GSC 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 and 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.

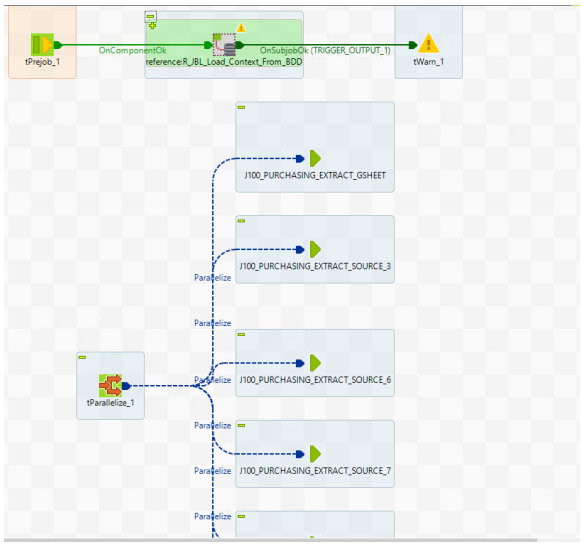
By utilizing Talend for data extraction, transformation, and loading (ETL), the web app ensures that data from diverse sources is collected, processed, and structured for analysis and reporting within Google BigQuery, enabling users to make informed decisions based on up-to-date and accurate data.

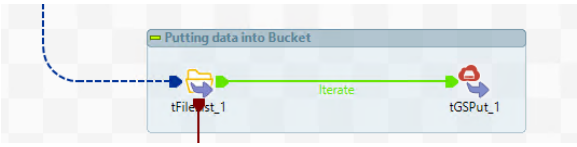
4.2 - Source Data Extraction

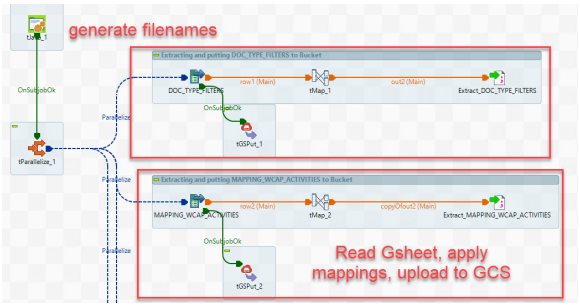
Main jobs for source extraction	• **F100_GSC_DATA_EXTRACT** • **J130_mainExtract_Files_Src_to_GCS** • **J120_GSC_mainExtract_BWQ_Data** • **J120_GSC_Extract_BW_data_Source_14_EXT** • **J110_GSC_mainExtract_Gsheet_data**	**--to the top --**
	Job description by steps	**Job design**

Responsible & contact points:

- Application Owner - Mathieu Pourqué
- Delivery Manager - Donia Rachidi
- Project Manager - Vitaly
- Data Architect - Joao Fonseca
- Tableau Dev - Maxime Marboeuf
- Data Engineer - Virgil Lissassi ; replaced by Matteo Menghetti

	1. In the job F100_PUR_DATA_EXTRACT: a. A parallel execution of all data sources is launched The flow triggers the extraction of all sources needed for the Global Supply chain dashboard.	

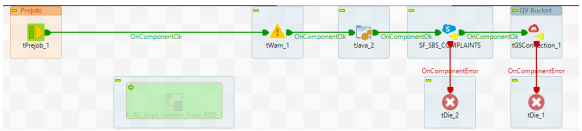
Main jobs for source extraction	• **J110_GSC_mainExtract_Gsheet_data**	**--to the top --**
	Job description by steps	**Job design**
	1. Parallel extraction of all Google Sheet sources: a. A parallel extraction of all Google Sheet is launched b. Simple transformation are done to align with specifications 2. Upload files to GCS a. The files are uploaded into the *GSC* folder of the target bucket	

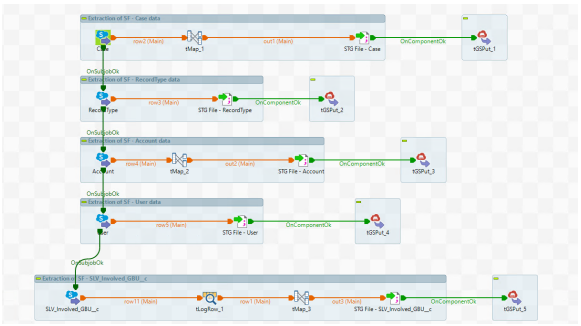
Main jobs for source extraction	• **J130_GSC_mainExtract_Files_Src_to_GCS**	--to the top --
	Job description by steps	**Job design**
	1. Filename creation: a. A specific filename following our standard naming convention is created to properly upload files into Cloud Storage	

