In [1]:
```
import findspark
findspark.init()
findspark.find()
```

Out[1]:
```
'C:\\Users\\Admin\\anaconda3\\envs\\SparkEnvironment\\Lib\\site-packages\\pyspark'
```

In [2]:
```
from pyspark.sql import SparkSession
from pyspark.sql.types import *

spark = (SparkSession
        .builder
        .appName("SparkConceptApp")
        .master("local[4]")
        .getOrCreate()
        )
sc = spark.sparkContext
spark
```

Out[2]:

SparkSession - in-memory

SparkContext

Spark UI (http://192.168.85.95:4040)

Version

v3.5.0

Master

local[4]

AppName

SparkConceptApp

In [3]:
```
tzxiZoneRdd = sc.textFile("C:\DataFiles\TaxiZones.csv",4)
tzxiZoneRdd.collect()
```

```
 '15,Queens,Bay Terrace/Fort Totten,Boro Zone',
 '16,Queens,Bayside,Boro Zone',
 '17,Brooklyn,Bedford,Boro Zone',
 '18,Bronx,Bedford Park,Boro Zone',
 '19,Queens,Bellerose,Boro Zone',
 '20,Bronx,Belmont,Boro Zone',
 '21,Brooklyn,Bensonhurst East,Boro Zone',
 '22,Brooklyn,Bensonhurst West,Boro Zone',
 '23,Staten Island,Bloomfield/Emerson Hill,Boro Zone',
 '24,Manhattan,Bloomingdale,Yellow Zone',
 '25,Brooklyn,Boerum Hill,Boro Zone',
 '26,Brooklyn,Borough Park,Boro Zone',
 '27,Queens,Breezy Point/Fort Tilden/Riis Beach,Boro Zone',
 '28,Queens,Briarwood/Jamaica Hills,Boro Zone',
 '29,Brooklyn,Brighton Beach,Boro Zone',
 '30,Queens,Broad Channel,Boro Zone',
 '31,Bronx,Bronx Park,Boro Zone',
 '32,Bronx,Bronxdale,Boro Zone',
 '33,Brooklyn,Brooklyn Heights,Boro Zone',
 '34,Brooklyn,Brooklyn Navy Yard,Boro Zone',
```

In [4]:
```
tzxiZoneWithColsRdd = tzxiZoneRdd.map(lambda zone:zone.split(","))
tzxiZoneWithColsRdd.collect()
```

Out[4]:
```
[['1', 'EWR', 'Newark Airport', 'EWR'],
 ['2', 'Queens', 'Jamaica Bay', 'Boro Zone'],
 ['3', 'Bronx', 'Allerton/Pelham Gardens', 'Boro Zone'],
 ['4', 'Manhattan', 'Alphabet City', 'Yellow Zone'],
 ['5', 'Staten Island', 'Arden Heights', 'Boro Zone'],
 ['6', 'Staten Island', 'Arrochar/Fort Wadsworth', 'Boro Zone'],
 ['7', 'Queens', 'Astoria', 'Boro Zone'],
 ['8', 'Queens', 'Astoria Park', 'Boro Zone'],
 ['9', 'Queens', 'Auburndale', 'Boro Zone'],
 ['10', 'Queens', 'Baisley Park', 'Boro Zone'],
 ['11', 'Brooklyn', 'Bath Beach', 'Boro Zone'],
 ['12', 'Manhattan', 'Battery Park', 'Yellow Zone'],
 ['13', 'Manhattan', 'Battery Park City', 'Yellow Zone'],
 ['14', 'Brooklyn', 'Bay Ridge', 'Boro Zone'],
 ['15', 'Queens', 'Bay Terrace/Fort Totten', 'Boro Zone'],
 ['16', 'Queens', 'Bayside', 'Boro Zone'],
 ['17', 'Brooklyn', 'Bedford', 'Boro Zone'],
 ['18', 'Bronx', 'Bedford Park', 'Boro Zone'],
 ['19', 'Queens', 'Bellerose', 'Boro Zone'],
```

In [5]:
```
print("after reading file = "+str(tzxiZoneRdd.getNumPartitions()))
print("after applying map  = "+str(tzxiZoneWithColsRdd.getNumPartitions()))
```

```
after reading file = 4
after applying map  = 4
```

In [9]:
```
taxiZonePairRdd = tzxiZoneWithColsRdd.map(lambda zoneRow:(zoneRow[1],1))
distinctZoneRdd = taxiZonePairRdd.distinct()
distinctZoneRdd.collect()
```

```
  Cell In[9], line 3
    //distinctZoneRdd.collect()
    ^
SyntaxError: invalid syntax
```

In [10]:
```
boroughCountRdd = taxiZonePairRdd.reduceByKey(lambda value1,value2:value1+valu
filteredZoneRdd = boroughCountRdd.filter(lambda row:row[1]>10)
filteredZoneRdd.collect()
```

Out[10]:
```
[('Queens', 69),
 ('Manhattan', 69),
 ('Staten Island', 20),
 ('Brooklyn', 61),
 ('Bronx', 43)]
```

In [ ]:
```

```