

## 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-manufacturerinfo-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.
  - **Maintenance personnel employment place** (*maintenance-personnel-airport-lookup.csv*) such that for each person from the maintenance personnel (i.e., *reporteurID* from table *TechnicalLogBookOrders*), your ETL also provides information at which airport this person works.
- 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.
  - 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.
- 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).
  - o For example, aircraft dimension table with information about to the corresponding aircraft model.
- Load fact tables of your star schemas, enabling the calculation of all the metrics needed to retrieve the required KPIs.

### *ETL process quality.*

In addition, you should pay additional attention to the quality of the ETL process, improving (but not limited) the following quality factors:

- *Performance*, mainly focusing on the execution time. For example, by means of parallelizing or assigning more resources to data processing tasks.
- *Reliability*, including but not limited to *robustness*<sup>1</sup> or *recoverability*<sup>2</sup>. For example, by means of creating recovery/checkpoints in the ETL process flow.

### *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

---

<sup>1</sup> Robustness: the degree to which the process operates as intended despite unpredictable or malicious input.

<sup>2</sup> Recoverability: the degree to which the process can recover the data directly affected in case of interruption or failure