DW Project Week 3: ETL

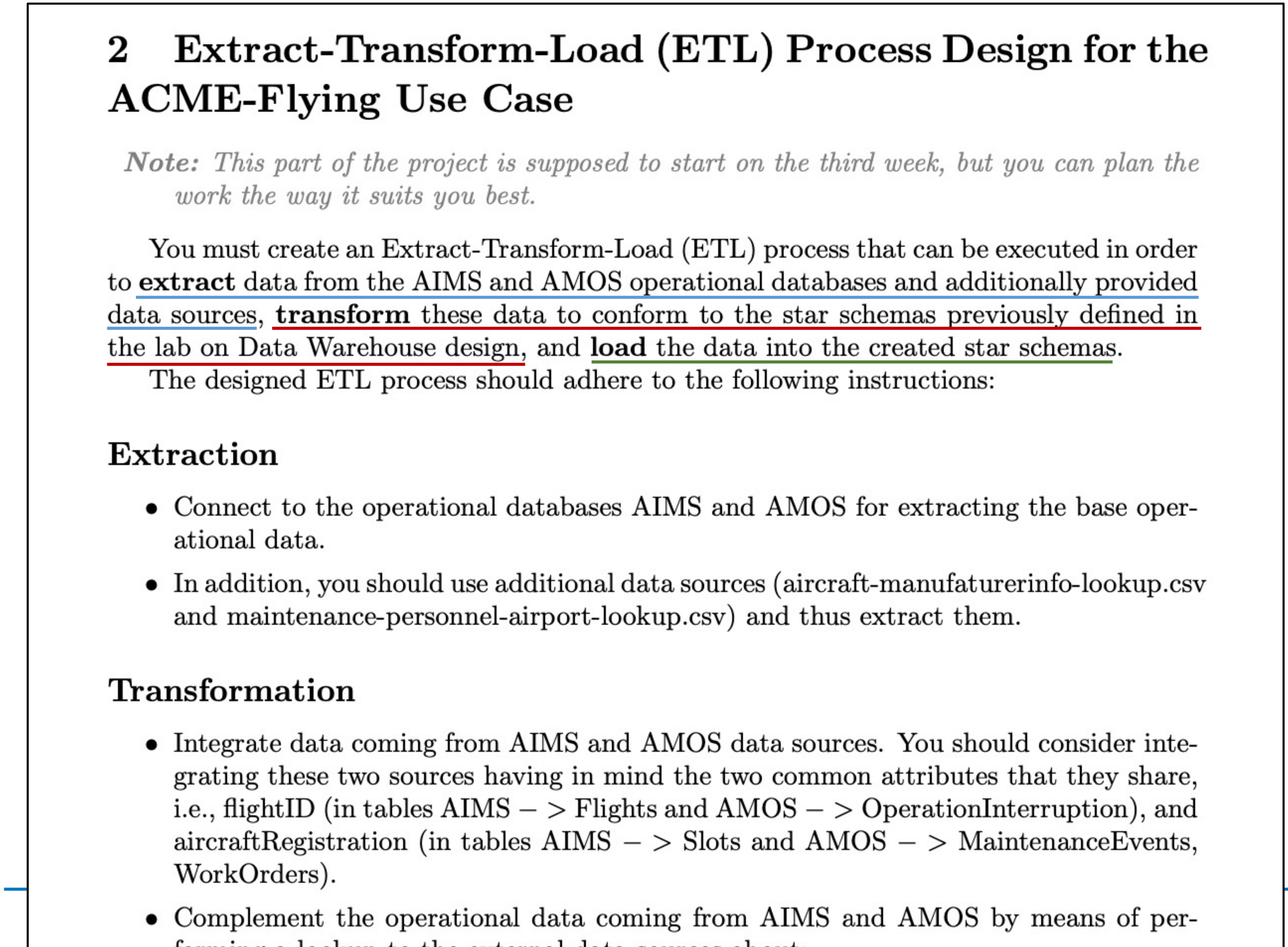
Design an ETL flow in Talend Open Studio

Week 3: ETL

Prerequisites

- Connect to PostgreSQL using DBeaver (Week 1)
  - You will find two databases: AMOS and AIMS
  - Understand the domain
- Design a multidimensional model using Indyco Builder (Week 2)
  - Create independent Star schemas -> Integrate them into a Constellation Schema
  - Propose a logical database schema (i.e., a set of CREATE TABLE statements) corresponding to that multidimensional schema -> Create the tables in Postgres
- Tutorial: ETL Design using Talend Open Studio (Week 3)
  - Follow the instructions in the tutorial to get familiar with TOS

Week 3: ETL (Statement)



2 Extract-Transform-Load (ETL) Process Design for the ACME-Flying Use Case

Note: This part of the project is supposed to start on the third week, but you can plan the work the way it suits you best.

You must create an Extract-Transform-Load (ETL) process that can be executed 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.

The designed ETL process should adhere to the following instructions:

