CNIT 372 Group Project

The goal of this project is to create a working database and program support for a real application.

You wouldn't have to create a front-end for the application (such as a website), but the backend (PL/SQL) would have to work flawlessly.

All groups should have a minimum size of three and a maximum size of four.

Theme: YouTube

Your group project will be centered around the overarching theme of YouTube. You have a range of example options provided below from which you can select. Additionally, feel free to propose alternative topics that are closely related to the YouTube platform.

Datasets:

On Brightspace, we have provided some datasets about YouTube (downloaded from https://www.kaggle.com/). In your project, you can choose to use as many as those datasets and combine them with your own datasets.

How do you collect your own datasets?

- You can download your watch history data from Google. Follow the instructions here (https://support.google.com/accounts/answer/3024190?hl=en).
- You can also use YouTube Data API (https://developers.google.com/youtube/v3/getting-started) to retrieve related data.
- You can also generate dummy data about YouTube usage. In that case, you will need to provide descriptions of all the data fields (columns) you created.
- Feel free to use other publicly available datasets about YouTube.

Option A: Usage of YouTube as a Content Consumer

People use YouTube for a wide variety of purposes, and the platform has evolved into a multi-faceted platform with diverse use cases. Some people use it as a search engine, others use it for entertainment, fitness, to name a few.

Putting on the hat of a regular YouTube user, how do you use YouTube? Do you like the way you use YouTube? In this project, you can use the knowledge learned in this class to analyze your YouTube usage, such as your watching habit, interests, and behavior patterns, to name a few.

Option B: YouTube Data Analysis as a Content Creator

YouTube is one of the most popular and premier video-sharing platforms by Google that allows content creators, be it individuals or companies to create a channel and upload

videos of all different kinds. With YouTube in place, anyone can create content and share it with the world for entertainment, business, or welfare purposes as well.

Putting on the hat of a content creator, how can you use the available YouTube data to inform plans for your channel? In this project, you can use the knowledge learned in this class to analyze the performance of certain channels or types of channels, identify trending or popular content for various user groups, and monitor the performance of your channel, to name a few.

Option C: YouTube Activity Analysis as a Researcher, Journalist, …

YouTube has more than 2.70 billion monthly active users as of 2023. YouTube Premium and Music together have more than 80 million subscribers in the world. More than ¼ of the world's population uses YouTube every month (source).

YouTube has also been widely used in different geographical areas, by people with different ideological beliefs, for various purposes ranging from advertising products to political campaigns. Research has shown that the recommendation system of YouTube can be biased and trap users in filter bubbles of misinformation, radical or extreme content (Tomlein et al., Srba et al.).

Putting on the hat of a curious intellectual, what do you want to know about the design, use, and impact of YouTube algorithms, content, and platform?

Requirements

- You will pick one of the above options or define your own project focus.
- You will come up with your own interesting questions about your selected option.
- Your submission needs to include a standalone readme file to provide instructions on how to download and import the necessary data to one's Oracle database.
- The data need to be separated into at least 3 related tables, and a total of at least 2000 rows.

Notes:

1. If YouTube watch history data is needed, you can assume that the user of your project will import their own watch history data for analysis. You are **not** expected to include your own YouTube usage history data files in any submissions.
2. When you present your results in your report and presentation, focus on what questions you are asking, why these questions are meaningful/important, and how you are answering them with the knowledge learned in this class. Feel free to **mask** your data when showing the results.

The following requirements apply to all options. Your project must include:

a) No less than 10 questions that you are answering.
b) The solutions to the questions must be in PL/SQL. They must include:
    a. Functions or procedures (for all questions)

b. Cursors, whenever appropriate, but in at least one question
c. Triggers (for at least one question)
d. At least one package that is suitable for handling your applications and all your questions (and answers). All of your functions or procedures must be included in some package in a meaningful way.

c) The answers don't have to be long, but they should correspond to the questions.

Milestones

You will complete your project through three milestones before submitting your final report.

Milestone 1: Form a Team and Outline a Goal. (30)

- Write down the names of the people in your group, and your option choice.
- Describe what data you plan to use, what insights can be derived from this data, and why you think the listed insights are important.
- Describe how you plan to design the database to store this data and explain your rationales.
  - Outline the tables that you plan to work with, describe the column names, their corresponding data types and constraints, and relationships between the tables.
- Describe your teamwork: how did you make the decision about the project goal, list the contribution of each team member.

