished reading transactions
---------T1---------------
+-------------+----------+----------------+-----------+--------------------+
|transactionID|customerID|transactionTotal|numberItems|     transactionDesc|
+-------------+----------+----------------+-----------+--------------------+
|            1|     28256|       834.78644|          1|wutildyjlctwwmptmdtp|
|            2|      9111|       346.77942|          8|dvrlhqevebvaecsvs...|
|            3|     47455|       214.94264|          9|colgqswbkrfiyzavu...|
|            4|     10168|       218.38982|          3|chngxyrnmvpykopuj...|
|            5|     34281|       662.36084|          2|prpymfraxvtckwpcz...|
|            6|     12183|       245.67786|          7|iorvvgrfwsvufwnir...|
|            7|     23232|       904.40204|          7|ybtvalwdfepkvnvrw...|
|            8|     43594|        954.8459|          1|     lyiihyrptkwpsgg|
|            9|     39975|        981.4259|          2|juelplojkwjgdcaeb...|
|           10|     27848|       329.01047|          1|  mvgwkltalutlzsqowg|
|           11|      9815|         666.321|          6|ignjrvadazqevzugl...|
|           17|     16970|       365.13303|          9|    tbwctzbilnejzmgg|
|           18|     32696|        797.7584|          8|dwbzigkeafvzvmizh...|
|           19|     28653|       404.69742|          8|cosyvyutpfmzlwvjl...|
|           21|     41856|        641.6384|          7|  dszrzfryztihnnaecr|
|           22|     28863|       306.34467|          3|nuahejtmwjtqeomfz...|
|           23|     41071|        358.7515|          6|fhivhdbuomkrhdcrr...|
|           24|     49950|        551.4389|         10|fimbqsfjjkpacwams...|
|           26|     38396|       453.33633|         10|obhlidmorghkmplgy...|
|           27|      7290|       893.82495|          9|        sippscefhepg|
+-------------+----------+----------------+-----------+--------------------+
only showing top 20 rows
```

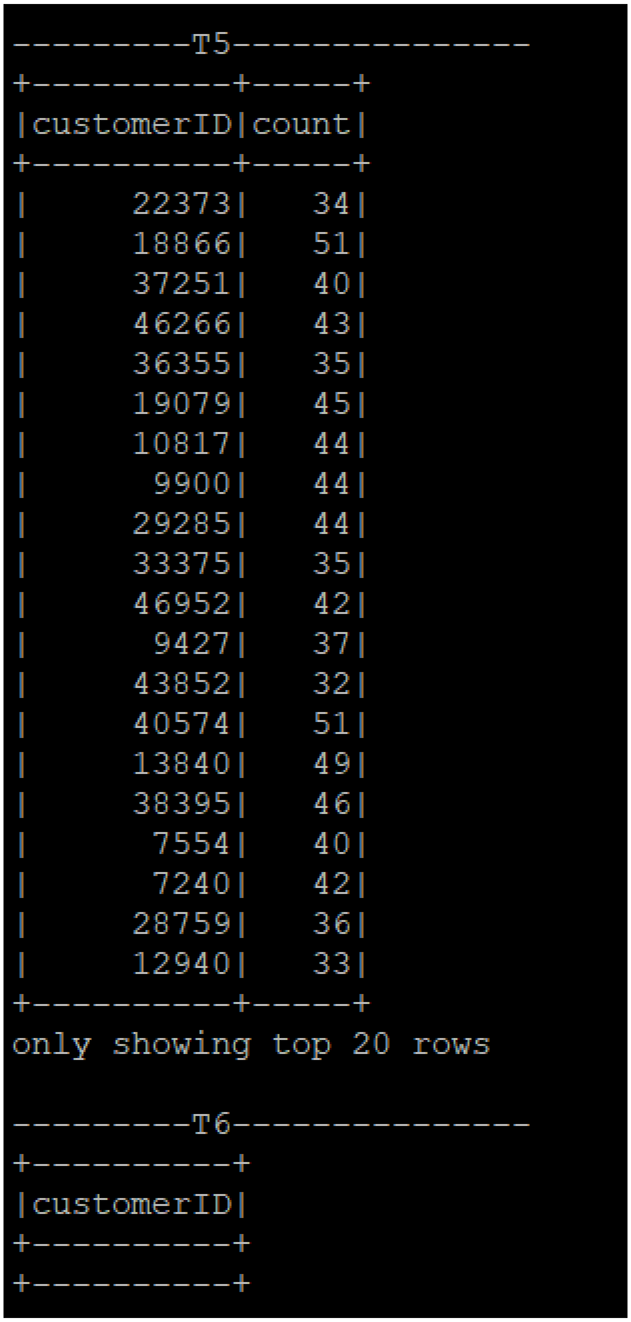


```
---------T2----------------
+-----------+--------------------+-----------------+---------+---------+
|numberItems|           sum_total|        avg_total|min_total|max_total|
+-----------+--------------------+-----------------+---------+---------+
|          1|2.4305913654687566E8|599.8868058169326|200.00473|999.99664|
|          6| 2.420895597910257E8|599.4022055561006|200.00273|999.99634|
|          3|2.4186831539495182E8|599.7215860981049|200.00125| 999.9991|
|          5|2.4228166001807228E8|599.9654802268112| 200.0009|  999.999|
