THIẾT KẾ VÀ THU THẬP DATASET

PHÂN TÍCH XU HƯỚNG NỘI DUNG TRÊN TIKTOK VÀ YOUTUBE

Học phần: ADY201m – AI, Data Science with Python & SQL
Sinh viên thực hiện:

- QE200009 – Dương Thị Mỹ Tâm
- QE200083 – Trần Vân Khánh

REPORT 2: DATA COLLECTION & PRE-PROCESSING

1. KIẾN TRÚC HỆ THỐNG DỮ LIỆU (SYSTEM ARCHITECTURE)

Để đảm bảo tính tự động hóa và khả năng mở rộng (Scalability), nhóm đã xây dựng một luồng dữ liệu (Data Pipeline) hoàn chỉnh chạy trên nền tảng Docker. Kiến trúc bao gồm 3 lớp chính:

Ingestion Layer (Thu thập): Sử dụng Python scripts.

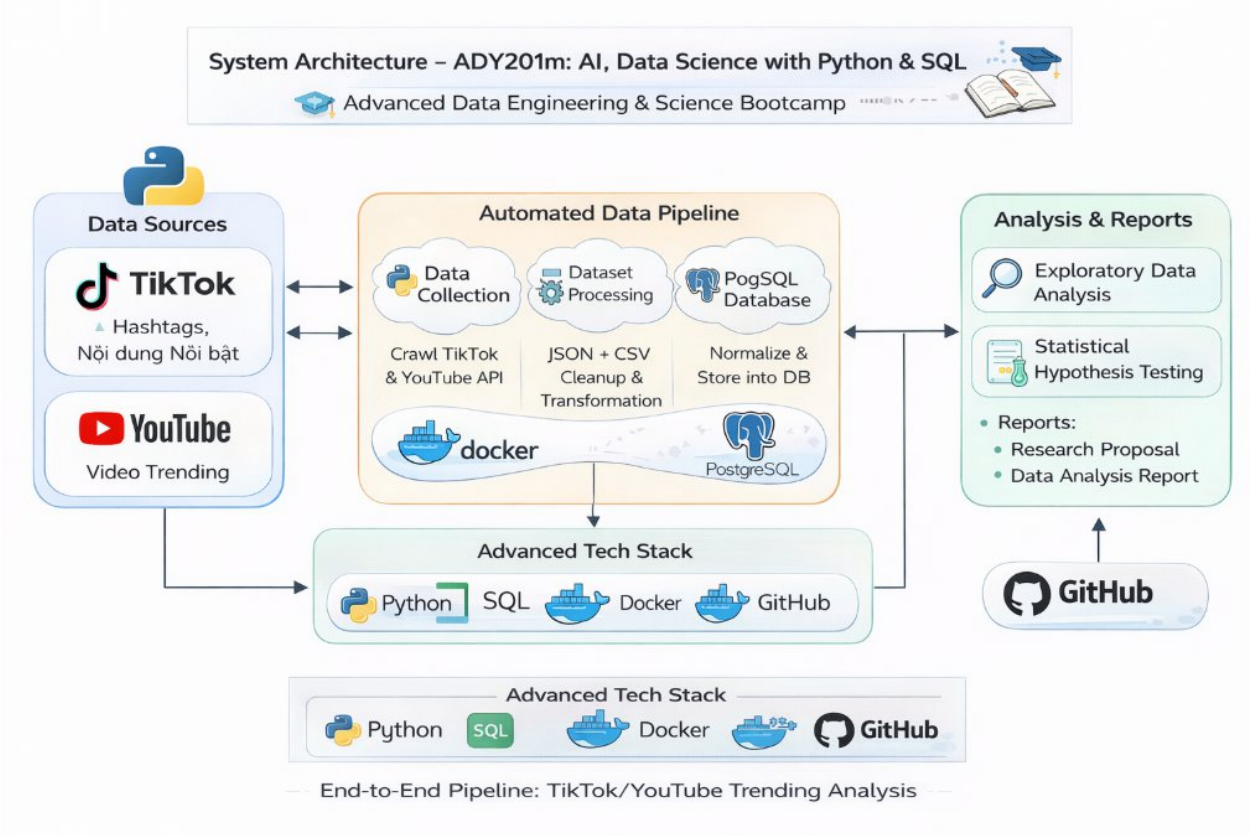
Youtube_api.py: Kết nối YouTube Data API v3 để lấy dữ liệu chính thống.

