HottestTemperature

November 29, 2017

In [1]:
```
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)
```

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.2.0
      /_/

Using Python version 3.6.3 (default, Oct  6 2017 12:04:38)
SparkSession available as 'spark'.
```

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

In [3]:
```
sorted_temp_df = df.sort(df.temperature_surface.desc())
```

In [5]:
```
sorted_temp_df.select(sorted_temp_df.Geohash,sorted_temp_df.temperature_surface).show(
```

```
+------------+-------------------+
|     Geohash|temperature_surface|
+------------+-------------------+
```

```
|d75zuxsuqtpb|           320.9536|
|d59d5yttuc5b|           320.9536|
|d59zxv5vmd5z|           320.8286|
+------------+-------------------+
only showing top 3 rows
```

```
In [ ]: import pyspark.sql.functions as sf
        from pyspark.sql import Column as col
        max_temp = df.select(sf.max(df.temperature_surface).alias("max_temperature_surface"))
        max_temp_itr = max_temp.toLocalIterator()
        max_temp_list = [float(x.max_temperature_surface) for x in max_temp_itr]
        max_temp_list

Out[ ]: [330.67431640625]

In [ ]: geo_itr = df[df.temperature_surface.isin(max_temp_list)].collect()
        geohash_list = [row.Geohash for row in geo_itr]
        geohash_list

In [74]: # Creating an SQL 'table'
         df.createOrReplaceTempView("FEATURE_DF")

         # What's the maximum value?
         MaxTempValues = spark.sql("SELECT Geohash,temperature_surface FROM FEATURE_DF WHERE te

         MaxTempValues

Out[74]: [Row(Geohash='d75zuxsuqtpb', temperature_surface=320.95361328125),
          Row(Geohash='d59d5yttuc5b', temperature_surface=320.95361328125)]
```