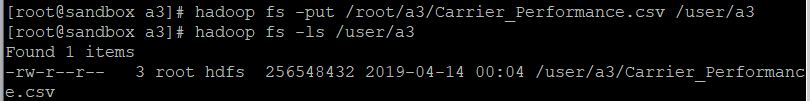
Assignment 3 CIND 719 Chantal Sylvestre

Loading the data to HDFS under directory ‘/user/a3’



-------------------------

-- PySpark Scripts

-------------------------

**Question 1:**

**Loading and exploring the dataset:**

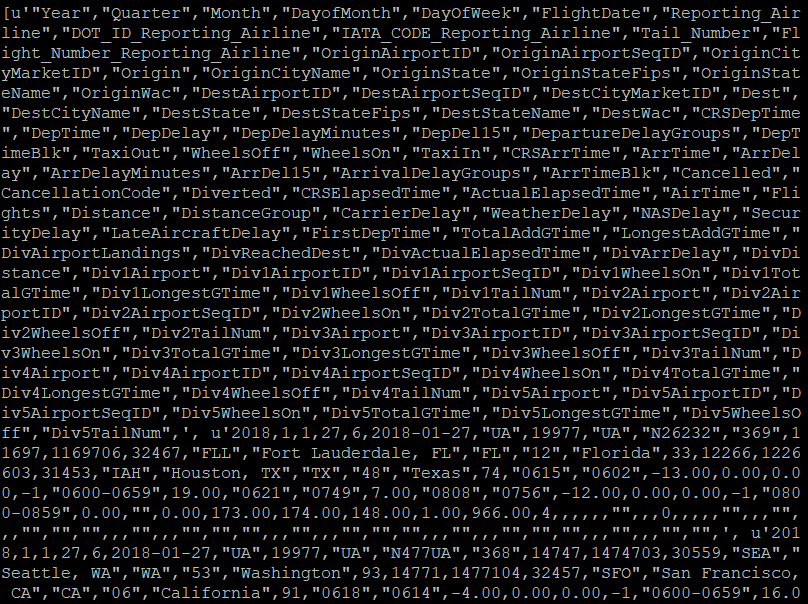
carrier\_raw=sc.textFile("/user/a3/Carrier\_Performance.csv")



carrier\_raw.count()



carrier\_raw.take(3)



**Cleaning up the dataset, splitting by comma and removing double quotes and the header line:**

carrier\_clean=carrier\_raw.map(lambda x :x.split(','))

carrier\_clean.take(3)



def clean(x):

return([xx.replace('"','') for xx in x])

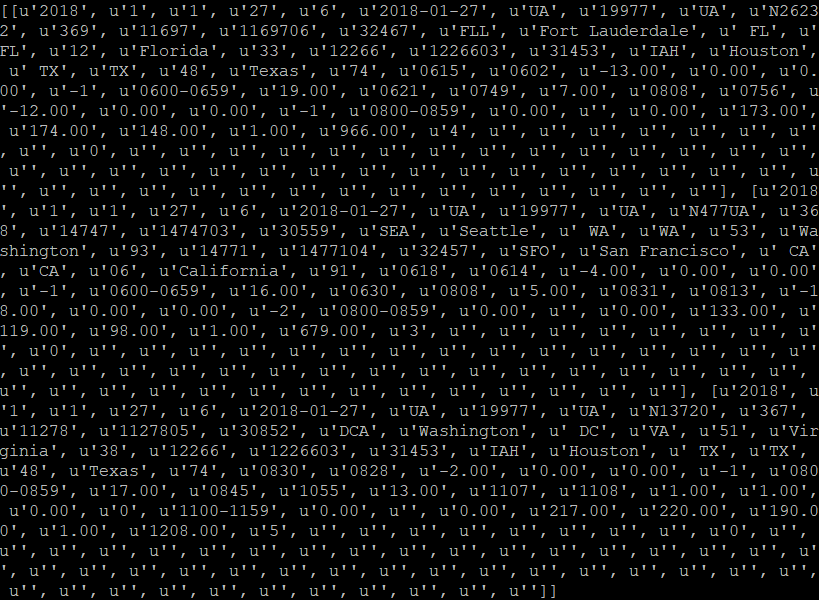
carrier\_clean=carrier\_clean.map(clean)

carrier\_cleanh=carrier\_clean.filter(lambda x : x[0]!='Year')

carrier\_cleanh.count()



carrier\_cleanh.take(3)



**Creating the taxi in RDD, removing 0s and creating float for time field:**

carrier\_taxi\_in=carrier\_cleanh.map(lambda x :(x[48], x[41]))

carrier\_taxi\_in.take(3)



carrier\_taxi\_inc=carrier\_taxi\_in.filter(lambda x : x[1]!='')

carrier\_taxi\_inc.count()



