Citi Bike data analysis (2019)

Description

Below analysis answers many questions raised by citi bike community like Where do Citi Bikers ride? When do they ride? How far do they go? Which stations are most popular? Busiest bike?

Dataset: https://s3.amazonaws.com/tripdata/index.html (https://s3.amazonaws.com/tripdata/index.html)

Technologies

- Hadoop Map Reduce,
- Pig
- Hive
- Neo4j

Data analysis using Hadoop Mapreduce

In [3]:

```
import pandas as pd
import matplotlib
```

- Top five most popular start stations?

In [4]:

```
top_5_start_station = pd.read_csv("top5StartStation/part-r-00000", index_col=
"station_name", delim_whitespace=True, names =["station_name", "popularity"])
top_5_start_station
```

Out[4]:

	popularity
station_name	
Grove St PATH	20563
Hamilton Park	9537
Sip Ave	8339
Newport PATH	7350
Harborside	6989

In [5]:

```
top_5_start_station_graph = top_5_start_station.plot(kind='bar')
top_5_start_station_graph
```

Out[5]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x18be0d98c88>
```

- Top five most popular end stations?

In [6]:

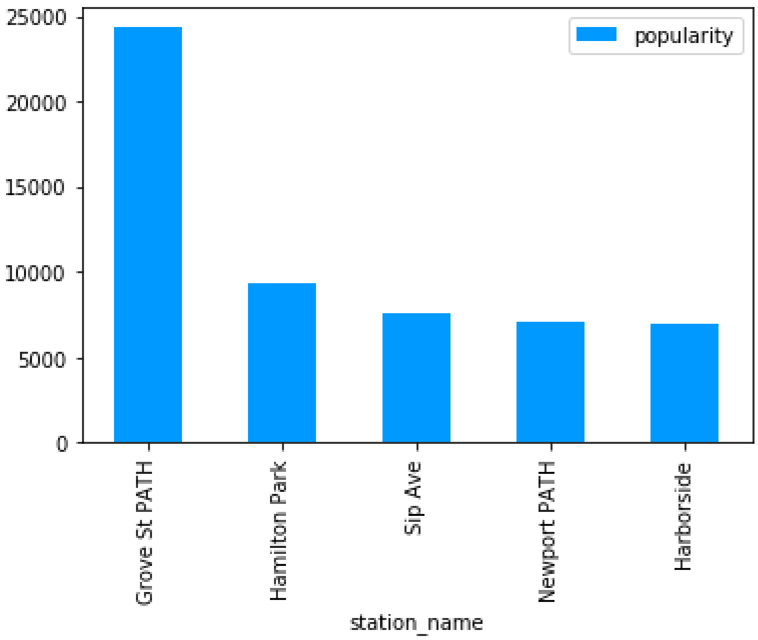
```
top_5_end_station = pd.read_csv("top5EndStation/part-r-00000", index_col="stat
ion_name", delim_whitespace=True, names =["station_name", "popularity"])
top_5_end_station
```

Out[6]:

	popularity
station_name	
Grove St PATH	24336
Hamilton Park	9280
Sip Ave	7571
Newport PATH	7134
Harborside	6947

In [7]:
```
top_5_end_station_graph = top_5_end_station.plot(kind='bar', color='#0099ff')
top_5_end_station_graph
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x18be10da6a0>
```



- Top five most popular journey?

In [8]:
```
top_5_journey = pd.read_csv("top5Journey/part-r-00000", engine='python', index
_col="journey", sep = "\"\s", names =["journey", "popularity"])
top_5_journey
```

Out[8]:

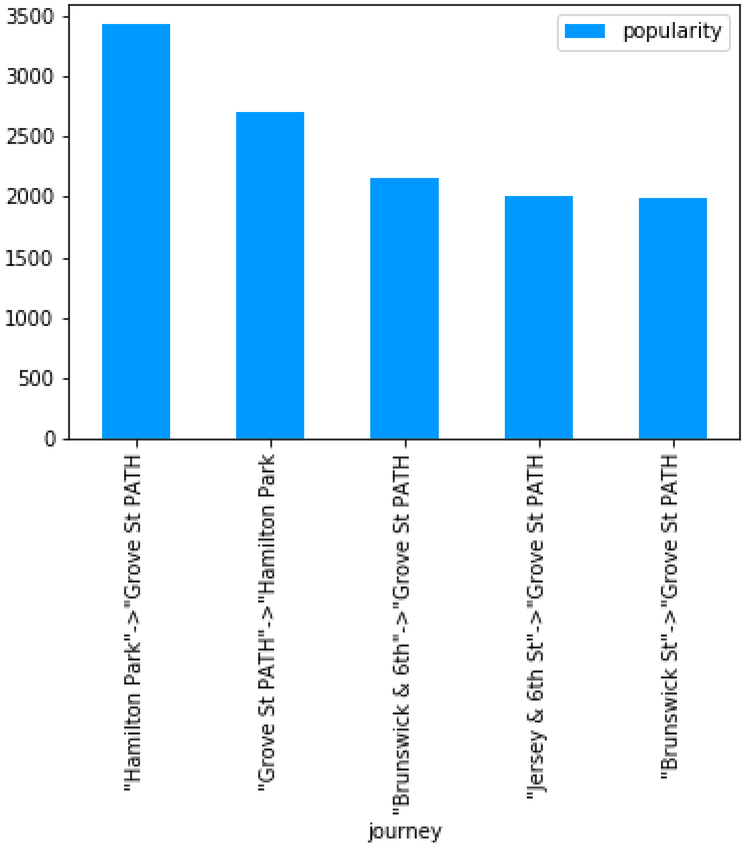
	popularity
journey	
"Hamilton Park"->"Grove St PATH	3432
"Grove St PATH"->"Hamilton Park	2695
"Brunswick & 6th"->"Grove St PATH	2162
"Jersey & 6th St"->"Grove St PATH	2012
"Brunswick St"->"Grove St PATH	1990

In [9]:

```
top_5_journey_graph = top_5_journey.plot(kind='bar', color='#0099ff')
top_5_journey_graph
```

Out[9]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x18be1275080>
```



- Bike ride distribution according to age

In [10]:

