# DW Project, Block 2: ETL

## Design an ETL flow in Talend Open Studio

# ETL

Prerequisites

- Connect to **PostgreSQL** using DBeaver
  - You will find two databases: AMOS and AIMS
  - Understand the domain, explore the data
    - Data are not perfect!!! Make your own assumptions!
  - 2 additional CSV files
    - maintenance airport, aircraft model and manufacturer
- Use a provided **multidimensional model**
  - 2 independent Star schemas with conformed dimension
  - A logical database schema (i.e., a set of CREATE TABLE statements) corresponding to that multidimensional schema
    - Create the tables in **Oracle**
- Tutorial: **ETL Design** using Talend Open Studio
  - Follow the instructions in the tutorial to get familiar with TOS

# ETL (Statement)

LAB ON EXTRACT-TRANSFORM-LOAD PROCESS DESIGN FOR THE ACME-FLYING USE CASE

You must create an Extract-Transform-Load (ETL) process that executes in order to ***extract*** data from the AIMS and AMOS operational databases and additionally provided data sources, ***transform*** these data to conform to the star schemas previously defined in the lab on Data Warehouse design, and ***load*** the data into the created star schemas. In addition, you need to improve the quality of the designed ETL process.

The designed ETL process should adhere to the following instructions:

*Extraction.*

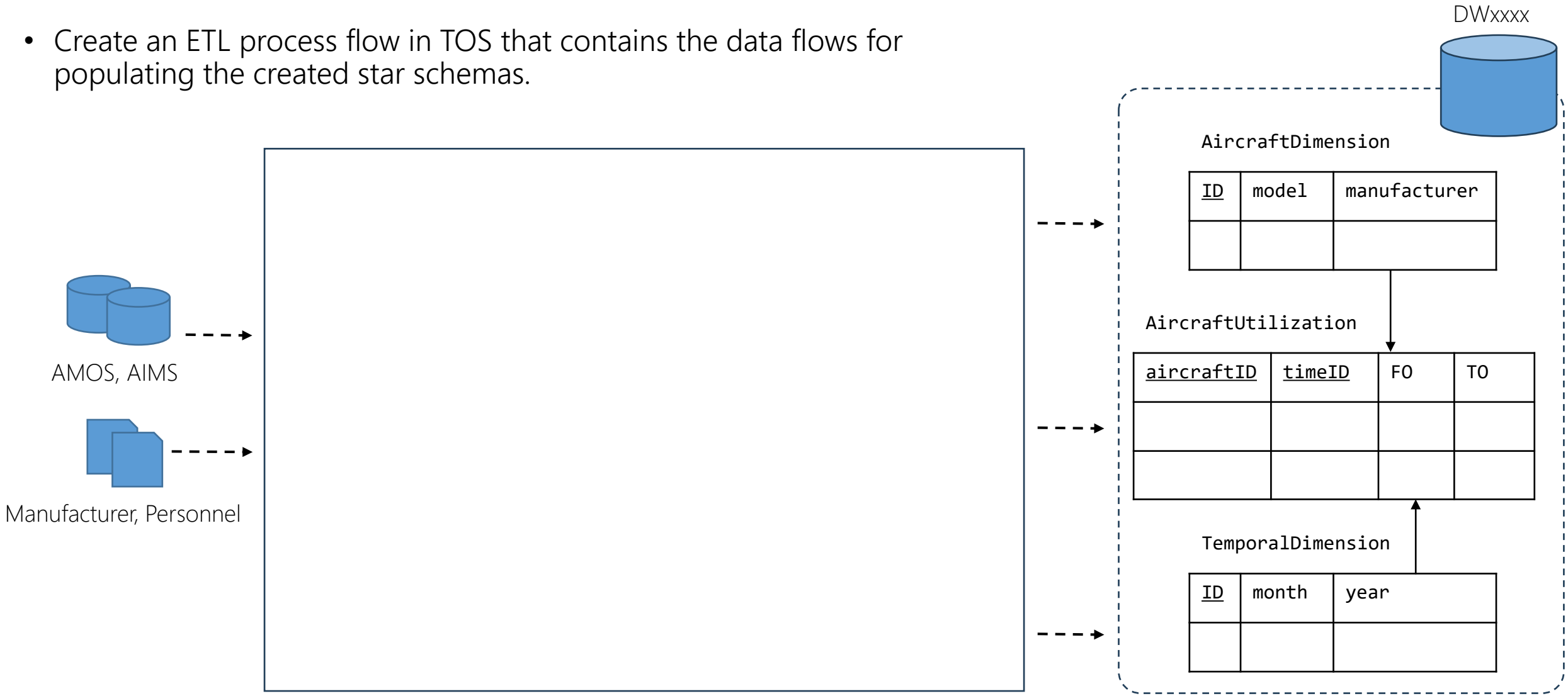
- Connect to the operational databases AIMS and AMOS for extracting the base operational data.
- In addition, you should use an additional data sources (*aircraft-manufaturerinfo-lookup.csv* and *maintenance-personnel-airport-lookup.csv*).

