Purchasing Dashboard 4. Data Loading with Talend

4.1 - Talend Integration

Source data integration with Talend ETL tool

- **Xtract Server (WBP):**
  - Talend connects to the Xtract server (*ACEW1PQLIKAPP02*) trigger extraction job using HTTP calls.
  - Xtract generates results as CSV files stored in memory.
  - Talend retrieves files, apply some transformation and then loads them into Google Cloud Storage.
- **Google Sheets (Hedges, wap solid fuels, CO2):**
  - Talend integrates with Google Sheets to extract some static files that will be use to apply mappings or lookup to BW tables.
  - It retrieves this data from Google Sheets.

Data Transformation and Loading to Google BigQuery:

- Once data from all sources is available in Google Cloud Storage as files, Talend proceeds with data transformation and loading.
- Talend performs data transformations as needed, including cleansing, mapping, and structuring the data for consistency.
- The transformed data is loaded into various stages, operational data stores (ODS), and data mart tables within Google BigQuery.
- These tables are organized to facilitate efficient querying and reporting for energy optimization purposes.

By utilizing Talend for data extraction, transformation, and loading (ETL), the web app ensures that data from diverse sources is collected, processed, and structured for analysis and reporting within Google BigQuery, enabling users to make informed decisions based on up-to-date and accurate data.

4.2 - Source Data Extraction

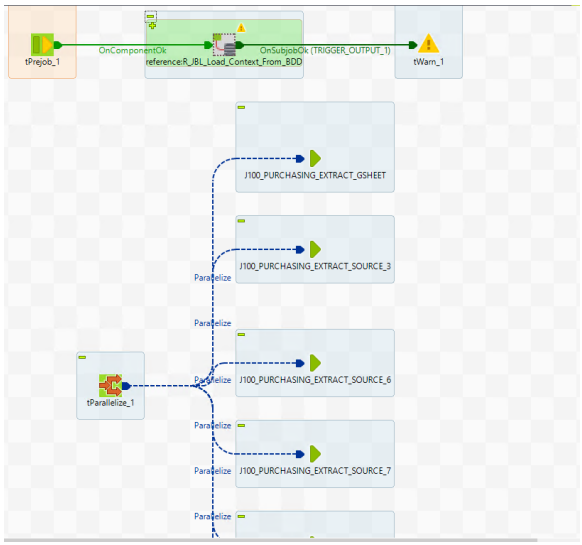
Responsible & contact points:

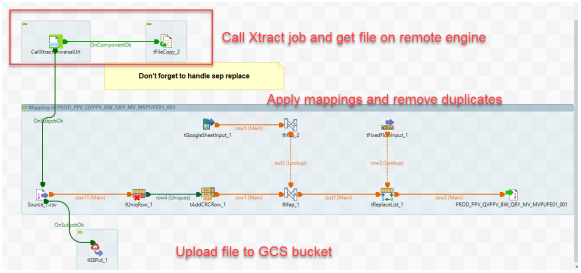
- Application Owner - Sylvie Severini
- Delivery Manager - Donia Rachdi
- Project Manager - Vitaly Alt ukhov
- Data Architect - Joao Fonseca
- Tableau Dev - Maxime Marboeuf
- Data Engineer - Virgil Lissassi; replaced by Matteo Menghetti

Main jobs for source extraction	• **F100_PUR_DATA_EXTRACT** • **J100_PURCHASING_EXTRACT_SOURCE_3** • **J100_PURCHASING_EXTRACT_SOURCE_6** • **J100_PURCHASING_EXTRACT_SOURCE_7** • **J100_PURCHASING_EXTRACT_SOURCE_11** • **J100_PURCHASING_EXTRACT_SOURCE_13**	**--to the top --**
	Job description by steps	**Job design**

