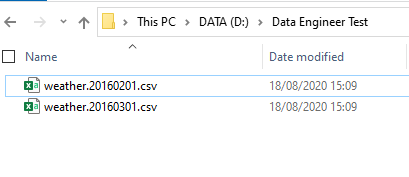
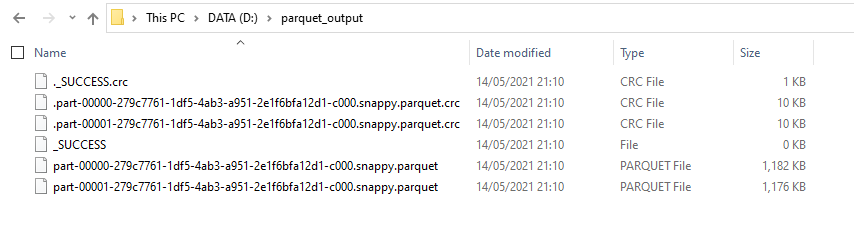
**Assumptions and Test**

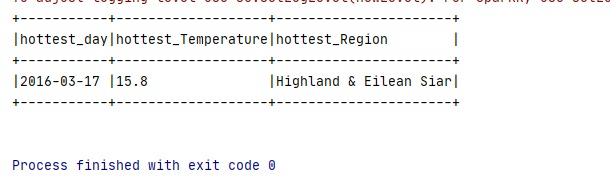
* Solution is designed keeping Big data problem in mind
* Pyspark is used for solution
* Data files are placed at D:\Data Engineer Test



* Parquet files are generated at D:\parquet\_output

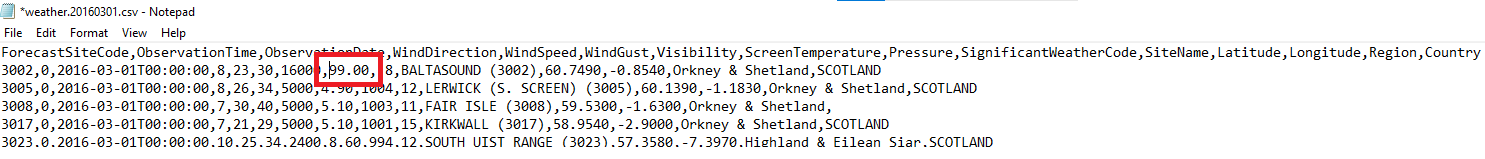


* Parquet Row size is set for 2 MB
* Analysed output generated as dataframe



**Additional Test**

* Update screen temperature to 99.00 in weather.20160301.csv



* And validate the updated output