|          9| 2.421811565728996E8|600.2874181986497|200.00037|  999.998|
|          4|2.4271784542421067E8| 599.992201970684|200.00273| 999.9999|
|          8|2.4214331226374197E8| 600.449605137357|200.00014|999.99915|
|          7| 2.426890662877996E8|600.1302350870676|200.00156|999.99854|
|         10|2.4259469792411867E8|599.9636401329511|200.00072| 999.9988|
|          2|2.4228210301148045E8| 600.469166397713|200.00291|999.99615|
+-----------+--------------------+-----------------+---------+---------+

---------T3----------------
+----------+-----+
|customerID|count|
+----------+-----+
|      7340|   76|
|     22373|   70|
|     18866|  108|
|     37251|   88|
|     46266|   77|
|      3749|  100|
|     32592|   88|
|      6654|   94|
|      6466|   91|
|     36355|   67|
|     19079|   89|
|     31261|   76|
|     10817|   96|
|      9900|   84|
|     29285|   80|
|     23364|   90|
|     17420|   86|
|     33375|   76|
|     46952|   77|
|     18051|   74|
+----------+-----+
only showing top 20 rows
```



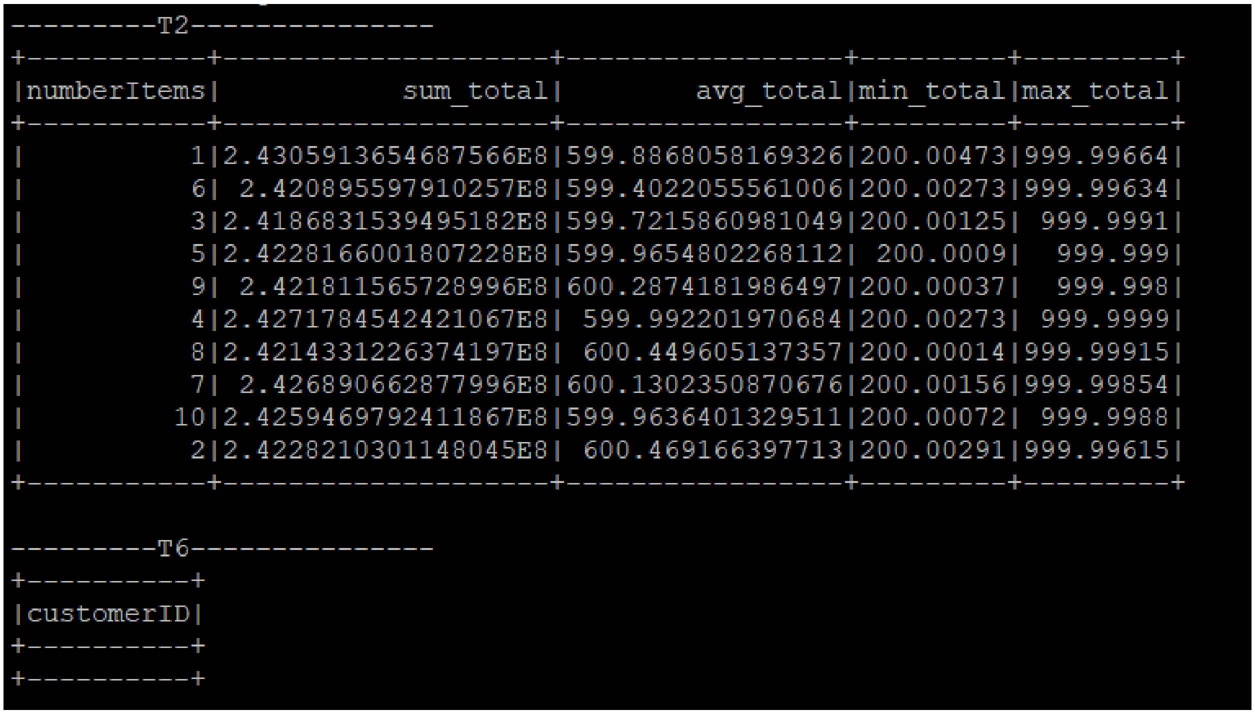
```
---------T4---------------
+-------------+----------+----------------+-----------+--------------------+
|transactionID|customerID|transactionTotal|numberItems|     transactionDesc|
+-------------+----------+----------------+-----------+--------------------+
|            1|     28256|       834.78644|          1|wutildyjlctwwmptmdtp|
|            5|     34281|       662.36084|          2|prpymfraxvtckwpcz...|
|            7|     23232|       904.40204|          7|ybtvalwdfepkvnvrw...|
|            8|     43594|        954.8459|          1|     lyiihyrptkwpsgg|
|            9|     39975|        981.4259|          2|juelplojkwjgdcaeb...|
|           11|      9815|         666.321|          6|ignjrvadazqevzugl...|
|           18|     32696|        797.7584|          8|dwbzigkeafvzvmizh...|
|           21|     41856|        641.6384|          7|  dszrzfryztihnnaecr|
|           27|      7290|       893.82495|          9|        sippscefhepg|
|           28|     34928|        603.2173|          7|    bxolvblvpdeqvbgi|
|           32|     15654|         656.321|          5|xcjoqwvfauxtxfxub...|
|           38|     11533|       982.57776|         10|lvkyucwgmojokfbrp...|
|           40|     15927|       998.82324|          8|ciepubgdezrervxqh...|
|           42|     41750|       961.60394|          8|jjxcjkwirmoegkjmd...|
|           43|      3181|       624.93646|         10|qmhtnsrjwprcymqfy...|
|           44|      9590|        887.0737|          4|vqdhufeprcpeotluv...|