*Transformation.*

- Integrate data coming from AIMS and AMOS data sources. You should consider integrating these two sources having in mind the two common attributes that they share, i.e., *flightID (*in tables AIMS -> *Flights* and AMOS -> *OperationInterruption), and aircraftRegistration (*in tables AIMS -> *Slots and* AMOS -> *MaintenanceEvents, WorkOrders*).
- Complement the operational data coming from AIMS and AMOS by means of performing a lookup to the external data sources about:
    - **Aircraft manufacturer information** (aircraft-manufacturerinfo-lookup.csv) such that with each aircraft registration code, your ETL also provides its manufacturer registration code, the aircraft model and manufacturer.

# ETL data flow

- Create an ETL process flow in TOS that contains the data flows for populating the created star schemas.

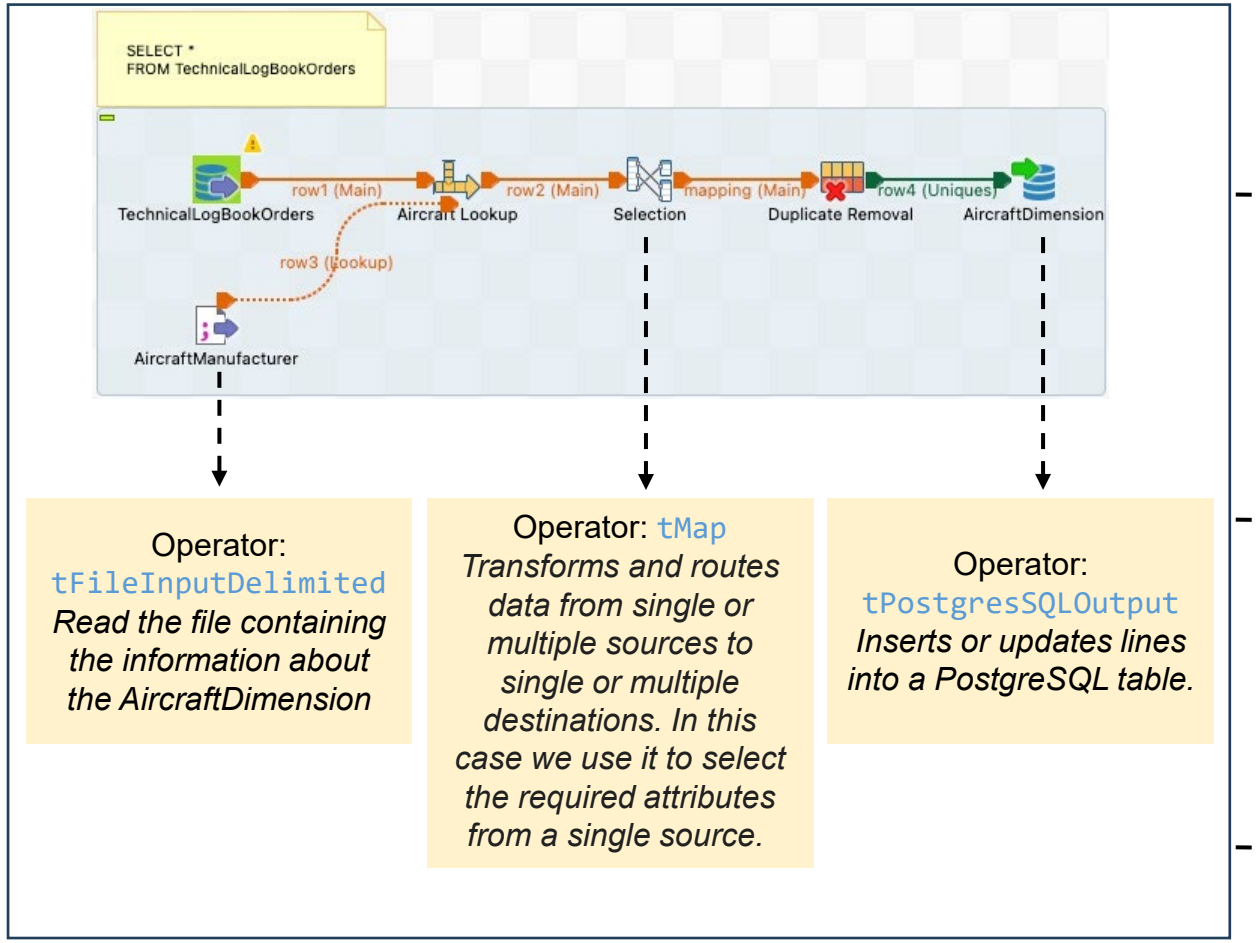


# ETL data flow

- Create an ETL process flow in TOS that contains the data flows for populating the created star schemas.


AMOS, AIMS

Manufacturer, Personnel




Operator: `tFileInputDelimited`
*Read the file containing the information about the AircraftDimension*

Operator: `tMap`
*Transforms and routes data from single or multiple sources to single or multiple destinations. In this case we use it to select the required attributes from a single source.*

Operator: `tPostgresSQLOutput`
*Inserts or updates lines into a PostgreSQL table.*

