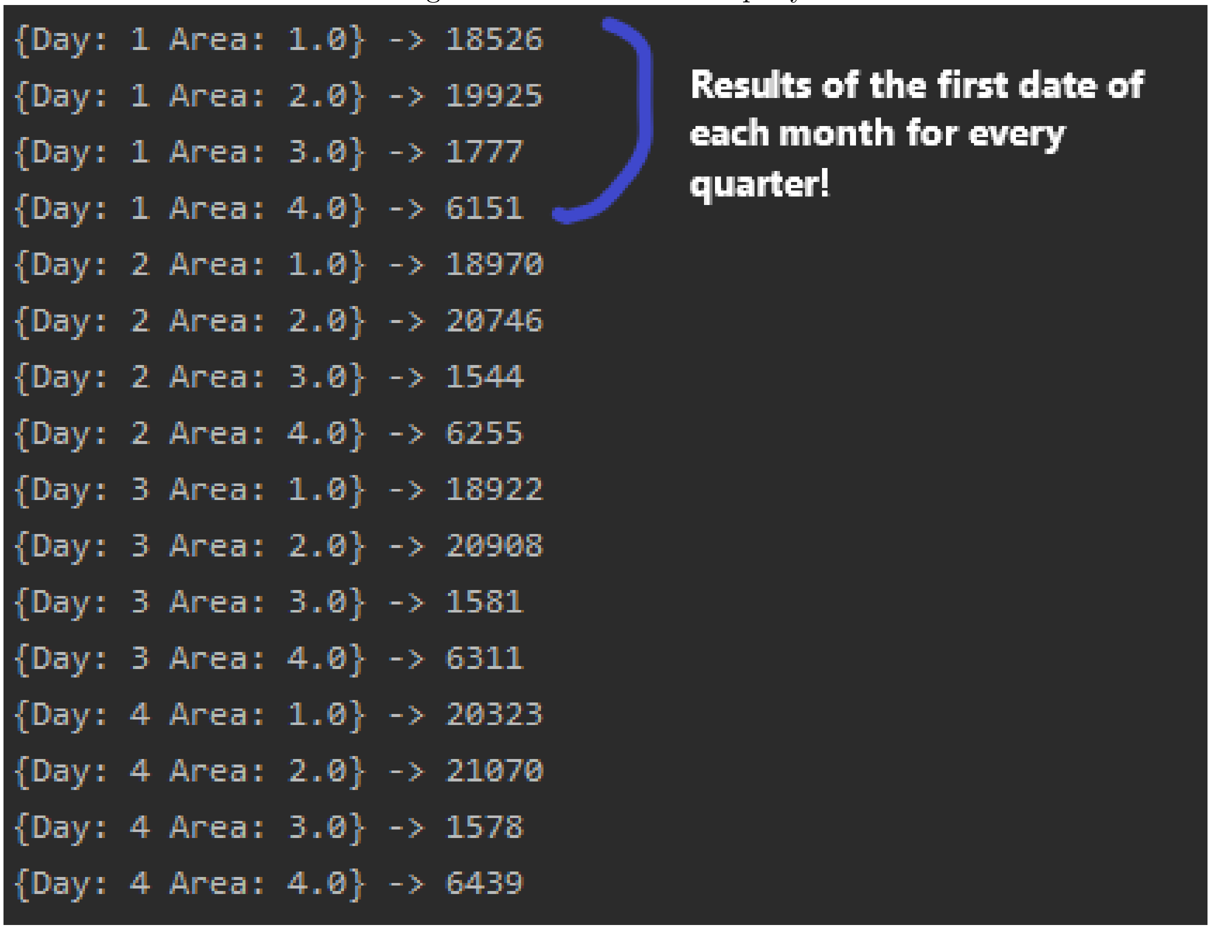
Distributed Systems **Spark-Hadoop-Docker**

Γιαννακόπουλος Αθανάσιος
Χατζηλυγερούδη Γεωργία
Τρογκάνη Παναγιώτα

First Part - Spark

Here we run some simple queries, using Spark.

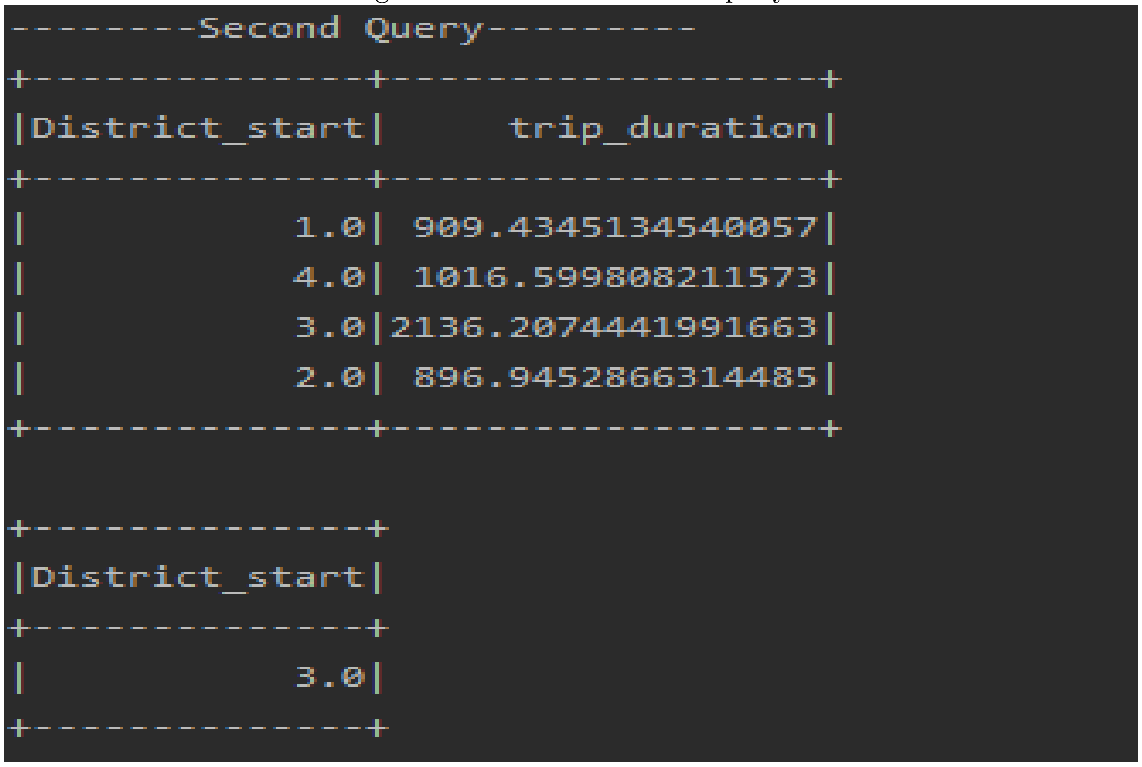
Figure 1: Answer for first query.



```
{Day: 1 Area: 1.0} -> 18526
{Day: 1 Area: 2.0} -> 19925
{Day: 1 Area: 3.0} -> 1777
{Day: 1 Area: 4.0} -> 6151
{Day: 2 Area: 1.0} -> 18970
{Day: 2 Area: 2.0} -> 20746
{Day: 2 Area: 3.0} -> 1544
{Day: 2 Area: 4.0} -> 6255
{Day: 3 Area: 1.0} -> 18922
{Day: 3 Area: 2.0} -> 20908
{Day: 3 Area: 3.0} -> 1581
{Day: 3 Area: 4.0} -> 6311
{Day: 4 Area: 1.0} -> 20323
{Day: 4 Area: 2.0} -> 21070
{Day: 4 Area: 3.0} -> 1578
{Day: 4 Area: 4.0} -> 6439
```

