Second Project: Extraction, Transformation & Load

Carlos Gascon Dominguez, Mohana Fathollahi

November 2021

Assumptions:

- ☐ Minimum Flight hour considered as 1.5 minute and longest flight hour considered as18 hours. These values are considered as lower and upper bound to count flight cycles.
- ☐ Lower and upper bounds of delay are based on project material, and are considered from 15 minute to 6 hours. Used to count the number of delays.
- ☐ To calculate duration of scheduled and unscheduled out of service, based on the project statement, “Maintenance and Revision” are considered as scheduled so unscheduled service time of them will be zero. “Aircraft on ground and Delay” are considered as Unscheduled, therefore, scheduled time of service for them will be zero.
- ☐ Cancelled flights replaced by one and uncancelled flights replaced by zero to make final calculations easier.
- ☐ Use of external data files to check the correctness of IATA and ATA codes.

Optimization:

- Some join operations have been done in queries instead of Pentaho, because when we did in Pentaho it was very time consuming.
- Using a unique row (hash set) instead of two operations “sort row” and “unique row” is not very time consuming, actually it makes minor changes in time.
- One of the most time consuming operations is the Cartesian product. We changed the location of it to test which place is better and needs less time compared to other options.
- Some recovery points are defined. However their connections are not set up due to a memory error of the JVM we did not manage to solve. The original idea was to set them in points after a considerable amount of computations were made.
- We have experimented with the cache size in some of the operations. However it was not translated into big differences when improving time performance.
- We also have tried different combinations when giving several threads to one step.
- We have tried to maximize the workload to be done in the ETL to avoid increasing the number of transfers between the ETL and the source engine.