Youtube Data Harvesting and Warehousing

|  |  |
| --- | --- |
| **Project Title** | **YouTube Data Harvesting and Warehousing using SQL and Streamlit** |
| **Skills take away From This Project** | **Python scripting, Data Collection, Streamlit, API integration, Data Management using SQL** |
| **Domain** | **Social Media** |

**Problem Statement:**

The problem statement is to create a Streamlit application that allows users to access and analyze data from multiple YouTube channels. The application should have the following features:

1. Ability to input a YouTube channel ID and retrieve all the relevant data (Channel name, subscribers, total video count, playlist ID, video ID, likes, dislikes, comments of each video) using Google API.
2. Ability to collect data for up to 10 different YouTube channels and store them in the data lake by clicking a button.
3. Option to store the data in a MYSQL or PostgreSQL.
4. Ability to search and retrieve data from the SQL database using different search options, including joining tables to get channel details.

**Approach:**

1. **Set up a Streamlit app:** Streamlit is a great choice for building data visualization and analysis tools quickly and easily. You can use Streamlit to create a simple UI where users can enter a YouTube channel ID, view the channel details, and select channels to migrate to the data warehouse.
2. **Connect to the YouTube API:** You'll need to use the YouTube API to retrieve channel and video data. You can use the Google API client library for Python to make requests to the API.
3. **Store and Clean data :** Once you retrieve the data from the YouTube API, store it in a suitable format for temporary storage before migrating to the data warehouse. You can use pandas DataFrames or other in-memory data structures.
4. **Migrate data to a SQL data warehouse:** After you've collected data for multiple channels, you can migrate it to a SQL data warehouse. You can use a SQL database such as MySQL or PostgreSQL for this.
5. **Query the SQL data warehouse:** You can use SQL queries to join the tables in the SQL data warehouse and retrieve data for specific channels based on user input. You can use a Python SQL library such as SQLAlchemy to interact with the SQL database.
6. **Display data in the Streamlit app:** Finally, you can display the retrieved data in the Streamlit app. You can use Streamlit's data visualization features to create charts and graphs to help users analyze the data.

Overall, this approach involves building a simple UI with Streamlit, retrieving data from the YouTube API, storing the data SQL as a warehouse, querying the data warehouse with SQL, and displaying the data in the Streamlit app.