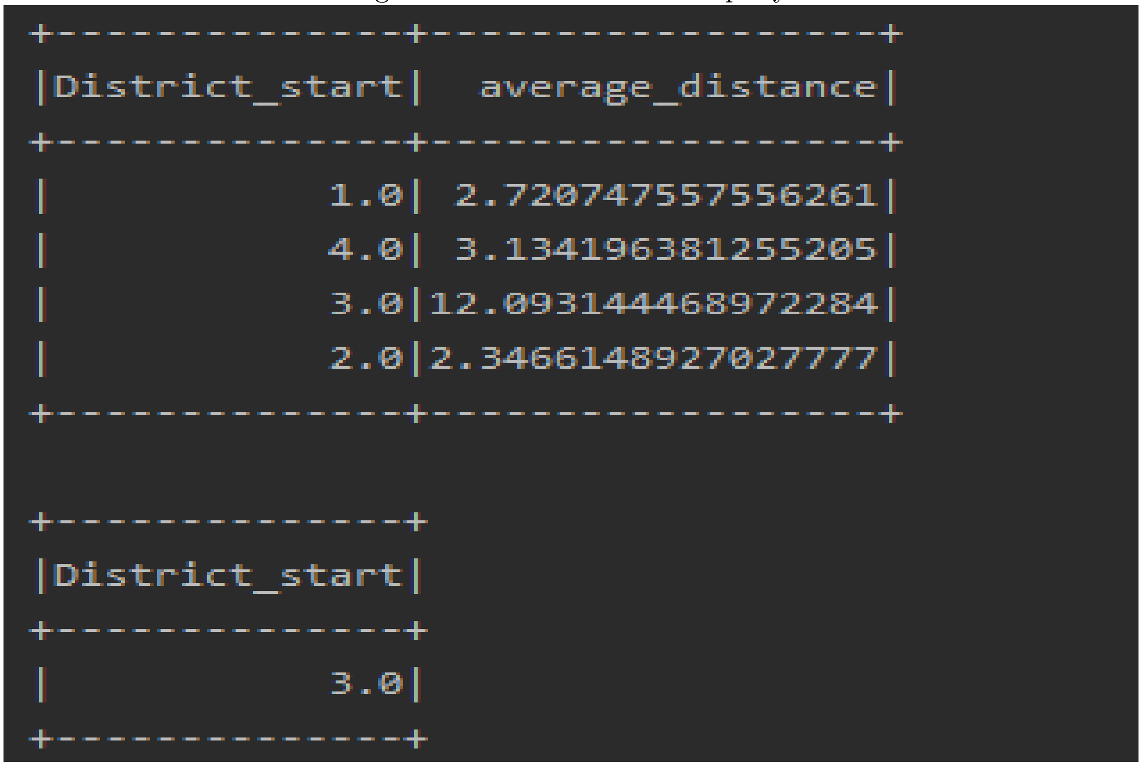
Results of the first date of each month for every quarter!

Figure 2: Answer for second query.



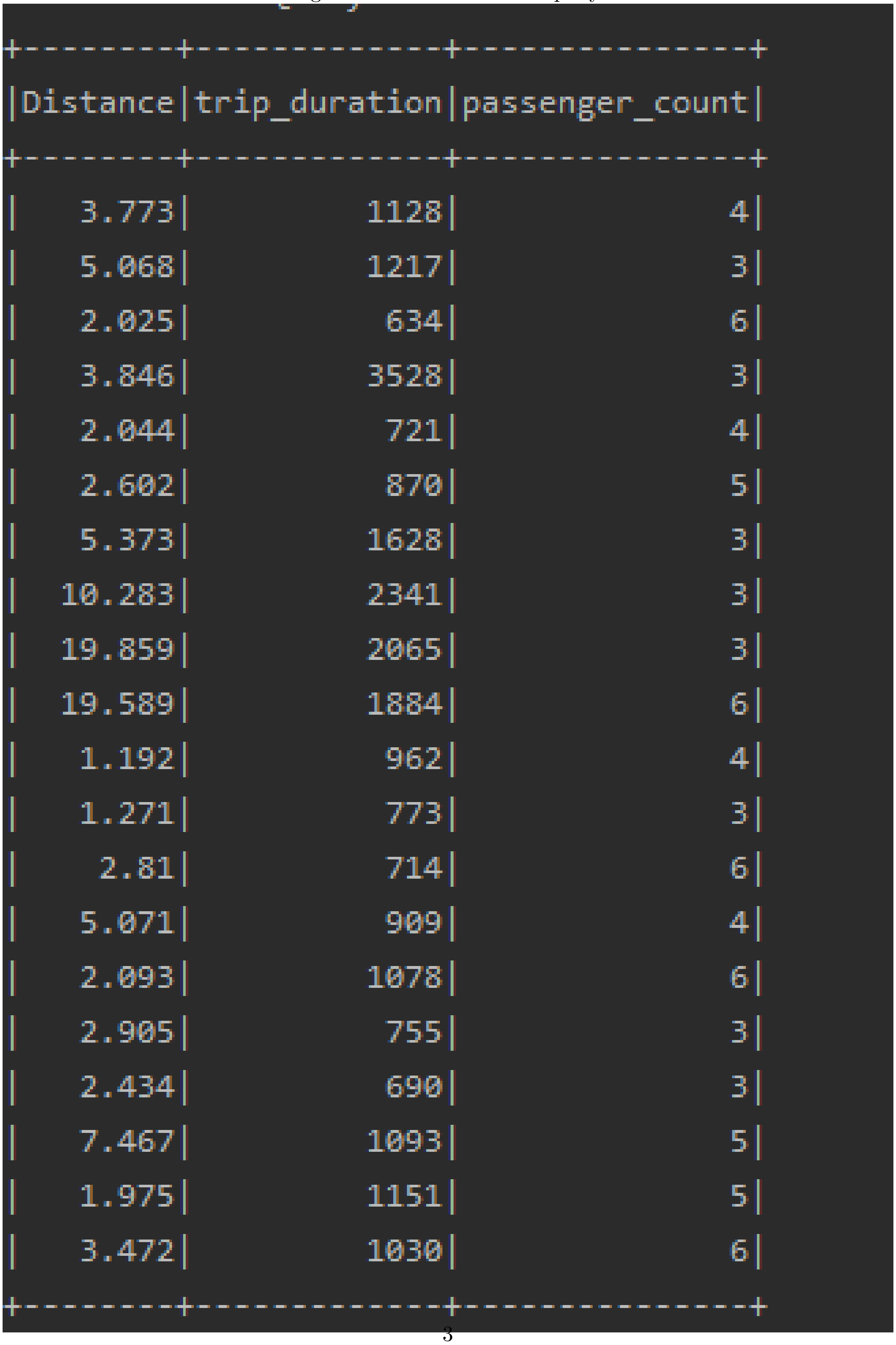
```
---------Second Query----------
+--------------+------------------+
|District_start|     trip_duration|
+--------------+------------------+
|           1.0| 909.4345134540057|
|           4.0| 1016.599808211573|
|           3.0|2136.2074441991663|
|           2.0| 896.9452866314485|
+--------------+------------------+


+--------------+
|District_start|
+--------------+
|           3.0|
+--------------+
```

Figure 3: Answer for second query.



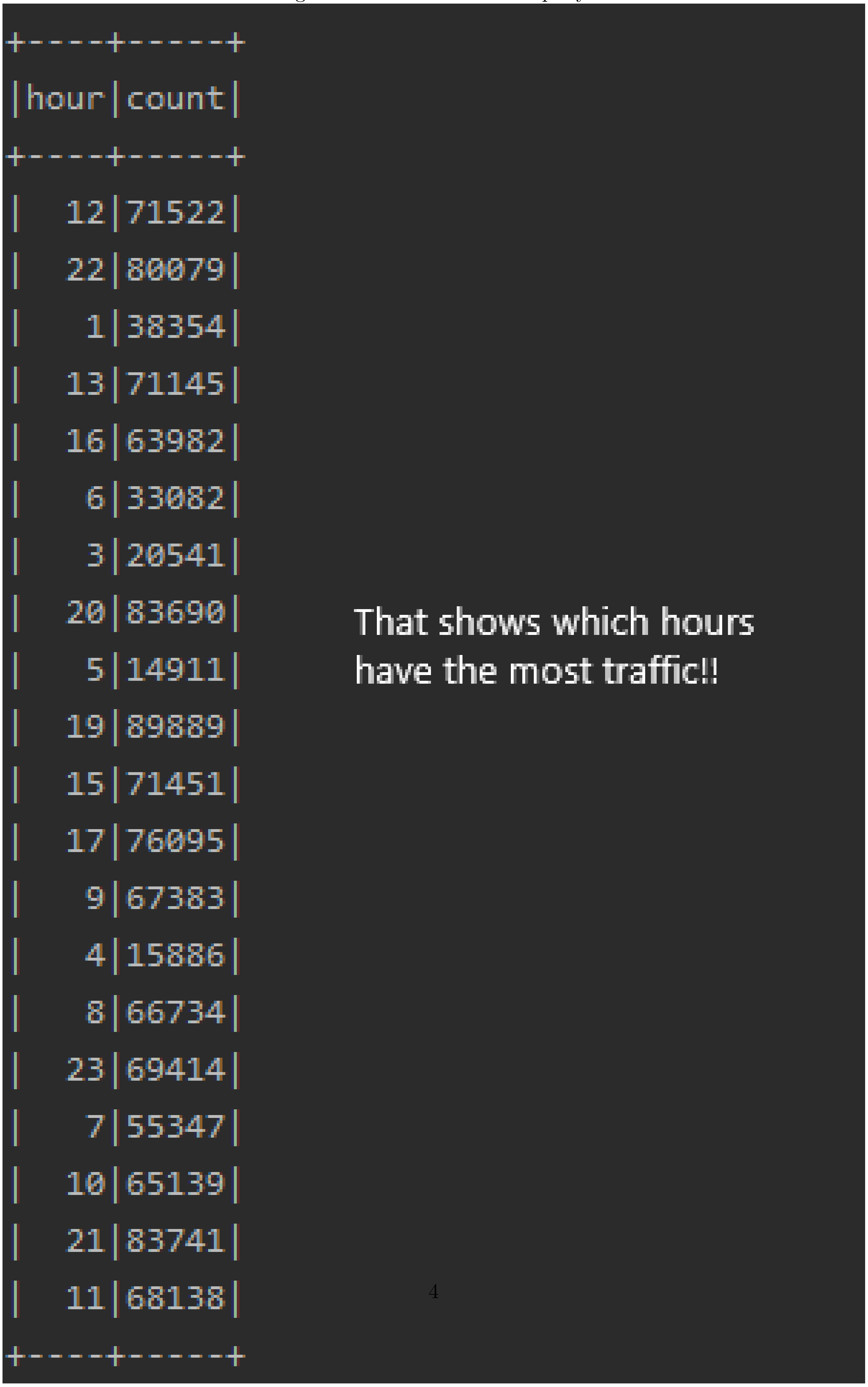
```
+--------------+------------------+
|District_start|  average_distance|
+--------------+------------------+
|           1.0| 2.720747557556261|
|           4.0| 3.134196381255205|
|           3.0|12.093144468972284|
|           2.0|2.3466148927027777|
+--------------+------------------+


+--------------+
|District_start|
+--------------+
|           3.0|
+--------------+
```

