

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 Văn 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.

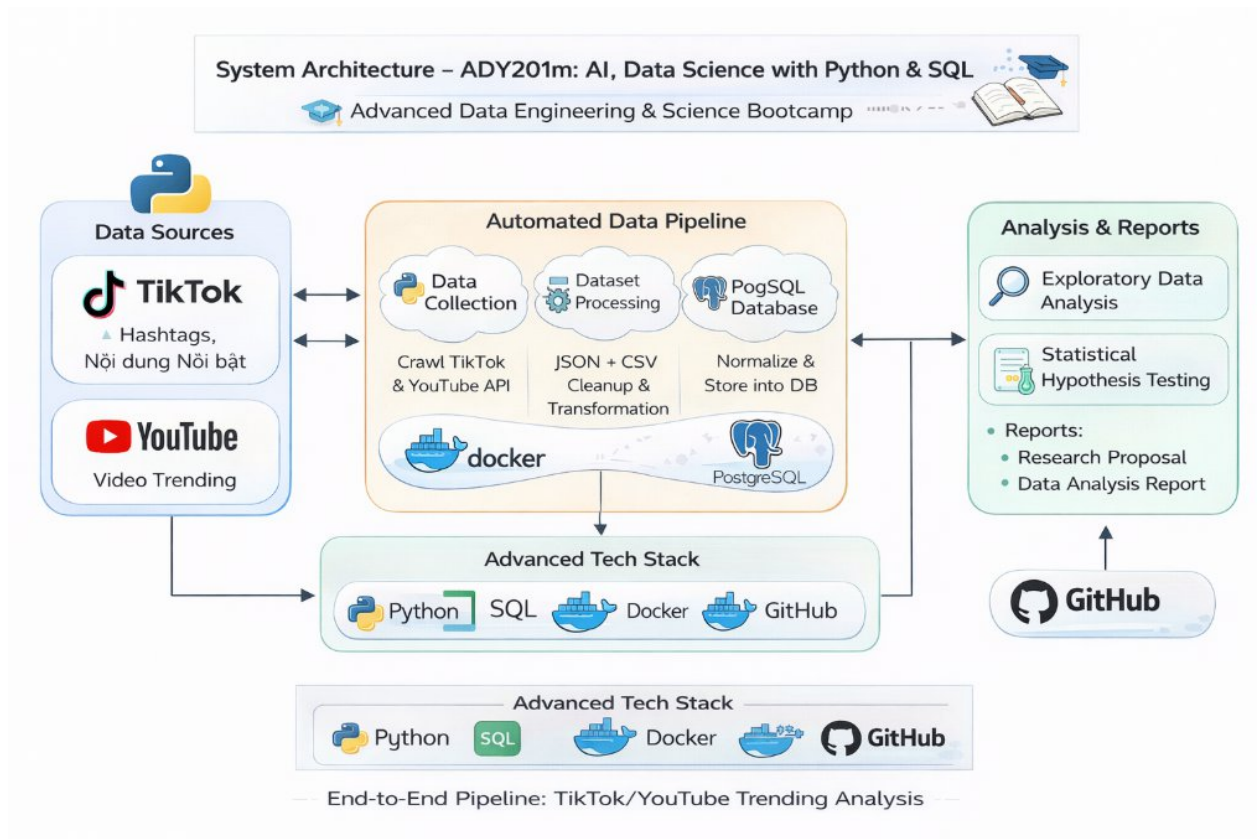
Youtube_api.py: Kết nối YouTube Data API v3 để lấy dữ liệu chính thống.

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 (Nhân 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.