	2. Parallel GCS upload a. Each Google Sheet table is read b. Simple transformation are done to align with specifications resulting in a CSV stored in the *Tmp folder* c. Each file is uploaded into the *GSC* folder of the target bucket	

Main jobs for source extraction	• **J120_GSC_Extract_BW_data_Source_14_EXT**	**--to the top --**
	Job description by steps	**Job design**

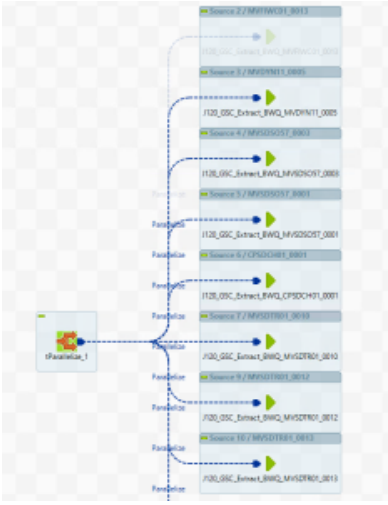
1. Login to Salesforce:
    a. The job starts by establishing a connection to Salesforce using the account sbs-is-appli-qlikview.support@solvay.com.crm



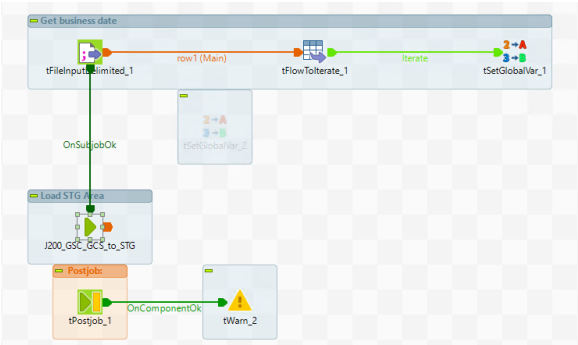


	2. Extraction of the Salesforce objects a. The list of Salesforce objects is extracted in sequence b. Depending on the source some mappings can be applied c. Each source is uploaded to GCS as a CSV file	

Main jobs for source extraction	• **J120_GSC_mainExtract_BWQ_Data_Source**	**--to the top --**
	Job description by steps	**Job design**

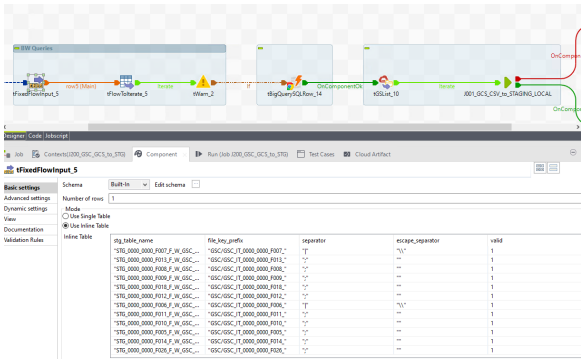
	1. Parellel execution of Xtract jobs: a. The job triggers in parallel all the jobs needed to extract data from BW.	

4.2 - Load Staging and ODS

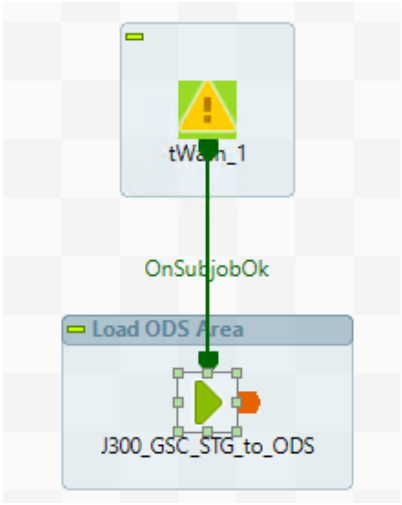
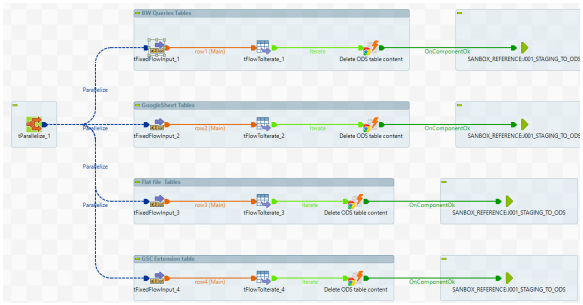
Main jobs for loading Staging	• **F200_GSC_Gcs_To_Stg** • **J200_GSC_Gcs_To_Stg**	**--to the top --**
	Job description by steps	**Job design**
	1. First the job retrieves the value of the Business date from the file business_date.csv file. This is to target only the files of the current execution in GCS. 2. The flow is used to call the job *J200_GSC_Gcs_To_Stg.*	