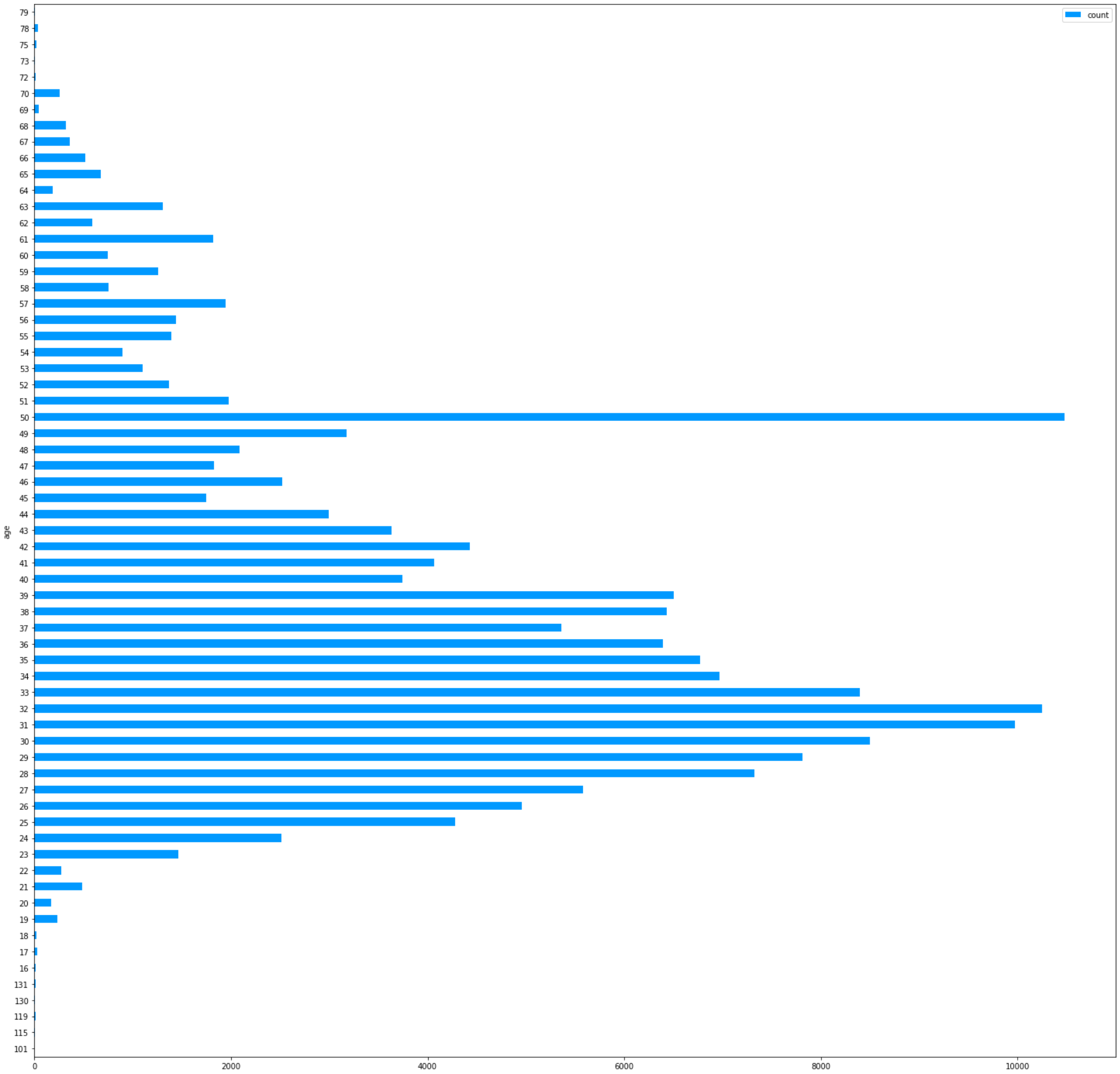
```
ride_dist_on_age = pd.read_csv("RideDistributionOnAge/part-r-00000", index_col
="age", delim_whitespace=True , names =["age", "count"])
ride_dist_on_age.head()
```

Out[10]:

	count
age	
101	1
115	6
119	14
130	4
131	11

In [11]:
```
# ride_dist_on_age.plot.pie(subplots=True, figsize=(30,30))
ride_dist_on_age.plot.barh(figsize=(25,25),color='#0099ff')
```

Out[11]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x18be130dac8>
```



- Bike ride distribution based on day hour

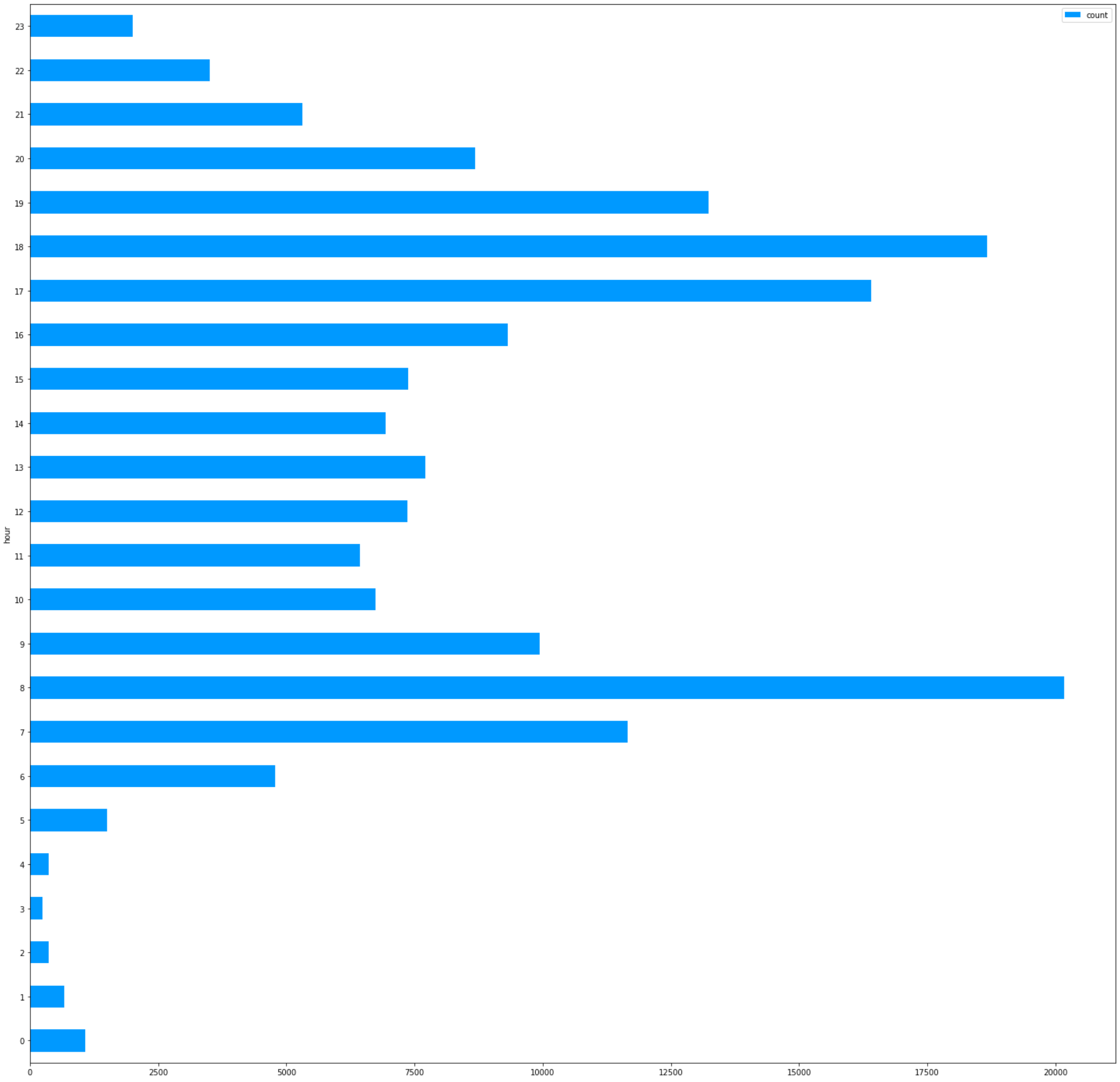
In [12]:
```
ride_dist_on_hour = pd.read_csv("RideDistributionOnDayTime/part-r-00000", inde
x_col="hour", delim_whitespace=True , names =["hour", "count"])
ride_dist_on_hour.head()
```

Out[12]:

	count
hour	
0	1085
1	665
2	365
3	248
4	366

In [13]:
```
ride_dist_on_hour.plot.barh(figsize=(25,25), color='#0099ff')
```

Out[13]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x18be1463128>
```



