

# Mecano - 3 - Technical - Data Pipeline

- High-Level Design architecture (HLD)
- Low-Level Design architecture (LLD)
- Architecture Data Flow
  - DataPrep Flow
- - A detailed description of the Talend flows can be found [here](#).
  - Conceptual Data Model
- Steps descriptions
  - DataSource
    - Description
    - Tools
    - Access rights
    - Source
      - Location
      - Format
    - Destination
      - Location
      - Format
      - Sizing
    - Assessment
    - Scheduling
    - Timing
    - Criticality
    - Logging
- [cs-ew1-prj-data-dm-industrial-prod-staging](#)

## High-Level Design architecture (HLD)

[Link.](#)

## Low-Level Design architecture (LLD)

[Link.](#)

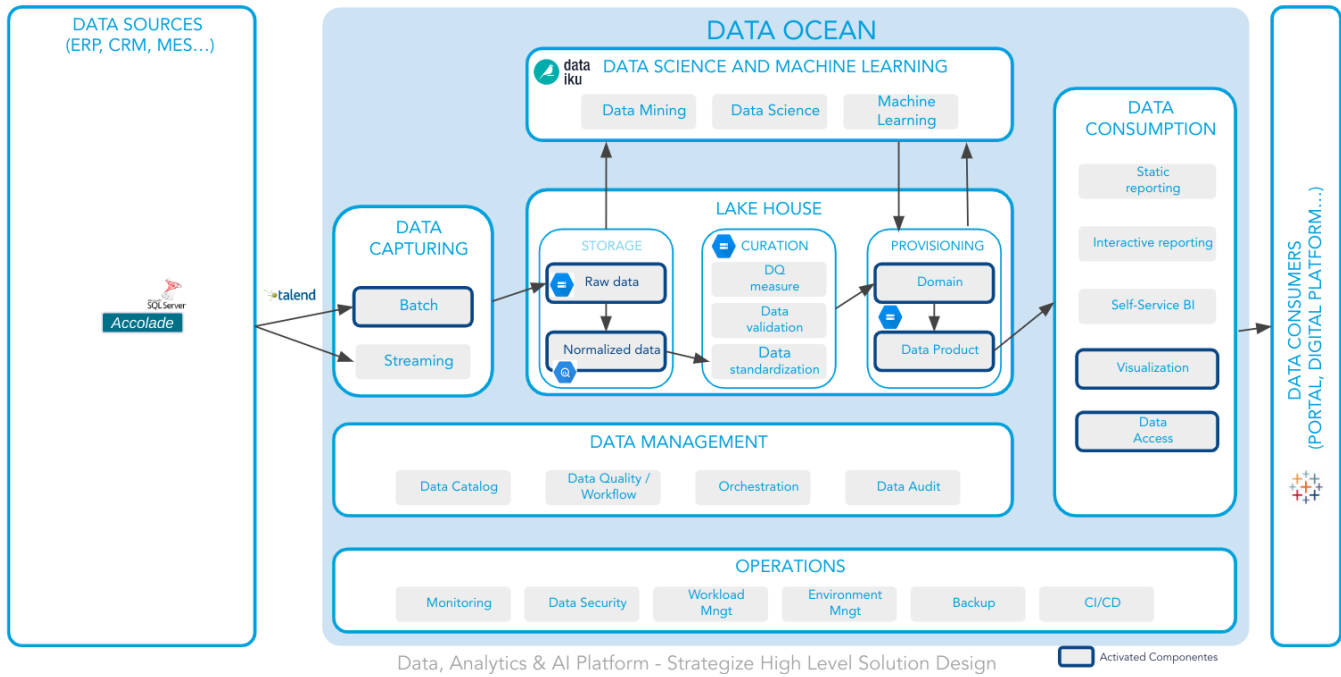
## Architecture Data Flow

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

DA&DA - Domain Mapping INDUSTRIAL

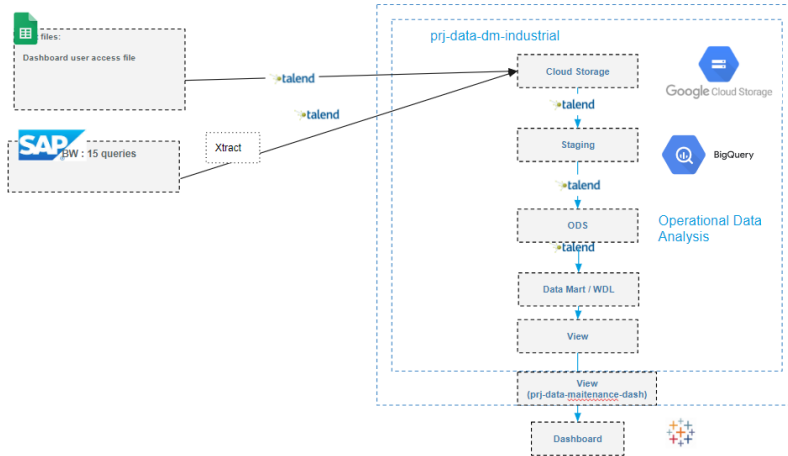
## DataPrep Flow

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



According to Data Ocean Blueprint and macro architecture

### Data Architecture - MECANO



BW --> GCP Data Ocean INDUSTRIAL (prj-data-dm-industrial-[env]) --> GCP Product MECANO (prj-data-pmo-maintenance-dash-[env]) --> Output Tableau

**A detailed description of the Talend flows can be found here.**

### Conceptual Data Model

The data model is accessible in GenMyModel.

### Steps descriptions

Data lineage can be found in the Google Sheet here below

Note: Table on project prj-data-dm-industrial-[env]

## DataSource

### Description

*ECC (PF1/WP1)*

*BW is the application where from where we collect information. The data is extracted using specific BW queries with detail in [the link](#)*

### Tools

*Talend to collect the data from the source and store on GCP/Google Big Query.*

### Access rights

*You can use Xtract server (ACEW1DTLNDENG02) to manually extract and monitor.*

*Only DA&AI DataEng's and Data Architects can access data on GCP/Google Big Query.*

### Source

#### Location

*Xtract server (ACEW1DTLNDENG02)*

#### Format

*BW queries extracted via CSV file*

### Destination

#### Location

*Extracted data will be stored on GCP/Cloud storage and Google Big Query.*

#### Format

*The format of the data saved in the databank*

### Sizing

*Expected data volume for :*

- incremental process from source to staging (as of Jan 8th 2024)*

*SELECT \* FROM STG.log\_files WHERE meta\_run\_id = '874d0cf9-94c4-4739-90ef-2d4d7f416d4c'*