1. In the job F100_PUR_DATA_EXTRACT:
   a. A parallel execution of all data sources is launched

We will analyze in detail on job but the logic in the other is simimar

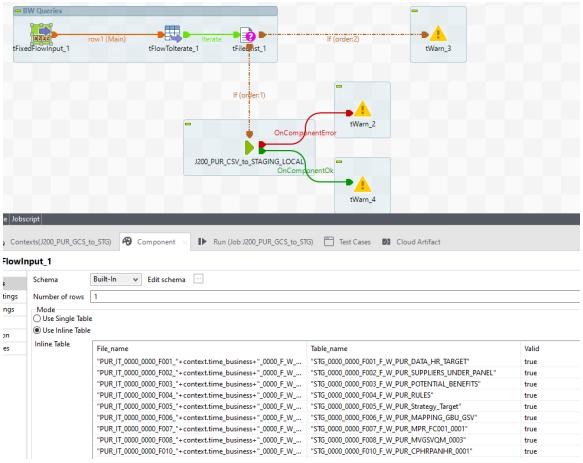


	1. The jobs triggers the extract universal jobs. In some cases we trigger loop over a list of parameters and trigger multiple extractions 2. We read the CSV file obtained by extract and we : a. Remove duplicates b. Apply mappings as indicated in the specifications and create a file ready to be uploaded to stagin 3. We upload the file into the bucket.	

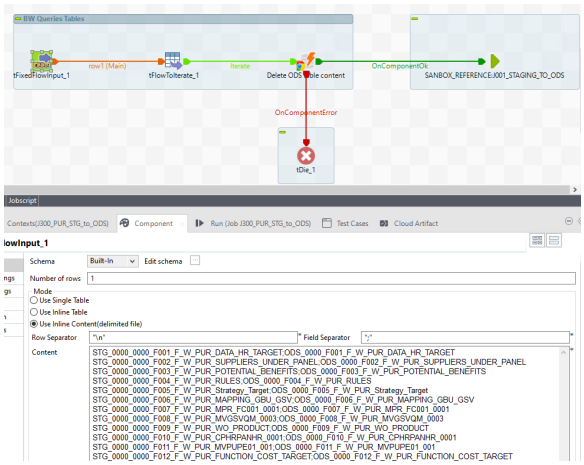
Main jobs for loading Staging	• **F200_PUR_GCS_TO_STG** • **J200_PUR_GCS_TO_STG**	**--to the top --**
	Job description by steps	**Job design**
	1. The flow is used to call the job *J200_PUR_GCS_TO_STG*:	

2. Looping over the files:

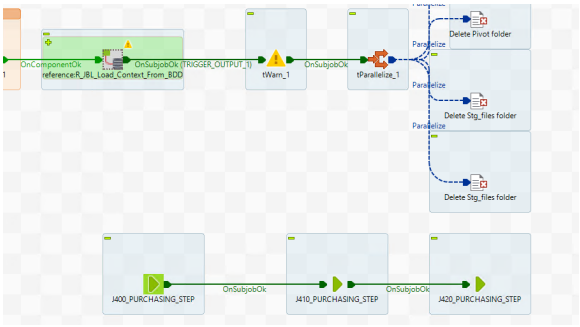
    a. The previous step generated a list of files in the *Stg_file* folder. The job iterates over them and for each file it calls the job J200_PUR_CSV_TO_STAGING_LOCAL.



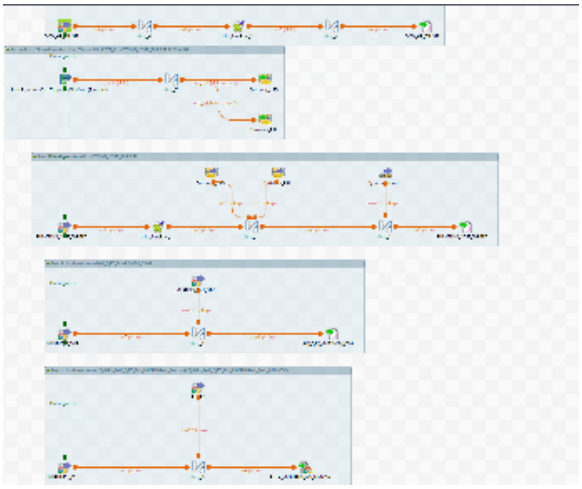
i. The Fixed Flow Input component contains the list of all files with the corresponding staging table name.