Tiktok_crawler.py: Sử dụng thư viện Playwright để giả lập trình duyệt và lấy dữ liệu từ TikTok (do TikTok không cung cấp API công khai).

Storage Layer (Lưu trữ - Data Lake): Sử dụng MinIO (Object Storage) để lưu trữ dữ liệu thô dưới dạng file JSON. Đây là nơi chứa các file youtube_trending_*.json và tiktok_dataset_*.json nguyên bản.

Processing & Serving Layer (Xử lý & Kho dữ liệu)

PostgreSQL: Cơ sở dữ liệu quan hệ lưu trữ dữ liệu sạch (Structured Data) phục vụ cho các truy vấn SQL và phân tích sau này.



2. THU THẬP DỮ LIỆU (DATA COLLECTION)

Nhóm đã thực hiện thu thập dữ liệu với chi tiết như sau:

2.1. Dữ liệu YouTube

Nguồn: YouTube Data API v3.

File dữ liệu: youtube_trending_20260119_230758.json

Số lượng mẫu: 5939 (Top Trending tại thời điểm thu thập).

Đặc điểm: Dữ liệu có cấu trúc tốt (Structured), các trường số liệu (views, likes) đã là số nguyên (Integer).

Các trường quan trọng: video_id, title, publishedAt (UTC Time), view_count, like_count.

2.2. Dữ liệu TikTok

Nguồn: Web Scraping (Playwright).

File dữ liệu: tiktok_dataset_merged (1).json

Số lượng mẫu: 4257 video.

Thách thức xử lý:

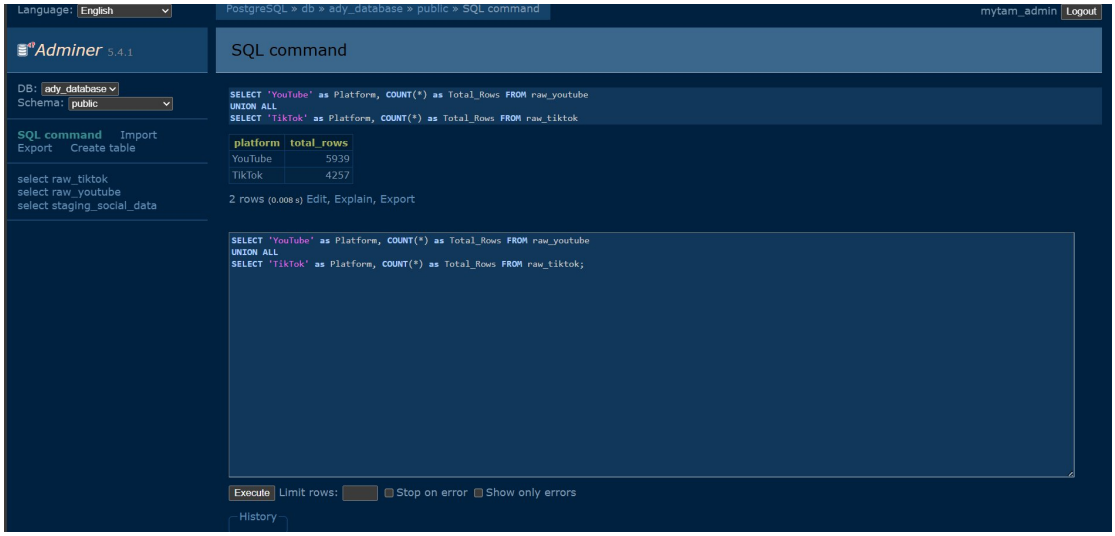
Dữ liệu likes và comments ở dạng rút gọn (ví dụ: "10.5K", "1.2M") → Cần bước xử lý chuyển đổi.

Dữ liệu thời gian ở dạng Unix Timestamp hoặc Text → Cần đồng bộ.

Các trường quan trọng: id, desc (caption), stats (likes, shares, comments), is_trending, has_clickbait (Nhãn nhị phân 0/1).

3. QUY TRÌNH XỬ LÝ DỮ LIỆU (ETL PROCESS)

Quy trình ETL (Extract - Transform - Load) được thực hiện để làm sạch và chuẩn hóa dữ liệu trước khi đưa vào Database.



