DFS TD - Synthesis Filtration Raw Data

Technical documentation for the data coming from tests on the equipment Filtration

Summary

Sum-up

Equipment / Scale	Filtration 25L	Filtration 170L	Filtration 170L Korea	2500L
Data Sources	ELN, Raw Data on file share	ELN, Raw Data on file share	ELN, Raw Data on file share	ELN, Raw Data on Google Drive
Raw Data File type	xlsx	xlsx	xlsx	xlsx
Scale Name on ELN	FR-25L	FR-170L	KR-170L	FR-2500L
Data Collection	Talend: R011_Download_Synthesis_gDrive_Filtration	Talend: J010_Download_Synthesis_LabServers Python : download_filtration_170L.py	Talend: J010_Download_Synthesis_LabServers Python : download_filtration_170L_KR.py	Talend: R011_Download_Synthesis_gDrive_Filtration
Parse	Python: parse_filtration_25L.py	Python: parse_filtration_170L.py	Python: parse_filtration_170L.py	Python: parse_filtration_2500L.py
Compute	Python: compute_filtration_25L.py	Python: compute_filtration_170L.py	Python: compute_filtration_170L.py	Python: compute_filtration_2500L.py
BigQuery	*Target tables* • *raw_data_synthesis.FiltrationDetails* • *raw_data_synthesis.FiltrationSummary*			
Mapping spreadsheet	Data Mapping			

Data Sources

- ELN
- Raw Data on Google Drive

Data Collection

The talend jobs **J010_Download_Synthesis_LabServers** and **J011_Download_Synthesis_gDrive_Filtration** extract the raw data files listed on the ELN table **filtration_raw_data_link** for which the field “filtration_plate_equipment_name” is the scale name, i.e. “FR-170L”. For information of how these job works, check the following page :

Talend - Jobs - Synthesis - Download - Filtration (needs to be created)

Schema using Google Drive

Examples

Talend jobs

- R011_Download_gDrive_Drying
- R012_Download_gDrive_Filtration

Tmp Folder

- D:\DATA\[ENV]\RnI\Silica\tmp\Synthesis\DryingTesla
- D:\DATA\[ENV]\RnI\Silica\tmp\Synthesis\Filtration2500L

Schema with file share

Exemples

Lab servers source

- \\FRPH2-labpc-backup\labo\W-522649\DATAS DATALAKE
- \\FRPH2-LABPC-BACKUP\LABO\W-509931

Python files

- download_filtration170L.py
- download_synthesis25L.py

Output folders

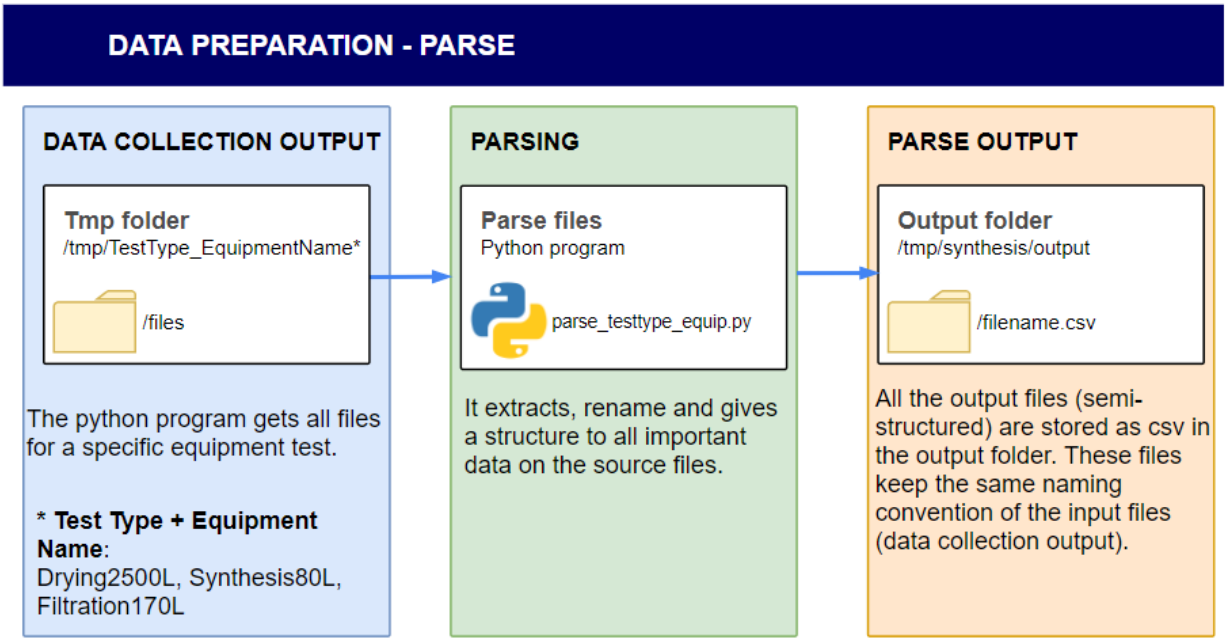
- D:\DATA\[ENV]\RnI\Silica\tmp\Synthesis25L
- D:\DATA\[ENV]\RnI\Silica\tmp\Synthesis170L

Please refers to the DFS TD - Synthesis - Norms and Conventions for the output filename convention on the Data Collection section

Data Preparation

Parse

The parsing python scripts extracts from the raw data files the needed columns.



