Data Quality Engineer Code Test

Instructions:

This code test is meant to get an understanding of your data quality pipeline design skills. The solution code files and any data output files need to be posted to a public GitHub repository owned by you. Please commit directly to the master branch. Commits can be made as often as needed.

System/Tool requirements:

- Any system with 4GB and above RAM can be used.
- You can use any data analysis tool to perform this task. Preferably, Python or R. This needs to be installed and functioning in your system, in case not already present, before you start the code test.

Problem Statement:

Zomato, an online food-ordering platform, receives data files daily which comprises data regarding existing restaurants or new restaurants. The files are location specific. You need to design a data processing/ingestion pipeline which runs daily to process these files and come up with required output files.

The data files should be read as per the following column data types:

Field Name	Data Type	Notes
url	char	
address	char	
name	char	Not null
rate	string	
votes	integer	
phone	integer	Not null
location	string	Not null
rest_type	string	
dish_liked	list	It is a comma-separated list in the data.

cuisines	string	It is a comma-separated list in the data.
reviews_list	array/list	

The following modules need to be incorporated into the utility which serves various functions as described below. Anything above and beyond what is explained here will be a good-to-have functionality.

1. File check module - Reads file daily from the given source location and applies few initial checks on the file like:
    - Check if it's a new file or not. We do not want to reprocess already processed files.
    - It should be a non-empty file.
    - File extensions are proper. Please consider .csv in this case.

    Note - There is a possibility of more than one file coming in any given day. Please make sure your code handles this sufficiently.

2. Data quality check module -
    1. Required Data Quality Checks:
        a. This module should take care of all possible kinds of data cleaning and validation activities that can be done to make sure the output data is in properly structured, readable and storable format. To achieve this you can do various checks including but not limited to-
            1. Descriptive fields like address, reviews_list can be cleaned by removing commas, period, exclamations or any other special/junk characters etc.
            2. Data in the phone field can be validated for correct phone numbers.
                - Any preceding "+" or spaces should be removed.
                - Ensure phone numbers are correctly formatted.
                - The field data can be split and stored in two separate fields e.g. contact number 1 and contact number 2 for easy readability and access.
            3. Checks for null values in fields which you think should not have null values.
            4. Validation of location field for correctness of data by looking up to the Areas_in_blore.csv file.
    2. Your Own Data Quality Check: Please come up with one data quality check idea of your own and incorporate it into the code. You can use any of the provided fields in the dataset.

3. Output files module - After the cleaning/validation operation, the data needs to be written into files having below output format.

- Capture all the clean records to a .out file.
- Capture all the bad records in a .bad file.
- For the .bad file create a metadata file which will contain the following fields:
    1. Type_of_issue - this is a short keyword for the type of non-conformity.
    2. Row_num_list - list of all the row numbers which have the issue.

  For e.g. if there are null records found in the dataset then type_of_issue field will have the value “null” and row_num_list will contain the list of all the row numbers which have the issue.

Files You’ll Need:

- data_file_20210527182730.csv
- data_file_20210528182554.csv
- data_file_20210528182844.csv
- Areas_in_blore.csv