3.1. Data Transformation (Chuyển đổi dữ liệu)

Metric Normalization (Chuẩn hóa số liệu):

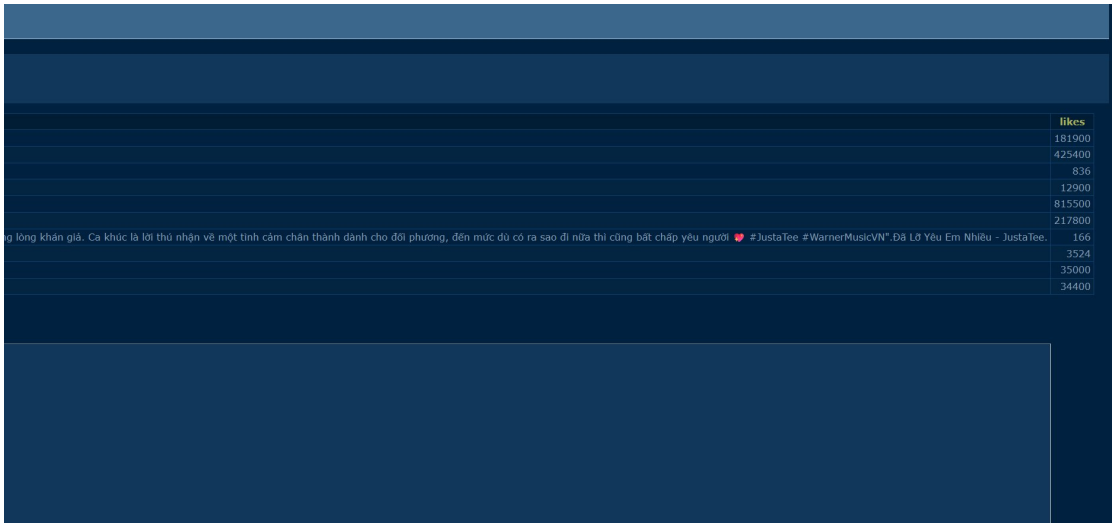
Vấn đề: TikTok trả về dữ liệu dạng chuỗi ("1.5K").

Giải pháp: Viết hàm clean_metric() để loại bỏ ký tự "K", "M" và nhân với hệ số tương ứng (1,000 hoặc 1,000,000).

Kết quả: "1.5K" → 1500 (Integer).

Data Enrichment (Làm giàu dữ liệu):

Thêm cột platform: Gán nhãn "YouTube" hoặc "TikTok" để phân biệt nguồn.



3.2. Data Loading (Tải vào Database)

Dữ liệu sạch được lưu vào bảng staging_social_data trong PostgreSQL với Schema (Cấu trúc) như sau:

Tên cột	Kiểu dữ liệu	Mô tả
content_id	VARCHAR	Mã định danh video/hashtag
platform	TEXT	Nền tảng (YouTube/TikTok)
title	TEXT	Tiêu đề hoặc Caption của video
likes	INTEGER	Số lượt thích (đã chuẩn hóa thành số)
upload_date	TIMESTAMP	Thời gian đăng tải video

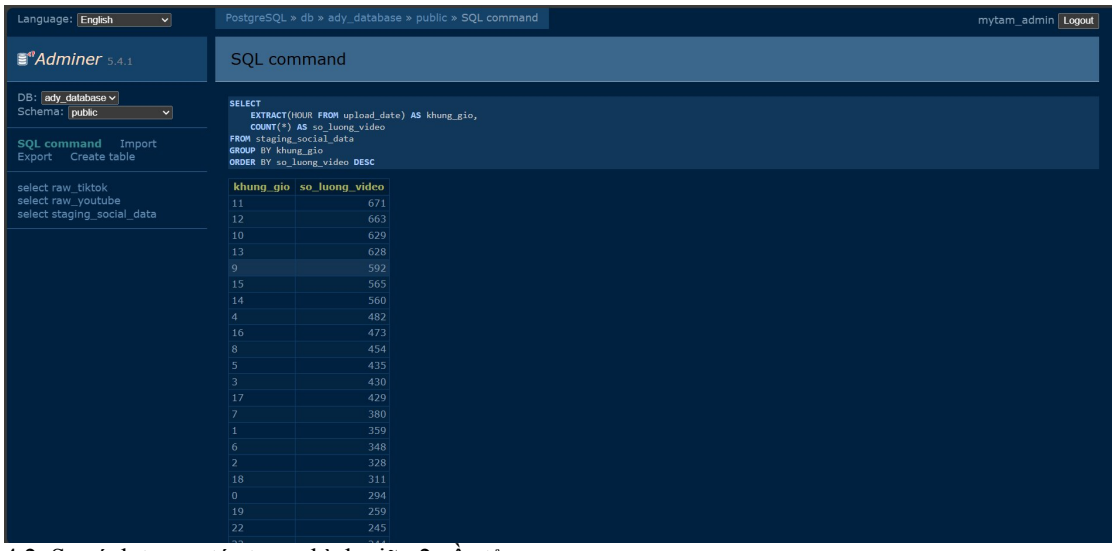
	likes	upload_date
	44	2026-01-26 23:49:12
	38	2026-01-26 23:49:07
	129	2026-01-26 20:29:26
y of her love to move forward 💔" #dylanwang #王鹤棣 #wanghedi #dylanwangedit #wanghediedit".Lyric Edit 🎶	35	2026-01-26 20:27:48
	0	2026-01-26 17:00:43
	21	2026-01-26 15:12:07
	53	2026-01-26 14:54:08
	63	2026-01-26 14:11:31
	49	2026-01-26 13:59:32
nhạc nền - Mie.	1087	2026-01-26 13:37:39
	51	2026-01-26 12:28:19
	40	2026-01-26 12:07:59
	11000	2026-01-26 12:04:57
	60	2026-01-26 12:03:47
	569	2026-01-26 11:54:52