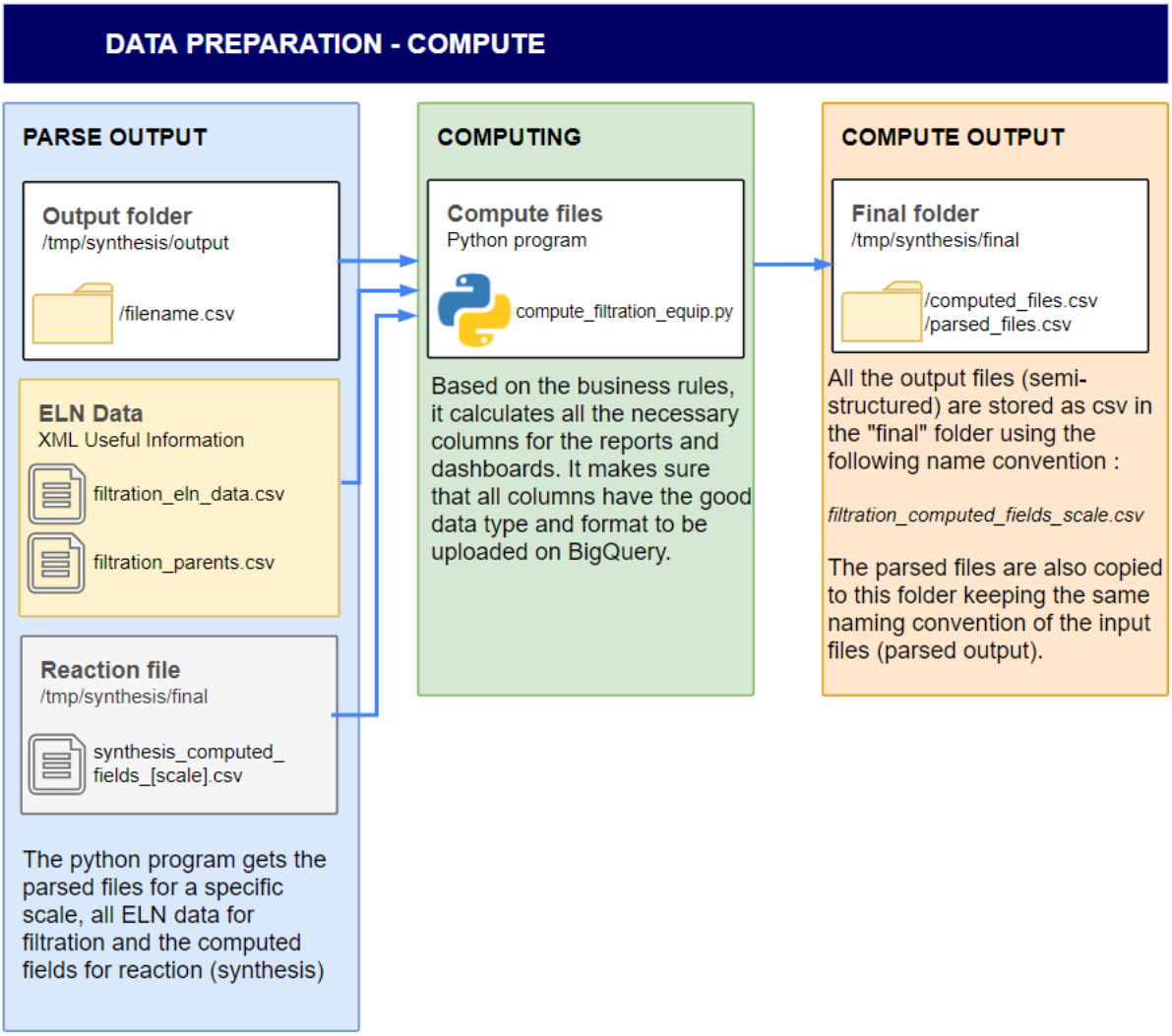
Columns List

For each sample, the script extracts the many fields from the raw data files and outputs a .csv file. For the mapping details, please refers to the sheet "Parse Mapping " on the Filtration Mapping spreadsheet (link to the spreadsheet on the Sum-up section).

Compute

The compute python script uses as input the parsed .csv files previously created . It computes the new columns and values from raw data and regenerates new files.

If the output files already exist the script will **NOT** replace them.



In the beginning of the script , it extracts many columns or values from ***filtration_eln_data, filtration_parents*** *and* ***synthesis_computed_fields_[scale]*** *files* . Those values are used in later computations as constants.

For each sample, it creates two different files that will be used to create new tables on BigQuery :

FiltrationDetails

The **first table** is composed of the columns previously extracted from the raw data files and the new columns calculated during the execution.

Dataset : raw_data_synthesis

For the columns details, please refers to the sheets " **Details Mappings** " on the **Filtration Mapping** spreadsheet (link to the spreadsheet on the Sum-up section).

FiltrationSummary

The **second table** is composed of the new values computed from raw data. This is a atomic table and it aggregates the values by **unique_id, study_id and sample_id** which represents one line per data raw file.

Dataset : raw_data_synthesis

For the columns details, please refers to the sheets " **Summary Mapping** " on the **Filtration Mapping** spreadsheet (link to the spreadsheet on the Sum-up section).

Presentation

The details and summary files are created as tables on **BigQuery** unifying all scales in the same tables. A Talend job is responsable to push all this data to a dataset called **raw_data_synthesis.**