Cấu trúc thư mục bài nộp

Bài nộp của nhóm gồm 3 folder **Slides**, **Code**, và **Dashboard** như sau:

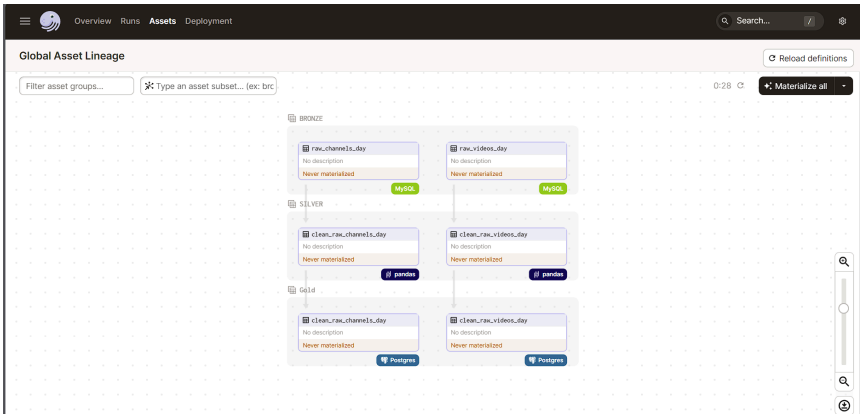
- **Slides**: folder chứa slide thuyết trình
  - Slides.pdf: slides ở định dạng pdf
- **Code**: folder chứa code dùng để phân tích

  Quá trình chạy các file

  1. **Youtube-CollectingData.ipynb**
     a. Mục tiêu: dùng để thu thập dữ liệu về video và channel trên Youtube.
     b. Output: các file csv chứa dữ liệu về video (*videos_day_1.csv*, *videos_day_2.csv, videos_day_3.csv...*) và các file csv chứa dữ liệu về channel (*channels_day_1.csv, channels_day_2.csv, channels_day_3.csv...*) được lưu ở folder **Youtube-CollectingData-Data**
  2. **ProccessingConflictData.ipynb**
     a. Mục tiêu: dùng để đồng nhất dữ liệu giữa các file csv video do có những file bị thiếu thông tin về 'duration', 'description' và 'tag'.
     b. Input: các file csv video trong folder **Youtube-CollectingData-Data**
     c. Output: file *dataNew.csv* trong folder **ProcessingConflictData-Data**
  3. **Youtube-Preprocessing_1.ipynb**
     a. Mục tiêu: dùng để gán thể loại (Genre) cho các video
     b. Input: file *dataNew.csv* trong folder **ProcessingConflictData-Data**
     c. Output: file *finalData.csv* trong folder **Youtube-Preprocessing-Data**
  4. **Youtube-Preprocessing_2.ipynb**
     a. Mục tiêu: dùng để gán thể loại (Genre) cho các video nhưng còn thiếu ở lần Preprocessing_1
     b. Input: file *finalData.csv* trong folder **Youtube-Preprocessing-Data**
     c. Output: file *finalData_1.csv* trong folder **Youtube-Preprocessing-Data**

5. **Thư mục Pipeline**
    a. Mục tiêu:
    - Run docker: docker compose , docker build

    - Import Data vào MySQL: file import_data.py
    - Run Pipeline: Run dagster dev -f assets/one_table.py



    Click Materialize all để thực hiện các assets đưa dữ liệu vào PostgreSQL.

    - Bởi vì sử dụng localhost nên sẽ xuất ra file để các thành viên khác thực hiện phân tích. File
    b. Input: *finalData_1.csv* trong folder **Youtube-Preprocessing-Data** và *channels_day_1.csv, channels_day_2.csv, channels_day_3.csv, channels_day_4.csv, channels_day_5.csv* trong folder **Youtube-CollectingData-Data**
    c. Output: File *clean_raw_channels.csv* và *clean_raw_videos.csv*

6. **Youtube-Analysis-Ranked.ipynb**
    a. Mục tiêu: Phân cấp rank cho channel và video
    b. Input: file *clean_raw_channels.csv* và file *clean_raw_videos.csv* trong **Thư mục Pipeline** ở bước trên.

c. Output: file *finalData_ranked.csv* lưu trong folder **Youtube-Analysis-Ranked-Data**, cũng là file data chuẩn cuối cùng được dùng cho các quá trình phân tích.

7. **duration_and_view.ipynb**
   a. Mục tiêu: Tìm mối tương quan giữa thời lượng video (Duration) và số lượt xem (View)
   b. Input: file *finalData_ranked.csv* trong folder **Youtube-Analysis-Ranked-Data**
   c. Output: Biểu đồ scatter plot thể hiện mối tương quan giữa thời lượng video (Duration) và số lượt xem (View)

8. **NLP_Titles.ipynb**
   a. Mục tiêu: Phân tích các tiêu đề (Title) video bằng tiếng Anh và đếm số lượng từ xuất hiện nhiều nhất (bao gồm từ mang tính tích cực và tiêu cực)
   b. Input: file *finalData_ranked.csv* trong folder **Youtube-Analysis-Ranked-Data**
   c. Output:
      + Biểu đồ Treemap và biểu đồ Word Cloud thể hiện các từ xuất hiện nhiều nhất trong các tiêu đề tiếng Anh.
      + file *word_count.csv* trong folder **NLP_Titles_Data**, được dùng cho Dashboard

9. **AnalysisTitlePodcast.ipynb**
   a. Mục tiêu: Phân tích độ độ dài title, một số khía cạch về format của title đối với video podcast như: Title có có chứa tên channel, title được đặt ở dạng câu hỏi, Trích xuất keyword đối với mỗi title podcast và xác định xu hướng dùng từ tích cực và tiêu cực
   b. Input: file *finalData_ranked.csv* trong folder **Youtube-Analysis-Ranked-Data**
   c. Output:
      + Biểu đồ tròn thể hiện các phần trăm độ dài title.
      + Biểu đồ cột cho thấy tỉ lệ số số video có tên channel trong title theo các phân khúc
      + Biểu đồ tròn tỉ lệ channel tồn tại video có title chứa tên channel theo các phân khúc
      + Biểu đồ cột cho thấy tỉ lệ số số video có title là câu hỏi
      + Biểu đồ tròn tỉ lệ channel tồn tại video có title là câu hỏi
      + Biểu đồ tròn tỷ lệ sử dụng từ tiêu cực trong title theo các phân khúc.

10. **File chromedriver.exe** để chạy trình duyệt Chrome ảo

- **Dashboard**: folder chứa dashboard powerBI
  - youtube_dashboard.pbix
    - Cách chạy file(với Windows):
      1. Cài đặt phần mềm Power BI Desktop từ Microsoft Store

2. Click chuột phải lên file youtube_dashboard.pbix -> Chọn Open with -> Chọn Power BI Desktop