**Python/****PySpark Exercise**

There is a source S3 bucket where RAW log files (in json format) are arriving in intervals of 10 minutes. The files contain CPC (cost per click – “money value” of the keyword) estimations and search volumes per keyword. We want to process the RAW data and create an aggregate table.  
  
Please answer the following questions:

1. Write a python/PySpark script that loads the attached files (RAW log files) and creates output files to support the agg table (schema is in the table below). You should support the following features:
   1. Agg data in daily granularity (each partition has its own file).
   2. Create a function that gets a dataframe and writes the output file. Try to save files in parquet format, otherwise write a csv file.
   3. Handle exceptions and raise errors if needed.
2. Describe how you would manage the queue of log files? How would you prevent loading the same file more than once? How do you make sure your process is resilient to failures and at the same time manages to process all files?

Agg table schema:

|  |  |  |
| --- | --- | --- |
| **fact\_daily\_agg\_avg\_cpc\_estimation** | | |
| **column name** | **type** | **comments** |
| keyword | string |  |
| avg\_searches\_monthly\_volume | double |  |
| avg\_cpc | double |  |
| dt | string | partition column (yyyymmdd) |

GOOD LUCK!