Raw Water Torrelavega - 3 - Technical - Data Preparation

High Level Project Architecture

Here is a suggested template:

https://drive.google.com/file/d/1R9Y2e5TI_pmrghmnEquAOS-1Lq7Pdofw1O61ZHr4g1s/view

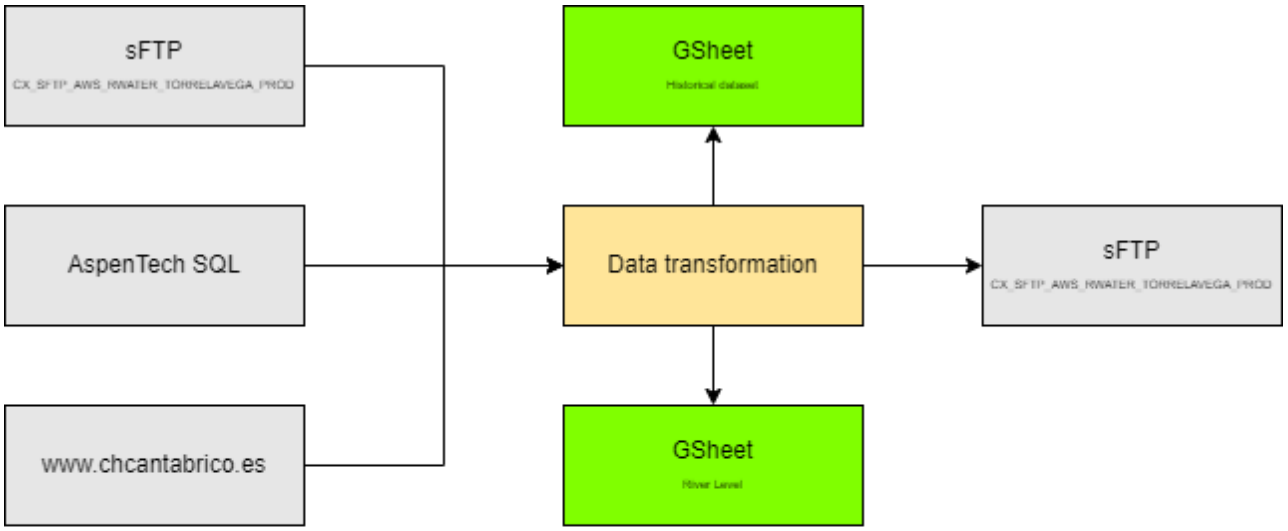
Architecture Data Flow

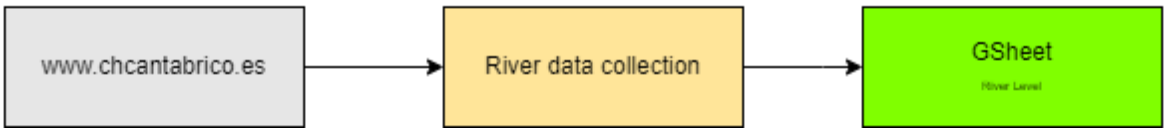
Here is a suggested template for Data Model + Data Mapping :

https://docs.google.com/spreadsheets/d/1bD8AIgsNUI2sgANoOEKTuHBlkxhsNVTD8cmOPYEloLw

DataPrep Flow

> *Schema showing the different STEPS of the application flow - with the data involved at each step*





Steps descriptions

> *Describe the data and process involved at each step*

Data Transformation

Description

> *Data from Sinfin, MES and CHCantabrico are aggregated and sent to Spanish Government through sFTP sevrer. S ince October 23, 2023, all the data comes directly from MES. So Webscraping step is therefore no longer applicable.*

Tools

> *WebMethods provides the sFTP server. Dataiku transforms and sends data to sFTP.*

Access rights

> - *sFTP: Dataiku connection is defined. Connection name: CX_SFTP_AWS_RWATER_TORRELAVEGA_PROD.*
> - *CHCantabrico: web scraping. No rights required. We have ceased web scraping since October 23, 2023, upon Manuel Valsco's request (WO0000000483037). Indeed, the data has been directly loaded from the MES since that day.*
> - *MES: Service account should be provided.*
> - *GSheets; service account should is stored in Dataiku Folder.*

Source

Location

- *sFTP: Data is stored on the AWS server. We retrieve this data directly to Dataiku.*
- *MES*
- *GSheets*

Format

- *sFTP: csv files.*
- *CHCantabrico: pandas DataFrame We no longer use web scraping data. For more accuracy, all the data is collected directly from the MES since October 23, 2023.*
- *MES: SQL table*

Destination

Location

- *Dataiku: https://dss.solvay.com/projects/RAW_WATER_TORRELAVEGA/managedfolder/5pHpj002/view/*
- *GSheets: 10fomPtvKsrliSpzfscny3j0YDEnk_7LOrymEyEnxVS8*

Format

csv files.

Sizing

Expected data volume for :

- *full process*
- *incremental process*

Full process: doesn't exist.

River level: 671 data point, 15.9+ KB

Incremental process:

Assessment

MES data: data preparation is done only when new data is available.

sFTP data: data preparation is done only when new data is uploaded. If some time stamps are missing, they are interpolated. If no data is uploaded for more than a day, email notification is sent to developers and SINFIN representatives on a daily basis.

Chcantabrico data: Data is appended to the historical data in the Dataiku dataset. We collect data from the website. When data is not available directly, the data is extracted from another page on the website. Not needed anymore since October 23, 2023. The data comes directly from MES.
Finally we try later that day.

Scheduling

Extraction is done every hour. Similarly, reporting is done when the transformation is finished.

Timing

The average time expected for :

- *full process*
- *incremental process*

Full process doesn't exist. Incremental process takes around 1.5 minutes.

Criticality

> *High*

Logging

> *Dataiku logging is stored for 2 days.*

River Data Collection

Description

> *Data from Chcantabrico.es is collected and stored on GSheet*

Tools

> *Dataiku*

Access rights

> *Dataiku collects the data. Google Service Account is stored in the Dataiku folder.*

Source

Location

> *Website. Source page may change, but quite unlikely. If so, the project must be updated.*

Format

> *pd DataFrame*

Destination

Location

> *GSheets: 1Sja2VlbUmya2Fa340mD3uT9DEIh3wjq_vTxu26IkVQc*

Format

> *table*

Sizing

> *Expected data volume for :*
>
> - *full process*
> - *incremental process*

River level: 671 data point, 15.9+ KB

Assessment

> *Chcantabrico data: Data is appended to the historical data in the Dataiku dataset. We collect data from the website. When data is not available directly, the data is extracted from another page on the website.* *Not needed anymore since October 23, 2023. The data comes directly from MES.*
> *Finally we try later that day.*

Scheduling

> *Extraction is done every 15 minutes.*

Timing

> *The average time expected for :*
>
> - *full process*
> - *incremental process*

Full process takes 20 seconds.

Criticality

> *High*

Logging

> *Dataiku logging.*