Extraction

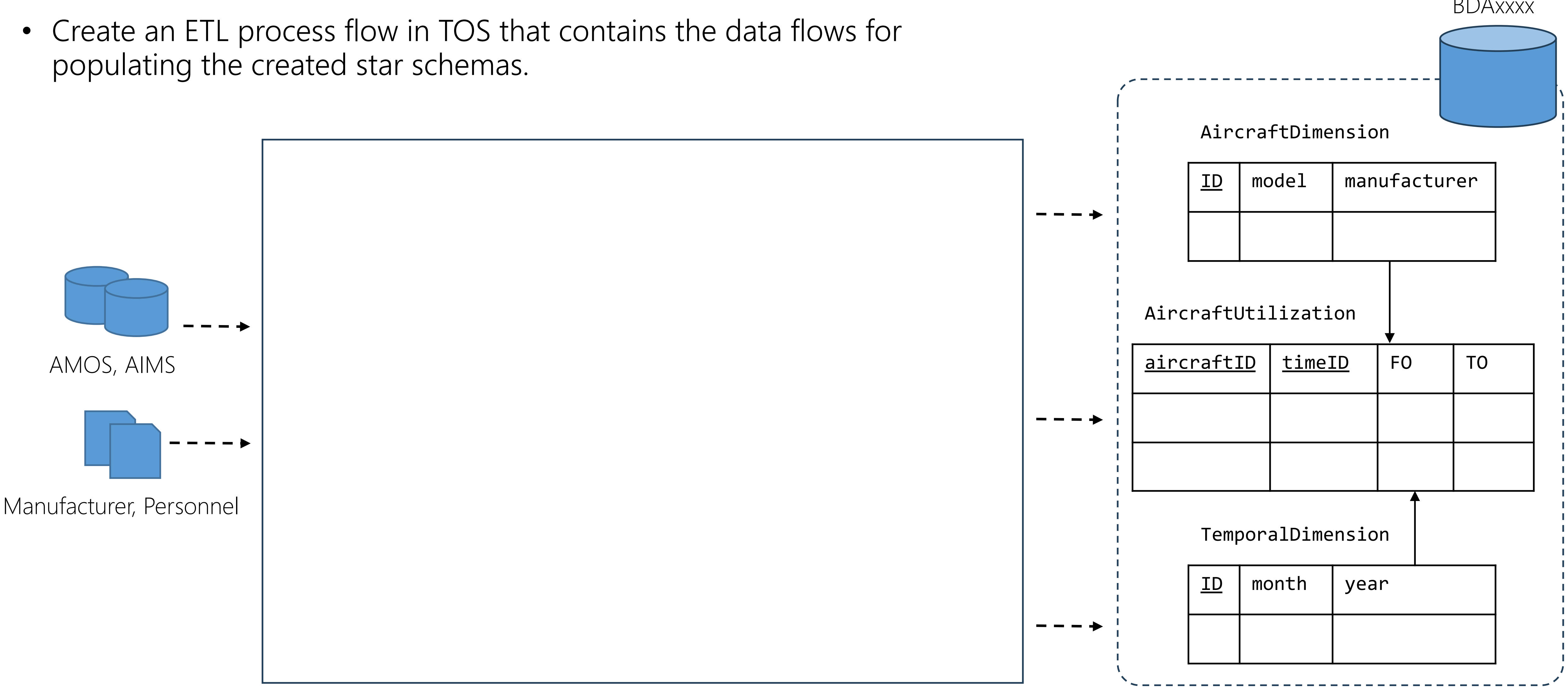
- Connect to the operational databases AIMS and AMOS for extracting the base operational data.
- In addition, you should use additional data sources (aircraft-manufaturerinfo-lookup.csv and maintenance-personnel-airport-lookup.csv) and thus extract them.

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:

Week 3 | ETL data flow

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



AircraftDimension

ID	model	manufacturer

AircraftUtilization

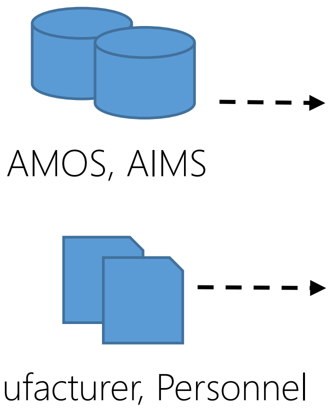
aircraftID	timeID	FO	TO

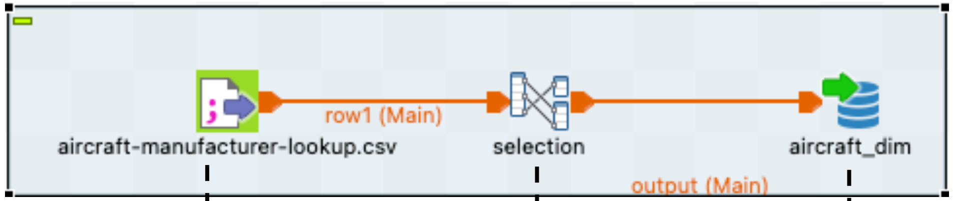
TemporalDimension

ID	month	year

Week 3 | ETL data flow

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

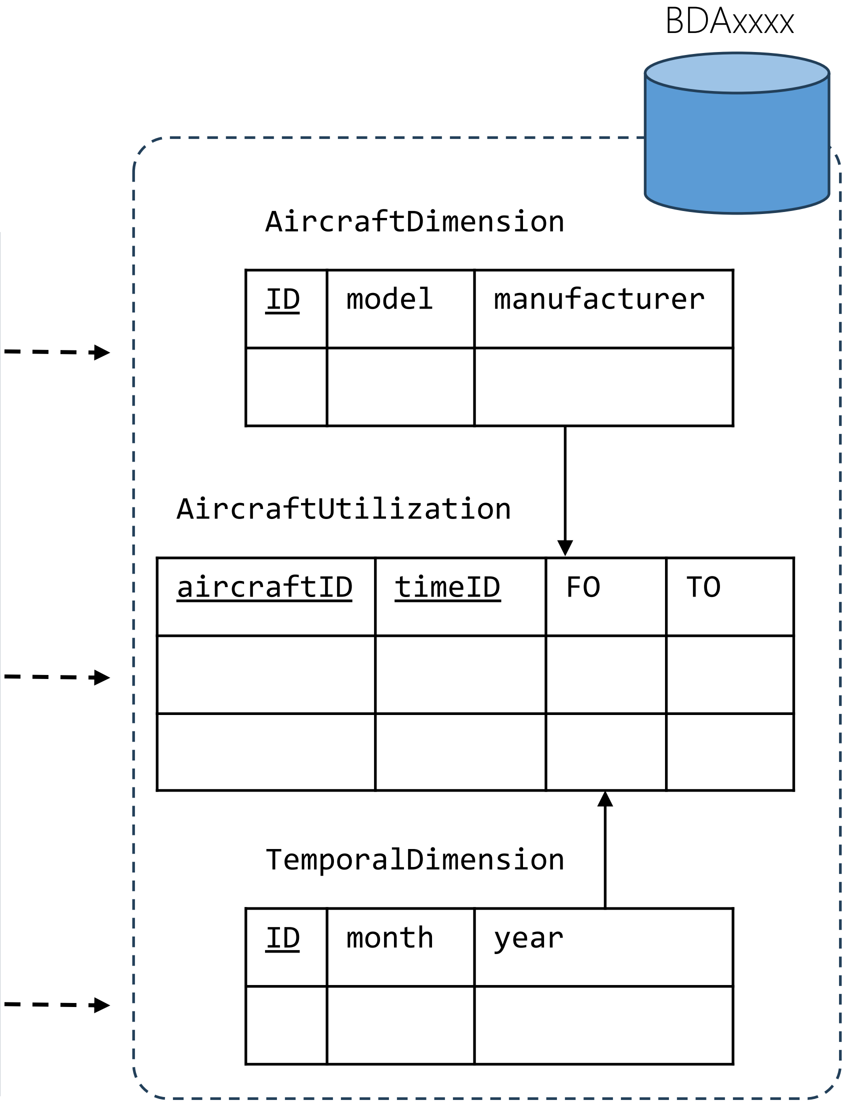




Operator: tFileInputDelimited
Read the file containing the information about the AircraftDimension

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.