|           45|     47064|       964.49286|          9|       duszyyhjnxjii|
|           46|     48334|       715.87616|          3|      osgznuikyqarld|
|           47|     47462|        624.2488|          8|xgqcdqowfytpurooz...|
|           50|     23759|        945.9975|          7|dzemenmwcxrukbrpl...|
+-------------+----------+----------------+-----------+--------------------+
only showing top 20 rows
```



```
---------T5---------------
+----------+-----+
|customerID|count|
+----------+-----+
|     22373|   34|
|     18866|   51|
|     37251|   40|
|     46266|   43|
|     36355|   35|
|     19079|   45|
|     10817|   44|
|      9900|   44|
|     29285|   44|
|     33375|   35|
|     46952|   42|
|      9427|   37|
|     43852|   32|
|     40574|   51|
|     13840|   49|
|     38395|   46|
|      7554|   40|
|      7240|   42|
|     28759|   36|
|     12940|   33|
+----------+-----+
only showing top 20 rows

---------T6---------------
+----------+
|customerID|
+----------+
+----------+
```

Here is what actually gets printed out since we only report T2 and T6

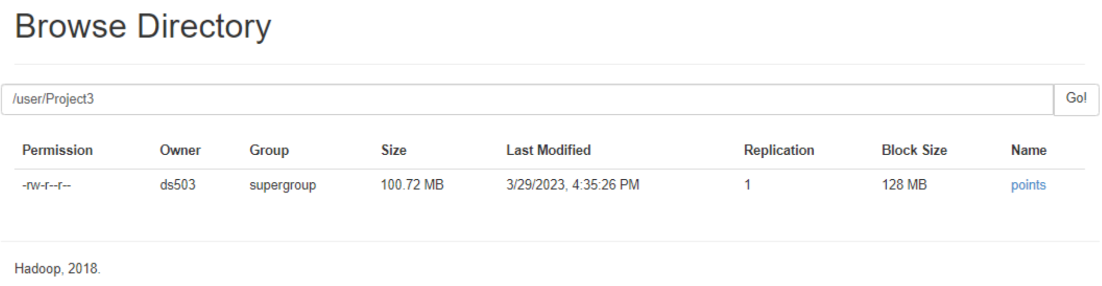


```
---------T2---------------
+-----------+--------------------+-----------------+---------+---------+
|numberItems|           sum_total|        avg_total|min_total|max_total|
+-----------+--------------------+-----------------+---------+---------+
|          1|2.4305913654687566E8|599.8868058169326|200.00473|999.99664|
|          6| 2.420895597910257E8|599.4022055561006|200.00273|999.99634|
|          3|2.4186831539495182E8|599.7215860981049|200.00125| 999.9991|
|          5|2.4228166001807228E8|599.9654802268112| 200.0009|  999.999|
|          9| 2.421811565728996E8|600.2874181986497|200.00037|  999.998|
|          4|2.4271784542421067E8| 599.992201970684|200.00273| 999.9999|
|          8|2.4214331226374197E8| 600.449605137357|200.00014|999.99915|
|          7| 2.426890662877996E8|600.1302350870676|200.00156|999.99854|
|         10|2.4259469792411867E8|599.9636401329511|200.00072| 999.9988|
|          2|2.4228210301148045E8| 600.469166397713|200.00291|999.99615|
+-----------+--------------------+-----------------+---------+---------+

---------T6---------------
+----------+
|customerID|
+----------+
+----------+
```

Question 2

The dataset is created in CreatePointDataset.java. This will create a file called points that will be placed in hdfs



Browse Directory

/user/Project3 Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	ds503	supergroup	100.72 MB	3/29/2023, 4:35:26 PM	1	128 MB	points

Hadoop, 2018.