- Trip durations sorted groaup by startstation (Secondary sort)

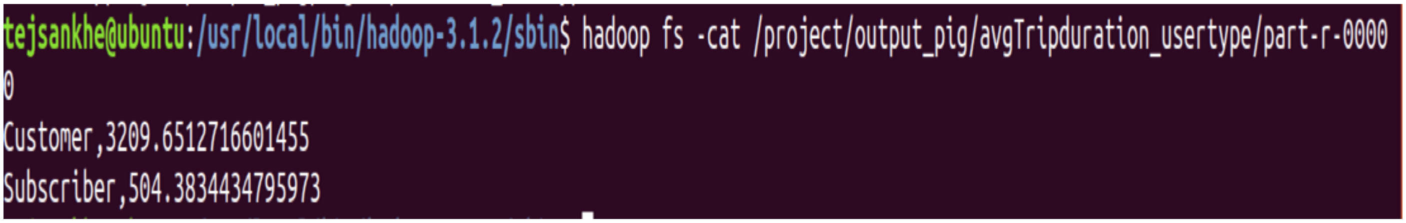
```
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=3359}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=3590}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=4218}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=4446}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=7437}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=7458}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=8786}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=15434}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=15472}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=15492}
CompositeKeyWritablePart4{stationName="'5 Corners Library'", tripduration=15668}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=103}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=109}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=130}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=131}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=136}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=142}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=156}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=158}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=163}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=164}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=165}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=166}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=169}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=170}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=171}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=173}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=175}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=179}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=199}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=199}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=203}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=206}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=207}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=207}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=209}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=222}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=223}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=224}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=228}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=233}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=233}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=241}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=251}
CompositeKeyWritablePart4{stationName="'Astor Place'", tripduration=254}
```

```
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=255}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=258}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=264}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=269}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=271}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=276}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=280}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=280}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=282}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=284}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=286}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=289}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=289}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=290}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=292}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=294}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=296}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=296}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=296}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=298}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=298}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=299}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=300}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=301}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=301}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=301}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=302}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=303}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=303}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=303}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=304}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=304}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=304}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=304}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=304}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=305}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=305}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=305}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=306}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=307}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=308}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=308}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=309}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=310}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=310}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=310}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=310}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=310}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=311}
```

```
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=313}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=313}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=313}
CompositeKeyWritablePart4{stationName='"Astor Place"', tripduration=314}
```

Data analysis using Pig

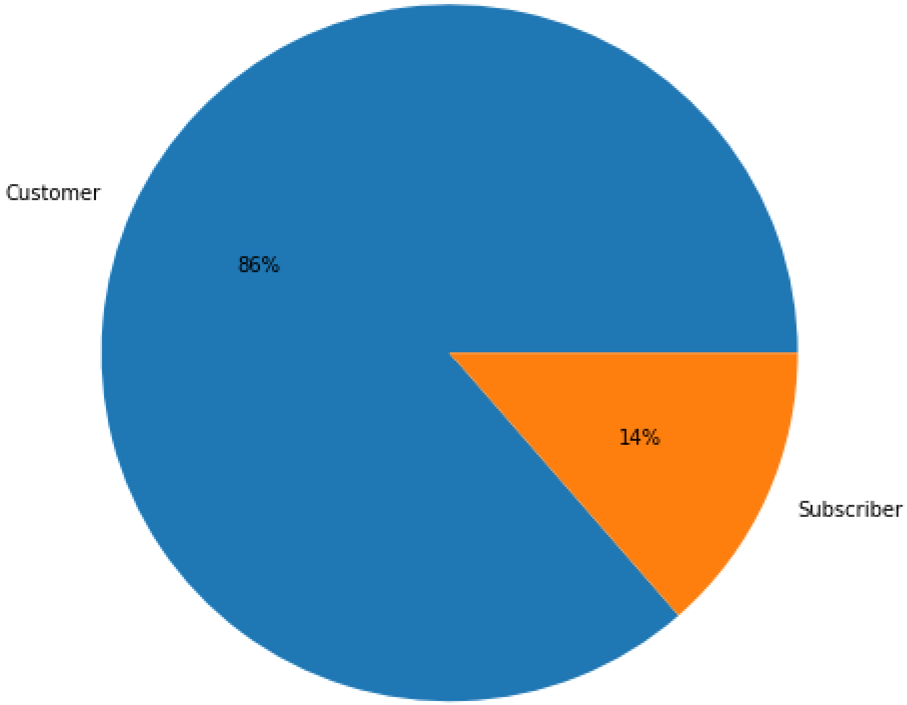
- Average trip duration based on user type



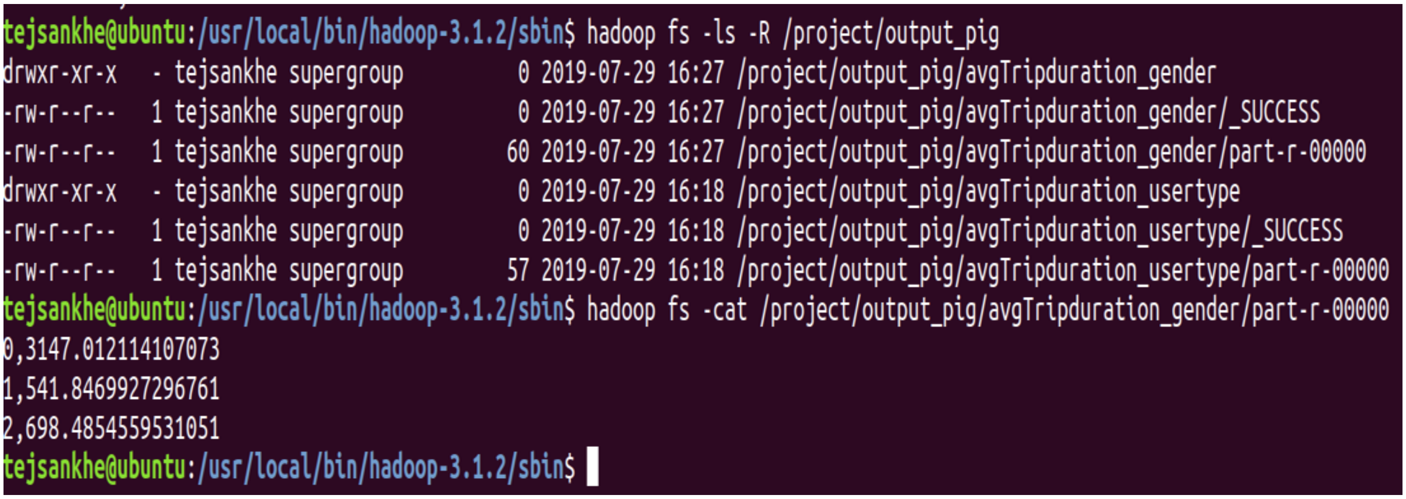
```
tejsankhe@ubuntu:/usr/local/bin/hadoop-3.1.2/sbin$ hadoop fs -cat /project/output_pig/avgTripduration_usertype/part-r-0000
0
Customer,3209.6512716601455
Subscriber,504.3834434795973
```