- **J200_GSC_Gcs_To_Stg**

1. The job launches a parallel load per source type into the Staging tables



	a. A list of the possible CSV file with the associated Staging table name, and separator is used to create a sequence b. Each staging table is truncated before a new upload c. The job gets all the files starting with the desired prefix and for each of them calls the standard job J001_GCS_to_STAGING_LOCAL	

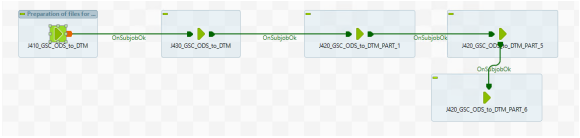
Main jobs for loading the ODS	• **F300_GSC_STG_TO_ODS** • **J300_GSC_STG_TO_ODS**	**--to the top --**
	Job description by steps	**Job design**
	1. The flow is used to call the job *J300_GSC_STG_TO_ODS*	
		 STG_Table_name / ODS_Table_name "STG_0000_0000_F005_F_W_GSC_Mvfiwc01_0013" / "ODS_0000_F005_F_W_GSC_Mvfiwc01_0013" "STG_0000_0000_F006_F_W_GSC_Mvdyn11_0005" / "ODS_0000_F006_F_W_GSC_Mvdyn11_0005" "STG_0000_0000_F007_F_W_GSC_Mvsdtr01_0010" / "ODS_0000_F007_F_W_GSC_Mvsdtr01_0010" "STG_0000_0000_F008_F_W_GSC_Mvsdtr01_0012" / "ODS_0000_F008_F_W_GSC_Mvsdtr01_0012" "STG_0000_0000_F009_F_W_GSC_Mvsdtr01_0013" / "ODS_0000_F009_F_W_GSC_Mvsdtr01_0013" "STG_0000_0000_F010_F_W_GSC_Cpsdso51_0001" / "ODS_0000_F010_F_W_GSC_Cpsdso51_0001" "STG_0000_0000_F011_F_W_GSC_Mvsdso57_0003" / "ODS_0000_F011_F_W_GSC_Mvsdso57_0003" "STG_0000_0000_F012_F_W_GSC_Mvsdso57_0001" / "ODS_0000_F012_F_W_GSC_Mvsdso57_0001" "STG_0000_0000_F013_F_W_GSC_Cpsdch01_0001" / "ODS_0000_F013_F_W_GSC_Cpsdch01_0001" "STG_0000_0000_F014_F_W_GSC_Mvfiwc01_0014" / "ODS_0000_F014_F_W_GSC_Mvfiwc01_0014"

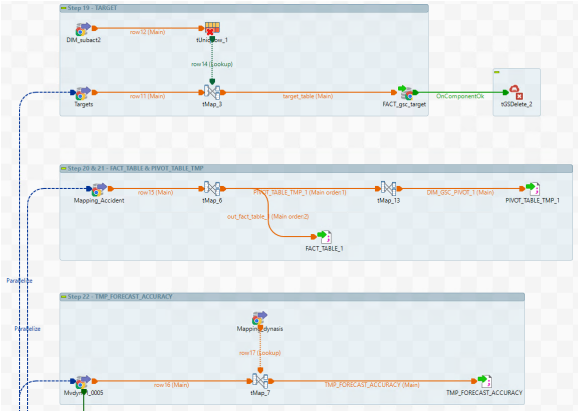
1. Parallel loading of the ODS table by source type
    a. A parallel loading of ODS table is done by source type (BW, Google Sheet, Salesforce)
2. Load in sequence all ODS tables of a source
    a. Given a source type the list of staging and ODS tables is hard coded.
    b. ODS table is truncated
    c. Th standard job J001_STAGING_TO_ODS is called to load the ODS table

