Team project description

Roaya Helal, Minchan Choi

Deliverables for ETL Project:

Write up containing the following:

Introduction describing the data you collected and why it is worth looking at.

3-5 sentences describing difficulties of cleaning and transforming your data set.

An explanation of why you chose either Mongo or Postgres to upload into.

We collected Netflix catalogue files to make a database which is accumulative so when there is a new file, we simply add new data to an initial database. The main problem of this project is to control duplicates in new data. Thus, we used two solutions to figure it out. First solution is to make a new defined function which add a mask which indicates duplicated titles. Second solution is to use existed merge attribution to discern duplicates. We chose Postgres because we wanted our database to have fixed field (title, rating, release year)

2. Working Jupyter Notebook

3. SQL or NoSQL queiries and schema.

We only created the database on PGAdmin, the schema was done in jupyter notebook using sql alchemy