Figure 4: Answer for third query.



```
+--------+-------------+---------------+
|Distance|trip_duration|passenger_count|
+--------+-------------+---------------+
|   3.773|         1128|              4|
|   5.068|         1217|              3|
|   2.025|          634|              6|
|   3.846|         3528|              3|
|   2.044|          721|              4|
|   2.602|          870|              5|
|   5.373|         1628|              3|
|  10.283|         2341|              3|
|  19.859|         2065|              3|
|  19.589|         1884|              6|
|   1.192|          962|              4|
|   1.271|          773|              3|
|    2.81|          714|              6|
|   5.071|          909|              4|
|   2.093|         1078|              6|
|   2.905|          755|              3|
|   2.434|          690|              3|
|   7.467|         1093|              5|
|   1.975|         1151|              5|
|   3.472|         1030|              6|
+--------+-------------+---------------+
```

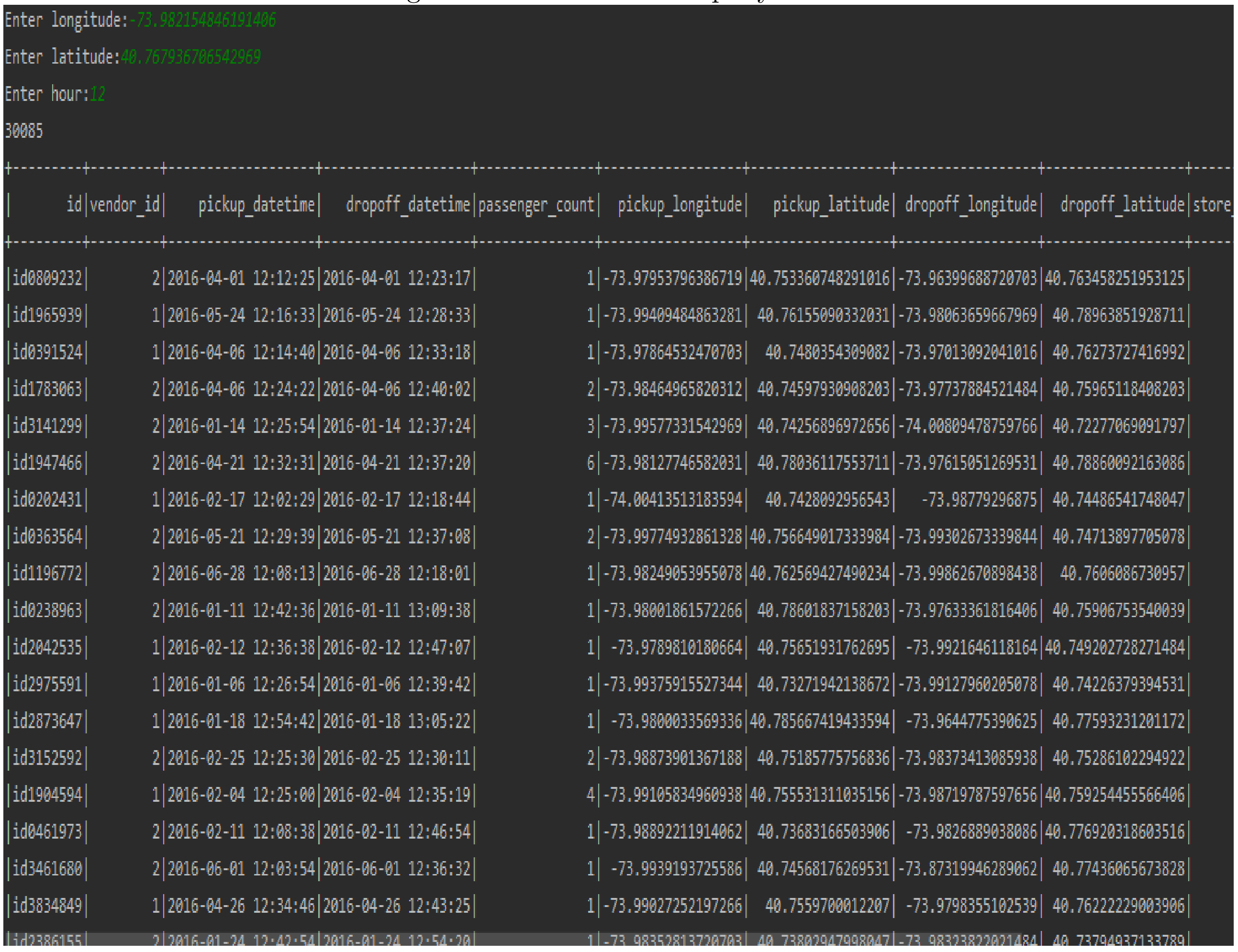
Figure 5: Answer for fourth query.



```
+----+-----+
|hour|count|
+----+-----+
|  12|71522|
|  22|80079|
|   1|38354|
|  13|71145|
|  16|63982|
|   6|33082|
|   3|20541|
|  20|83690|
|   5|14911|
|  19|89889|
|  15|71451|
|  17|76095|
|   9|67383|
|   4|15886|
|   8|66734|
|  23|69414|
|   7|55347|
|  10|65139|
|  21|83741|
|  11|68138|
+----+-----+
```

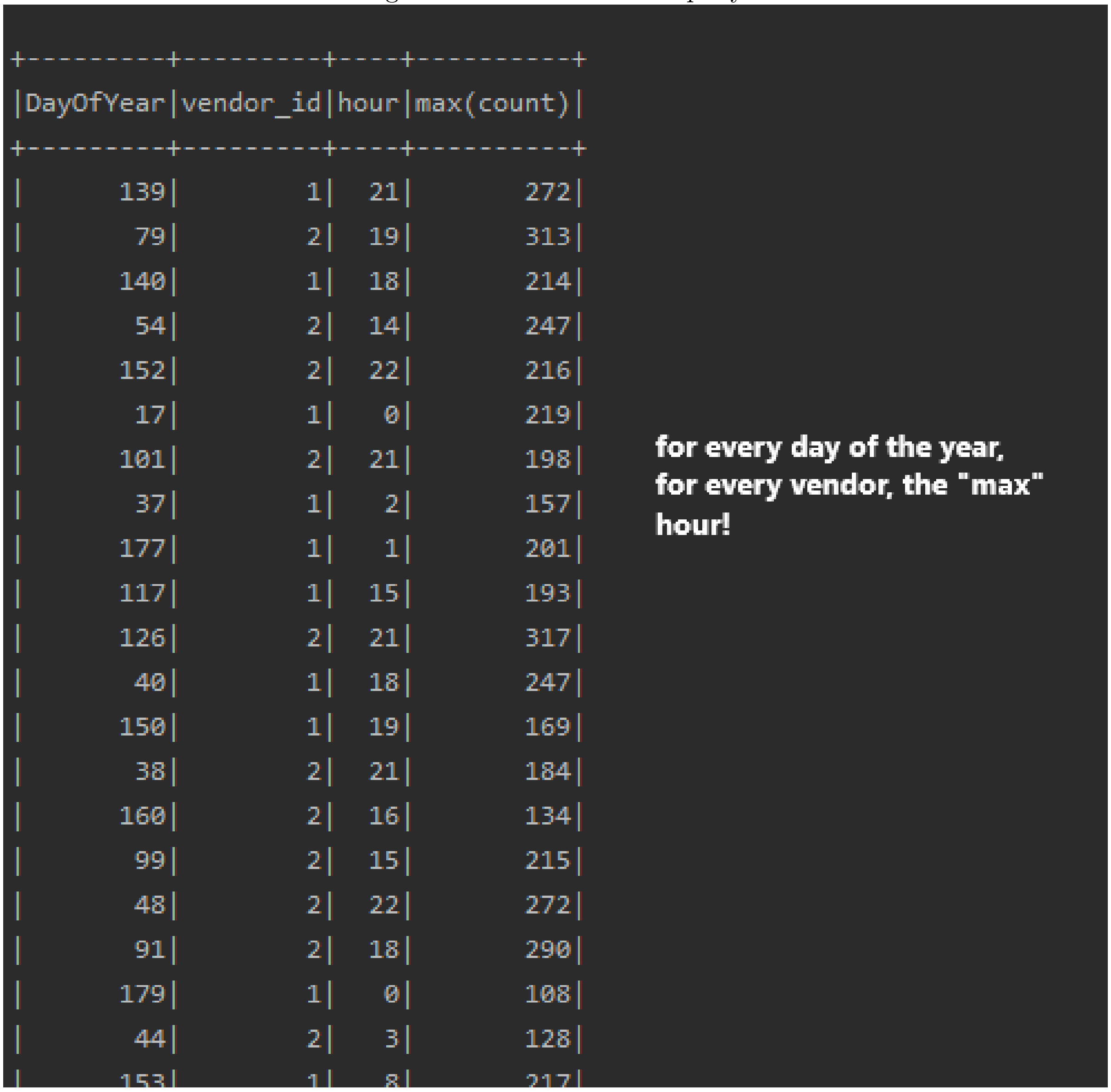
That shows which hours have the most traffic!!

