Contents

BUSINESS REQUIREMENTS

Convert the weather data into parquet format. Set the raw group to appropriate value you see fit for this data.

The converted data should be quarriable to answer the following question.

- Which date was the hottest day?

- What was the temperature on that day?

- In which region was the hottest day?

The weather data is provided separately

DATA TYPE and INFORMATION

Use the weather data from client (w1.csv and w2.csv). It includes two different months' data. There are 15 fields for that. Please find below data types for csv files.

Column Name	Data Type
ForecastSiteCode	IntegerType
ObservationTime	IntegerType
ObservationDate	DateType
WindDirection	IntegerType
WindSpeed	IntegerType
WindGust	IntegerType
Visibility	IntegerType
ScreenTemperature	FloatType
Pressure	IntegerType
SignificantWeatherCode	IntegerType
SiteName	StringType
Latitude	DoubleType
Longitude	DoubleType
Region	StringType
Country	StringType

C2P

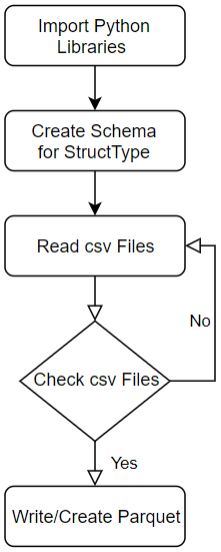
“c2p.py” has been created for converting the file csv to parquet.

Some libraries have been installed for running this python code.

- Shutil, os, sys, pathlib has been used file systems operations.
- Json has been used for creating response from method.
- Pyspark has been used for converting csv to parquet.

Csv files has been converted parquet file using pyspark.

Before that, we needed to read csv file using “spark.read.option” and csv data types.



C2P TEST CASES

6 different test cases/classes have been created for unit test. Also, some default values have been created for unit test:

cls.**testCsvFile** = "test.csv"

cls.**outdir** = "out"

cls.**maxTemp** = 15.8

cls.**hottestDate** = '2016-03-17'

cls.**hottestRegion** = 'Highland & Eilean Siar'

cls.**create_a_csv_file**(cls, cls.testCsvFile, cls.textList)

test00_return_error_when_file_not_exists

If the csv file doesn’t exist in folder, you can see the error on the output line. It has been written on the line using “assertEqual”

test01_convert_csv_correctly

When I check the converting csv to parquet file correctly in that test case, “test.csv” file has been created as temp for checking the data which has been correctly. If everything goes correctly, this test case will be successful.

test02_convert_weather_file

This test case for only checking csv to parquet file process. If “test02_convert_weather_file” complete correctly, unit test case passes this step.

test03_run_sql_for_hottest_day

This test case checks the “Which date was the hottest day?”. If it works correctly, hottest day has been put the output line using below sql.

```
SELECT ObservationDate as hday, max(ScreenTemperature) as temp
FROM parquetFile
GROUP BY hday
ORDER BY temp desc
LIMIT 1
```

test04_run_sql_for_temperature_on_hottest_day

This test case checks the “What was the temperature on that day?”. If it works correctly, temperature has been put the output line.

test05_run_sql_for_region_of_hottest_day

This test case checks the “In which region was the hottest day?”. If it works correctly, region has been put the output line using below sql.

```
SELECT ObservationDate as hday, Region as region, max(ScreenTemperature) as temp
FROM parquetFile
GROUP BY hday, region
ORDER BY temp desc
LIMIT 1
```