AircraftDimension

ID	model	manufacturer

AircraftUtilization

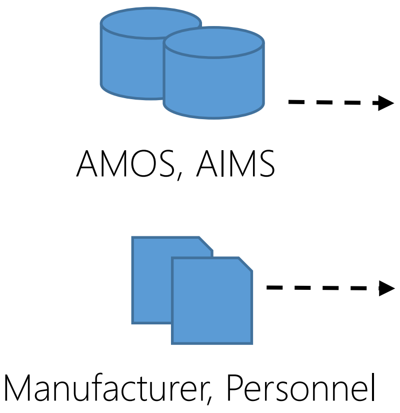
aircraftID	timeID	FO	TO

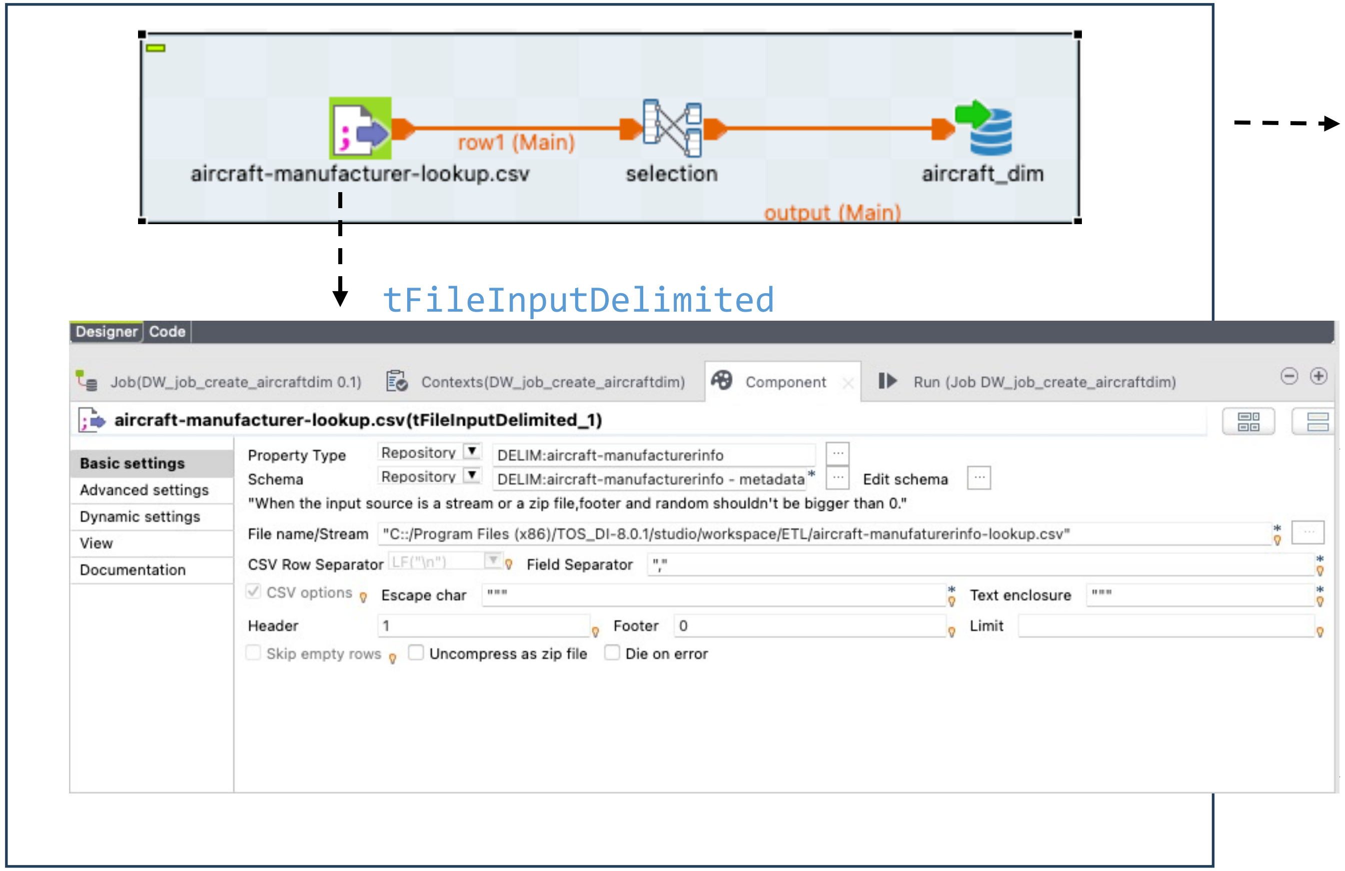
TemporalDimension

ID	month	year

Week 3 | ETL data flow

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





aircraft-manufacturer-lookup.csv(tFileInputDelimited_1)

Property Type Repository ▼ DELIM:aircraft-manufacturerinfo

Schema Repository ▼ DELIM:aircraft-manufacturerinfo - metadata* Edit schema

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

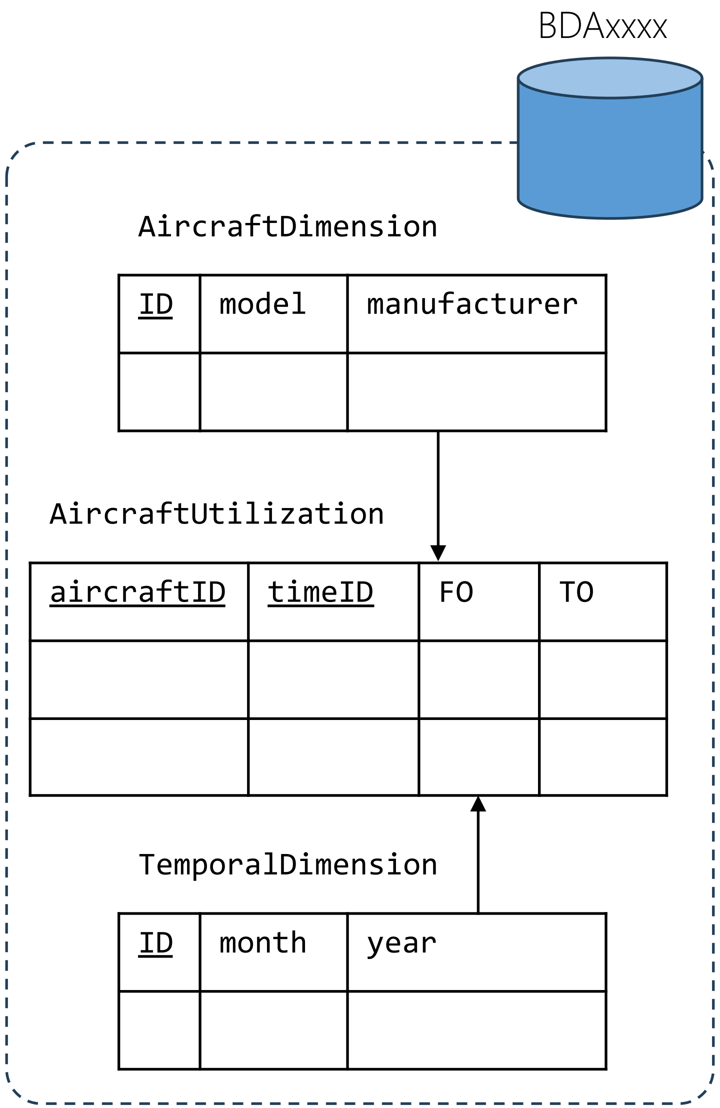
File name/Stream "C::/Program Files (x86)/TOS_DI-8.0.1/studio/workspace/ETL/aircraft-manufaturerinfo-lookup.csv"

