Lightning

November 29, 2017

In [18]:
```
from pyspark.sql.types import StructType, StructField, FloatType, LongType, StringType
from pyspark.shell import spark

feats = []
f = open('features.txt')
for line_num, line in enumerate(f):
    if line_num == 0:
        # Timestamp
        feats.append(StructField(line.strip(), LongType(), True))
    elif line_num == 1:
        # Geohash
        feats.append(StructField(line.strip(), StringType(), True))
    else:
        # Other features
        feats.append(StructField(line.strip(), FloatType(), True))

schema = StructType(feats)
```

In [19]:
```
df = spark.read.format('csv').option('sep', '\t').schema(schema).load('inputs/nam_2015
```

In [20]:
```
import pyspark.sql.functions as sf
from pyspark.sql import Column as col
with_prefix_column = df.withColumn("Prefix4",df.Geohash.substr(0,4))
#with_prefix_column.select(with_prefix_column.Geohash,with_prefix_column.Prefix4).sho
```

In [21]:
```
lightning_count_values = with_prefix_column.groupBy('Prefix4').agg(sf.sum(df.lightning
sorted_values = lightning_count_values.sort(lightning_count_values.num_times_lightning
sorted_values.select(sorted_values.Prefix4,sorted_values.num_times_lightning).show(n=5
```

```
+-------+------------------+
|Prefix4|num_times_lightning|
+-------+------------------+
|   9g3m|             713.0|
|   9g0g|             711.0|
|   9g3y|             677.0|
|   9err|             671.0|
|   9ery|             659.0|
+-------+------------------+
```

```
only showing top 5 rows
```