I used docker to read from hdfs so I am reading from:
"hdfs://localhost:9000/user/Project3/points"

I designed my grids to be from 0 - 249999. This is because I found it easier to use 0-indexing. My rows were from 0 - 499 and the columns were also 0-499. Therefore the grid ids look like this:

0 1 2 … 499

500 501 502 … 999
…
249500 … 249999

I created a map-reduce job that found the grid and its points count.

From there, I then used flatMap for each of the grids to create:
(neighborGrid, gridID+count)

This means that for each of the grids I sent its count to all of the neighbors. I also sent the count to itself which I will explain later.

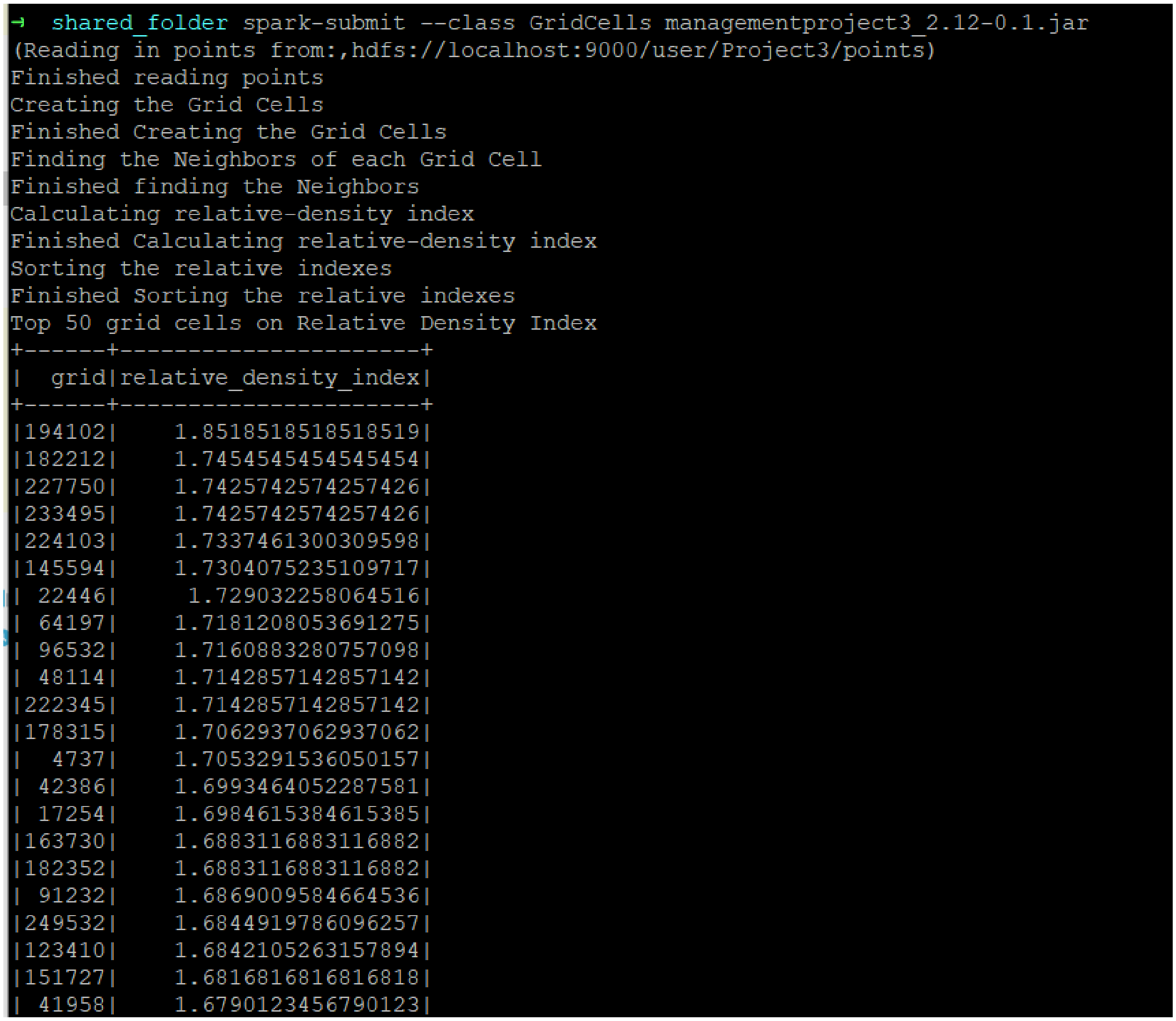
I then did a group by key so that each grid will have the counts of their neighbors + themself. With this information I calculated the relative-density index.

I then sorted the RDD by the density and I took the first 50 rows and printed them.

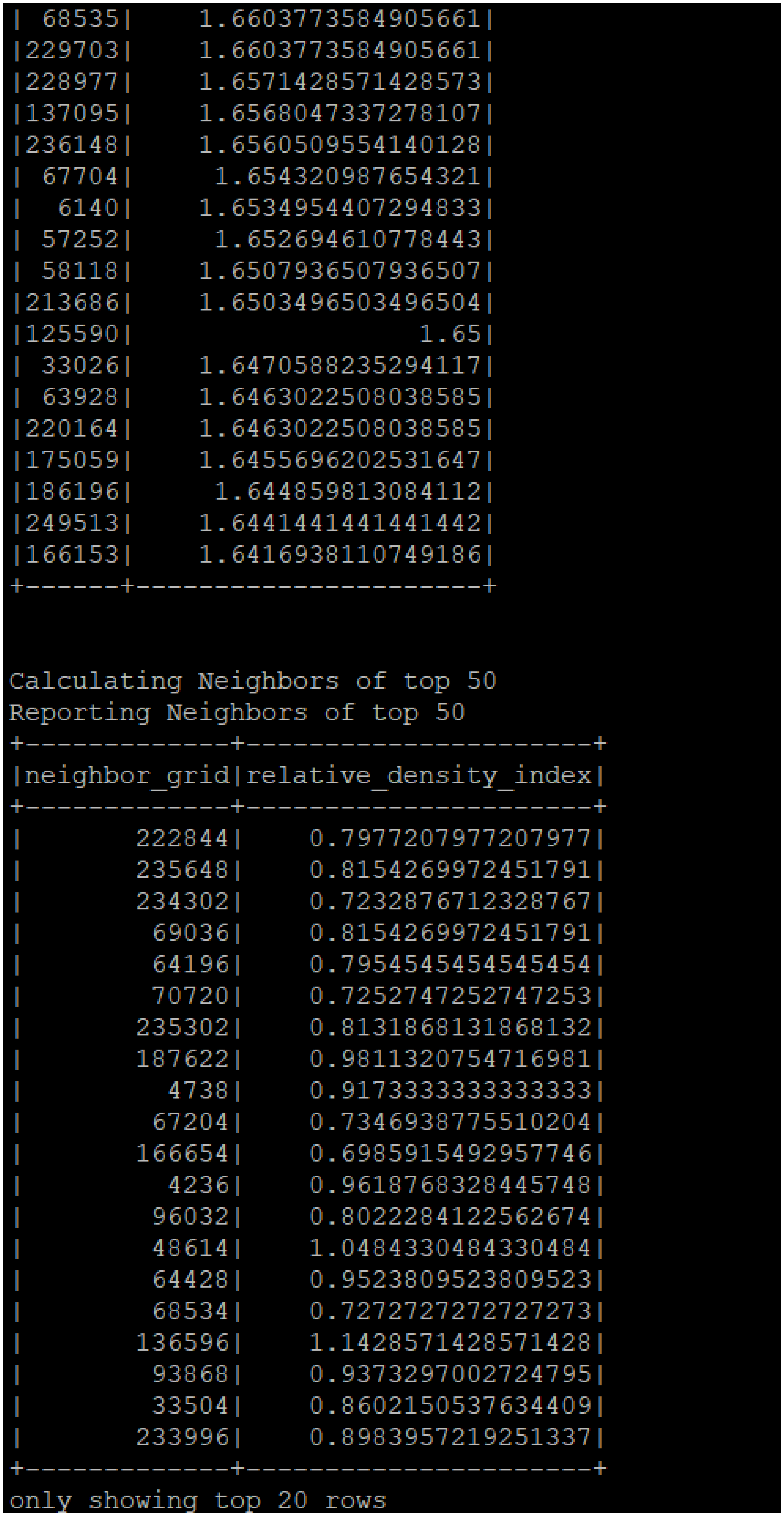
To calculate the neighbors of the top 50, I mapped each of the top50 grids and got their neighbors. Then I reduced it so that we didn't have duplicate neighbors. I joined the RDD with the density index and the neighbor RDD to get the RDD of neighbors of the top 50.

