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

Javier Herrer Torres ([javier.herrer.torres@estudiantat.upc.edu](mailto:javier.herrer.torres@estudiantat.upc.edu))

Max Ticó Miñarro ([max.tico@estudiantat.upc.edu](mailto:max.tico@estudiantat.upc.edu))

# ASSUMPTIONS

The given Business Rules were translated into Data Quality evaluations only for those which made sense including the ETL. This is, there were some rules related to attributes or tables which weren’t related to any of the metrics to compute.

# DECISIONS

One ETL per table was created following the separation of concerns principle and also allowing modularization, cleaner architecture and better maintainability.

Moved computations and transformations to the extraction queries, relying on the cluster for the heavy lifting and improving the ETL performance. However, Data Quality was done through the Talend operators as it allowed logging the DQ evaluation results, this is, records which got excluded due to DQ issues.

For the metrics extracted from the Flights table, some attributes were transformed into integers with 1 or 0 value so that the aggregation was done SUMing over that column and computing metrics as delays, flight cycles, cancellation and flight hours.