Proj-02-12-H: Hasan Evci, Mohana Fathollahi 30/11/2022

DW Lab 2 - ETL Process Design for the ACME-Flying Use Case

Assumptions:

- Some cleaning approaches have been applied on queries over databases before feeding data into main flow. For example: there were some negative duration values in maintenanceEvents which we removed in the query. We also apply additional filtering in the queries, such as checking for not null - values or distinct values and other.
- In the maintenance lookup file, there exist reporteurid's of people whose role are 'PIREP' which is contradictory because pilots cannot have the role 'MAREP' simultaneously. Therefore, these incorrect data have been removed.
- To calculate Flight cycle we consider flight hours that exist between 15 minutes and 18 hours.
- Cancelled flights are replaced by one and uncancelled flights replaced by zero to make further calculations easier.
- We internally used the component 'tLogRow' to see whether output flows are correct and removed them at the end to avoid additional computation.
- As a valid delay, we count spans between 15 min. and 6 hours. In our flow, a delay is computed by subtracting the scheduledDeparture from actualDeparture.
- To extract from and load to our Oracle Database, it was necessary to set the field of the *Oracle schema* to "\"NAME.SURNAME\"". It was not possible to include this in the Metadata, but only as a built-in component.

Optimizations:

- Multithreading was added whenever meaningful, i.e. parallelizable flows. For us, this was the case for the control flows and the data flow of AircraftUtilization. Thus, multithreading was not added when it did not decrease the runtime significantly.
- In order to parallelize the control flow, i.e. to execute data flows (e.g. dimensions) in parallel, we use 3 jobs for the whole control flow. One subjob for loading all dimensions in parallel, one subjob for loading both fact tables in parallel, and one "global" job that orchestrates that dimensions are loaded before fact tables.
- Memory Assignment: When we assigned more memory to the dimensions (e.g. Xmx4096M), the runtime increased. Only for the largest job, AircraftUtilization, assigning more memory leads to better performance.

Reliability (Recovery points):

- In general, we add recovery points after expensive join-operations.
- We do not include any recovery points in the job for Month- and Temporal-dimension because no Join-operation is used but rather a Unite-operation which is not as expensive.
- Because tHashOutput only loads data to the cache memory, it is useful to have recovery points before them in the job of AircraftUtilization.