In [14]:
```
matplotlib.pyplot.pie(x=[3209.65,504.38], autopct='%1.0f%%', labels=["Customer","Subscriber"], radius = 2)
```

Out[14]:
```
([<matplotlib.patches.Wedge at 0x18be21170f0>,
  <matplotlib.patches.Wedge at 0x18be21177f0>],
 [Text(-2.0027942693243954, 0.9103928354075295, 'Customer'),
  Text(2.002794354561516, -0.9103926478923925, 'Subscriber')],
 [Text(-1.0924332378133066, 0.49657791022228875, '86%'),
  Text(1.0924332843062812, -0.49657780794130496, '14%')])
```



- Average trip duration based on gender



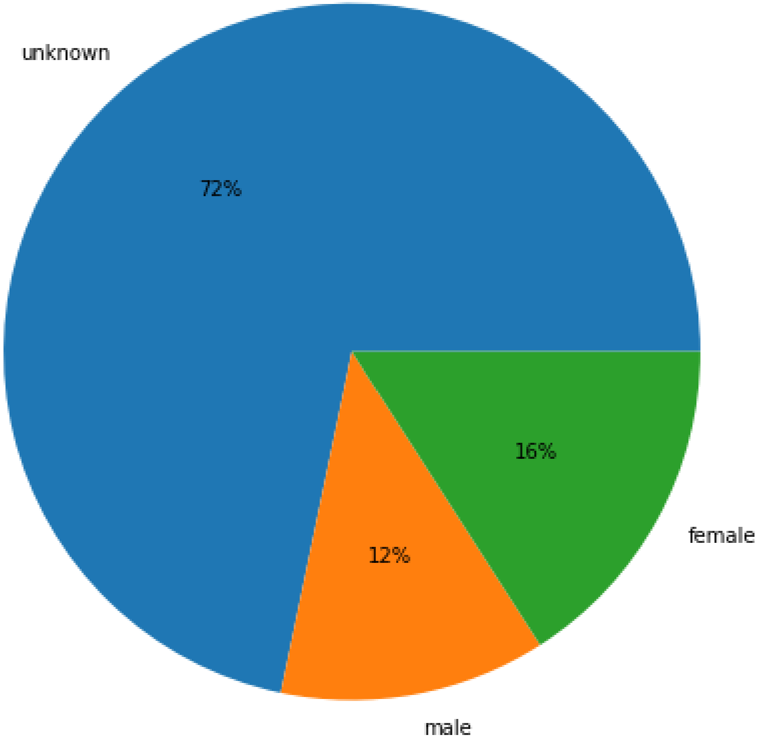
```
tejsankhe@ubuntu:/usr/local/bin/hadoop-3.1.2/sbin$ hadoop fs -ls -R /project/output_pig
drwxr-xr-x   - tejsankhe supergroup          0 2019-07-29 16:27 /project/output_pig/avgTripduration_gender
-rw-r--r--   1 tejsankhe supergroup          0 2019-07-29 16:27 /project/output_pig/avgTripduration_gender/_SUCCESS
-rw-r--r--   1 tejsankhe supergroup         60 2019-07-29 16:27 /project/output_pig/avgTripduration_gender/part-r-00000
drwxr-xr-x   - tejsankhe supergroup          0 2019-07-29 16:18 /project/output_pig/avgTripduration_usertype
-rw-r--r--   1 tejsankhe supergroup          0 2019-07-29 16:18 /project/output_pig/avgTripduration_usertype/_SUCCESS
-rw-r--r--   1 tejsankhe supergroup         57 2019-07-29 16:18 /project/output_pig/avgTripduration_usertype/part-r-00000
tejsankhe@ubuntu:/usr/local/bin/hadoop-3.1.2/sbin$ hadoop fs -cat /project/output_pig/avgTripduration_gender/part-r-00000
0,3147.012114107073
1,541.8469927296761
2,698.4854559531051
tejsankhe@ubuntu:/usr/local/bin/hadoop-3.1.2/sbin$
```

In [15]:

```
matplotlib.pyplot.pie(x=[3147.0121,541.8469,698.4854], autopct='%1.0f%%', labe
ls=["unknown","male","female"], radius=2)
```

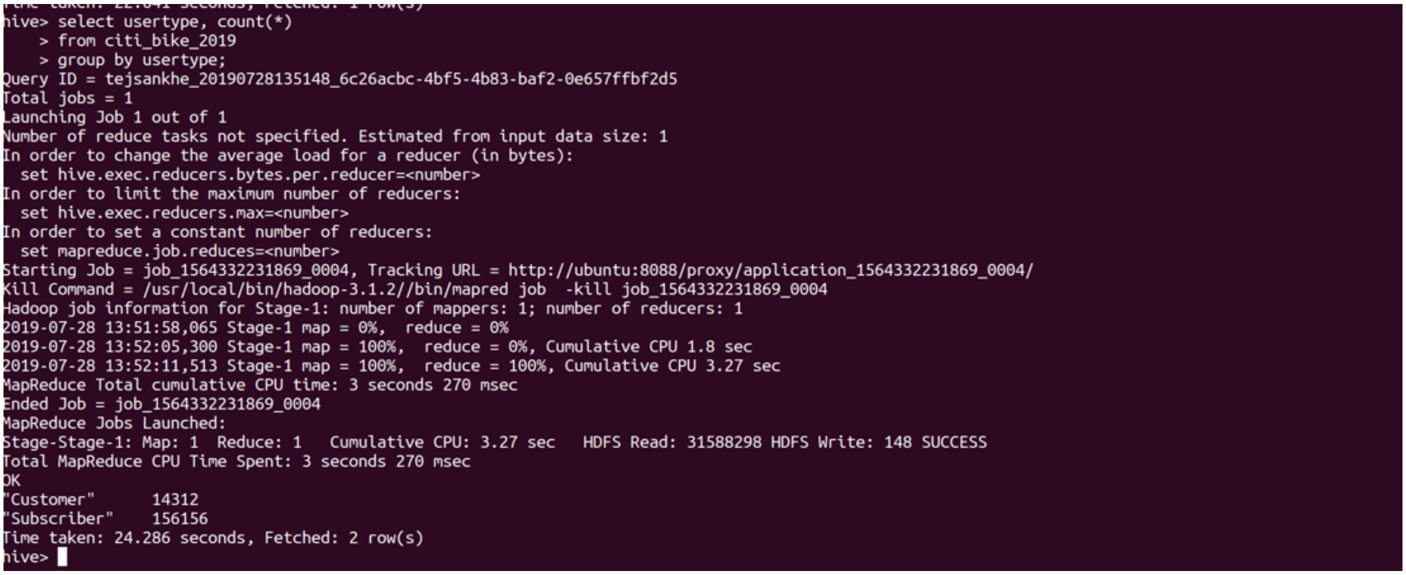
Out[15]:

```
([<matplotlib.patches.Wedge at 0x18be1d1a048>,
  <matplotlib.patches.Wedge at 0x18be1d1a780>,
  <matplotlib.patches.Wedge at 0x18be1d1ae80>],
 [Text(-1.3878671374359546, 1.7069929141110485, 'unknown'),
  Text(0.3992544348500846, -2.1634684874646406, 'male'),
  Text(1.9305169657373258, -1.0550375562037353, 'female')],
 [Text(-0.7570184386014297, 0.9310870440605717, '72%'),
  Text(0.2177751462818643, -1.1800737204352585, '12%'),
  Text(1.0530092540385412, -0.5754750306565828, '16%')])
```



