TRAVANCORE ANALYTICS

CSV File Validation

Created By : Baiju M P

Created on: 29-Nov-2021

What needs to be done

Wanted to validate uploaded csv file with the following constraints. The validation logic should traverse the whole csv file and report all errors.

- The fields will be separated by commas (,)
- Typical csv file size will be in range of 500 MB
- Every record should have equal number of columns. The columns in first row will be taken as reference. All rows with less or greater than the reference column count should be reported.
- The data type of all cells in any column should be same. All other cases should be reported. If the first field is empty for any column, the very first non-empty value will be taken as reference to determine data type of the column.
- There should not be any NaN data in any cells. The data should be either empty or should have a valid value.
- The logic should be scalable to add more complex data types.

How validation works

Here is the validation data flow

- All valid data types needs to be populated in a string array. Any new data types can be added to this array in future.
- Traverse through the columns to get the very first non-empty data. This data will be checked for match with every items in regex array to find most closed match, and that pattern will be treated as data type of the column.
- Using **MatchesPatternValidation** function of pandas_schema library, set validation pattern for every columns of the table data. In addition to that, a lambda function is added in pandas_schema to validate for null check.
- Invoke **validate**() of pandas_schama library which in turn validate the csv against the given data types for every columns and report errors in an array.
- In another logic validate column counts of every rows and append errors to already existing errors list
- Print all errors in another csv file

Validation metrics

- Pandas_schema library of python will be used to validate field data types, as it is powered by pandas and is performance oriented.
- For a 50MB sized file, it takes around 15-17 seconds for data validation with 20% of incorrect data. For 500MB sized file, the time taken will be around 150-170 seconds.

Thank you