Main jobs for loading the ODS	• **F300_PUR_STG_TO_ODS** • **J300_PUR_STG_TO_ODS**	**--to the top --**
	Job description by steps	**Job design**
	1. The flow is used to call the job *J300_PUR_STG_TO_ODS*: 2. Looping over the list of staging tables: a. Truncate the corresponding ODS table b. Call the job J001_STAGING_TO_ODS to insert data in the table	

Main jobs for building DM	• **F400_PUR_PREP_DATA** ○ **F400_PURCHASING_STEP** ○ **F410_PURCHASING_STEP** ○ **F420_PURCHASING_STEP**	**--to the top --**
	Job description by steps	**Job design**

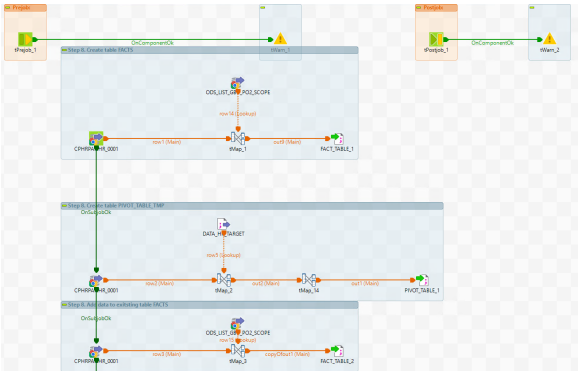
	1. This Talend flow first delete files that remain in the remote engine from previous executions 2. It call in sequesce 3 subjobs which read the different ODS tables needed to build the dimension and pivot table.	

Prepare DM file	**F400_PURCHASING_STEP**	**--to the top --**
	Job description by steps	**Job design**

	1. Reqd data from DATA_HR_TARGET: • The job reads data from the ODS table ODS_0000_F001_F_W_PUR_DATA_HR_TARGET and apply the necessqry mapping 2. Calculation of rate cost: • The job reads this Googel Sheet is order to extact the various currency rate per year and saves them in memory. 3. Compute Function cost target: • The job read data from the ODS table ODS_0000_F012_F_W_PUR_FUNCTION_COST_TARGET and apply mapping to build the FUNCTION_COST_TARGET This job builds the first pieces of data necessary to build the FUNCTION cost target	

Main jobs for source extraction	• **F410_PURCHASING_STEP**	**--to the top --**
	Job description by steps	**Job design**
	The job reads in sequence a list of ODS tables and apply the requested mappings	

1. ODS_0000_F010_F_W_PUR_CPHRPANHR_0001
    a. The table is read twice
        i. Reads ODS data applying filter *Bus__Unit__Ent__Grp__Key='FPUR'* and getting only the last



two year of data.

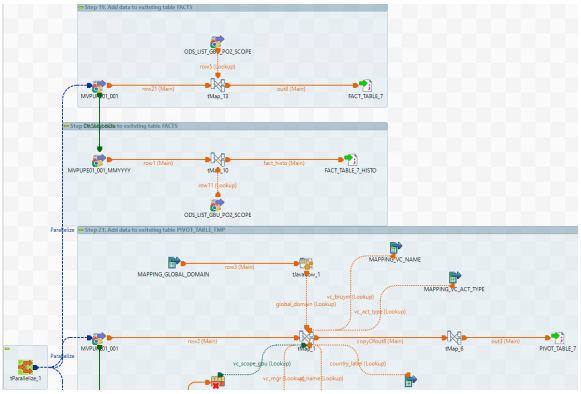
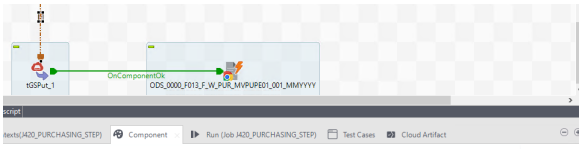
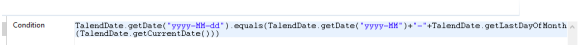
ii. Reads ODS data applying filter *SPLIT(Job_Function,'/')[OFFSET(0)]in('PROC','SC')* and