DWxxxx

AircraftDimension

| ID | model | manufacturer |
|---|---|---|
| | | |

AircraftUtilization

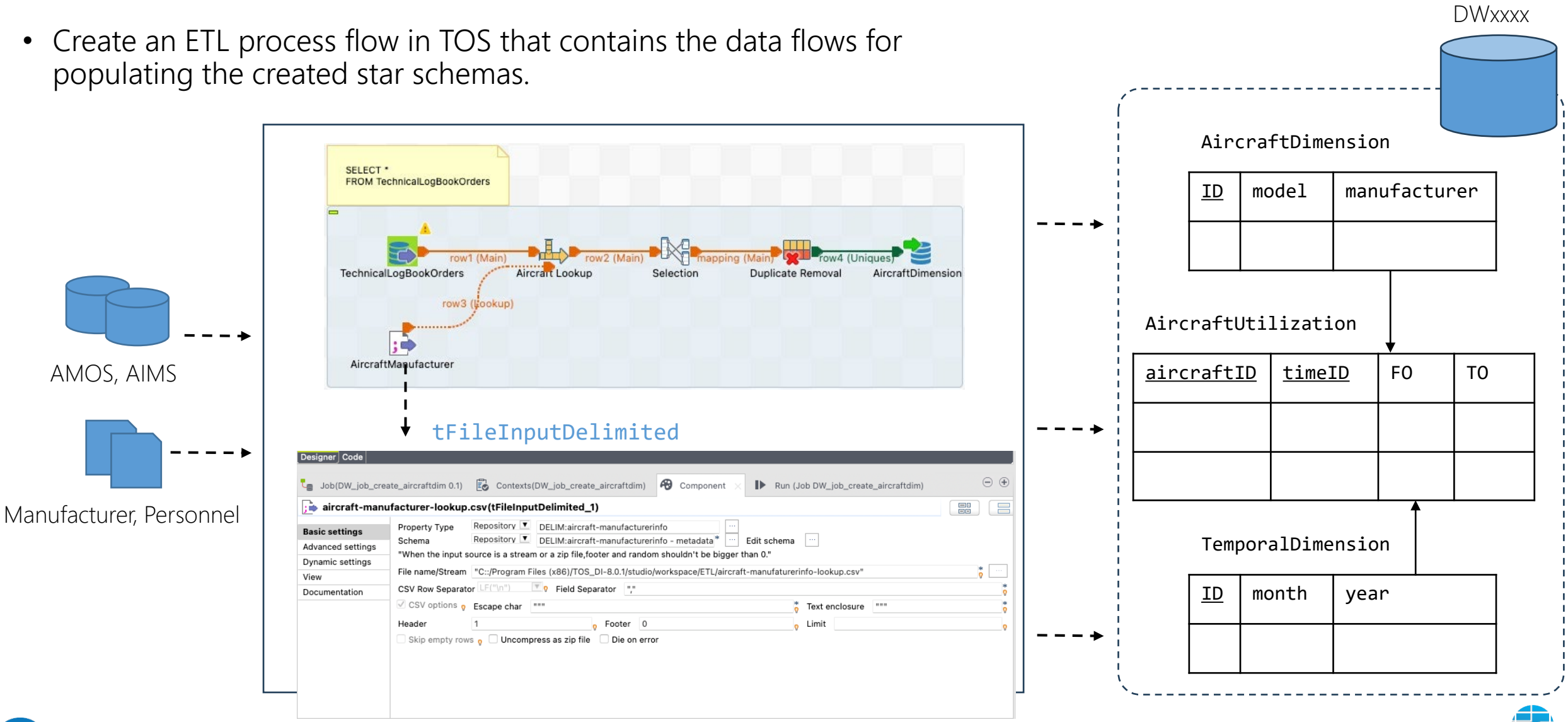
| aircraftID | timeID | FO | TO |
|---|---|---|---|
| | | | |
| | | | |