Milestone 2: List the Questions. (40)

- Submit 10 questions that your project will answer based on the data provided.
- For each question, explain what knowledge it will use from this class, and why it is important/useful for whom.
- Please make sure the questions are of interest to somebody who would be using your application.
  - A good example: *Which flavors of ice cream are most popular and during what time?* This question can help the supply chain management team of an ice cream company to optimize future order plans and customize their ice cream supply for different customers.
  - Try not to ask questions like *list all the first names that start with the letter T.*
- Questions should be meaningful and non-repetitive. Each of the answers should demonstrate something that you learned in the course. For example, do not submit questions that all require a count of something.
- Describe your teamwork: how did you come up with the 10 questions, list the contribution of each team member.

Milestone 3: Your Solutions to half of the questions. (30)

Set up a GitHub repository and manage your teamwork there.

You will label all the questions that can be answered by the knowledge you have learned in this course so far and provide your solutions to those questions.

You will submit:

- A link to your GitHub repository.
- In your repository, there should be
    a. A script for importing data into an empty database.
    b. The project description from Milestone 1 and the question list from Milestone 2.
    c. The solutions you have implemented, labeled by the corresponding questions.
- Describe your teamwork: how did you come up with the 10 questions, list the contribution of each team member.
-

Final Report (10)

Your final submission to Brightspace should be a link to your GitHub repository.

The repository should include:

- A readme file to briefly introduce the theme and purpose of your project, and instructions for using your code.
- Your final project report (a pdf file). More instructions below.
- Your code (organized in folders with meaningful names, as you see fit).

Your final report should contain the following sections:

- Background: application motivation and explanation.
- Database Description: describe the tables and the columns in each table, and the relationships between tables.
- Solutions: all your questions and the corresponding answers.
    - Each question should be formatted as a question (with a question mark, not statement or instruction) and also describe its business value (what problem it can solve or help solve, as you did in Milestone 2).
    - Each answer should include
        - A description of your answer/insights from your query results
        - the SQL or PL/SQL code
        - the query results
- Team: describe your team members, and the contributions made by each member (who worked on which parts of the project).

Final Presentation (10)

Prepare a 15 min presentation that contains:

- Team member introduction
- Your application motivation and explanation.
- The data and tables your team worked with.
- The questions that you are answering and the answers themselves.
- The structure of your code (package spec), and how it maps to the questions.
- Every team member should present some part of your project.

    - Special circumstances need to be discussed with the instructor asap.
- The presentation needs to be in person in the last three lecture sessions in the classroom. The order of presentation will be randomly assigned.

Submit your presentation slides (or pdf) to Brightspace.

Project peer-assessment. (4)

- You will be assigned to review up to three projects.
- You will evaluate them based on the rubrics provided to you.
- Your evaluation quality will count towards your final grade.

Evaluation

All group members will share the same grade by default, except that the peer assessment grade will be assigned individually based on your assessment efforts. In case there is a disagreement or concern about the individual contribution, please contact the instructor **as soon as possible.** Accommodations for last-minute appeals will not be guaranteed.

Appendix: YouTube Data Download

I provided some screenshots in case you need some guidance for downloading your YouTube data (Figure 1 Export YouTube Data and Specify the Data Format and What Data to Export and Figure 2 Choose JSON or HTML data format.)

Note that the size of history records may vary for different individuals.

You can use a free online converter to convert HTML or other data formats to prepare for SQL Developer imports. You can also provide your own data processing scripts. No matter how you choose to do this, document the data import steps in detail.

FAQs for downloading your own YouTube watch history data:

Disconnected Tables

The YouTube data has very few columns and is difficult to deal with!

Please feel free to connect the watch history data with the other datasets provided on Brightspace. You can also add new columns for connecting the YouTube data from the same user in your CSV file.

