Short Summary

Last week, we integrated and explored multiple datasets using Python's pandas library, identifying key entities, attributes, and relationships. Data inconsistencies such as mismatched types and null values were noted.

Initially, we defined the business process as "**every time a user places an order,**" but this week we corrected it to "**every product for every order a user makes.**" The focus shifted from order_id to a newly defined **line_id**. The conceptual model was revised accordingly, and Kimball's dimensional modeling was still applied to implement the logical and physical models in SQL.

- We designed the ETL/ELT pipeline in Kestra, creating a Git repository with seeds/ and models/ directories similar to a **tutorial we found of Kestra UI**.
- Source files are converted to CSV with python
- Using *io.kestra.plugin.dbt.cli.DbtCLI* Kestra can find and read the files on our repository.
- CSV files are staged using queries on file names ***stg_*.sql*** where we rename fields, cast data types, and standardize format.
- Results are transformed according to the schema using queries on file names ***_dim.sl** and ***_fact.sql*.
- Finally, we export the dim and fact tables as *outputFiles on Kestra UI* using python.

Before we move on, we need consultation to finalize project decisions, including confirming assumptions, where to store output, task modularization, removal of hardcoding, and whether datasets will be updated.

Evidence of progress