Big Data Analytics

Assignment 1

Due Date: 11:59 PM, Monday, 25-April-2022

> "The will to win, the desire to succeed, the urge to reach your full potential... these are the keys that will unlock the door to personal excellence."
>
> - Confucius

Introduction:

YouTube has been collecting data on their video streaming and the activities of user. They have recently shown their interest in the proper analysis of this data. The most important thing what they want to know is "what videos are users most interested in?"

For the time being they have been following a very difficult way to query their data.

1- They have directories of JSON log files for the user activities.
2- They have a directory that contains Metadata of these videos in form of JSON files.

We want you build a PostgreSQL database for performing a good **Play Analysis** of these videos.

You will be doing the following things,

1- Optimization of tables by following a star schema (using fact and dimension tables) so that we can run queries.
2- Creating a database schema of these optimized tables and writing an ETL (Extract Transform and Load) layer to load data from JSON log files to this database.

Dataset:

In the **data** folder you will find two directories,

- youtube_data contain metadata of videos and youtubers in many sub directories.
- log_data contain metadata of events.

To read these files, you can use **pandas.read_json** function of pandas library.

Instructions:

Following are the steps you need to follow to complete this assignment:

Step 1: Creation of Tables.

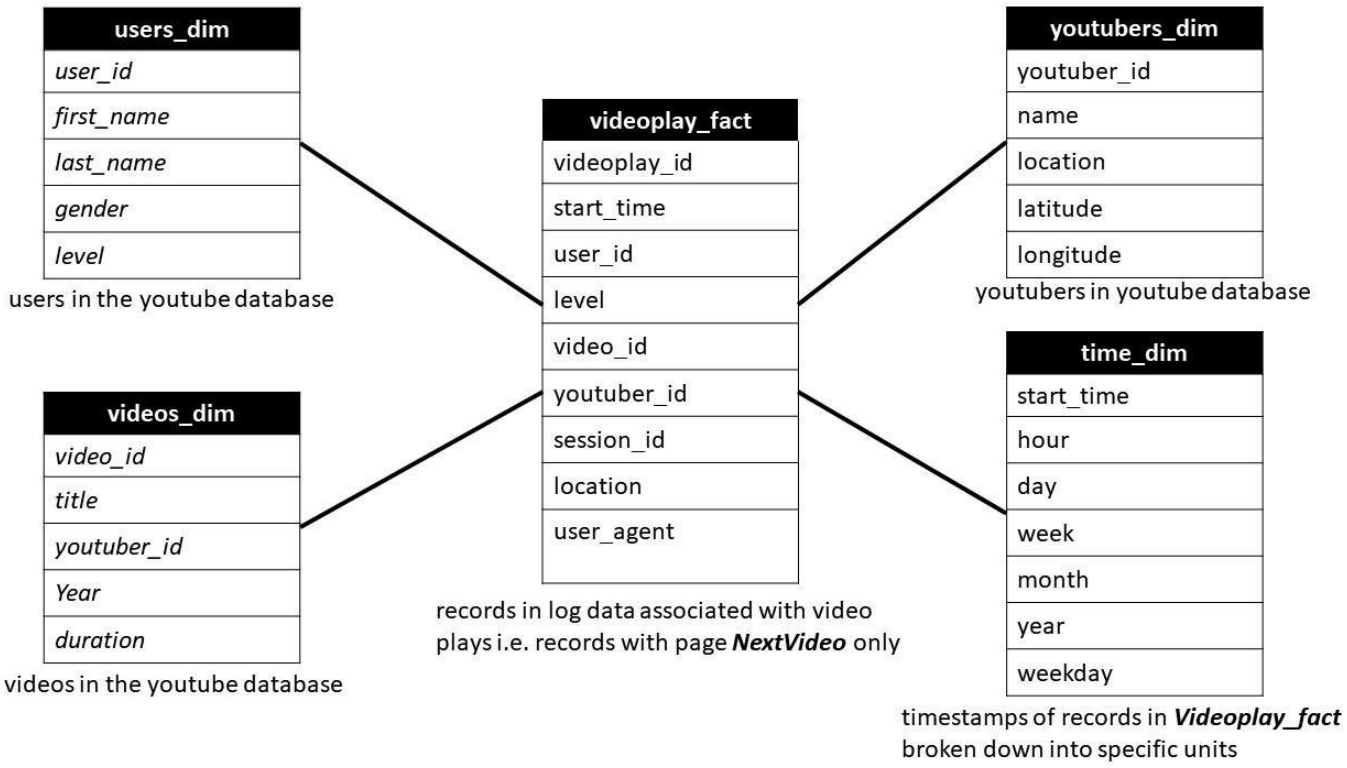
- You should write all "Create Table" statements in *Create_Table_queries.py* to create each table.
- You should write DROP table statements in *Create_Table_queries.py* to drop each table if it exists already.
- Run *Create_Tables.py* to create your database and tables.
- Run *Test_table_schemas.ipynb* to confirm that the tables and correct columns have been created successfully. Make sure to click "Restart kernel to close the connection to the database after running this notebook

Step 2: Building an Extract Transform Load (ETL) pipeline

Build an Extract Transform Load (etl) Pipeline in *extract_transform_load.ipynb* for each table. In this notebook you will be writing code to extract data from JSON files and load it properly into fact and dimension tables. You can run *test_tables.ipynb* to check that your data has been properly loaded or not.

Schema for Videos Play Analysis:

You are required to create a query-optimized schema on the given video play analysis using youtube videos and log data set. Star schema that you need to create is given below.



Note: You should **only make changes** to the following two files. Please don't change other files

1. ***Create_Table_queries.py***
2. ***extract_transform_load.ipynb***

PART2:

- Complete the tasks given in notebook in Part B folder.
- ETL pipeline for preprocessing the files
- Complete the Apache Cassandra coding portion of your project.

Please add proper comments with your code. Submission guidelines will be shared soon.