4. KIỂM TRA & ĐÁNH GIÁ (VALIDATION)

Để kiểm chứng độ tin cậy của dữ liệu trong Database, nhóm thực hiện các câu lệnh SQL (Validation Queries):

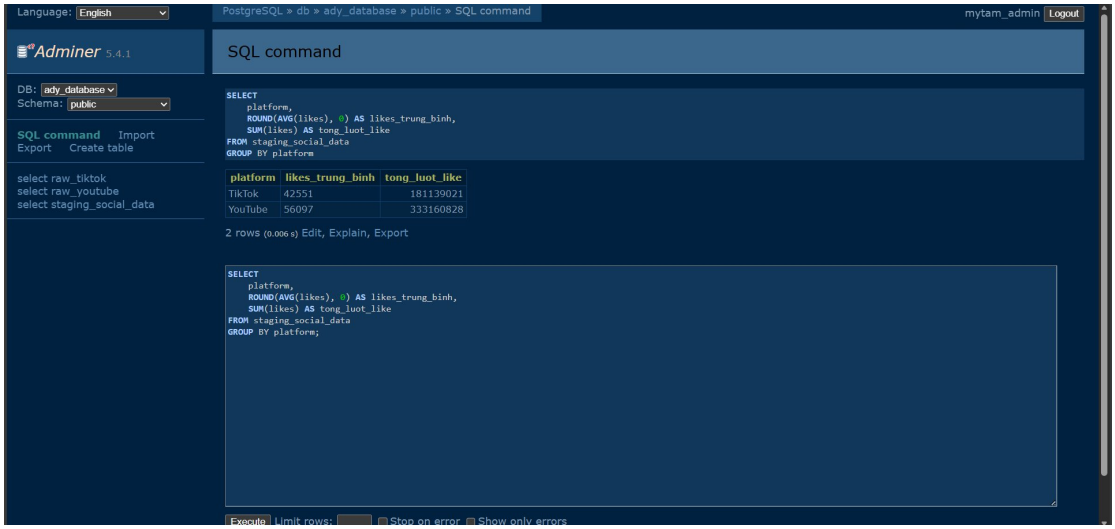
4.1. Kiểm tra phân bố Video theo khung giờ (Golden Hour)

Mục đích: Xác định khung giờ nào có nhiều video được đăng nhất.



4.2. So sánh tương tác trung bình giữa 2 nền tảng

Mục đích: Kiểm tra xem TikTok hay YouTube có tỷ lệ tương tác (Likes) cao hơn.


Language: English
PostgreSQL » db » ady_database » public » SQL command
mytam_admin Logout
Adminer 5.4.1
DB: ady_database
Schema: public
SQL command Import
Export Create table
select raw_tiktok
select raw_youtube
select staging_social_data
SQL command
SELECT
platform,
ROUND(AVG(likes), 0) AS likes_trung_binh,
SUM(likes) AS tong_luot_like
FROM staging_social_data
GROUP BY platform
platform likes_trung_binh tong_luot_like
TikTok 42551 181139021
YouTube 56097 333160828
2 rows (0.006 s) Edit, Explain, Export
SELECT
platform,
ROUND(AVG(likes), 0) AS likes_trung_binh,
SUM(likes) AS tong_luot_like
FROM staging_social_data
GROUP BY platform;
Execute Limit rows: Stop on error Show only errors