Open-ended Capstone Step 4: Data Exploration

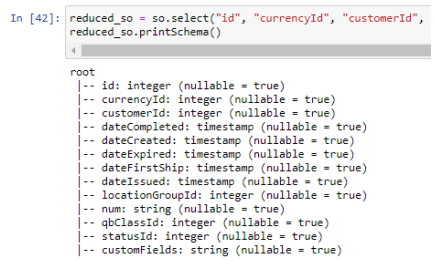
The main goal of the project is to be able to evaluate the success of each product in the different marketplaces (units sold per product per sales channel). At this phase of the project, we explored part of our dataset to start identifying what cleaning and transformation steps will be required for production.

In order to achieve this, Jupyter and Pandas were used as the exploration tool. Exploration was performed on the Sales Order table, which is one of the key tables of our MySQL database and the one that will be the core element of the Dimensional Star Schema development. This table represents the sales transactions at order level.

These are the main insights derived from the exploration:

- Initial number of sales transactions in 2021 is: 2,319 (up to date)

- Checking the data types of the Pandas dataframe, we found that the date fields are read as object data type (mixed numeric and non-numeric values). Since we will be using Spark for transformation in production and MySQL tables will be read directly from the database, we will have access to the schema. For example, the schema for the sales order table pulled using PySpark is:



- There are some fields that have NaN values (currencyId and dateCompleted). At this stage of the project (using Pandas), these were corrected either by replacing values or by filtering rows out.

  For production, depending on how these are read in Spark and the specific column, a transformation will be possibly performed for correction as well.

- Order status column (statusId) was validated to only include fulfilled orders.

- Transactions were categorized by Sales Channel based on defined conditions.

- We found that 5 channels (out of 15) represent almost 85% of the total sales in 2021 (number of orders per channel).

- After cleaning and transformation, number of relevant sales transactions is: 2,143

Transformation, storage and compression of storage files

- Data will be programmatically extracted once a day. This will be done using Pyspark thru a JDBC connection to the MySQL database.
- Transformation will be performed using Pyspark. Once data is cleaned and enriched, it will be added to the Fact and Dimensional tables.
- File format to use is Parquet, as it is a columnar data store, ideal for OLAP.
- Fact and Dimensional tables will be saved in Tower Server.