Figure 6: Answer for fifth query.



```
Enter longitude:-73.982154846191406
Enter latitude:40.767936706542969
Enter hour:12
30085
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------+-----
|       id|vendor_id|    pickup_datetime|   dropoff_datetime|passenger_count|  pickup_longitude|   pickup_latitude| dropoff_longitude|  dropoff_latitude|store_
+---------+---------+-------------------+-------------------+---------------+------------------+------------------+------------------+------------------+-----
|id0809232|        2|2016-04-01 12:12:25|2016-04-01 12:23:17|              1|-73.97953796386719|40.753360748291016|-73.96399688720703|40.763458251953125|
|id1965939|        1|2016-05-24 12:16:33|2016-05-24 12:28:33|              1|-73.99409484863281| 40.76155090332031|-73.98063659667969| 40.78963851928711|
|id0391524|        1|2016-04-06 12:14:40|2016-04-06 12:33:18|              1|-73.97864532470703|  40.7480354309082|-73.97013092041016| 40.76273727416992|
|id1783063|        2|2016-04-06 12:24:22|2016-04-06 12:40:02|              2|-73.98464965820312| 40.74597930908203|-73.97737884521484| 40.75965118408203|
|id3141299|        2|2016-01-14 12:25:54|2016-01-14 12:37:24|              3|-73.99577331542969| 40.74256896972656|-74.00809478759766| 40.72277069091797|
|id1947466|        2|2016-04-21 12:32:31|2016-04-21 12:37:20|              6|-73.98127746582031| 40.78036117553711|-73.97615051269531| 40.78860092163086|
|id0202431|        1|2016-02-17 12:02:29|2016-02-17 12:18:44|              1|-74.00413513183594|  40.7428092956543|    -73.98779296875| 40.74486541748047|
|id0363564|        2|2016-05-21 12:29:39|2016-05-21 12:37:08|              2|-73.99774932861328|40.756649017333984|-73.99302673339844| 40.74713897705078|
|id1196772|        2|2016-06-28 12:08:13|2016-06-28 12:18:01|              1|-73.98249053955078|40.762569427490234|-73.99862670898438|  40.7606086730957|
|id0238963|        2|2016-01-11 12:42:36|2016-01-11 13:09:38|              1|-73.98001861572266| 40.78601837158203|-73.97633361816406| 40.75906753540039|
|id2042535|        1|2016-02-12 12:36:38|2016-02-12 12:47:07|              1| -73.9789810180664| 40.75651931762695| -73.9921646118164|40.749202728271484|
|id2975591|        1|2016-01-06 12:26:54|2016-01-06 12:39:42|              1|-73.99375915527344| 40.73271942138672|-73.99127960205078| 40.74226379394531|
|id2873647|        1|2016-01-18 12:54:42|2016-01-18 13:05:22|              1| -73.9800033569336|40.785667419433594| -73.9644775390625| 40.77593231201172|
|id3152592|        2|2016-02-25 12:25:30|2016-02-25 12:30:11|              2|-73.98873901367188| 40.75185775756836|-73.98373413085938| 40.75286102294922|
|id1904594|        1|2016-02-04 12:25:00|2016-02-04 12:35:19|              4|-73.99105834960938|40.755531311035156|-73.98719787597656|40.759254455566406|
|id0461973|        2|2016-02-11 12:08:38|2016-02-11 12:46:54|              1|-73.98892211914062| 40.73683166503906| -73.9826889038086|40.776920318603516|
|id3461680|        2|2016-06-01 12:03:54|2016-06-01 12:36:32|              1| -73.9939193725586| 40.74568176269531|-73.87319946289062| 40.77436065673828|
|id3834849|        1|2016-04-26 12:34:46|2016-04-26 12:43:25|              1|-73.99027252197266|  40.7559700012207| -73.9798355102539| 40.76222229003906|
|id2386155|        2|2016-01-24 12:42:54|2016-01-24 12:54:20|              1|-73.98352813720703| 40.73802947998047|-73.98323822021484| 40.73794937133789|
```

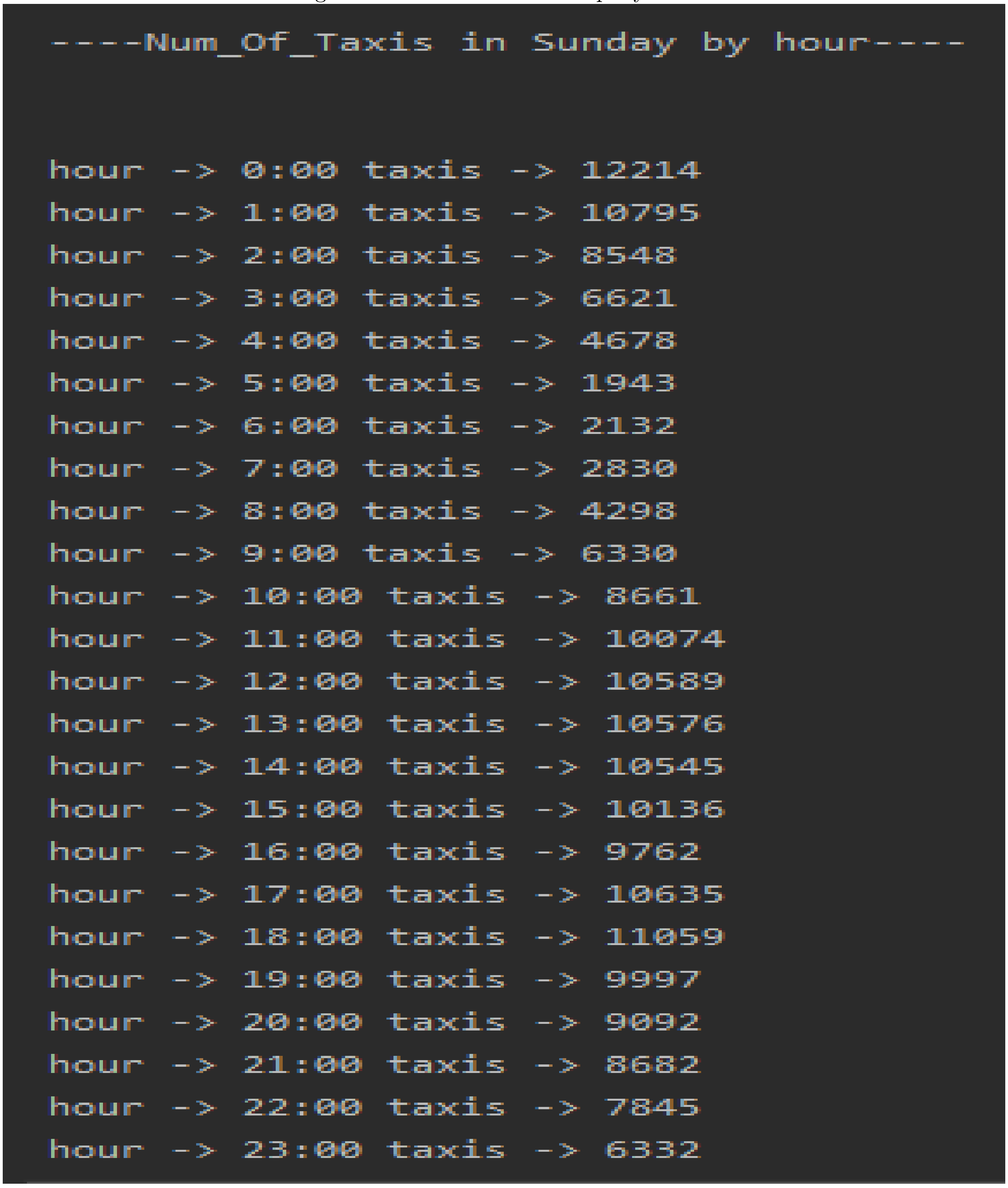
Figure 7: Answer for sixth query.



```
+---------+---------+----+----------+
|DayOfYear|vendor_id|hour|max(count)|
+---------+---------+----+----------+
|      139|        1|  21|       272|
|       79|        2|  19|       313|
|      140|        1|  18|       214|
|       54|        2|  14|       247|
|      152|        2|  22|       216|
|       17|        1|   0|       219|
|      101|        2|  21|       198|
|       37|        1|   2|       157|
|      177|        1|   1|       201|
|      117|        1|  15|       193|
|      126|        2|  21|       317|
|       40|        1|  18|       247|
|      150|        1|  19|       169|
|       38|        2|  21|       184|
|      160|        2|  16|       134|
|       99|        2|  15|       215|
|       48|        2|  22|       272|
|       91|        2|  18|       290|
|      179|        1|   0|       108|
|       44|        2|   3|       128|
|      153|        1|   8|       217|
```