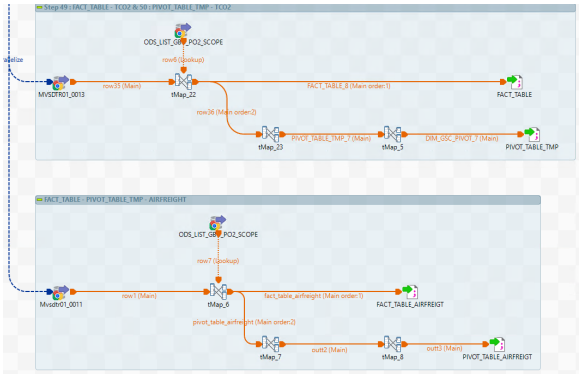
	Given the fact that staging and ods are identical and tables are always truncated, this step is unnecessary.	

4.2 - Prepare DM files

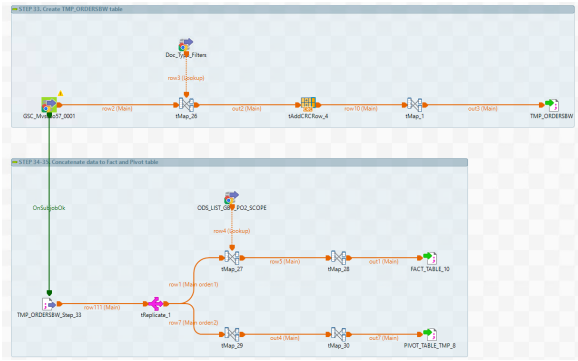
Main jobs for building DM	• **F400_GSC_ODS_to_Prep_DM_Files** **F400_PURCHASING_STEP** ◦ **F410_GSC_ODS_to_DTM** ◦ **F420_GSC_ODS_to_DTM_Part_1** ◦ **F420_GSC_ODS_to_DTM_Part_5** ◦ **F420_GSC_ODS_to_DTM_Part_6** ◦ **F430_GSC_ODS_to_DTM**	**--to the top --**
	Job description by steps	**Job design**

	1. This job is used to call a series of subjob that creates files needed to create the dimensions and fact tables	

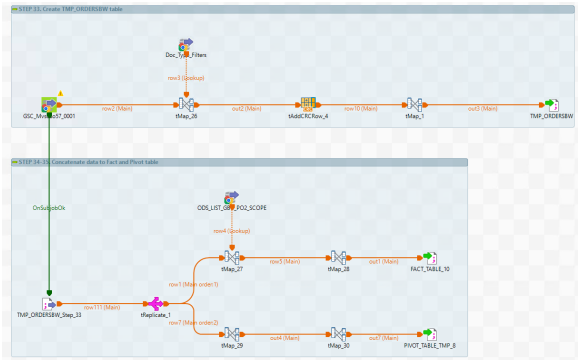
Prepare DM file	• **F410_GCS_ODS_to_DTM**	**--to the top --**
	Job description by steps	**Job design**
	The job computes steps 19-28 of the specification document 1. The job reads data from the ODS table and apply the necessary mappings; 2. The job output PIVOT and FACT files which are stored on the remote engine and will be uploaded to to the DM in the J500 jobs. 3. The jobs creates the table **FACT_gsc_target**	

Main jobs for source extraction	• **F430_GCS_ODS_to_DTM**	**--to the top --**
	Job description by steps	**Job design**
	The job computes steps 41-49 of the specification document 1. The job reads data from the ODS table and apply the necessary mappings; 2. The job output PIVOT and FACT files which are stored on the remote engine and will be uploaded to to the DM in the J500 jobs. 3. The job also builds the files necessary to create the airfreight KPI following these specifications	

Main jobs for source extraction	• **F420_GSC_ODS_to_DTM_PART_1**	**--to the top --**
	Job description by steps	**Job design**

	The job computes steps 33,34, and 35-49 of the specification document 1. The job reads data from the ODS table and apply the necessary mappings; 2. The job output PIVOT and FACT files which are stored on the remote engine and will be uploaded to to the DM in the J500 jobs.	