TemporalDimension

| ID | month | year |
|---|---|---|
| | | |

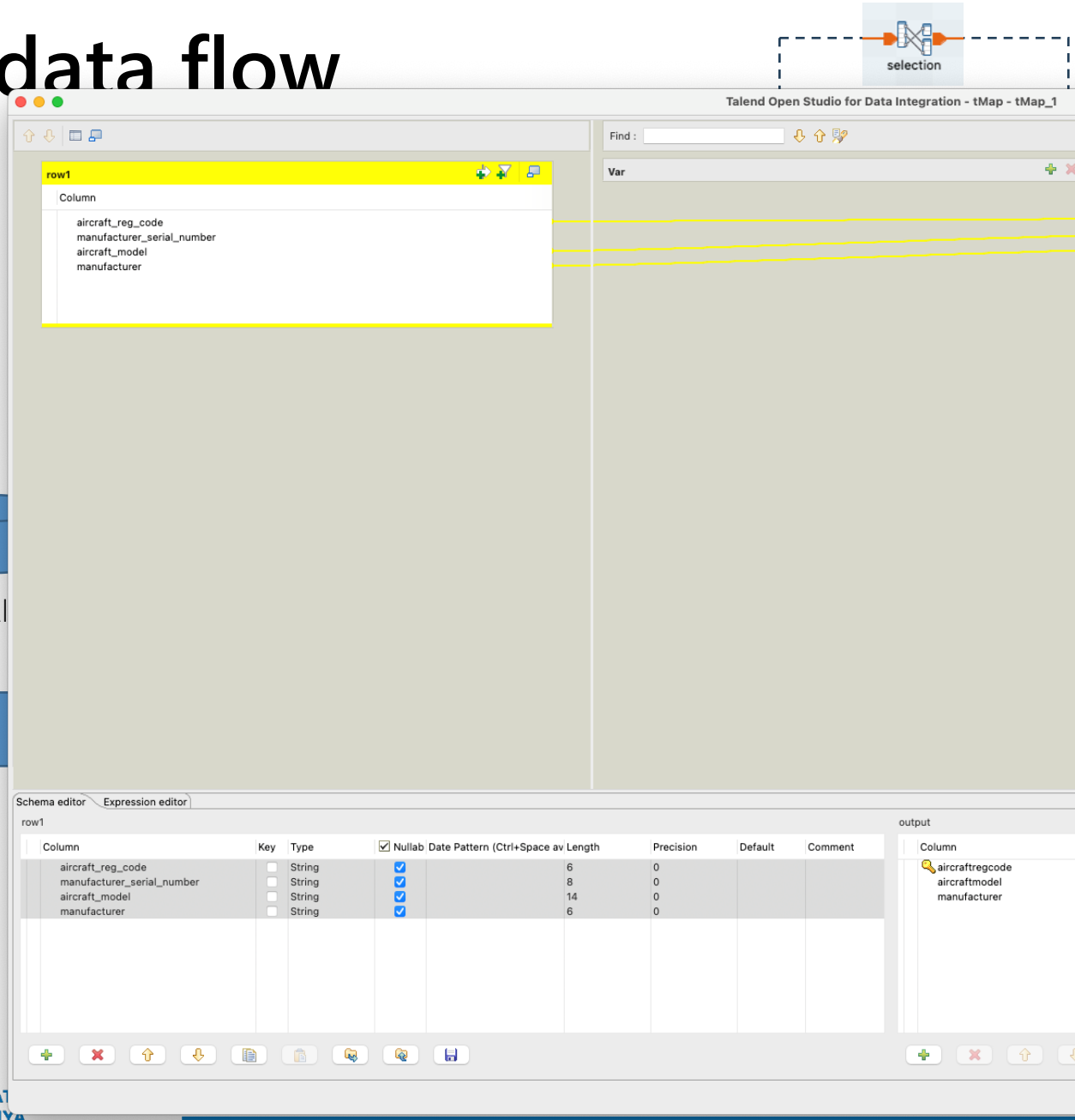
# ETL data flow

- Create an ETL process flow in TOS that contains the data flows for populating the created star schemas.