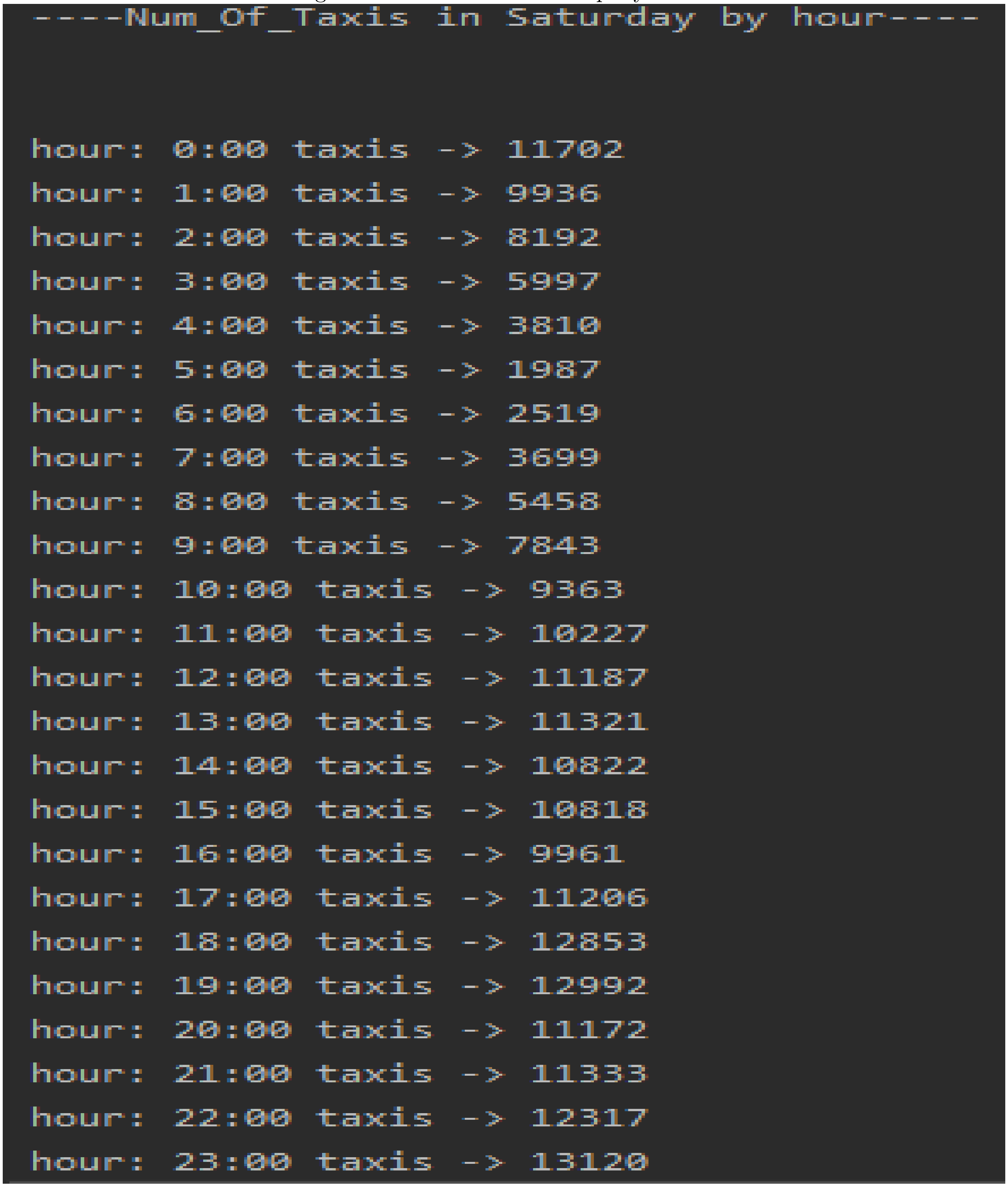
for every day of the year, for every vendor, the "max" hour!

Figure 8: Answer for seventh query.



```
----Num_Of_Taxis in Sunday by hour----


hour -> 0:00 taxis -> 12214
hour -> 1:00 taxis -> 10795
hour -> 2:00 taxis -> 8548
hour -> 3:00 taxis -> 6621
hour -> 4:00 taxis -> 4678
hour -> 5:00 taxis -> 1943
hour -> 6:00 taxis -> 2132
hour -> 7:00 taxis -> 2830
hour -> 8:00 taxis -> 4298
hour -> 9:00 taxis -> 6330
hour -> 10:00 taxis -> 8661
hour -> 11:00 taxis -> 10074
hour -> 12:00 taxis -> 10589
hour -> 13:00 taxis -> 10576
hour -> 14:00 taxis -> 10545
hour -> 15:00 taxis -> 10136
hour -> 16:00 taxis -> 9762
hour -> 17:00 taxis -> 10635
hour -> 18:00 taxis -> 11059
hour -> 19:00 taxis -> 9997
hour -> 20:00 taxis -> 9092
hour -> 21:00 taxis -> 8682
hour -> 22:00 taxis -> 7845
hour -> 23:00 taxis -> 6332
```

Figure 9: Answer for seventh query.

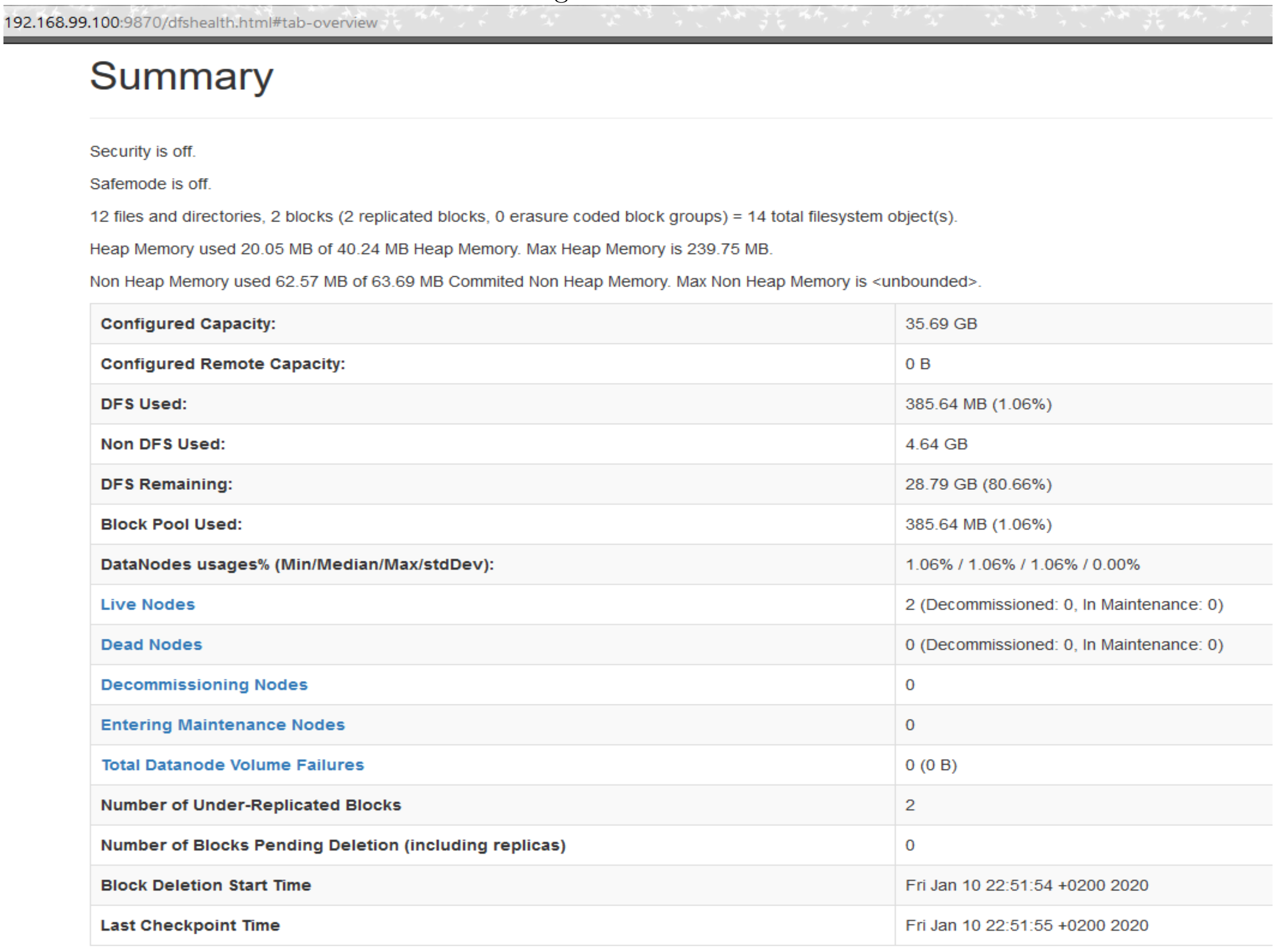


```
----Num_Of_Taxis in Saturday by hour----



hour: 0:00 taxis -> 11702
hour: 1:00 taxis -> 9936
hour: 2:00 taxis -> 8192
hour: 3:00 taxis -> 5997
hour: 4:00 taxis -> 3810
hour: 5:00 taxis -> 1987
hour: 6:00 taxis -> 2519
hour: 7:00 taxis -> 3699
hour: 8:00 taxis -> 5458
hour: 9:00 taxis -> 7843
hour: 10:00 taxis -> 9363
hour: 11:00 taxis -> 10227
hour: 12:00 taxis -> 11187
hour: 13:00 taxis -> 11321
hour: 14:00 taxis -> 10822
hour: 15:00 taxis -> 10818
hour: 16:00 taxis -> 9961
hour: 17:00 taxis -> 11206
hour: 18:00 taxis -> 12853
hour: 19:00 taxis -> 12992
hour: 20:00 taxis -> 11172
hour: 21:00 taxis -> 11333
hour: 22:00 taxis -> 12317
hour: 23:00 taxis -> 13120
```

