# Reference Document

This reference document contains the assumption made while writing the program.

The data is received in the csv format with file names starting with ‘weather.’ appended by the date.

The schema of the file is assumed to be the same for all the files.

In the script:

1. The program first reads the data from the csv files and writes it to the parquet format with partitions on the ‘Country’ column. The data from parquet files are then read and a new dataframe ‘df’ is created. This dataframe is then used to find the hottest temperature of the month and their corresponding details.
2. The program is run along with the config file. The config file currently contains the path of the csv files. This config file can later be used to set different variables like servers, environments, log file paths, parquet file paths, etc.
3. This program is logging enabled thus making it easy to track the execution.
4. The program gives the output for each month.
5. The program is saved in .py format.
6. The same program can be automated for the scenario: csv source file is received every month for previous month data.

How to run the program?

1. Any python IDE can be used to run the program.
2. Place the program file and .config file in the same location.
3. Provide the location of the source files in the config file.
4. The output of the program is printed on the console.