SummaryStatistics

November 30, 2017

In [42]:
```
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 [43]:
```
df = spark.read.format('csv').option('sep', '\t').schema(schema).load('inputs/mini-sa
```

In [44]:
```
len(df.columns)
```

Out[44]: 58

In [45]:
```
import pyspark.sql.functions as sf
summary_values = []
feature_names = []
for i in range(2,len(df.columns)):
    feature_names.append(df.columns[i])
    summary_values.append(df.select(sf.max(df.columns[i]).alias("Max"),
              sf.min(df.columns[i]).alias("Min"),
              sf.avg(df.columns[i]).alias("Avg"),
              sf.stddev(df.columns[i]).alias("Std_Dev")))
```

In [46]:
```
f = open("summary_stats.txt","w")
for i in range(0,len(feature_names)):
    f.write("Feature: " + feature_names[i] + "\n")
    df_summ = summary_values[i]
```

```
            summ_values = df_summ.select(df_summ.Max,df_summ.Min,df_summ.Avg,df_summ.Std_Dev)
            f.write("Max: " + (str)(summ_values[0].Max) + "\n")
            f.write("Min: " + (str)(summ_values[0].Min) + "\n")
            f.write("Average: " + (str)(summ_values[0].Avg) + "\n")
            f.write("Std.Dev: " + (str)(summ_values[0].Std_Dev) + "\n")
            f.write("\n")
        f.close()
```

In [ ]:
```
#approximately 32 mins to run this job
```