# ETL data flow

- Create




AMOS, AI

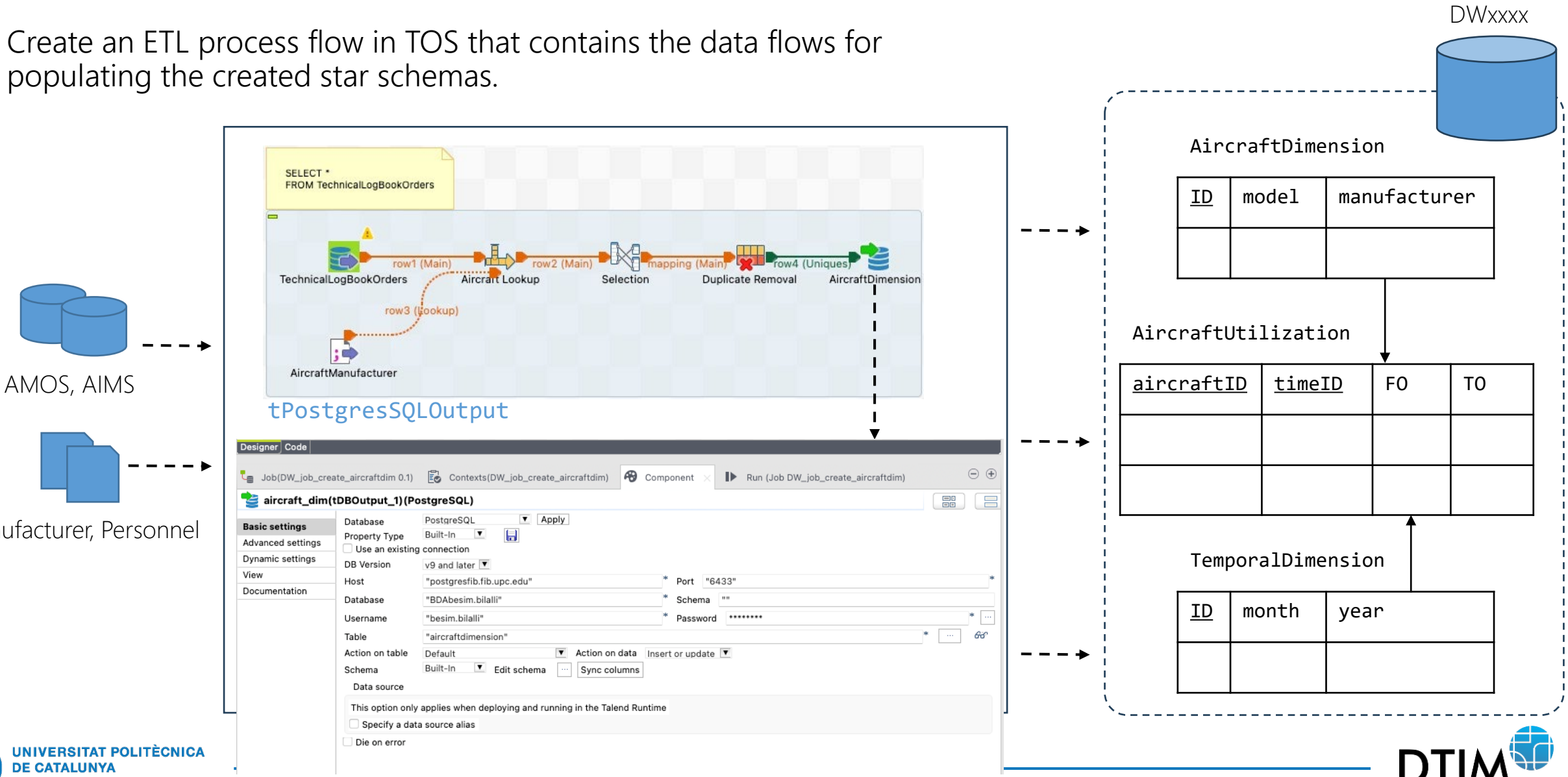
Manufacturer,

rer

TO

# ETL data flow

- Create an ETL process flow in TOS that contains the data flows for populating the created star schemas.



# ETL (Business rules)

## Business rules

Below are listed the business rules that one would expect to be true in the data. Nevertheless, neither the processes nor the DBMS enforced them. Thus, they may have been violated giving rise to quality problems.

### *AMOS database*

Identifiers:

BR1. *workPackageID* is an identifier of *WorkPackage*.
BR2. *workOrderID* is an identifier of *WorkOrders/ForecastedOrders/TechnicalLogBookOrders*.
BR3. *maintenanceID* is an identifier of *MaintenanceEvents/OperationInterruption*.
BR4. *file* is an identifier of *Attachments*.

Datatypes/Domains:

