

PBI API ETL FLOW DOCUMENTATION

Azure Datalake Gen2 folder with daily loaded comma delimited standard CSVs

Condition 1 (checks to see if a master file has already been created)
If master file already exists (Boolean)

If response = true

Delta Script

Condition 2: Compares processed in masterfile existing and list of unprocessed and counts comparison.
If File Count > 0

If response count > 0

Condition 3: checks to see if the file being processed contains data.
If file is **not** empty (Boolean)

If response = true

Condition 4: checks to see if a column in each file to process has a column "description"
If Column Description Exists (Boolean)

If response = true

Adds ETL transforming the column description using trim and regex replace which replaces all newlines and leading/trailing spaces.

If response = false

Iterates through each file and uses sparks DataFrame API to transform the dataframe to end desired masterfile.

If response count < 0

If response = false

Full Script

Condition 3: checks to see if the file being processed contains data.
If file is **not** empty (Boolean)

If response = true

Condition 3: checks to see if a column in each file to process has a column "description"
If Column Description Exists (Boolean)

If response = true

Adds ETL transforming the column description using trim and regex replace which replaces all newlines and leading/trailing spaces.

If response = false

Iterates through each file and uses sparks DataFrame API to transform the dataframe to end desired masterfile.

If response = false

Appends Each transformed File, joins Schemas, and writes master dataframe to a on master csv into Azure Datalake Gen2