CSV Row Separator LF("\n") Field Separator ","

☑ CSV options Escape char """ Text enclosure """

Header 1 Footer 0 Limit

□ Skip empty rows □ Uncompress as zip file □ Die on error



AircraftDimension

ID	model	manufacturer

AircraftUtilization

aircraftID	timeID	FO	TO

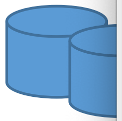
TemporalDimension

ID	month	year

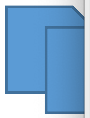
Week 3 | ETL data flow

tMap

- Create



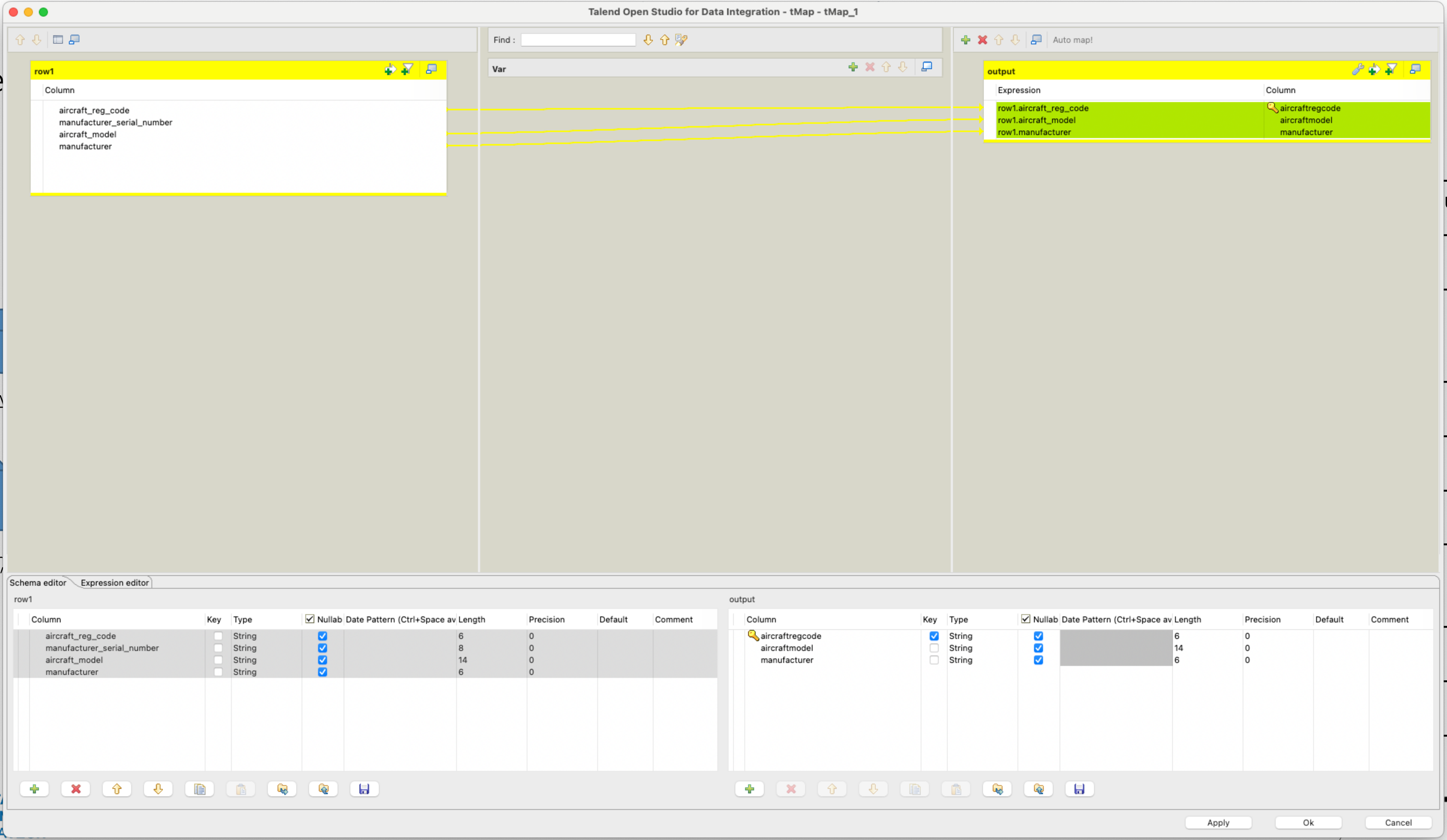
AMOS, A



Manufacturer,

urer

TO



Talend Open Studio for Data Integration - tMap - tMap_1

row1

Column
aircraft_reg_code
manufacturer_serial_number
aircraft_model
manufacturer

output

Expression	Column
row1.aircraft_reg_code	aircraftregcode
row1.aircraft_model	aircraftmodel
row1.manufacturer	manufacturer

row1

Column	Key	Type	☑ Nullab	Date Pattern (Ctrl+Space av	Length	Precision	Default	Comment
aircraft_reg_code	□	String	☑		6	0		
manufacturer_serial_number	□	String	☑		8	0		
aircraft_model	□	String	☑		14	0		
manufacturer	□	String	☑		6	0		

output

Column	Key	Type	☑ Nullab	Date Pattern (Ctrl+Space av	Length	Precision	Default	Comment
aircraftregcode	☑	String	☑		6	0		
aircraftmodel	□	String	☑		14	0		
manufacturer	□	String	☑		6	0		