Second Part - Spark-Hadoop-Docker

Here we made two datanodes and upload the data file.

Figure 10:



192.168.99.100:9870/dfshealth.html#tab-overview

Summary

Security is off.

Safemode is off.

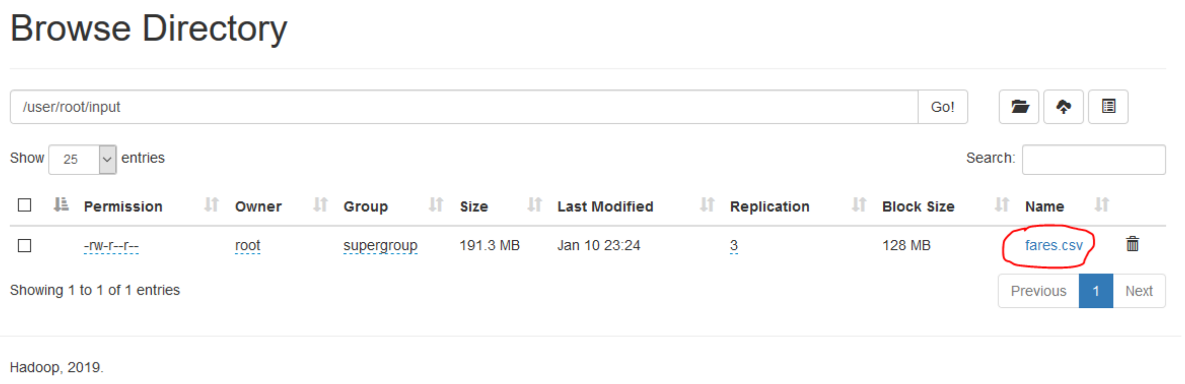
12 files and directories, 2 blocks (2 replicated blocks, 0 erasure coded block groups) = 14 total filesystem object(s).

Heap Memory used 20.05 MB of 40.24 MB Heap Memory. Max Heap Memory is 239.75 MB.

Non Heap Memory used 62.57 MB of 63.69 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	35.69 GB
Configured Remote Capacity:	0 B
DFS Used:	385.64 MB (1.06%)
Non DFS Used:	4.64 GB
DFS Remaining:	28.79 GB (80.66%)
Block Pool Used:	385.64 MB (1.06%)
DataNodes usages% (Min/Median/Max/stdDev):	1.06% / 1.06% / 1.06% / 0.00%
Live Nodes	2 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	2
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Fri Jan 10 22:51:54 +0200 2020
Last Checkpoint Time	Fri Jan 10 22:51:55 +0200 2020

Figure 11: The data file.



Browse Directory

/user/root/input Go!

Show 25 entries Search:

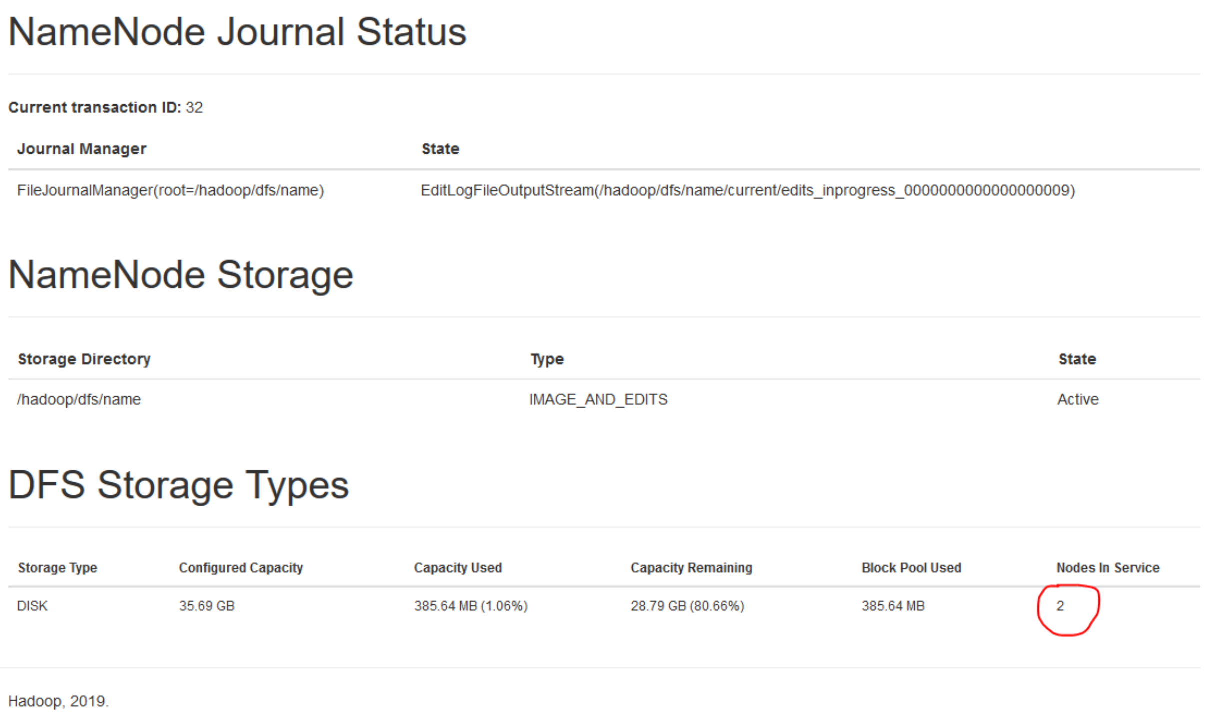
	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	-rw-r--r--	root	supergroup	191.3 MB	Jan 10 23:24	3	128 MB	fares.csv

Showing 1 to 1 of 1 entries

Previous 1 Next

Hadoop, 2019.

Figure 12:



NameNode Journal Status

Current transaction ID: 32

Journal Manager	State
FileJournalManager(root=/hadoop/dfs/name)	EditLogFileOutputStream(/hadoop/dfs/name/current/edits_inprogress_0000000000000000009)

