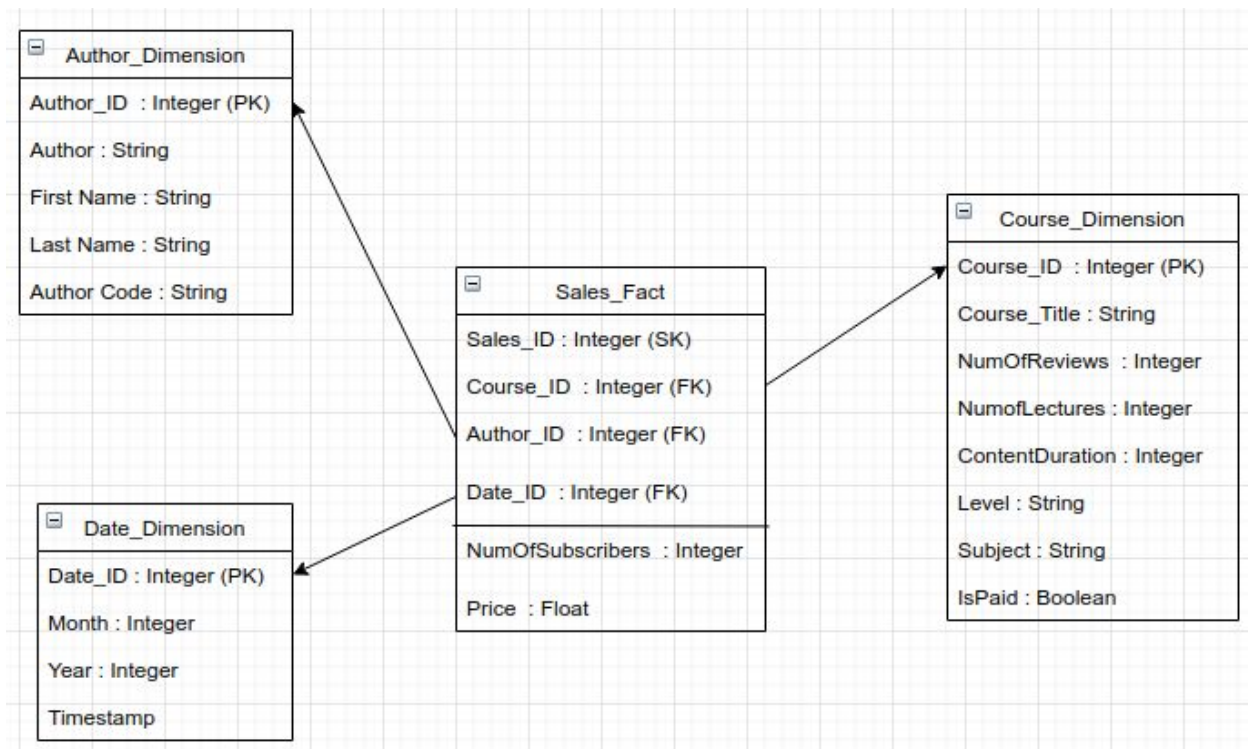


W-trainer

Nuwan Tilsera

Star Schema



This contains one fact table and 3 dimension tables. I have used primary keys and Surrogate keys here. Surrogate keys (SK) protect the system from sudden administrative changes. Another advantage of SK is, It enables us to add rows to dimensions that do not exist in the system now. SK's can improve query processing.

A fact table is a table that contains the measures of interest. Here the measure is Addictive type. We are mainly focussing on price and no of subscribers here in sales fact. Granularity is most important when it comes to facts. E.g: "Sales amounts, by product, by month."

Analytical Queries

Analytical queries are more focussed on sales of the courses. The sql file contains queries.

Challenges

First challenge was to find the best ETL(Extract, Transform, and Load) tool for Python. Between various options Pandas, Airflow, Petl, PySpark etc. Later I found SQLAlchemy (ORM) that is

much easier and flexible. It provides an efficient and high performance database.

When i started adding data into tables using python script, i got some SQL issues saying its violating UNIQUE constraint in courseId field. So I think of keeping a separate Primary key in those tables.

There were some issues when parsing datetime in python. It was resolved with a correct regex.

Improvements

Row and page compression

Fact table partitioning

DB Indexing

Invalidate JSON file