

BI Project

Loading Phase


Tong Xinze - Ringuet Nicolas - Chanson Alexandre

Introduction

In this phase we loaded the data into the previously created warehouse, this represents the bulk of the work of the data warehousing project.

We opted for a mixed approach using both ETL software (talend) and custom scripts written in python to improve performances. As we query an API for information we built a simple caching engine in python to avoid hitting the rate limits.

Architecture

 dw_architecture

We chose to shift our initial plan to use the million song dataset to using the spotify API, this approach enables us to obtain up to date information and to avoid matching ID between too systems. The informations we wanted on google searches of songs turned out to be quite expensive so we didn't include it in the warehouse (This would have been solved by asking money to our boss in a profesionnal environnement).

We plan to add data from twitter and other social media platforms to have additional features for the data mining process.

Master Job

 master_job

The master job is split accross the two technologies, python first and talend next handling the last phase. In simpler terms the E and T phases are done using python while talend handles the L phase.

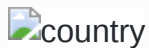
The scripts are available on [github](#) and a description of the talend jobs thereafter.

Python excecutor

 extracttransform

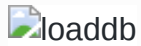
This is a simple job that runs all the python scripts necessary to extract and clean the data.

Extract & Load Country Codes

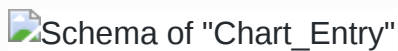


The country codes come from [github](#) and doesn't need much processing, its IDs are simply turned to lowercase and loaded in the database.

Load flat table



Star schema



The principal fact is translated to a star schema, the one to many relationship is handled using a bridge (Table Plays), this is the abstract table schema (primary keys underlined) types are specified in the SQL table creation script.

Create table SQL



The second fact that stores metrics about songs shares its dimension with the main fact, here track_id is the fact table's primary key and a foreign key linking to Track.

Conclusion

The loading was a complex operation yet it would have been simpler to use only python as talend only performs loading operation and most of the logic for API querring cannot be easily integrated in this tool. We learn that not all data is free and next time we want a specific data set such as google searches we will first check if we can obtain it and at what price point.