NameNode Storage

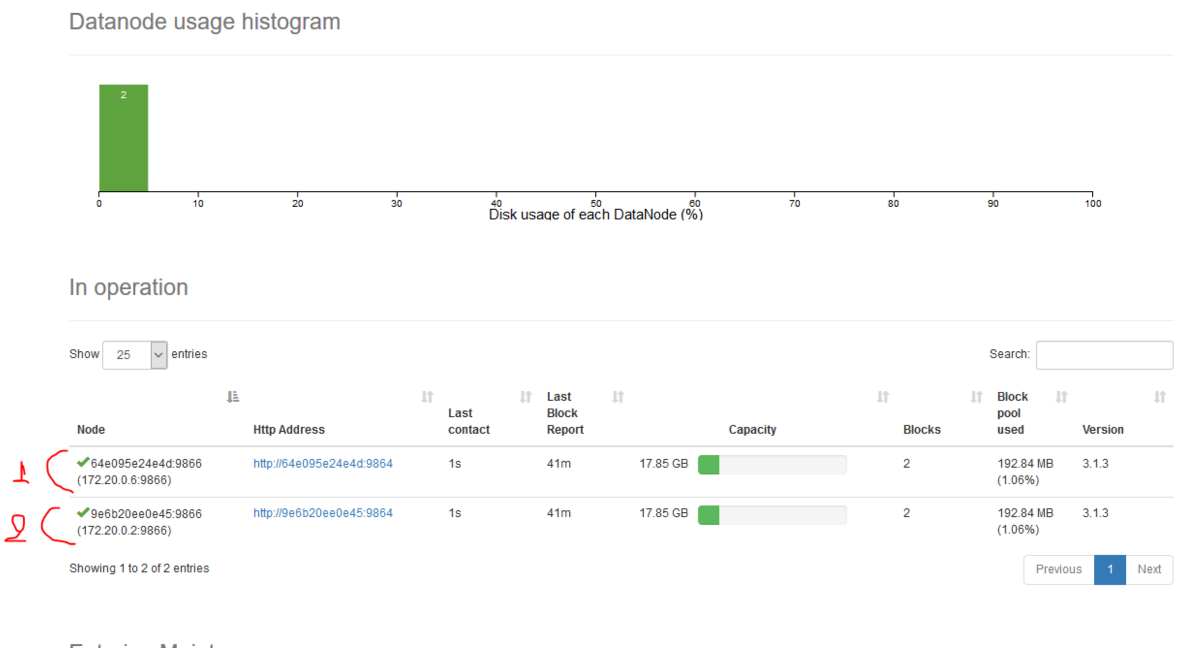
Storage Directory	Type	State
/hadoop/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	35.69 GB	385.64 MB (1.06%)	28.79 GB (80.66%)	385.64 MB	2

Hadoop, 2019.

Figure 13: The two nodes.



In operation

Show 25 entries Search:

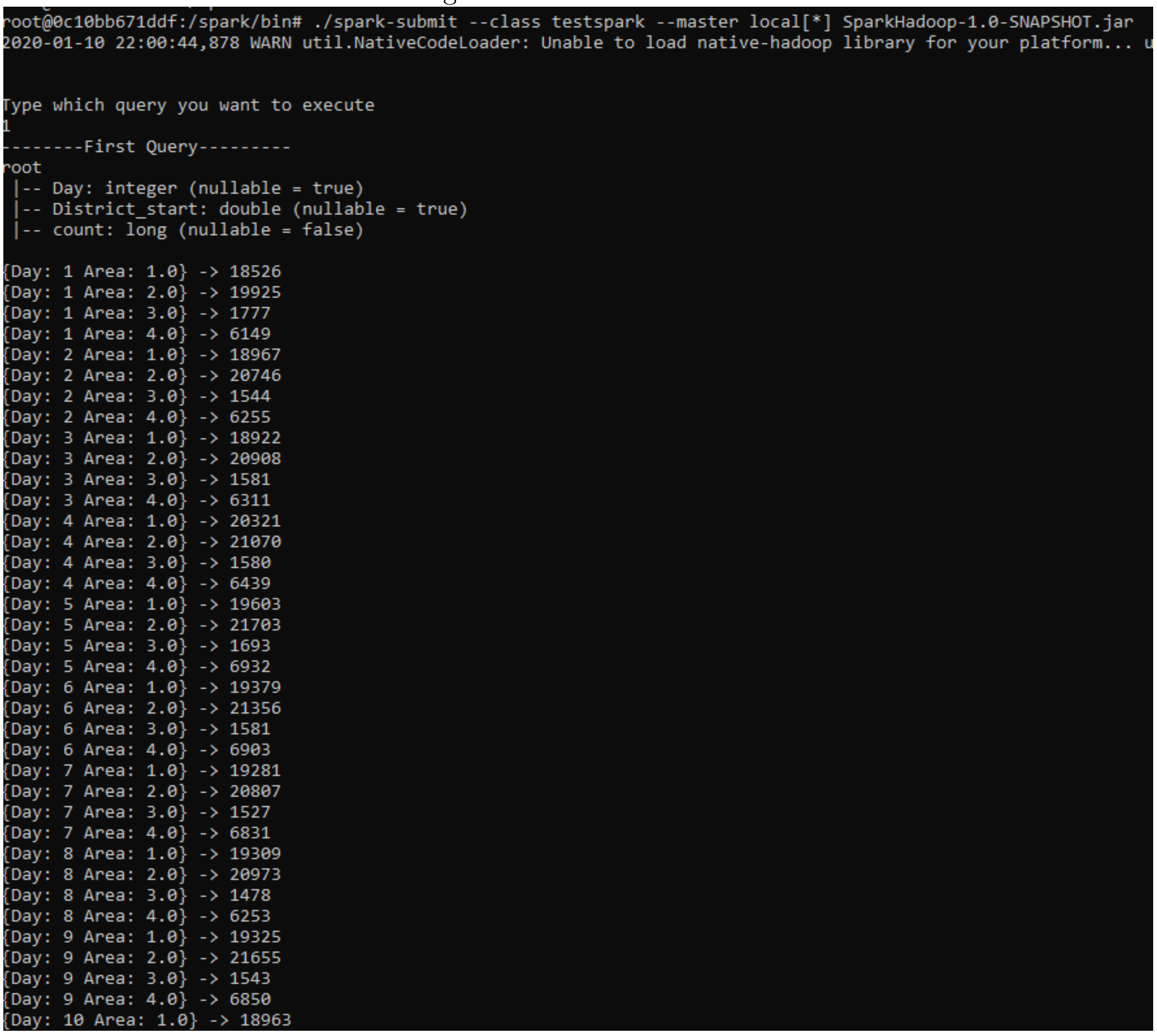
	Node	Http Address	Last contact	Last Block Report	Capacity		Blocks	Block pool used	Version
1	✓64e095e24e4d:9866 (172.20.0.6:9866)	http://64e095e24e4d:9864	1s	41m	17.85 GB	🟩	2	192.84 MB (1.06%)	3.1.3
2	✓9e6b20ee0e45:9866 (172.20.0.2:9866)	http://9e6b20ee0e45:9864	1s	41m	17.85 GB	🟩	2	192.84 MB (1.06%)	3.1.3

Showing 1 to 2 of 2 entries

Previous 1 Next

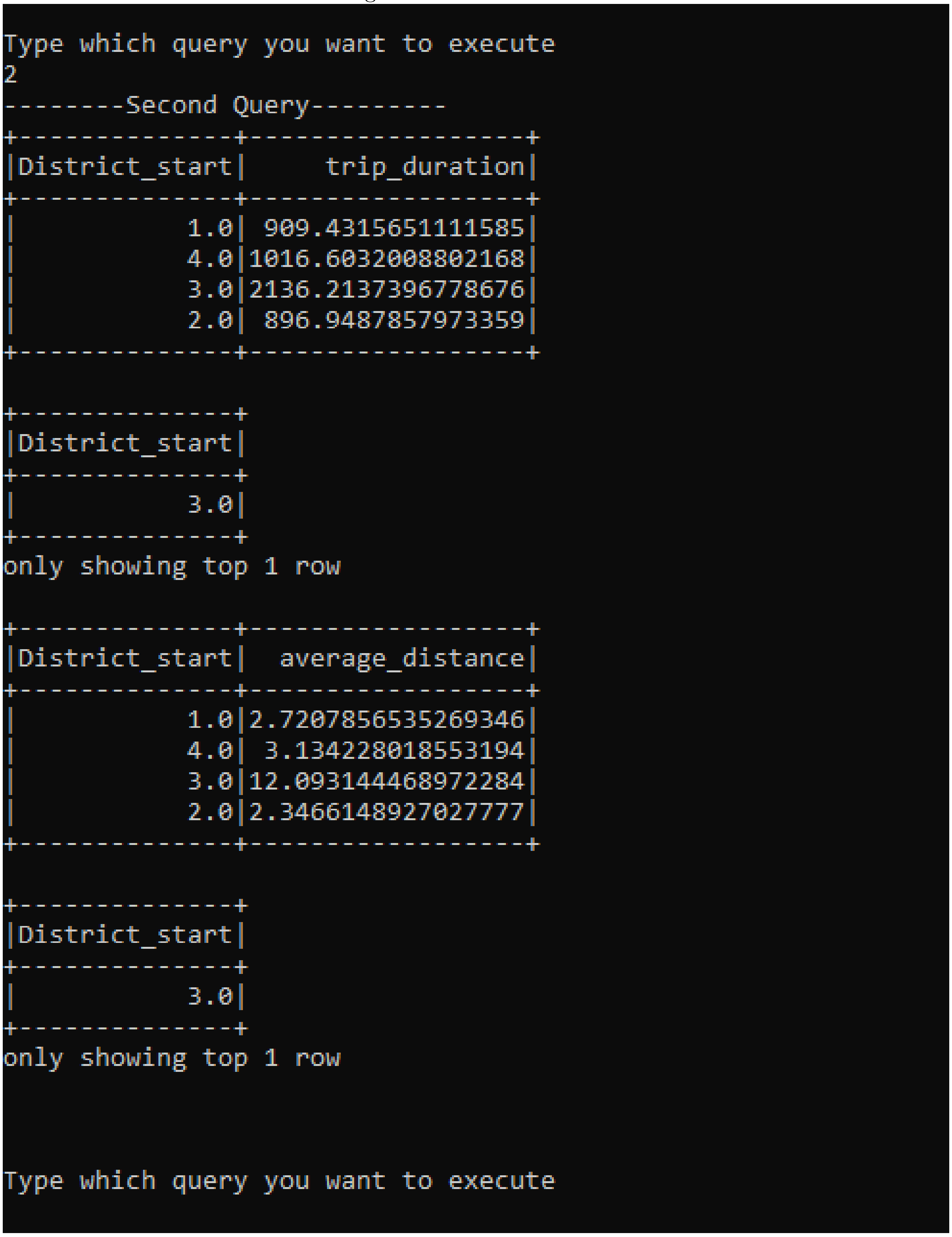
Run the previous queries again!