getting only the last two years of data.

2. ODS_0000_F007_F_W_PUR_MPR_FC001_0001:
    a. We read the entire table and apply the requested mappings
3. FUNCTION_COST_TARGET
    a. The file created in job J400_PURCHASING_STEP is parsed to adapt it to the dimension and fact schema

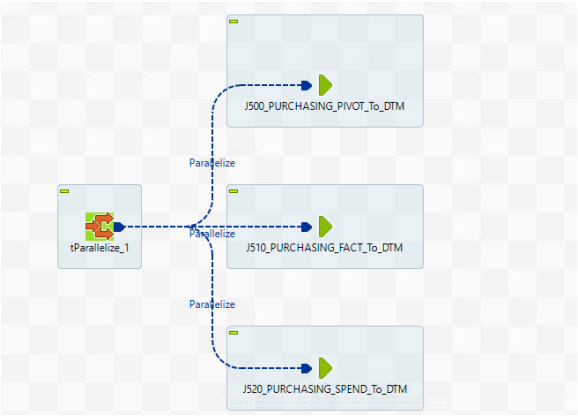
4. ODS_0000_F009_F_W_PUR_WO_PRODUCT
    a. We read the table applying the following filters on Creditor_Class__P__Key *not like ('% S%' OR '% X%')* and getting only the last two year of data.
5. ODS_0000_F008_F_W_PUR_MVGSVQM_0003:
    a. We read the table applying the following filters on excluding when *Vendor_Domain_Key contains 9* and getting only the last two year of data.

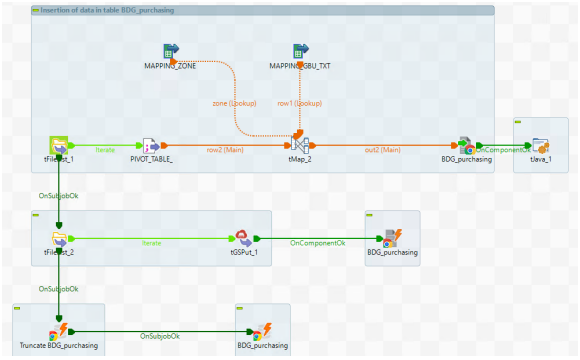
	This Talend job is the second step of building Dimension and Fact table. All data in FACT tables are also assigned to ECO or SCO.	

Main jobs for source extraction	• **F420_PURCHASING_STEP**	**--to the top --**
	Job description by steps	**Job design**
	The job reads in sequence a list of ODS tables and apply the requested mappings 1. ODS_0000_F011_F_W_PUR_MVPUPE01_001 a. We read the table applying the following filters on *VC_PROGRAM = 'Procurement'* and getting only the last two year of data.	  

	2. ODS_0000_F005_F_W_PUR_Strategy_Target a. We read the table applying the following filters for getting only the last two year of data. and where TARGET_EUR is not empty. 3. At the end of step there is a condition **that verifies if the current date is the last day of the month ,** if true it appends the content of the table ODS_0000_F013_F_W_PUR_MVPUPE01_001 to the table ODS_0000_F013_F_W_PUR_MVPUPE01_001_MMYYYY		

4.4 - Load to DM (calculations and transformations)

Main jobs for source extraction	• **F500_PUR_DATA_TO_DTM**	--to the top --
	Job description by steps	**Job design**
	This jobs reads the files stored in remote engine and computed in the previous and use them to populate dimensions and fact tables.	