UNIVERSITAT DE CATALUNYA BARCELONA

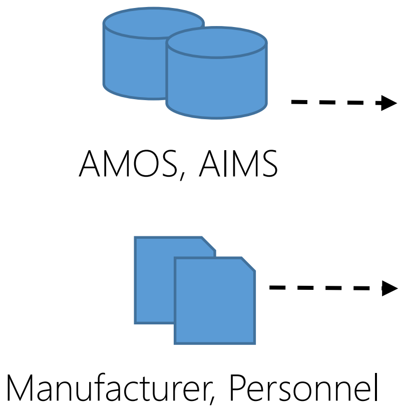
UPC

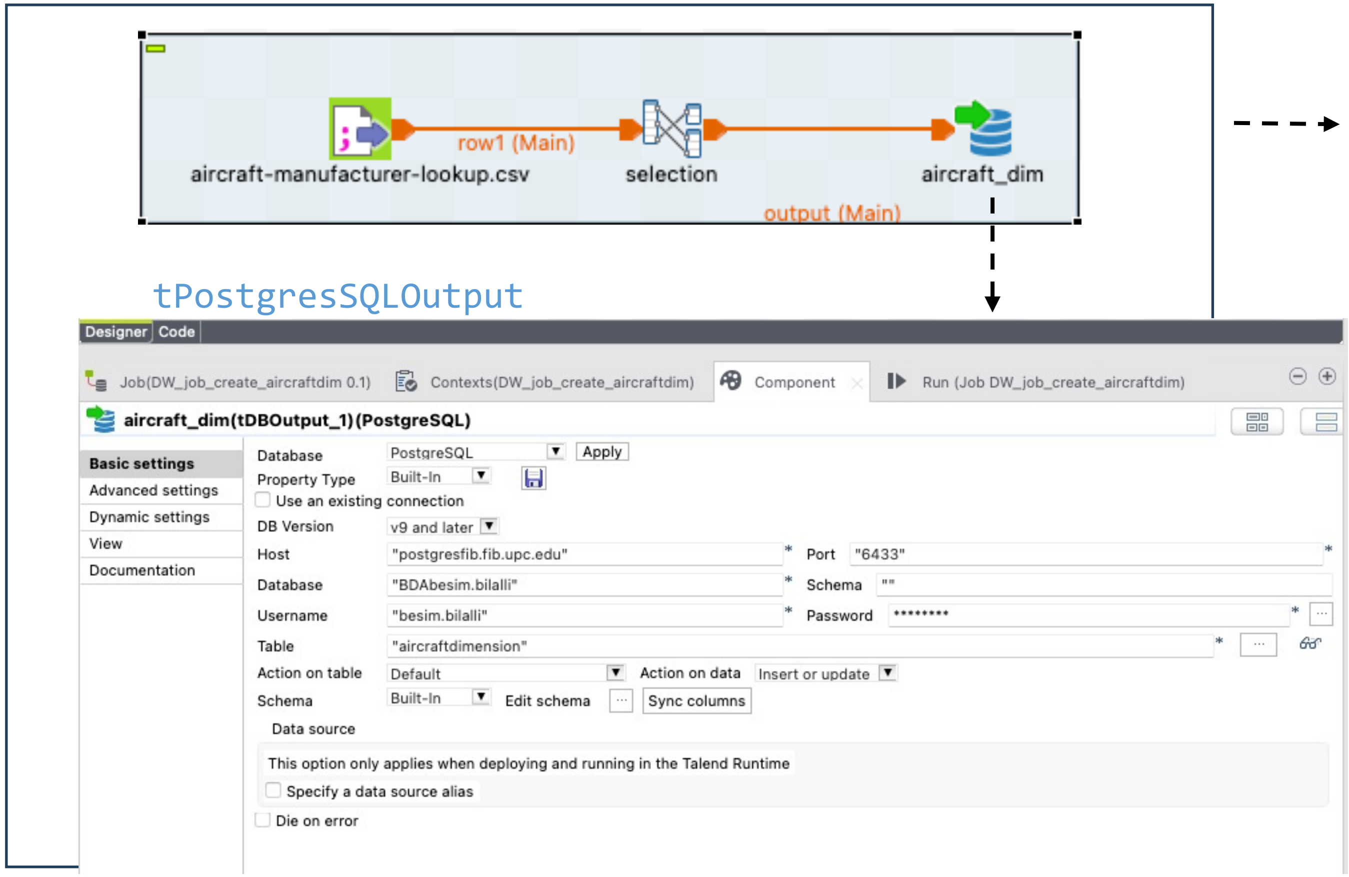
TIM

www.essi.upc.edu/dtim

Week 3 | ETL data flow

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

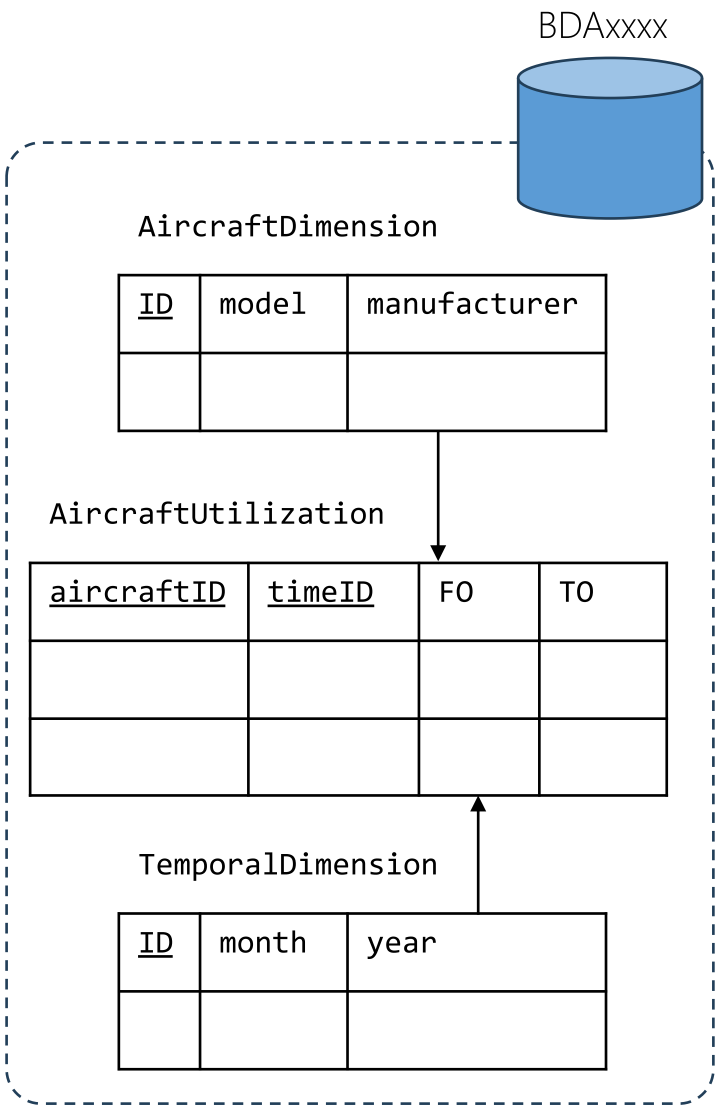




aircraft_dim(tDBOutput_1)(PostgreSQL)

