Import Libraries

In [3]:

```
import pandas as pd
import pyarrow as pa
import pyarrow.parquet as pq
import glob
import dask.dataframe as dd
```

Function to read CSV files

This can be more enhanced using parameters if needed rather than writing new code

In [4]:

```
def read_csv(filename):
    return pd.read_csv(
        filename
    )
```

Start of main analysis

Reading the entire forlder for CSV files

In [5]:

```
files = glob.glob("data/*.csv")
files
```

Out[5]:

```
['data\\weather.20160201.csv', 'data\\weather.20160301.csv']
```

Mapping all the CSV files to our function

In [6]:

```
dfs = list(map(read_csv, files))
dfs
```

Out[6]:

```
[       ForecastSiteCode  ObservationTime      ObservationDate  WindDirect
ion  \
 0                  3002                0  2016-02-01T00:00:00
12
 1                  3005                0  2016-02-01T00:00:00
10
 2                  3008                0  2016-02-01T00:00:00
8
 3                  3017                0  2016-02-01T00:00:00
6
 4                  3023                0  2016-02-01T00:00:00
10
 ...                 ...              ...                  ...
...
 93250              3797               23  2016-02-29T00:00:00
8
 93251              3866               23  2016-02-29T00:00:00
11
 93252              3872               23  2016-02-29T00:00:00
10
 93253              3876               23  2016-02-29T00:00:00
11
 93254              3882               23  2016-02-29T00:00:00
10

        WindSpeed  WindGust  Visibility  ScreenTemperature  Pressure  \
 0              8       NaN     30000.0                2.1     997.0
 1              2       NaN     35000.0                0.1     997.0
 2              6       NaN     50000.0                2.8     997.0
 3              8       NaN     40000.0                1.6     996.0
 4             30      37.0      2600.0                9.8     991.0
 ...          ...       ...         ...                ...       ...
 93250          7       NaN     25000.0                2.3    1025.0
 93251         14       NaN         NaN                6.4    1025.0
 93252          8       NaN     35000.0                5.7    1025.0
 93253          6       NaN         NaN                5.5    1025.0
 93254          2       NaN     18000.0                4.3    1025.0

        SignificantWeatherCode                     SiteName  Latitude  \
 0                           8             BALTASOUND (3002)   60.7490
 1                           7    LERWICK (S. SCREEN) (3005)   60.1390
 2                         -99              FAIR ISLE (3008)   59.5300
 3                           8              KIRKWALL (3017)   58.9540
 4                          11       SOUTH UIST RANGE (3023)   57.3580
 ...                       ...                           ...       ...
 93250                       8               MANSTON (3797)   51.3422
 93251                     -99      ST CATHERINES PT. (3866)   50.5770
 93252                       8         THORNEY ISLAND (3872)   50.8200
 93253                     -99              SHOREHAM (3876)   50.8360
 93254                       7  HERSTMONCEUX WEST END (3882)   50.8900

        Longitude                      Region   Country
 0        -0.8540           Orkney & Shetland  SCOTLAND
 1        -1.1830           Orkney & Shetland  SCOTLAND
 2        -1.6300           Orkney & Shetland       NaN
 3        -2.9000           Orkney & Shetland  SCOTLAND
 4        -7.3970      Highland & Eilean Siar  SCOTLAND
 ...          ...                         ...       ...
 93250     1.3461  London & South East England   ENGLAND
```

```
 93251    -1.2970  London & South East England   ENGLAND
 93252    -0.9200  London & South East England       NaN
 93253    -0.2920  London & South East England   ENGLAND
 93254     0.3190  London & South East England   ENGLAND

 [93255 rows x 15 columns],
         ForecastSiteCode  ObservationTime      ObservationDate  WindDirec
tion  \
 0                   3002                0  2016-03-01T00:00:00
8
 1                   3005                0  2016-03-01T00:00:00
8
 2                   3008                0  2016-03-01T00:00:00
7
 3                   3017                0  2016-03-01T00:00:00
7
 4                   3023                0  2016-03-01T00:00:00
10
 ...                  ...              ...                  ...
...
 101437              3797               23  2016-03-31T00:00:00
1
 101438              3866               23  2016-03-31T00:00:00
0
 101439              3872               23  2016-03-31T00:00:00
1
 101440              3876               23  2016-03-31T00:00:00
1
 101441              3882               23  2016-03-31T00:00:00
1

         WindSpeed  WindGust  Visibility  ScreenTemperature  Pressure  \
 0              23      30.0     16000.0              -99.0       NaN
 1              26      34.0      5000.0                4.9    1004.0
 2              30      40.0      5000.0                5.1    1003.0
 3              21      29.0      5000.0                5.1    1001.0
 4              25      34.0      2400.0                8.6     994.0
 ...           ...       ...         ...                ...       ...
 101437          5       NaN     22000.0                4.9    1019.0
 101438         10       NaN         NaN                8.4    1018.0
 101439          2       NaN     50000.0                3.5    1019.0
 101440          3       NaN         NaN                6.1    1019.0
 101441          2       NaN     35000.0                3.7    1019.0

         SignificantWeatherCode                     SiteName  Latitude  \
 0                            8             BALTASOUND (3002)   60.7490
 1                           12    LERWICK (S. SCREEN) (3005)   60.1390
 2                           11              FAIR ISLE (3008)   59.5300
 3                           15               KIRKWALL (3017)   58.9540
 4                           12       SOUTH UIST RANGE (3023)   57.3580
 ...                        ...                           ...       ...
 101437                       0               MANSTON (3797)   51.3422
 101438                     -99      ST CATHERINES PT. (3866)   50.5770
 101439                       0         THORNEY ISLAND (3872)   50.8200
 101440                     -99               SHOREHAM (3876)   50.8360
 101441                       0  HERSTMONCEUX WEST END (3882)   50.8900

         Longitude                       Region   Country
 0         -0.8540           Orkney & Shetland  SCOTLAND
 1         -1.1830           Orkney & Shetland  SCOTLAND
 2         -1.6300           Orkney & Shetland       NaN
```

```
3          -2.9000            Orkney & Shetland  SCOTLAND
4          -7.3970       Highland & Eilean Siar  SCOTLAND
...            ...                          ...       ...
101437      1.3461  London & South East England   ENGLAND
101438     -1.2970  London & South East England   ENGLAND
101439     -0.9200  London & South East England       NaN
101440     -0.2920  London & South East England   ENGLAND
101441      0.3190  London & South East England   ENGLAND

[101442 rows x 15 columns]]
```