Main jobs for source extraction	• **F420_GSC_ODS_to_DTM_PART_5**	--to the top --
	Job description by steps	**Job design**

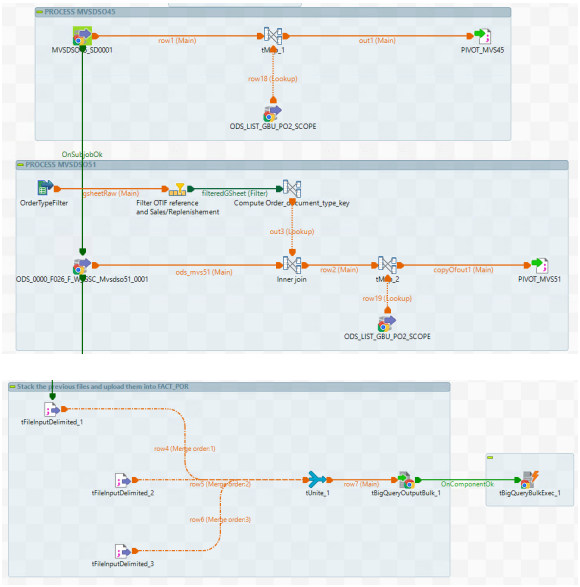
	The job computes steps 33,34, and 60-69 of the specification document 1. The job reads data from the ODS table and apply the necessary mappings; 2. The job output PIVOT and FACT files which are stored on the remote engine and will be uploaded to to the DM in the J500 jobs. The job uses the tUnpivotRow component (you can find it here).	

4.3 - Compute Perfect Order Rate

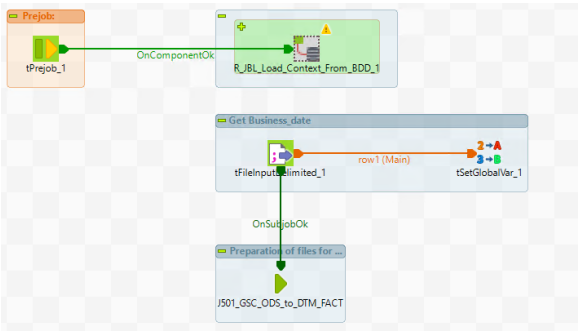
Main jobs for source extraction	• **F420_GSC_ODS_to_DTM_PART_6**	**--to the top --**
	Job description by steps	**Job design**

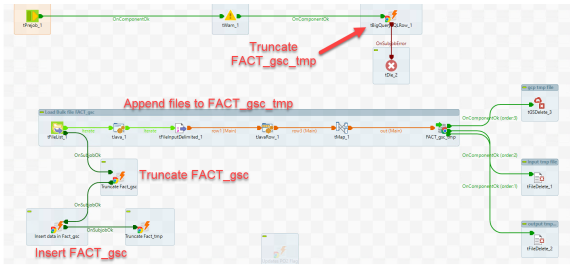
The job creates the table *FACT_POR* which is displayed by the Perfect Order rate dashbaord.

1. The job reads data from the ODS table and apply the necessary mappings. The main ODS tables are:
   a. ODS_0000_F025_F_W_GSC_MVSDSO45_SD0001
   b. ODS_0000_F026_F_W_GSC_Mvsdso51_0001
   c. ODS_0000_F020_F_W_GSC_Sf_Complaint_Case
2. Each step outputs a FACT file which is stored in the *Tmp* folder
3. The files are stacked to create a unique output bulk file which is used to create the FACT_POR table.