Database: PostgreSQL Apply
Property Type: Built-In
□ Use an existing connection
DB Version: v9 and later
Host: "postgresfib.fib.upc.edu" * Port: "6433" *
Database: "BDAbesim.bilalli" * Schema: ""
Username: "besim.bilalli" * Password: ******** *
Table: "aircraftdimension" *
Action on table: Default Action on data: Insert or update
Schema: Built-In Edit schema ... Sync columns

Data source
This option only applies when deploying and running in the Talend Runtime
□ Specify a data source alias
□ Die on error



AircraftDimension

<u>ID</u>	model	manufacturer

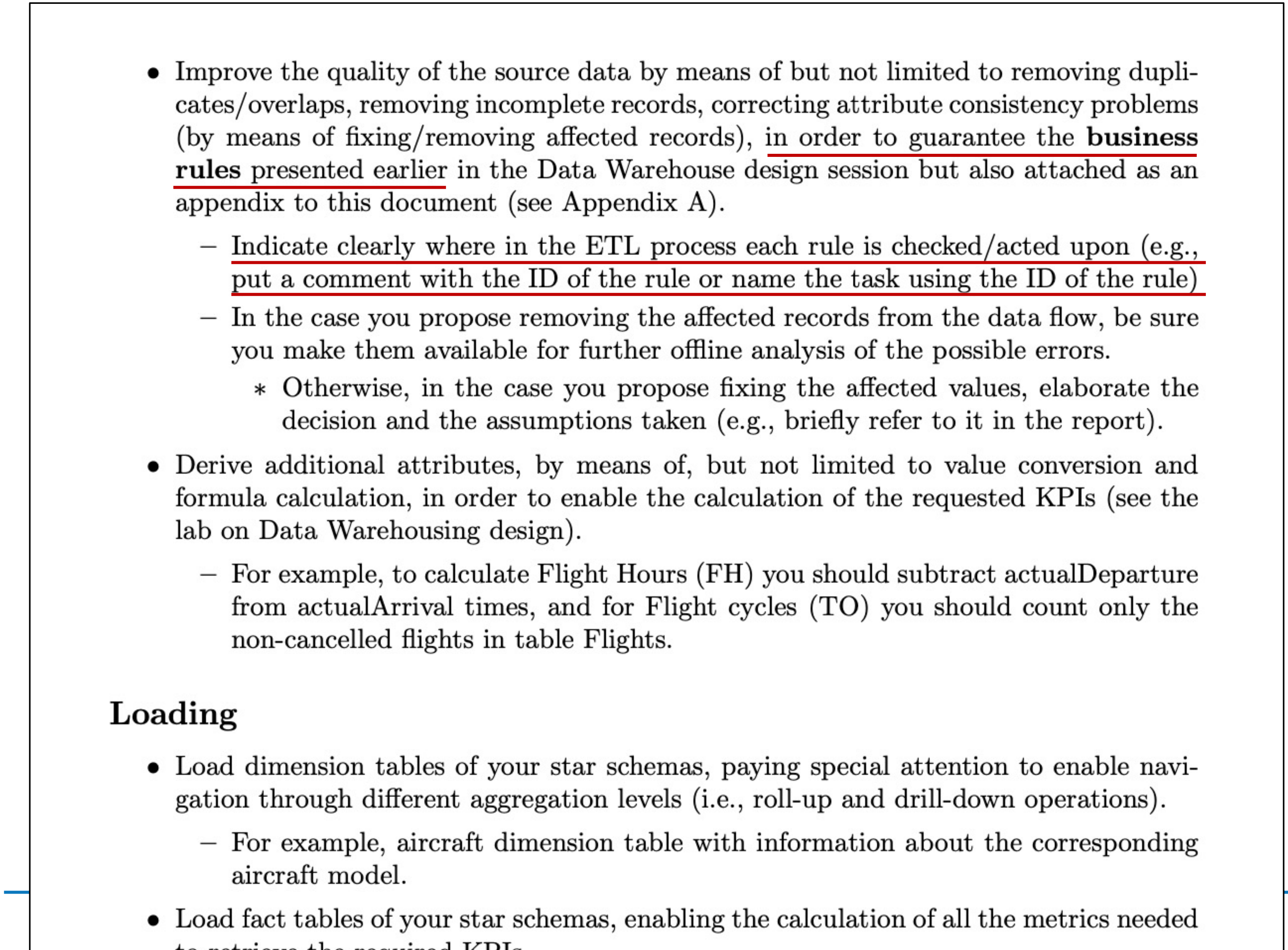
AircraftUtilization

<u>aircraftID</u>	<u>timeID</u>	FO	TO

TemporalDimension

<u>ID</u>	month	year

Week 3: ETL (Business rules)

