P3-D2

December 8, 2017

In [45]:
```
from pyspark.sql.types import StructType, StructField, FloatType, LongType, StringTyp
from pyspark.shell import spark
```

In [46]:
```
attrs = []
f = open('./GDELT-EVENTS-ATTRIBUTES.txt')

for line in f:
    tokens = line.split(',')
    if tokens[1].strip() == "INTEGER":
        attrs.append(StructField(tokens[0].strip(), IntegerType(), True))
    elif tokens[1].strip() == "STRING":
        attrs.append(StructField(tokens[0].strip(), StringType(), True))
    elif tokens[1].strip() == "FLOAT":
        attrs.append(StructField(tokens[0].strip(), FloatType(), True))

schema = StructType(attrs)
schema
```

Out[46]:
```
StructType(List(StructField(GLOBALEVENTID,IntegerType,true),StructField(SQLDATE,Integ
```

In [54]:
```
df = spark.read.format('CSV').option('sep', '\t').schema(schema).load('inputs/gdelt/2
```

In [55]:
```
df.take(1)
```

Out[55]:
```
[Row(GLOBALEVENTID=597122373, SQLDATE=20151110, MonthYear=201511, Year=2015, Fraction
```

In [56]:
```
df.count()
```

Out[56]:
```
73270827
```

In [57]:
```
df_2015 = spark.read.format('CSV').option('sep', '\t').schema(schema).load('inputs/gd
```

In [58]:
```
df_2015.take(1)
```

Out[58]:
```
[Row(GLOBALEVENTID=478037761, SQLDATE=20051025, MonthYear=200510, Year=2005, Fraction
```

In [59]:
```
df_2015.count()
```

Out[59]:
```
66370819
```