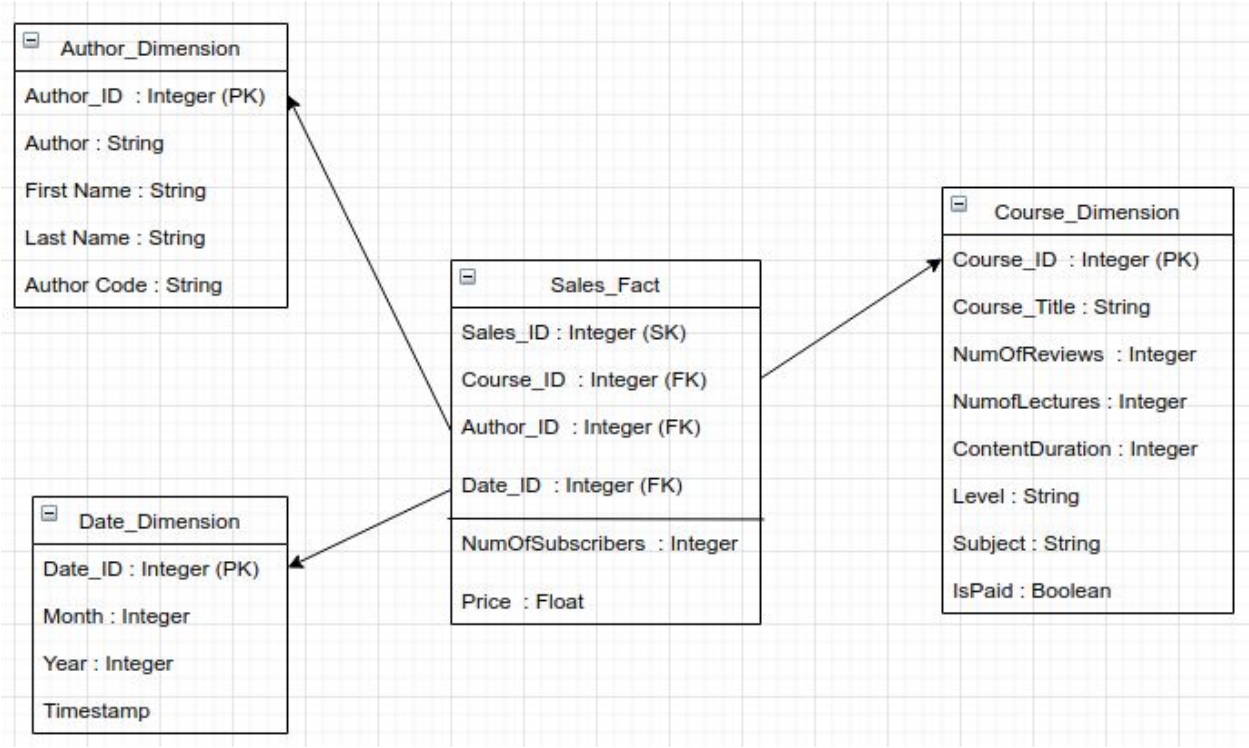
W-trainer

Nuwan TIssera

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 queries.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

1. Row and page compression
   Compression allows more rows to be stored on a page, but does not change the maximum row size of a table or index.
   Anyway, SQLalchemy library does not provide compression tools.
2. Fact table partitioning
   We can separate fact tables considering some fields to improve performance in large tables.
3. DB Indexing
   Using right and most wanted indexes in a data warehouse system can increase its speed and efficiency when it comes to a large dataset.
   E.g:
   CourseId_index001 = Index('CourseId_index001', Course.CourseId)
   CourseId_index001.create(bind=engine)
4. Invalidate JSON file
   As a robust measure, we can validate the JSON whether the JSON contains the correct (amount and field) of keys. So if it's not so, We can give a proper error. So it makes more sense.