Row	meta_nut_id	meta_file_name	meta_bucket_load_date	meta_stg_load_date	meta_status	meta_source_system	meta_periodicity	meta_extraction_type	meta_domain	meta_num_lines	meta_num_error
1	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F002_20 240108164747_0000_F_W_M EC_DBPMMND02_0001.csv	2024-01-08 16:48:05 UTC	2024-01-08 17:10:24 UTC	OK	BWH	random	full	IND	35316	0
2	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F001_20 240108164747_0000_F_W_M EC_DBPMMND01_0001.csv	2024-01-08 16:47:57 UTC	2024-01-08 17:10:40 UTC	OK	BWH	random	full	IND	14746	0
3	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F003_20 240108164747_0000_F_W_M EC_DBPMMND03_0001.csv	2024-01-08 16:48:08 UTC	2024-01-08 17:10:06 UTC	OK	BWH	random	full	IND	2769	0
4	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F012_20 240108164747_0000_F_W_M EC_MVPMOPR04_0002.csv	2024-01-08 17:00:42 UTC	2024-01-08 17:04:48 UTC	OK	BWH	random	full	IND	8985	0
5	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F013_20 240108164747_0000_F_W_M EC_MVPMOPR04_0001.csv	2024-01-08 17:02:23 UTC	2024-01-08 17:04:27 UTC	OK	BWH	random	full	IND	245891	0
6	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F008_20 240108164747_0000_F_W_M EC_MVPMOPR04_0002.csv	2024-01-08 17:00:17 UTC	2024-01-08 17:06:37 UTC	OK	BWH	random	full	IND	354480	0
7	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F011_20 240108164747_0000_F_W_M EC_MVPMOPR04_0005.csv	2024-01-08 17:00:32 UTC	2024-01-08 17:05:05 UTC	OK	BWH	random	full	IND	2031	0
8	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F006_20 240108164747_0000_F_W_M EC_MVPMOPR04_0001.csv	2024-01-08 16:57:10 UTC	2024-01-08 17:07:38 UTC	OK	BWH	random	full	IND	233924	0
9	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F007_20 240108164747_0000_F_W_M EC_MVPMOPR04_0001.csv	2024-01-08 16:57:23 UTC	2024-01-08 17:06:56 UTC	OK	BWH	random	full	IND	6054	0
10	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F004_20 240108164747_0000_F_W_M EC_MVPMND04_0001.csv	2024-01-08 16:50:11 UTC	2024-01-08 17:09:49 UTC	OK	BWH	random	full	IND	186258	0
11	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F005_20 240108164747_0000_F_W_M EC_MVPMND04_0002.csv	2024-01-08 16:55:25 UTC	2024-01-08 17:09:12 UTC	OK	BWH	random	full	IND	633927	83
12	87480f9-9464-4739-90af-264	MEC_IT_0000_0000_F014_20 240108164747_0000_F_W_M	2024-01-08 17:00:50 UTC	2024-01-08 17:03:27 UTC	OK	BWH	random	full	IND	18592	0