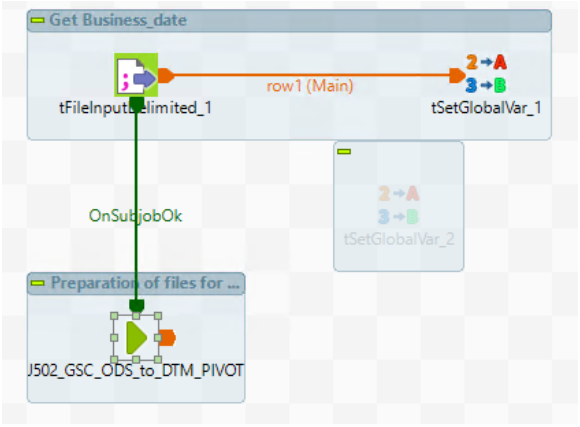
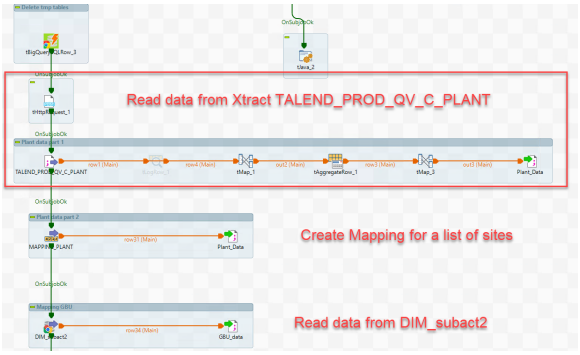
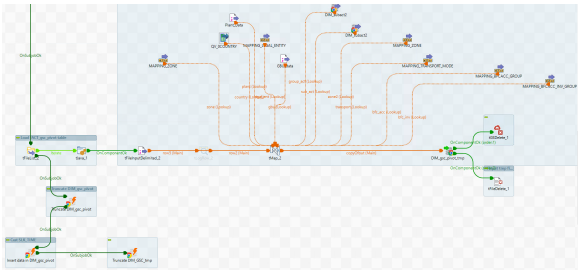


4.4 - Load to DM (calculations and transformations)

Main jobs for source extraction	• **F501_GSC_Files_To_FACT** • **501_GSC_ODS_to_DTM_FACT**	**--to the top --**
	Job description by steps	**Job design**
	This jobs is used to call the job J501_GSC_ODS_to_DTM_FACT	

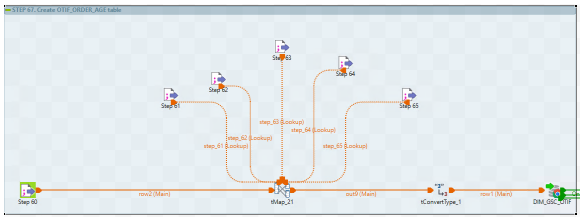
501_GSC_ODS_to_DTM_FACT	1. First the job truncates the table *DM. FACT_gsc_tmp* 2. The job reads all files whose name starts with FACT_FILE inside the InOut folder and uploads them into the FACT_gsc_tmp table. If files is successfully inserted the local file is deleted from the remote engine 3. The job truncates the table *DM. FACT_gsc* 4. The job copies data from *FACT_gsc_tmp* into *FACT_gsc* 5. The job truncates the table *DM. FACT_gsc_tmp*	

Main jobs for source extraction	• **F502_GSC_Files_To_PIVOT** • **J502_GSC_ODS_to_DTM_PIVOT**	**--to the top --**
	Job description by steps	**Job design**

	1. This jobs is used to call the job J502_GSC_ODS_to_DTM_PIVOT	
• **J502_GSC_ODS_to_DTM_PIVOT**	1. First the job truncates the table *DM.DIM_gsc_pivot_tmp* 2. The job gets data used to perform lookups **a.** The job executes the TALNED_PROD_QV_C_PLANT Xtract job **b.** Creates mapping for a list of sites **c.** Read data from DIM_subact2 **d.** Creates a mapping for some old GBU codes	 

	3. The job reads all files whose name starts with *PIVOT_* and *TPIVOT_* inside the InOut folder and uploads them into the *DIM_gsc_tmp* table. If files is successfully inserted the local file is deleted from the remote engine. Some joins are done using the sources previously extracted 4. The job truncates the table *DM.DIM_gsc_pivot* 5. The job copies data from *DIM_gsc_pivot_tmp* into *DIM_gsc_pivot* 6. The job truncates the table *DM.DIM_gsc_pivot_tmp*	

Main jobs for source extraction	• **F503_GSC_Files_To_OTIF** • **J503_GSC_ODS_**	**--to the top --**
	Job description by steps	**Job design**

	1. The flow is used to build the table *DIM_gsc_otif* 2. The job reads data from the Step_60.csv file computed in the job and joins it with the fiels Step_{61..65}. These are the output of job J420_GSC_ODS_to_DTM_PART_5 3. The combined file is used to overwrite the content of the table *DIM_gsc_otif*	

4.5 - Scheduling and Automation

TMC - PL_QV_TO_TABLEAU_GSC - Daily at 9 PM.

4.6 - Data Validation