SnowDepth

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_2015
```

In [3]:
```
import pyspark.sql.functions as sf
from pyspark.sql import Column as col
snow_values = df.groupBy('Geohash').agg(sf.min(df.snow_depth_surface).alias("min_snow_
                                        sf.avg(df.snow_depth_surface).alias("avg_sn
values_grtr_zero = snow_values.filter(snow_values.min_snow_depth_for_geohash > 0)
sorted_values_grtr_zero = values_grtr_zero.sort(values_grtr_zero.avg_snow_depth_for_ge
sorted_values_grtr_zero.select(sorted_values_grtr_zero.Geohash,sorted_values_grtr_zero
```

```
+------------+--------------------------+
|     Geohash|avg_snow_depth_for_geohash|
+------------+--------------------------+
|c41xurr50ypb|        1.4427825378580033|
|c1p5fmbjmkrz|        0.8555590688246522|
|c1gyqex11wpb|        0.5148891935595334|
+------------+--------------------------+
```