Data analysis using Hive

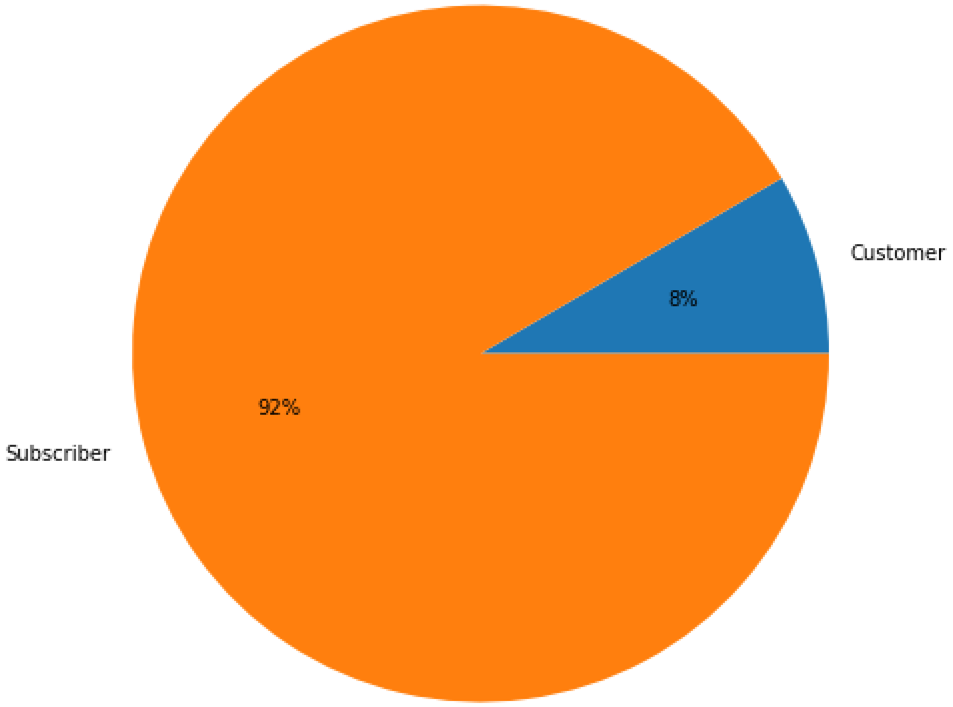
- Ride distribution based on user type



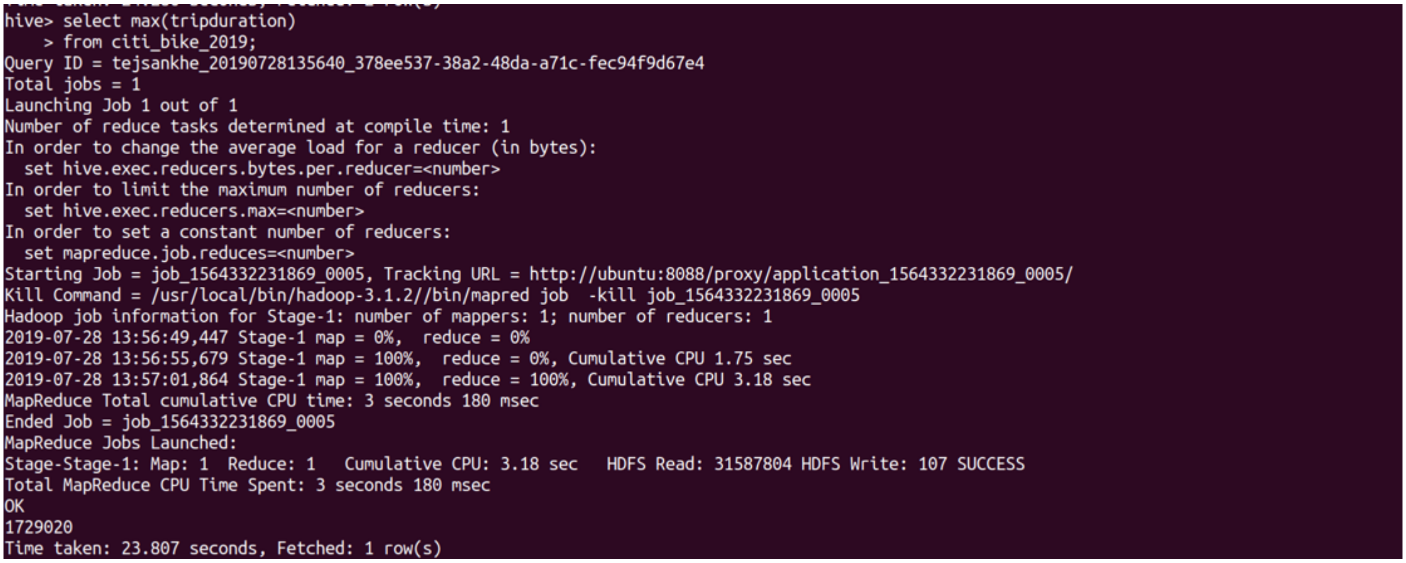
```
Time taken: 22.641 seconds, Fetched: 1 row(s)
hive> select usertype, count(*)
    > from citi_bike_2019
    > group by usertype;
Query ID = tejsankhe_20190728135148_6c26acbc-4bf5-4b83-baf2-0e657ffbf2d5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 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_1564332231869_0004, Tracking URL = http://ubuntu:8088/proxy/application_1564332231869_0004/
Kill Command = /usr/local/bin/hadoop-3.1.2//bin/mapred job  -kill job_1564332231869_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-07-28 13:51:58,065 Stage-1 map = 0%,  reduce = 0%
2019-07-28 13:52:05,300 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.8 sec
2019-07-28 13:52:11,513 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.27 sec
MapReduce Total cumulative CPU time: 3 seconds 270 msec
Ended Job = job_1564332231869_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.27 sec   HDFS Read: 31588298 HDFS Write: 148 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 270 msec
OK
"Customer"      14312
"Subscriber"    156156
Time taken: 24.286 seconds, Fetched: 2 row(s)
hive> █
```

```
In [16]: matplotlib.pyplot.pie(x=[14312, 156156], autopct='%1.0f%%', labels=["Customer"
         ,"Subscriber"], radius = 2)

Out[16]: ([<matplotlib.patches.Wedge at 0x18be1d565f8>,
           <matplotlib.patches.Wedge at 0x18be1d56cf8>],
          [Text(2.1239169159812756, 0.573565108778932, 'Customer'),
           Text(-2.1239169696823073, -0.5735649099234764, 'Subscriber')],
          [Text(1.1585001359897864, 0.3128536956975992, '8%'),
           Text(-1.1585001652812583, -0.3128535872309871, '92%')])
```



