# DATA ENGINEER PYTHON TEST

# Solution by Jason Frempong

Environment:

Development was done initially in pycharm (Python 3.7) then scripted in notepad++ and tested in a powershell terminal (Powershell V5 run in Admin mode). Spark was used as another option to query the parquet files (see “pyspark commands.txt” text file for commands). All development was done locally on a Win10 device in a local directory “C:\dlg” with Read/Write access.

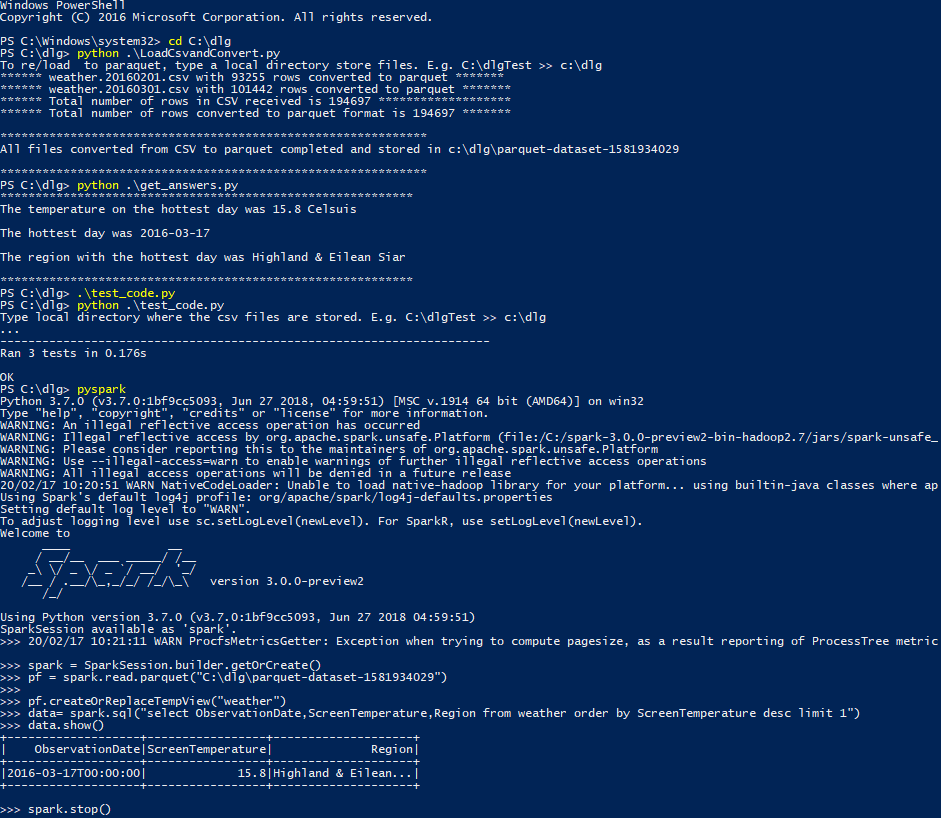
Several python libraries were installed and imported; these are listed above each script.

Scripts:

1. LoadCsvandConvert.py
   1. This script requires user input to specify directory of scripts and CSV files.
   2. It has a method to grad only the csv files and convert to parquet whilst checking counts as a sanity check. The script is repeatable and generated new parquet datasets each time. It also writes into a txt file with the latest dataset filename for down stream processes. Assuming new weather data is dropped into the same location, this process allows for new parquet file creation without overwriting previous data that may be in use.
   3. There are other methods that sets the file name of the latest dataset and another to produce the answers to the test questions (QueryParquetFiles.py expands on this).
2. QueryParquetFiles.py
   1. This script gets the location of the latest parquet dataset and queries it to get the answer needed. This script also supports unit testing.
3. get\_answers.py
   1. This script produces a report of the answers
4. test\_code.py
   1. This script unit tests my code with a few test cases.
5. pyspark commands.txt
   1. These are shell commands in pyspark to query the dataset via SQL. The directory of the dataset is needed.
6. dataset\_location.txt
   1. A text file that is updated each time “LoadCsvandConvert.py” is ran.

Review process:

1. Copy all files to a local directory
2. Open all scripts and review in editor of choice
3. Open cmd in admin mode
4. Change directory to the local directory and run LoadCsvandConvert.py
5. Input the same directory (without quotes) and hit return to run script.
6. Run get\_answers.py to get report of answers.
7. Run test\_code.py to unit test code
8. Assuming spark is installed and all env variables are set up:
   1. Run in cmd “pyspark” to start spark
   2. Copy and paste each line of code in “pyspark commands.txt” (change the directory of the dataset)



Next Steps:

Automate process with event based processing triggered when csv files are dropped into specified location.