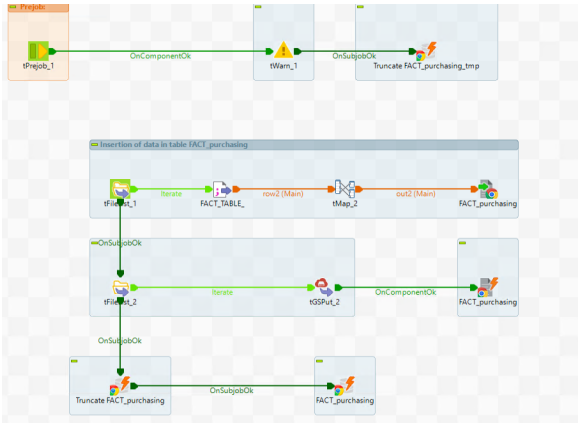
Main jobs for source extraction	• **J500_PURCHASING_PIVOT_TO_DTM**	--to the top --
	Job description by steps	**Job design**
	1. The job starts by truncating the table *BDG_purchasing_tmp*	

2. Read Pivot files:
   - The job iterqates over all files of the *Pivot* folder (whose name starts with *PIVOT_TABLE_*).
   - Each file is processed and a look up is done with the two following Google Sheets to get a uniform GBU text.
   - An BigQuery Bulk Output file is created

	3. Upload data in the *BDG_purchasing_tmp* table: a. The job upload all bulk files into Cloud Storage and each file is appended into the *BDG_purchasing_tmp* table 4. The job truncates the table *BDG_purchasing*. 5. . The job inserts copies data into from the *BDG_purchasing_tmp* table. During the copy the columns *mont_year* and *vc_year* are converted into string.	

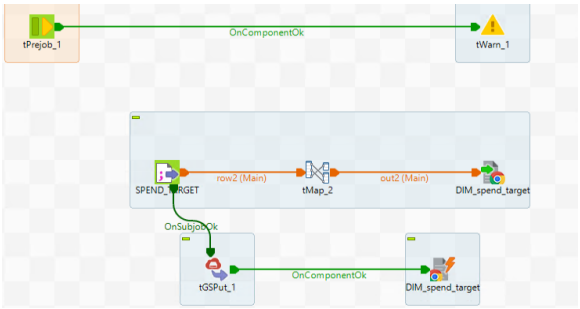
Main jobs for source extraction	• **J510_PURCHASING_FACT_TO_DTM**	**--to the top --**
	Job description by steps	**Job design**
	1. The job starts by truncating the table *FACT_purchasing_tmp*	

2. Read Pivot files:
   - The job iterqates over all files of the *Pivot* folder (whose name starts with *FACT_TABLE_*).
   - Each file is processed to have the desired schema of the fact table
   - An BigQuery Bulk Output file is created



	3. Upload data in the FACT_*purchasing_tmp* table: a. The job upload all bulk files into Cloud Storage and each file is appended into the *FACT_purchasing_tmp* table 4. The job truncates the table *FACT_purchasing*. 5. . The job inserts copies data into from the FACT_*purchasing_tmp* table	

Main jobs for source extraction	• **J520_PURCHASING_SPEND_TO_DTM**	**--to the top --**
	Job description by steps	**Job design**

	1. Read SPEND_TARGET_1 file: • The job reads the file SPEND_TARGET_1.csv in the folder *Spend_Targe* • A BigQuery Bulk Output file is created 2. Upload the file bulk file in Google Cloud Storage 3. The job uses the bulk file to overwrite the values in the table *DIM_spend_target*.	

4.5 - Scheduling and Automation

TMC - PL_QV_TO_TABLEAU_PURCHASING - Daily at 5 AM.

4.6 - Data Validation