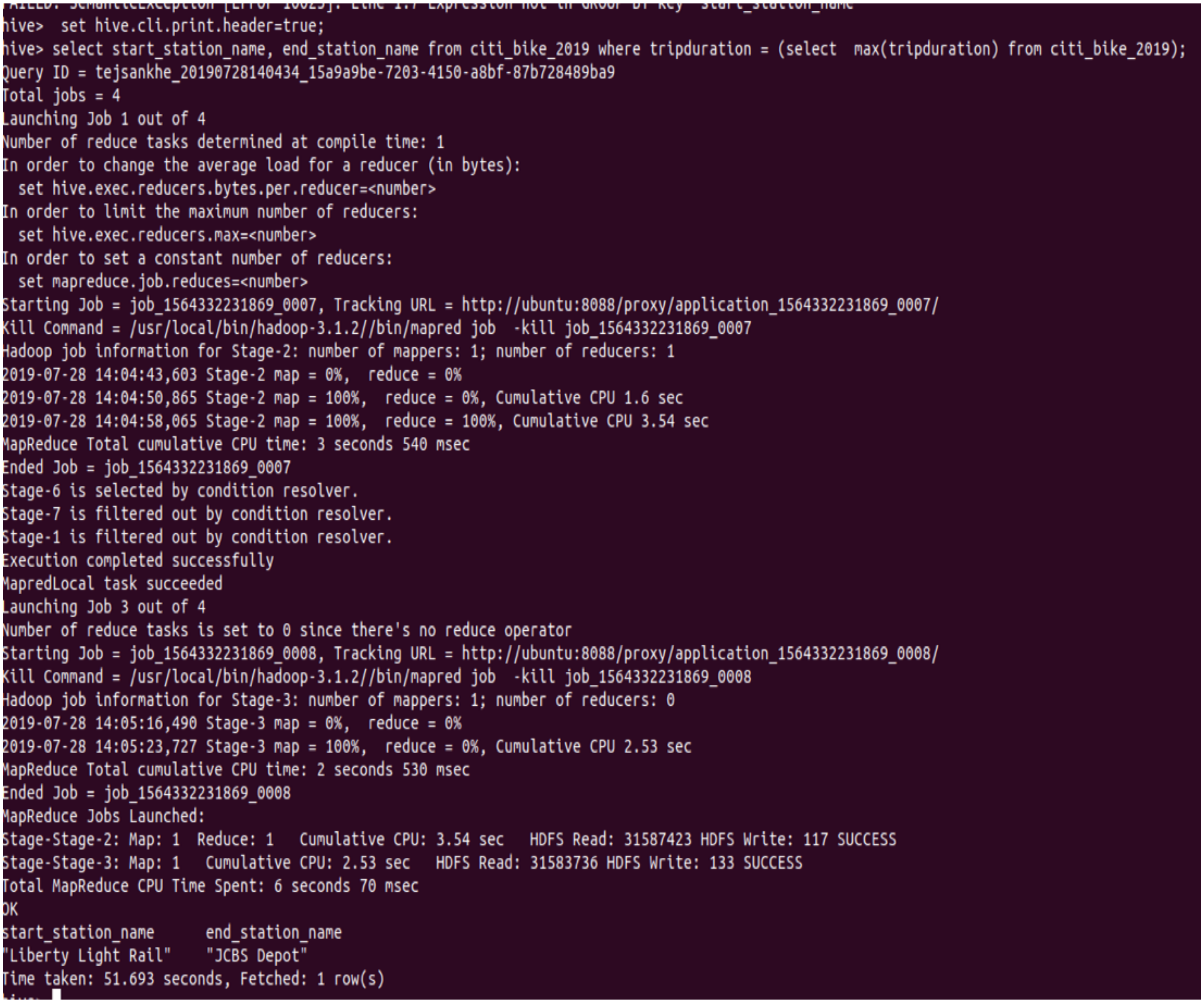
- Longest trip duration



```
hive> select max(tripduration)
    > from citi_bike_2019;
Query ID = tejsankhe_20190728135640_378ee537-38a2-48da-a71c-fec94f9d67e4
Total jobs = 1
Launching Job 1 out of 1
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_1564332231869_0005, Tracking URL = http://ubuntu:8088/proxy/application_1564332231869_0005/
Kill Command = /usr/local/bin/hadoop-3.1.2//bin/mapred job  -kill job_1564332231869_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-07-28 13:56:49,447 Stage-1 map = 0%,  reduce = 0%
2019-07-28 13:56:55,679 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.75 sec
2019-07-28 13:57:01,864 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.18 sec
MapReduce Total cumulative CPU time: 3 seconds 180 msec
Ended Job = job_1564332231869_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.18 sec   HDFS Read: 31587804 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 180 msec
OK
1729020
Time taken: 23.807 seconds, Fetched: 1 row(s)
```

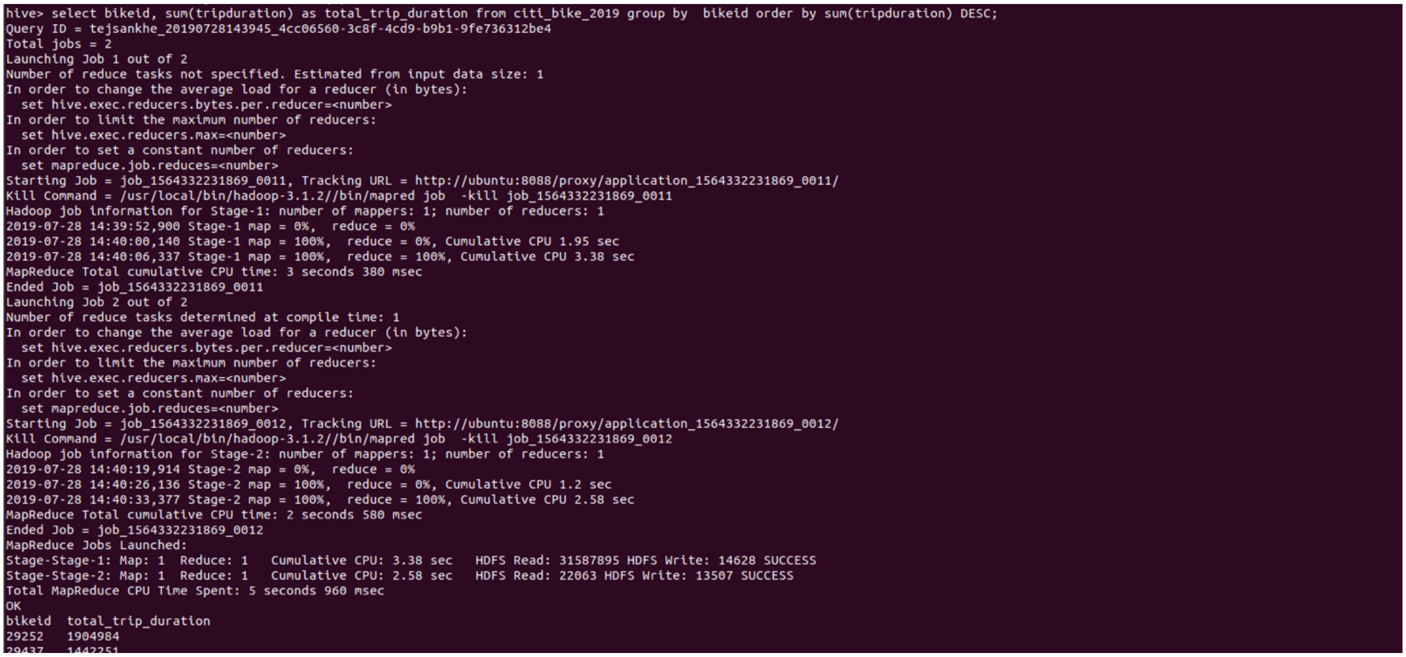
Longest trip duration is approx 20 hrs

- Longest trip duration to and from location (subquery)