Language: English

PostgreSQL » db » ady_database » public » SQL command

mytam_admin Logout

Adminder 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 'YouTube' as Platform, COUNT(*) as Total_Rows FROM raw_youtube  
UNION ALL  
SELECT 'TikTok' as Platform, COUNT(*) as Total_Rows FROM raw_tiktok;
```

platform	total_rows
YouTube	5939
TikTok	4257

2 rows (0.008 s) Edit, Explain, Export

```
SELECT 'YouTube' as Platform, COUNT(*) as Total_Rows FROM raw_youtube  
UNION ALL  
SELECT 'TikTok' as Platform, COUNT(*) as Total_Rows FROM raw_tiktok;
```

Execute Limit rows: Stop on error Show only errors

History

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.

DB: ady_database Schema: public

SQL command Import Export Create table

select raw_tiktok
select raw_youtube
select staging_social_data

```
SELECT platform, title, likes  
FROM staging_social_data  
WHERE platform = 'TikTok'  
LIMIT 10;
```

platform	title	likes
TikTok	Video TikTok từ RAP VIỆT (@rapvietmusic.official): "Những bản tình ca bất hủ, lâu rồi nghe lại vẫn hay như ngày nào #justatee #music #xuhuong #fyf" Justatee, 2AM	1.5K
TikTok	Video TikTok từ JustaTee (@chotchachi): "Khám phá hành trình JustaTee đối diện với cuộc sống làm bố. Cùng nghe nhạc và câu chuyện gia đình thú vị! #EZPAPA #JustaTee	1.5K
TikTok	Video TikTok từ Những Bản Nhạc Nghe Là Nghiện (@nhacngheien): "Nghe ngay bản cover 'Crying Over You' của Justatee để cảm nhận nỗi lòng. Tham gia cùng chúng tôi! #n	1.5K
TikTok	Video TikTok từ JustaTee (@chotchachi): "Khám phá ca khúc 'Forever Alone' của JustaTee, một bản tình ca đầy cảm xúc về nỗi cô đơn. #ForeverAlone #JustaTee #NhacJust	1.5K
TikTok	Video TikTok từ Ngô Đức Duy (@duyvy.real.channel): "Cuối cùng tôi cũng gặp Justatee! Những cảm xúc thật sự thú vị khi chạm mặt anh ấy. Hãy cùng xem những khoảnh k	1.5K
TikTok	Video TikTok từ Chúng Ta Của Hiện Tại (@chungta.cuahientai): "Justa Tee nhìn cuộc sống như văn thơ trong âm nhạc. Con mưa cuối, phiên bản hoàn hảo. #FanEnt #chung	1.5K
TikTok	Video TikTok từ Warner Music Vietnam (@warnermusicvn): "Chúc mừng bản R&B mộng mơ 'Đã Là Yêu Em Nhiều' của @chotchachi đã chính thức cán mốc 200 triệu lượt x	1.5K
TikTok	Video TikTok từ Vạch Là Tim Sao (@vachlatimsao): "Giám đốc âm nhạc JustaTee gây lo ngại với biểu hiện lạ ở miệng. Khán giả bày tỏ sự quan tâm đến sức khỏe của anh. #	1.5K
TikTok	Video TikTok từ Minh Thành (@beanbuncha): "She Never Know Remix - Quất Lên Em Oi #xuhuong #justatee" touliver remix. nhạc nền - Minh Thành.	1.5K
TikTok	Video TikTok từ babygirl9999 (@babygirl9999): "Enjoy the mesmerizing lyrics video of Thăng Diên by JustaTee, a trending song that will captivate you. Download "Thăng Di	1.5K

10 rows (0.003 s) Edit, Explain, Export

```
SELECT platform, title, likes  
FROM staging_social_data  
WHERE platform = 'TikTok'  
LIMIT 10;
```


	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.

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
    EXTRACT(HOUR FROM upload_date) AS khung_gio,
    COUNT(*) AS so_luong_video
FROM staging_social_data
GROUP BY khung_gio
ORDER BY so_luong_video DESC
```

khung_gio	so_luong_video
11	671
12	663
10	629
13	628
9	592
15	565
14	560
4	482
16	473
8	454
5	435
3	430
17	429
7	380
1	359
6	348
2	328
18	311
0	294
19	259
22	245
20	210


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 (0.006s) [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