I printed that out as well.

Here are the results:



```
➜  shared_folder spark-submit --class GridCells managementproject3_2.12-0.1.jar
(Reading in points from:,hdfs://localhost:9000/user/Project3/points)
Finished reading points
Creating the Grid Cells
Finished Creating the Grid Cells
Finding the Neighbors of each Grid Cell
Finished finding the Neighbors
Calculating relative-density index
Finished Calculating relative-density index
Sorting the relative indexes
Finished Sorting the relative indexes
Top 50 grid cells on Relative Density Index
+------+----------------------+
|  grid|relative_density_index|
+------+----------------------+
|194102|    1.8518518518518519|
|182212|    1.7454545454545454|
|227750|    1.7425742574257426|
|233495|    1.7425742574257426|
|224103|    1.7337461300309598|
|145594|    1.7304075235109717|
| 22446|     1.729032258064516|
| 64197|    1.7181208053691275|
| 96532|    1.7160883280757098|
| 48114|    1.7142857142857142|
|222345|    1.7142857142857142|
|178315|    1.7062937062937062|
|  4737|    1.7053291536050157|
| 42386|    1.6993464052287581|
| 17254|    1.6984615384615385|
|163730|    1.6883116883116882|
|182352|    1.6883116883116882|
| 91232|    1.6869009584664536|
|249532|    1.6844919786096257|
|123410|    1.6842105263157894|
|151727|    1.6816816816816818|
| 41958|    1.6790123456790123|
```



```
| 68535|    1.6603773584905661|
|229703|    1.6603773584905661|
|228977|    1.6571428571428573|
|137095|    1.6568047337278107|
|236148|    1.6560509554140128|
| 67704|     1.654320987654321|
|  6140|    1.6534954407294833|
| 57252|     1.652694610778443|
| 58118|    1.6507936507936507|
|213686|    1.6503496503496504|
|125590|                  1.65|
| 33026|    1.6470588235294117|
| 63928|    1.6463022508038585|
|220164|    1.6463022508038585|
|175059|    1.6455696202531647|
|186196|     1.644859813084112|
|249513|    1.6441441441441442|
|166153|    1.6416938110749186|
+------+----------------------+


Calculating Neighbors of top 50
Reporting Neighbors of top 50
+-------------+----------------------+
|neighbor_grid|relative_density_index|
+-------------+----------------------+
|       222844|    0.7977207977207977|
|       235648|    0.8154269972451791|
|       234302|    0.7232876712328767|
|        69036|    0.8154269972451791|
|        64196|    0.7954545454545454|
|        70720|    0.7252747252747253|
|       235302|    0.8131868131868132|
|       187622|    0.9811320754716981|
|         4738|    0.9173333333333333|
|        67204|    0.7346938775510204|
|       166654|    0.6985915492957746|
|         4236|    0.9618768328445748|
|        96032|    0.8022284122562674|
|        48614|    1.0484330484330484|
|        64428|    0.9523809523809523|
|        68534|    0.7272727272727273|
|       136596|    1.1428571428571428|
|        93868|    0.9373297002724795|
|        33504|    0.8602150537634409|
|       233996|    0.8983957219251337|
+-------------+----------------------+
only showing top 20 rows
```

I on;y showed the top 20 of the second table because there were too many too show but the RDD does exist.