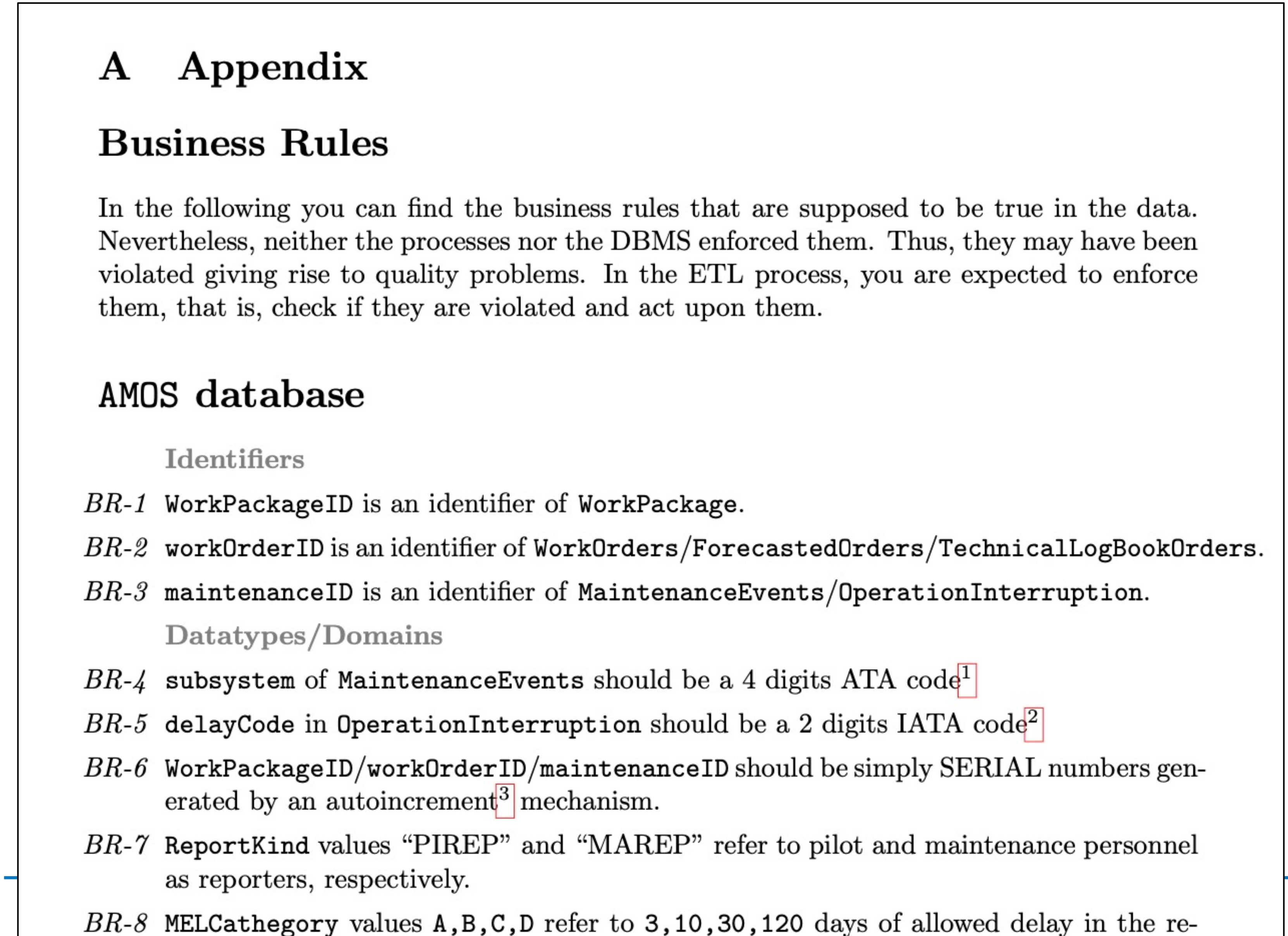


- Improve the quality of the source data by means of but not limited to removing duplicates/overlaps, removing incomplete records, correcting attribute consistency problems (by means of fixing/removing affected records), in order to guarantee the **business rules** presented earlier in the Data Warehouse design session but also attached as an appendix to this document (see Appendix A).
  - Indicate clearly where in the ETL process each rule is checked/acted upon (e.g., put a comment with the ID of the rule or name the task using the ID of the rule)
  - In the case you propose removing the affected records from the data flow, be sure you make them available for further offline analysis of the possible errors.
    - Otherwise, in the case you propose fixing the affected values, elaborate the decision and the assumptions taken (e.g., briefly refer to it in the report).
- Derive additional attributes, by means of, but not limited to value conversion and formula calculation, in order to enable the calculation of the requested KPIs (see the lab on Data Warehousing design).
  - For example, to calculate Flight Hours (FH) you should subtract actualDeparture from actualArrival times, and for Flight cycles (TO) you should count only the non-cancelled flights in table Flights.

Loading

- Load dimension tables of your star schemas, paying special attention to enable navigation through different aggregation levels (i.e., roll-up and drill-down operations).
  - For example, aircraft dimension table with information about the corresponding aircraft model.
- Load fact tables of your star schemas, enabling the calculation of all the metrics needed to retrieve the required KPIs

Week 3: ETL (Business rules)



A Appendix

Business Rules

In the following you can find the business rules that are supposed 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. In the ETL process, you are expected to enforce them, that is, check if they are violated and act upon them.

`AMOS` database

Identifiers

BR-1 `WorkPackageID` is an identifier of `WorkPackage`.

BR-2 `workOrderID` is an identifier of `WorkOrders`/`ForecastedOrders`/`TechnicalLogBookOrders`.

BR-3 `maintenanceID` is an identifier of `MaintenanceEvents`/`OperationInterruption`.

Datatypes/Domains

BR-4 `subsystem` of `MaintenanceEvents` should be a 4 digits ATA code[1]

BR-5 `delayCode` in `OperationInterruption` should be a 2 digits IATA code[2]

