

Data Engineer FAQ

I should use the dedicated project for the domain, in this case, the one for marketing. The list of Talend projects can be found [here](#).

If, at the time, the data architect does provide you with the domain name, you must ask for it. In order not to get blocked, you can use the common domain. Your job combines Talend jobs to comply with the Data Ocean architecture.

First, you must always store data files in a bucket, then the data from that file will be moved into Staging tables and then into the ODS tables.

The job show is a template you must use.

The Talend job J003_Extraction_till_ODS must be used, where you can change the data extraction component depending on your specific use case. You must use the GCP project of the desired domain you are working on. The full list can be found [here](#). The files must be named following the naming convention. The file names depend on the source from where information is extracted and the type of extraction.

You must use [this Google Sheet](#) to create your file name by providing information about the source. The file will also generate the name of the staging and ODS table you must create.

You should read [this page](#) to know more about the naming convention's details.

A list of reference jobs is available to connect to various sources (SAP, BW, databases...) and store the result in Google Cloud Storage.

Please read [this page](#) and find the most suitable job for you.

You should not start developing a custom job or change the ones provided, if none matches your need, raise your need in DE meetings of the chat forum.

You should use Terraform to build tables in the domain and data product project.

Please follow [these guidelines](#) to use Terraform and GitLab pipelines in your project.

If you are developing a data product, you should use (or create) a project under [this space](#).

We store the credentials to access our systems in a centralized Keepass file stored in this folder (\\ACEW1DTLNDENG02\\Keepass\\bda_tealend.kdbx). Check if there are available credentials for the source you are targeting.

If not, please reach out to the data architect of your project to understand how to connect and insert the new credentials in the Keepass.

You can go thru the decision tree [here](#)

1. SAP ECC (ERP Central Component) or ERP (Enterprise Resource Planning) such as PF1_020, WP1_400, and PI1_020 to keep the daily transaction from Financials (FI), Material Management (MM), Sales and Distribution (SD). It is the data source of BW
2. SAP BW (Business Warehouse), such as WBP, load data from various source, including SAP ECC, to generate report

The list of all SAP landscape is here

The logs are available in the remote engines, in the following folders:

- DEVELOPMENT : \\acew1dtlndeng02\\logs
- UAT : \\acew1tltlndeng01\\logs

Each execution generates a folder and the **folder names corresponds to the meta_run_id of the runs_jobs log table**.

The meta_source_system field identifies the source system of the data. You can find the standardized list by executing this query

```
SELECT * FROM `prj-data-dm-common-dev.DM.DIM_source_system`
```

Master data table are handled by the Data Hub project. For this, you must work on the DATA_HUB project.

The Data Hub project in GCP is prj-data-hub-dev