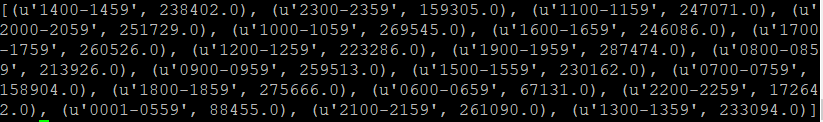
carrier\_taxi\_inc=carrier\_taxi\_inc.map(lambda fields :(fields[0], float(fields[1] )))

carrier\_taxi\_inc.take(3)



carrier\_taxi\_inc=carrier\_taxi\_inc.reduceByKey(lambda a, b:b+a)

carrier\_taxi\_inc.collect()



**Creating the taxi out RDD, removing 0s and creating float for time field:**

carrier\_taxi\_out=carrier\_cleanh.map(lambda x :(x[37], x[38]))

carrier\_taxi\_out.take(3)



carrier\_taxi\_outc=carrier\_taxi\_out.filter(lambda x : x[1]!='')

carrier\_taxi\_outc.count()



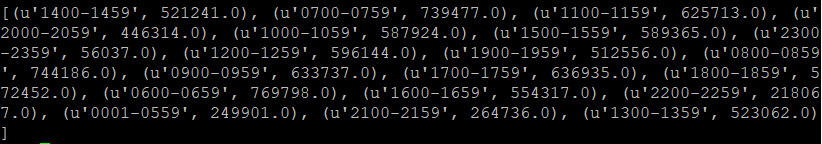
carrier\_taxi\_outc=carrier\_taxi\_outc.map(lambda fields :(fields[0], float(fields[1])))

carrier\_taxi\_outc.take(3)



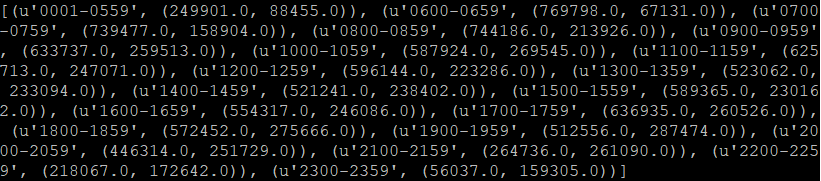
carrier\_taxi\_outc=carrier\_taxi\_outc.reduceByKey(lambda a, b:b+a)

carrier\_taxi\_outc.collect()



**Joining the two RDDs:**

carrier\_taxi\_outc.join(carrier\_taxi\_inc).sortByKey(1).collect()

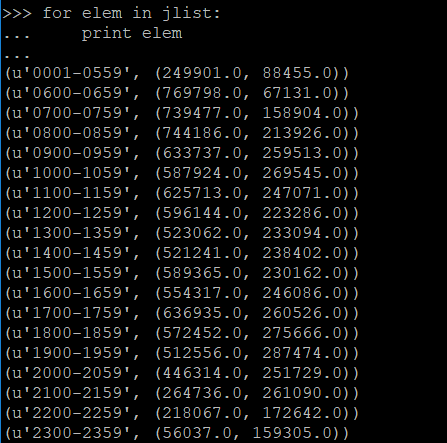


**Sorting in ascending order:**

taxi\_joined =carrier\_taxi\_out.join(carrier\_taxi\_in).sortByKey(1).collect()

for elem in taxi\_joined:

print elem



**Question 2:**

**Creating RDD of dalay times and removing 0s:**

carrier\_delayraw=carrier\_cleanh.map(lambda x :(x[8], x[9],x[34]))

carrier\_delay=carrier\_delayraw.filter(lambda x :x[2]!='0.00')

carrier\_delay.count()



carrier\_delay.take(3)



carrier\_delay=carrier\_delay.filter(lambda x :x[2]!='')

carrier\_delay.count()



carrier\_delay.take(3)



carrier\_delayf=carrier\_delay.map(lambda x :((x[0]+x[1]) ,float(x[2])))

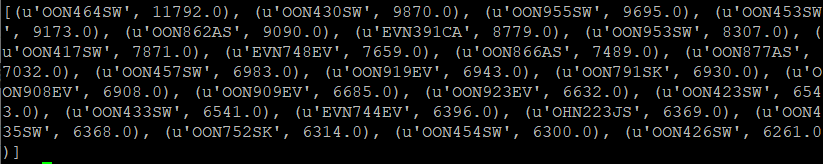
carrier\_delayf.take(3)



**Reducing by key value pairs by V:**

carrier\_delayf=carrier\_delayf.reduceByKey(lambda a, b:b+a)

carrier\_delayf.sortBy(lambda x:x[1], ascending = False).take(25)



**Sorting and taking first 25 lines:**

delay\_list = carrier\_delayf.sortBy(lambda x:x[1], ascending = False).take(25)

for elem in delay\_list:

print elem