Use the first table to create schema for the writer

In [7]:

```
table = pa.Table.from_pandas(dfs[0], preserve_index=False)
writer = pq.ParquetWriter('weather-rowgroups.parquet', table.schema)
```

In [8]:

```
table
```

Out[8]:

```
pyarrow.Table
ForecastSiteCode: int64
ObservationTime: int64
ObservationDate: string
WindDirection: int64
WindSpeed: int64
WindGust: double
Visibility: double
ScreenTemperature: double
Pressure: double
SignificantWeatherCode: int64
SiteName: string
Latitude: double
Longitude: double
Region: string
Country: string
```

Using Writer and the dataframes to create table

In [9]:

```
for df in dfs:
    table = pa.Table.from_pandas(df, preserve_index=False)
    writer.write_table(table)
writer.close()
```

Some analysis on the parquet file and its row groups to identify characteristics of our data structure

In [10]:

```
filename = "weather-rowgroups.parquet"
pq_file = pq.ParquetFile(filename)
```

In [11]:

```
data = []
for rg in range(pq_file.metadata.num_row_groups):
    rg_meta = pq_file.metadata.row_group(rg)
    data.append([rg, rg_meta.num_rows, rg_meta.total_byte_size])
data
```

Out[11]:

```
[[0, 93255, 537181], [1, 101442, 560608]]
```

In [12]:

```
# To get number of rows
pq_file.metadata.num_rows
```

Out[12]:

```
194697
```

In [13]:

```
# To get number of columns
pq_file.metadata.num_columns
```

Out[13]:

```
15
```

In [14]:

```
# To get metadata of column
rg_meta.column(7)
```

Out[14]:

```
<pyarrow._parquet.ColumnChunkMetaData object at 0x00000179450F1720>
  file_offset: 913160
  file_path:
  physical_type: DOUBLE
  num_values: 101442
  path_in_schema: ScreenTemperature
  is_stats_set: True
  statistics:
    <pyarrow._parquet.Statistics object at 0x00000179450F1C20>
      has_min_max: True
      min: -99.0
      max: 15.8
      null_count: 0
      distinct_count: 0
      num_values: 101442
      physical_type: DOUBLE
      logical_type: None
      converted_type (legacy): NONE
  compression: SNAPPY
  encodings: ('PLAIN_DICTIONARY', 'PLAIN', 'RLE')
  has_dictionary_page: True
  dictionary_page_offset: 810485
  data_page_offset: 811425
  total_compressed_size: 102675
  total_uncompressed_size: 103687
```

Find min and max statistics of a column for each row group

In [15]:

```
column = 7
data = []

for rg in range(pq_file.metadata.num_row_groups):
    rg_meta = pq_file.metadata.row_group(rg)
    data.append([rg, str(rg_meta.column(column).statistics.min), str(rg_meta.column(col
umn).statistics.max)])

print(data)
data_df = pd.DataFrame(data, columns=["rowgroup", "min", "max"])
```

```
[[0, '-99.0', '15.6'], [1, '-99.0', '15.8']]
```

In [42]:

```
# To find maximum value
max_temp = float(data_df['max'].max())
max_temp
```

Out[42]:

```
15.8
```

In [43]:

```
rg_meta.column(column).statistics.max
```

Out[43]:

15.8

Using the maximum tempreture to filter our tour data and columns to avoid fetching extra data and limit the load to what we really need.

In [44]:

```
df = dd.read_parquet("weather-rowgroups.parquet", columns=['ObservationDate', 'Region',
'ScreenTemperature'])
```

In [45]:

```
df = df[df.ScreenTemperature == max_temp]
```

In [46]:

```
df.compute()
```

Out[46]:

	ObservationDate	Region	ScreenTemperature
147768	2016-03-17T00:00:00	Highland & Eilean Siar	15.8

Result

Hottest day = 2016-03-17T00:00:00

Tempreture on that day = 15.8

Region = Highland & Eilean Siar