15 tables are updated and roughly 1.8mln rows are inserted globally (this number may slightly change). No table is supposed to be empty.

## Assessment

Check the log tables in GCP on table log\_tables and run\_jobs to check that there is no error loading from source to staging/ods

Check the surrogate key must be unique in the data mart layer

## Scheduling

Is there an automatic schedule ? Yes

At what frequency ? We have two Talend plans in the TMC:

- PL\_MECANO\_DASH : Every monday at 4am (Paris time)
- PL\_MECANO\_DAILY\_LOAD : Every tuesday, wednesday, thursday, friday at 4 am (Paris time)

What is the trigger ? TMC

## Timing

The average time expected for :

- PL\_MECANO\_DASH : around 40 minutes
- PL\_MECANO\_DAILY\_LOAD : less than 5 minutes

## Criticality

Medium

## Logging

Table table log\_tables, run\_jobs, log\_files, and reject\_files in `prj-data-dm-industrial-[environment].STG.[table]`

**Link to the Mapping document :**

**Mecano KT Call - Questions to Matteo/Sebastien - 15Jan2024 :**

**Requesting more info/documentation on below:**

**Step 0 :**

Source Details/BW ServerName:  
Source repository/Connection Details:

**Step1 :**

1. List of total Xtract jobs and their frequency? - Please update the wiki.
2. Naming convention to identify 'Xtract jobs' related to Mecano?
3. Xtract jobs and (CSV) files location?
4. R001\_MECANO 0.1 Job and Date Calculation script .
5. Google Cloud Storage Bucket Name and location?

## cs-ew1-prj-data-dm-industrial-prod-staging

Understand more about BW Query: - (Sai/Prabakar)

TALEND\_QVMECANO\_BW\_QRY\_MVPMNO04\_0001  
TALEND\_QVMECANO\_BW\_QRY\_MVPMNO04\_0002  
TALEND\_QVMECANO\_BW\_QRY\_MVPMOP04\_0001  
TALEND\_QVMECANO\_BW\_QRY\_MVPMOP04\_0002

*prj-data-dm-industrial :*

Source GCP Buckets --> *prj-data-dm-industrial.STG.* --> *prj-data-dm-industrial.ODS* --> *prj-data-dm-industrial.DM* . --> *prj-data-dm-industrial.DS\_MaintenanceDashboard* |  
||  
V

Reporting will be done from PRJ-DATA-MAINTENANCE-DASH-DEV project on DPL dataset (*prj-data-pmo-dash-dev.DPL*).

**Tableau** --> *prj-data-maintenance-dash.DPL* --> *prj-data-maintenance-dash.DataOcean* --> *prj-data-maintenance-dash.DM* *prj-data-dm-industrial.DS\_MaintenanceDashboard*