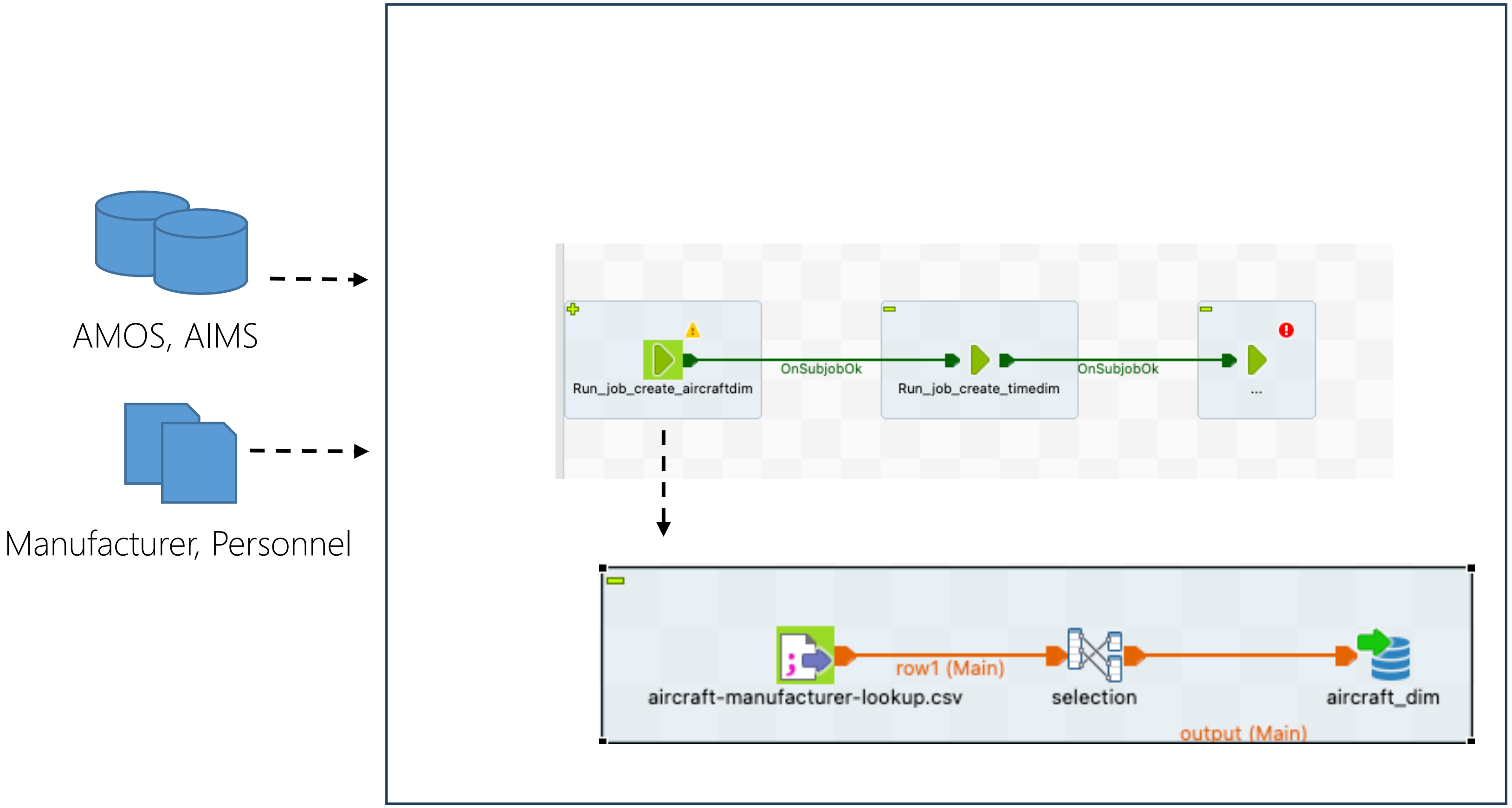
BR-6 `WorkPackageID`/`workOrderID`/`maintenanceID` should be simply SERIAL numbers generated by an autoincrement[3] mechanism.

BR-7 `ReportKind` values "PIREP" and "MAREP" refer to pilot and maintenance personnel as reporters, respectively.

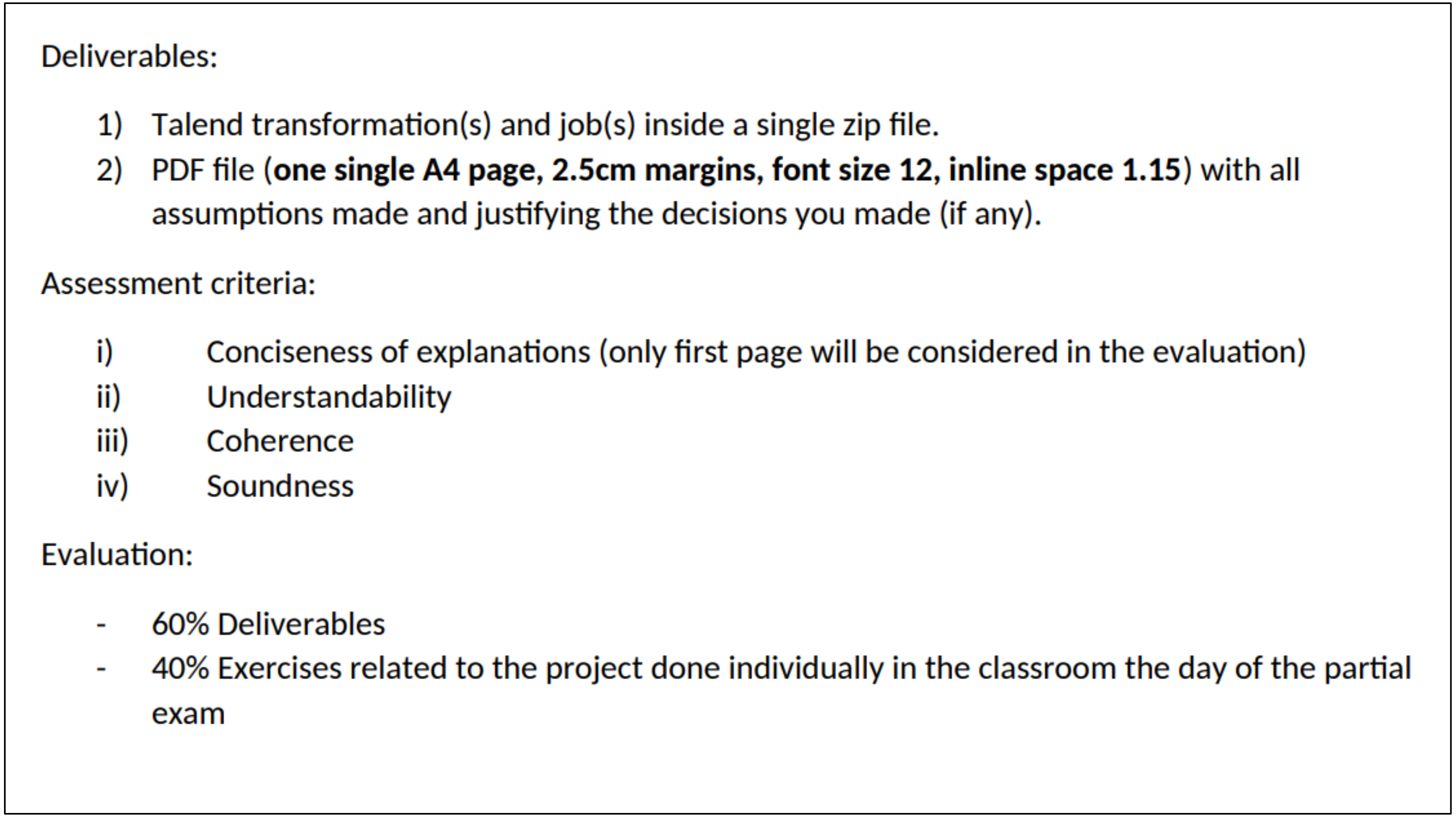
BR-8 `MELCathegory` values `A,B,C,D` refer to `3,10,30,120` days of allowed delay in the re-

Week 3 | ETL control flow

- Create a control flow that orchestrates the execution



Deliverables



Deliverables:

1) Talend transformation(s) and job(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