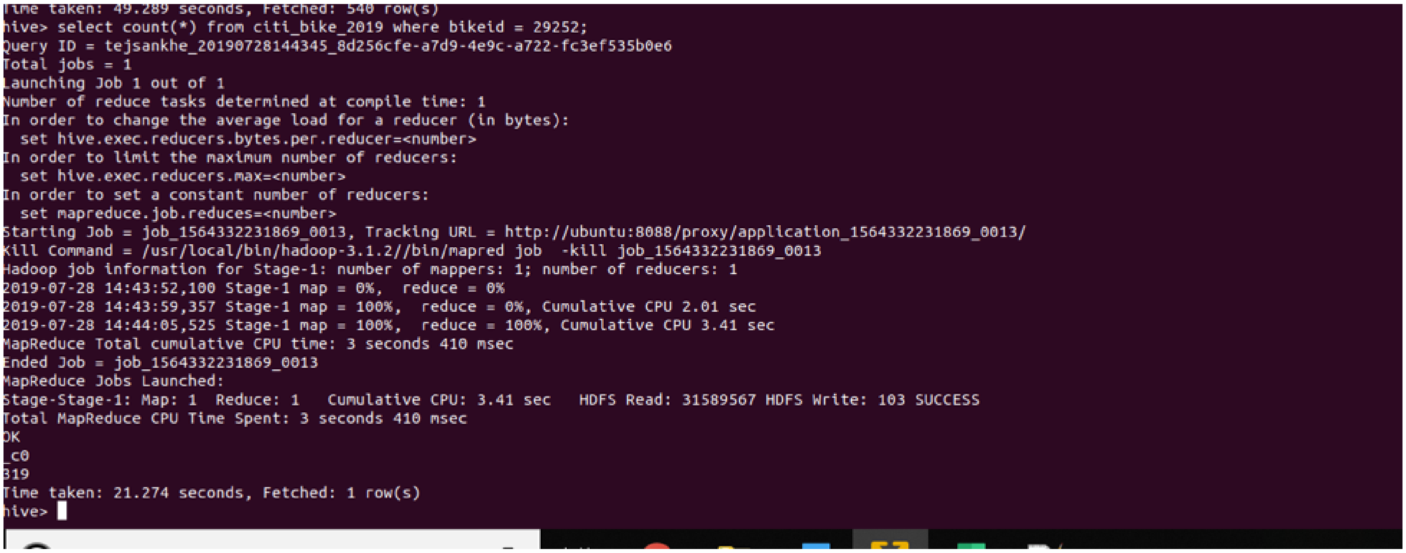


```
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'start_station_name'
hive>  set hive.cli.print.header=true;
hive> select start_station_name, end_station_name from citi_bike_2019 where tripduration = (select  max(tripduration) from citi_bike_2019);
Query ID = tejsankhe_20190728140434_15a9a9be-7203-4150-a8bf-87b728489ba9
Total jobs = 4
Launching Job 1 out of 4
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_1564332231869_0007, Tracking URL = http://ubuntu:8088/proxy/application_1564332231869_0007/
Kill Command = /usr/local/bin/hadoop-3.1.2//bin/mapred job  -kill job_1564332231869_0007
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-07-28 14:04:43,603 Stage-2 map = 0%,  reduce = 0%
2019-07-28 14:04:50,865 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.6 sec
2019-07-28 14:04:58,065 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 3.54 sec
MapReduce Total cumulative CPU time: 3 seconds 540 msec
Ended Job = job_1564332231869_0007
Stage-6 is selected by condition resolver.
Stage-7 is filtered out by condition resolver.
Stage-1 is filtered out by condition resolver.
Execution completed successfully
MapredLocal task succeeded
Launching Job 3 out of 4
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1564332231869_0008, Tracking URL = http://ubuntu:8088/proxy/application_1564332231869_0008/
Kill Command = /usr/local/bin/hadoop-3.1.2//bin/mapred job  -kill job_1564332231869_0008
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2019-07-28 14:05:16,490 Stage-3 map = 0%,  reduce = 0%
2019-07-28 14:05:23,727 Stage-3 map = 100%,  reduce = 0%, Cumulative CPU 2.53 sec
MapReduce Total cumulative CPU time: 2 seconds 530 msec
Ended Job = job_1564332231869_0008
MapReduce Jobs Launched:
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 3.54 sec   HDFS Read: 31587423 HDFS Write: 117 SUCCESS
Stage-Stage-3: Map: 1   Cumulative CPU: 2.53 sec   HDFS Read: 31583736 HDFS Write: 133 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 70 msec
OK
start_station_name      end_station_name
"Liberty Light Rail"    "JCBS Depot"
Time taken: 51.693 seconds, Fetched: 1 row(s)
hive>
```

- What is the busiest bike in NYC in 2019? How many times was it used?



```
hive> select bikeid, sum(tripduration) as total_trip_duration from citi_bike_2019 group by  bikeid order by sum(tripduration) DESC;
Query ID = tejsankhe_20190728143945_4cc06560-3c8f-4cd9-b9b1-9fe736312be4
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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_1564332231869_0011, Tracking URL = http://ubuntu:8088/proxy/application_1564332231869_0011/
Kill Command = /usr/local/bin/hadoop-3.1.2//bin/mapred job  -kill job_1564332231869_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-07-28 14:39:52,900 Stage-1 map = 0%,  reduce = 0%
2019-07-28 14:40:00,140 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.95 sec
2019-07-28 14:40:06,337 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.38 sec
MapReduce Total cumulative CPU time: 3 seconds 380 msec
Ended Job = job_1564332231869_0011
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_1564332231869_0012, Tracking URL = http://ubuntu:8088/proxy/application_1564332231869_0012/
Kill Command = /usr/local/bin/hadoop-3.1.2//bin/mapred job  -kill job_1564332231869_0012
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2019-07-28 14:40:19,914 Stage-2 map = 0%,  reduce = 0%
2019-07-28 14:40:26,136 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.2 sec
2019-07-28 14:40:33,377 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.58 sec
MapReduce Total cumulative CPU time: 2 seconds 580 msec
Ended Job = job_1564332231869_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.38 sec   HDFS Read: 31587895 HDFS Write: 14628 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 2.58 sec   HDFS Read: 22063 HDFS Write: 13507 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 960 msec
OK
bikeid  total_trip_duration
29252   1904984
29437   1442251
```