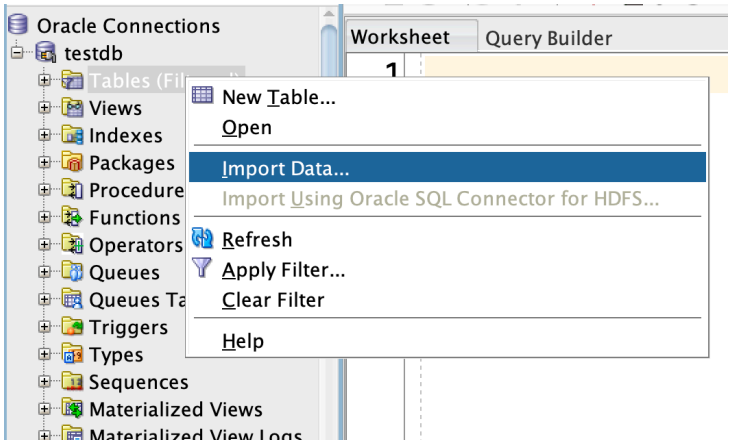
Your team may have data from multiple users, so one data preprocessing you can do is to concatenate the CSV files from different people into one (e.g. one aggregated watch-history.csv, with multiple people's data inside, and additional columns to distinguish the different people).

Support format for data import / converter tools

JSON format is not yet supported, so you will need to convert HTML/JSON data to .csv format and import the data as tables in SQL Developer.

You can find multiple resources online for converting data files, here is an example: https://www.convertcsv.com/

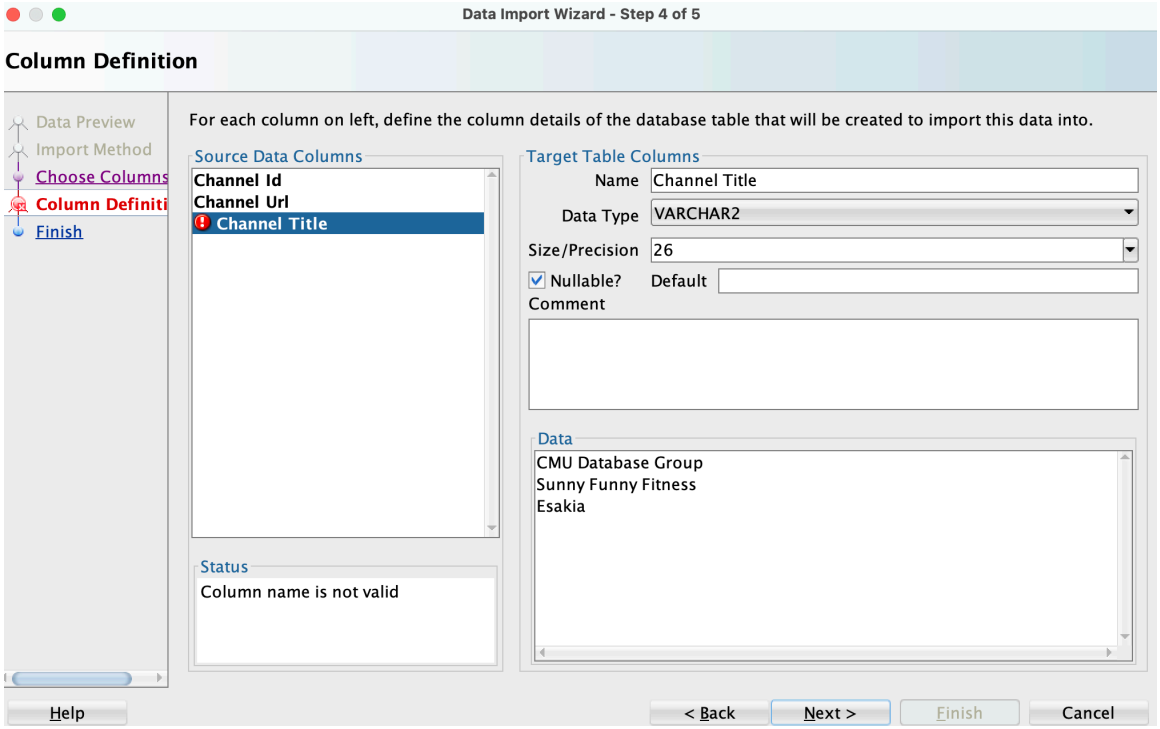
You can import data to tables directly in SQL developer GUI. (See below).



Data import process

You will be able to fix incorrect column names and specify data types and constraints for the table you created from the data import.