Figure 14: The two nodes.



```
root@0c10bb671ddf:/spark/bin# ./spark-submit --class testspark --master local[*] SparkHadoop-1.0-SNAPSHOT.jar
2020-01-10 22:00:44,878 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... u


Type which query you want to execute
1
--------First Query---------
root
 |-- Day: integer (nullable = true)
 |-- District_start: double (nullable = true)
 |-- count: long (nullable = false)

{Day: 1 Area: 1.0} -> 18526
{Day: 1 Area: 2.0} -> 19925
{Day: 1 Area: 3.0} -> 1777
{Day: 1 Area: 4.0} -> 6149
{Day: 2 Area: 1.0} -> 18967
{Day: 2 Area: 2.0} -> 20746
{Day: 2 Area: 3.0} -> 1544
{Day: 2 Area: 4.0} -> 6255
{Day: 3 Area: 1.0} -> 18922
{Day: 3 Area: 2.0} -> 20908
{Day: 3 Area: 3.0} -> 1581
{Day: 3 Area: 4.0} -> 6311
{Day: 4 Area: 1.0} -> 20321
{Day: 4 Area: 2.0} -> 21070
{Day: 4 Area: 3.0} -> 1580
{Day: 4 Area: 4.0} -> 6439
{Day: 5 Area: 1.0} -> 19603
{Day: 5 Area: 2.0} -> 21703
{Day: 5 Area: 3.0} -> 1693
{Day: 5 Area: 4.0} -> 6932
{Day: 6 Area: 1.0} -> 19379
{Day: 6 Area: 2.0} -> 21356
{Day: 6 Area: 3.0} -> 1581
{Day: 6 Area: 4.0} -> 6903
{Day: 7 Area: 1.0} -> 19281
{Day: 7 Area: 2.0} -> 20807
{Day: 7 Area: 3.0} -> 1527
{Day: 7 Area: 4.0} -> 6831
{Day: 8 Area: 1.0} -> 19309
{Day: 8 Area: 2.0} -> 20973
{Day: 8 Area: 3.0} -> 1478
{Day: 8 Area: 4.0} -> 6253
{Day: 9 Area: 1.0} -> 19325
{Day: 9 Area: 2.0} -> 21655
{Day: 9 Area: 3.0} -> 1543
{Day: 9 Area: 4.0} -> 6850
{Day: 10 Area: 1.0} -> 18963
```

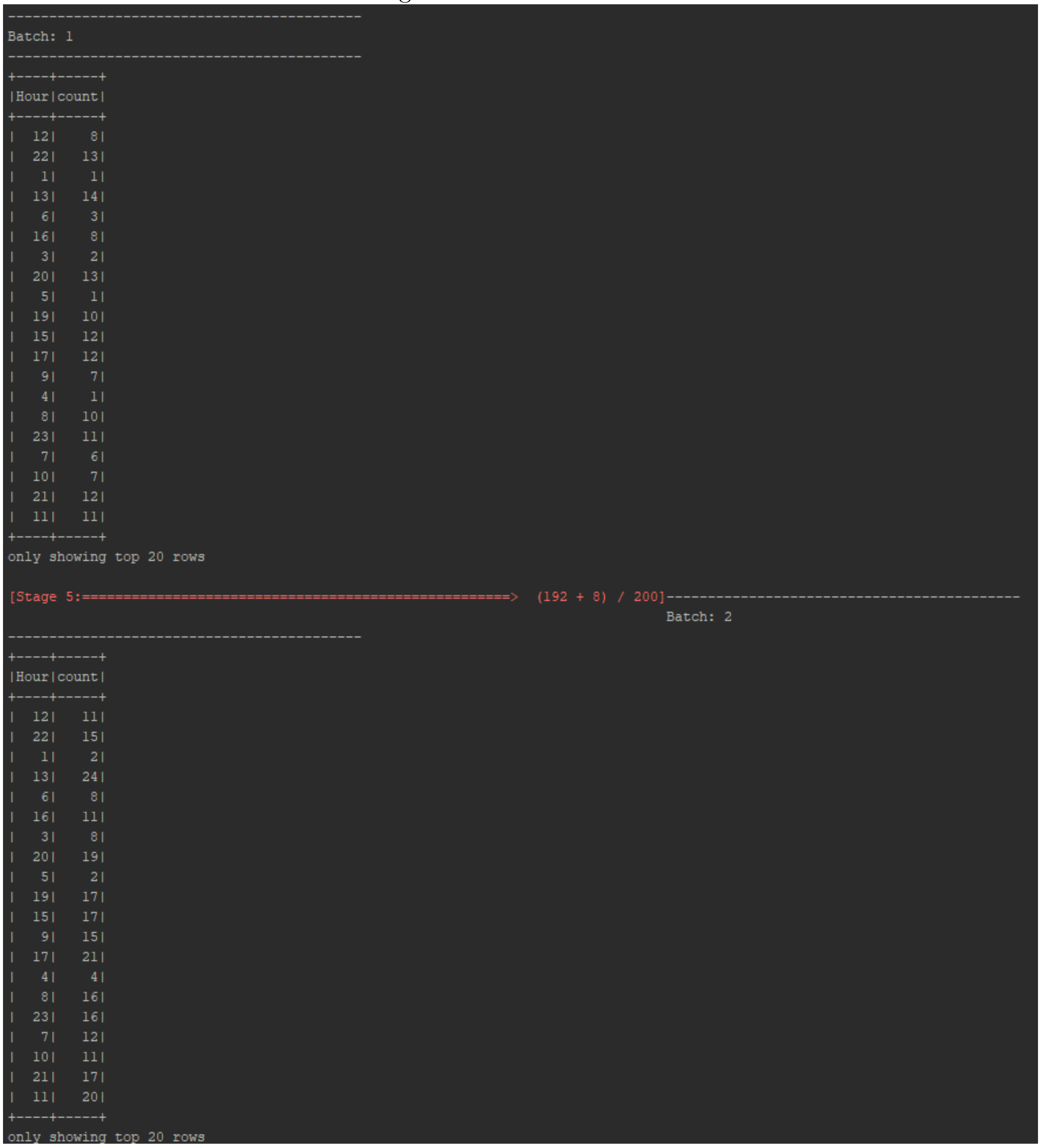
Figure 15: The two nodes.



```
Type which query you want to execute
2
--------Second Query---------
+--------------+------------------+
|District_start|     trip_duration|
+--------------+------------------+
|           1.0| 909.4315651111585|
|           4.0|1016.6032008802168|
|           3.0|2136.2137396778676|
|           2.0| 896.9487857973359|
+--------------+------------------+

+--------------+
|District_start|
+--------------+
|           3.0|
+--------------+
only showing top 1 row

+--------------+------------------+
|District_start|  average_distance|
+--------------+------------------+
|           1.0|2.7207856535269346|
|           4.0| 3.134228018553194|
|           3.0|12.093144468972284|
|           2.0|2.3466148927027777|
+--------------+------------------+

+--------------+
|District_start|
+--------------+
|           3.0|
+--------------+
only showing top 1 row



Type which query you want to execute
```

Third Part - Streaming

Figure 16: The two nodes.



```
-------------------------------------------
Batch: 1
-------------------------------------------
+----+-----+
|Hour|count|
+----+-----+
|  12|    8|
|  22|   13|
|   1|    1|
|  13|   14|
|   6|    3|
|  16|    8|
|   3|    2|
|  20|   13|
|   5|    1|
|  19|   10|
|  15|   12|
|  17|   12|
|   9|    7|
|   4|    1|
|   8|   10|
|  23|   11|
|   7|    6|
|  10|    7|
|  21|   12|
|  11|   11|
+----+-----+
only showing top 20 rows

[Stage 5:=====================================================>  (192 + 8) / 200]-------------------------------------------
                                                                                Batch: 2
-------------------------------------------
+----+-----+
|Hour|count|
+----+-----+
|  12|   11|
|  22|   15|
|   1|    2|
|  13|   24|
|   6|    8|
|  16|   11|
|   3|    8|
|  20|   19|
|   5|    2|
|  19|   17|
|  15|   17|
|   9|   15|
|  17|   21|
|   4|    4|
|   8|   16|
|  23|   16|
|   7|   12|
|  10|   11|
|  21|   17|
|  11|   20|
+----+-----+
only showing top 20 rows
```