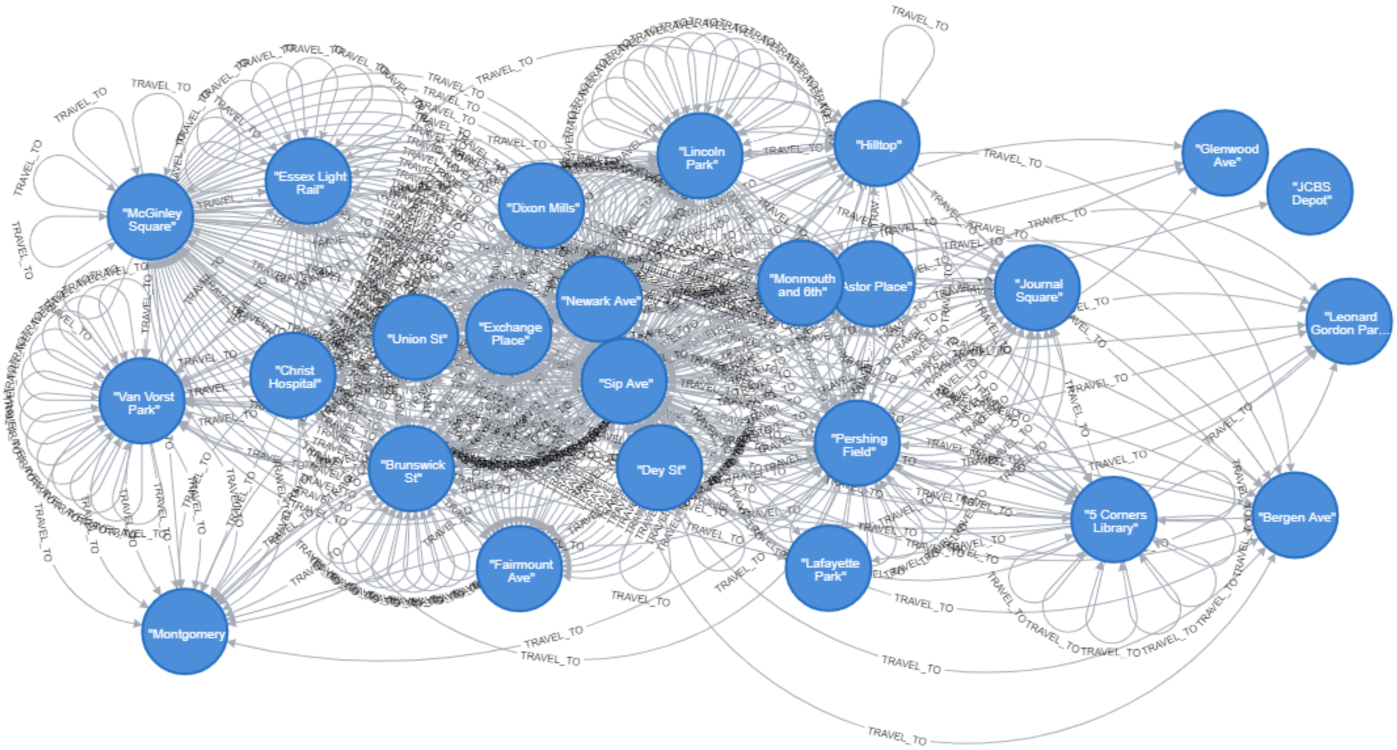


```
Time taken: 49.289 seconds, Fetched: 540 row(s)
hive> select count(*) from citi_bike_2019 where bikeid = 29252;
Query ID = tejsankhe_20190728144345_8d256cfe-a7d9-4e9c-a722-fc3ef535b0e6
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change 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_1564332231869_0013, Tracking URL = http://ubuntu:8088/proxy/application_1564332231869_0013/
Kill Command = /usr/local/bin/hadoop-3.1.2//bin/mapred job  -kill job_1564332231869_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-07-28 14:43:52,100 Stage-1 map = 0%,  reduce = 0%
2019-07-28 14:43:59,357 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.01 sec
2019-07-28 14:44:05,525 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.41 sec
MapReduce Total cumulative CPU time: 3 seconds 410 msec
Ended Job = job_1564332231869_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.41 sec   HDFS Read: 31589567 HDFS Write: 103 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 410 msec
OK
_c0
319
Time taken: 21.274 seconds, Fetched: 1 row(s)
hive> █
```

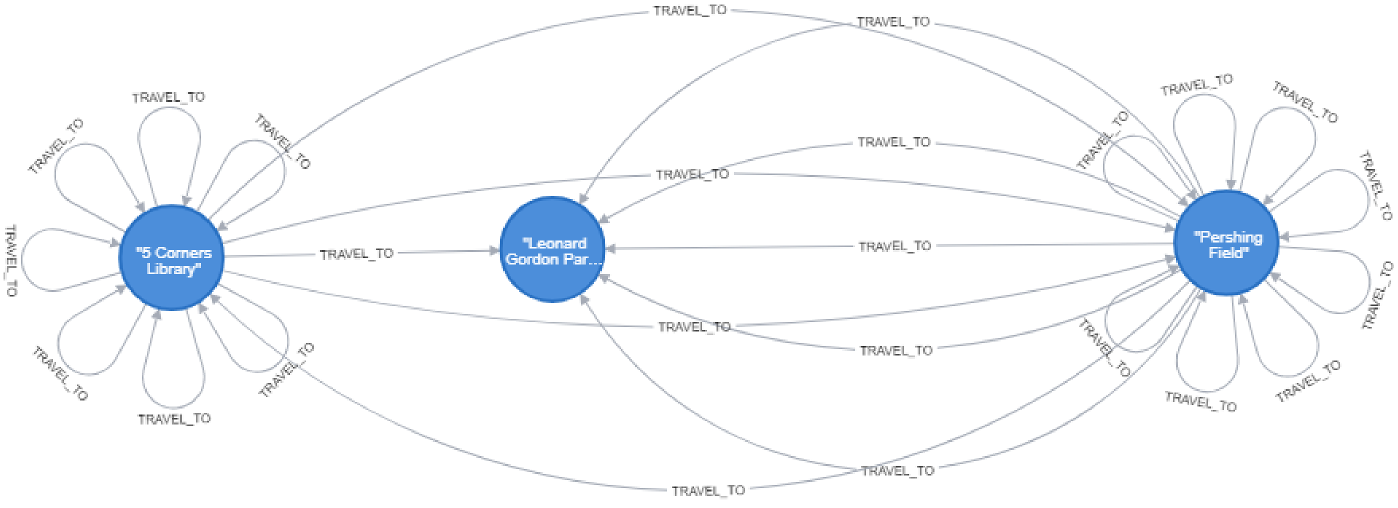
Busiest bike id is 29252 and was rode for 319 times

Data analysis using Neo4j

- Full graph



- Limit 5



References

- https://www.citibikenyc.com/system-data (https://www.citibikenyc.com/system-data)
- https://towardsdatascience.com/citi-bike-2017-analysis-efd298e6c22c (https://towardsdatascience.com/citi-bike-2017-analysis-efd298e6c22c)
- https://github.com/apurvapatkeshwar/NYCTaxiBigData (https://github.com/apurvapatkeshwar/NYCTaxiBigData)