Feel free to do whatever you think is relevant to the questions you are trying to answer.



Illegal characters

My own YouTube data did not incur any errors due to illegal characters during my import process. Here is some official guidance that you might find helpful: https://docs.oracle.com/en/cloud/saas/sales/22b/faiem/import-data-with-special-characters.html#s20073482

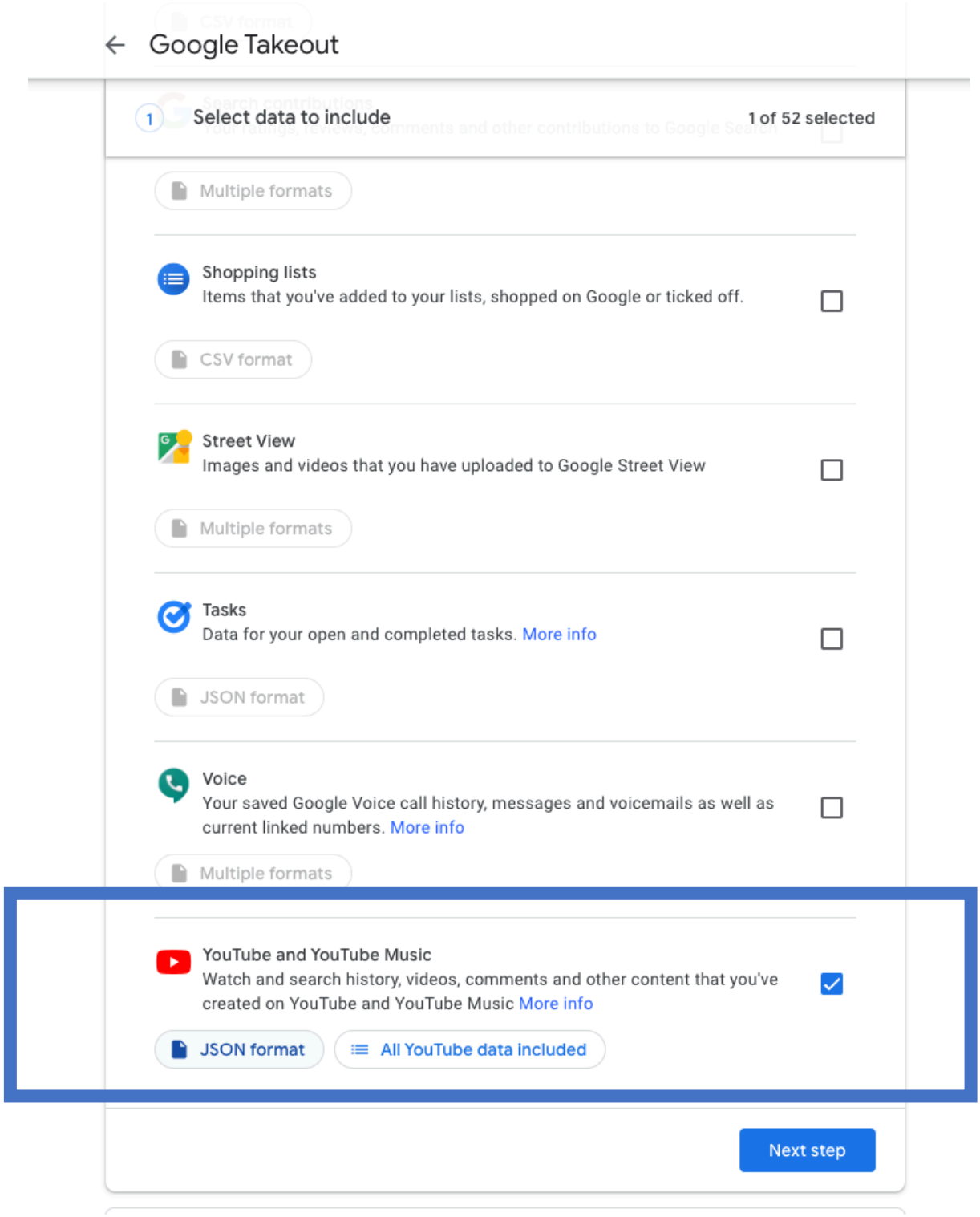


Figure 1 Export YouTube Data and Specify the Data Format and What Data to Export

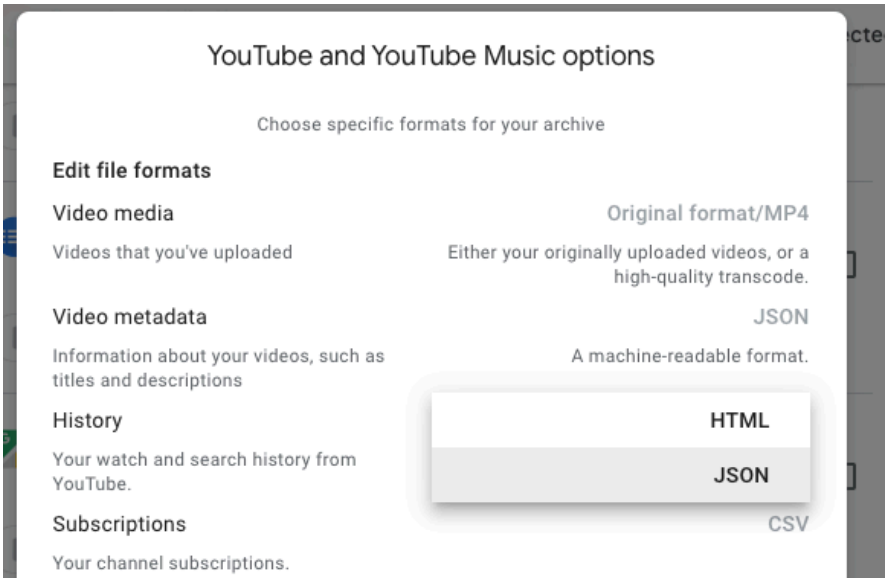


Figure 2 Choose JSON or HTML data format.

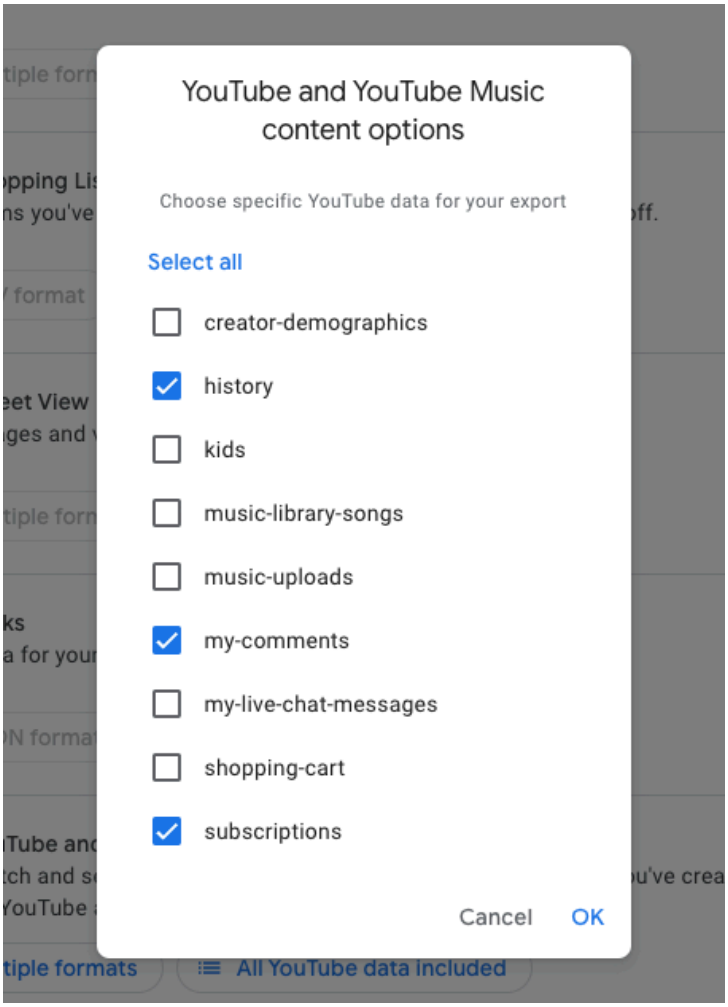


Figure 3 The data export can include different types of content.