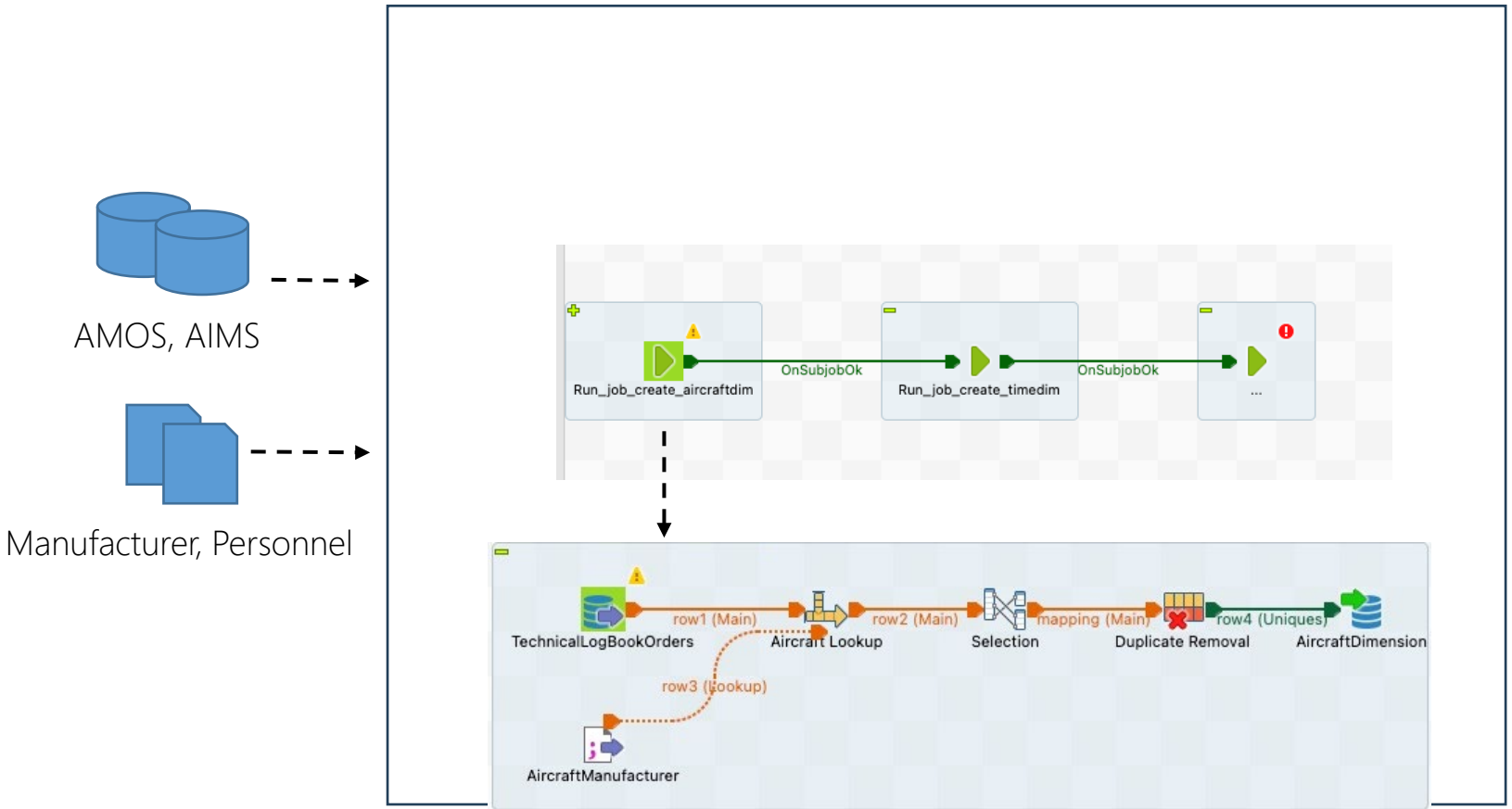
*BR5. ReportKind* values *"PIREP" and "MAREP"* refer to *pilot* and *maintenance* personnel as reporters, respectively.
*BR6. MELCathegory* values *A,B,C,D* refer to *3,10,30,120* days of allowed delay in the repairing of the problem in the aircraft, respectively.
*BR7. airport* in *MaintenanceEvents* must have a value.

Other business rules:

BR8. In *OperationInterruption*, *departure* must coincide with the date of the *flightID* (see bellow how it is composed).
BR9. The flight registered in *OperationInterruption*, must exist in the *Flights* of AIMS database, and be marked as "delayed" *(i.e., delayCode* is not null) with the same IATA delay code.
BR10. In *MaintenanceEvents*, the events of kind *Maintenance* that correspond to a *Revision*, are those of the **same aircraft** whose **interval is completely included in that of the** *Revision*. For all of them, the **airport must be the same**.

# ETL control flow

- Create a control flow that orchestrates the execution



# Deliverables

Deliverables:

1) Talend Open Studio project with data and control flow(s) inside a single zip file.
2) PDF file (**one single A4 page, 2.5cm margins, font size 12, inline space 1.15**) with all assumptions made and justifying the decisions you made (if any).

Assessment criteria:

i) Conciseness of explanations (only first page will be considered in the evaluation)
ii) Understandability
iii) Coherence
iv) Soundness

Evaluation:

- 60% Deliverables
- 40% Exercises related to the project done individually in the classroom the day of the partial exam