

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.

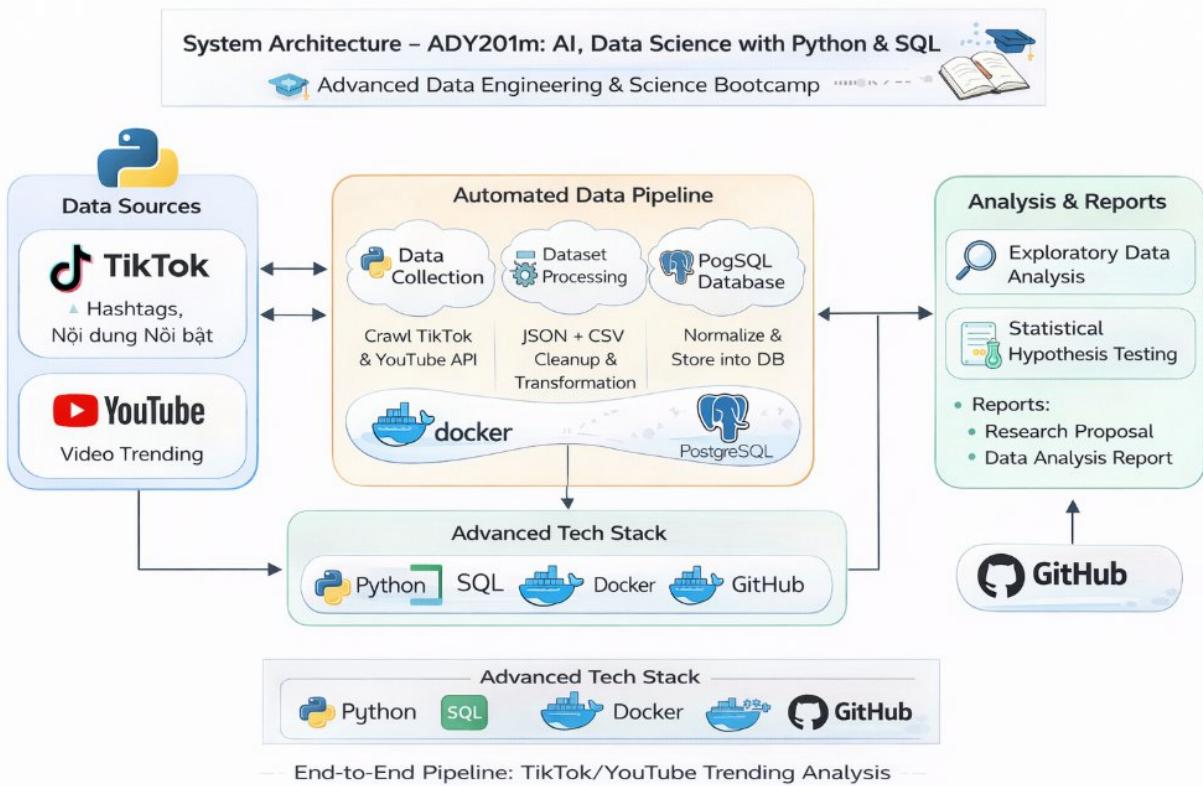
Youtube_api.py: Kết nối YouTube Data API v3 để lấy dữ liệu chính thống.

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:

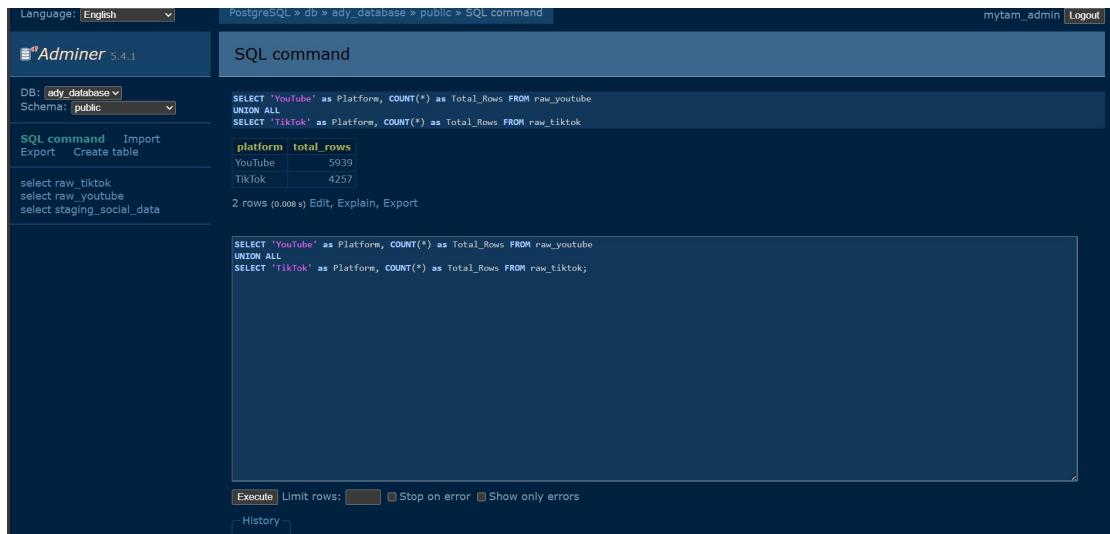
Dữ liệu likes và comments ở dạng rút gọn (ví dụ: "10.5K", "1.2M") → Cần bước xử lý chuyên đổi.

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.



The screenshot shows the Adminer 5.4.1 interface connected to the 'ady_database' schema. Two SQL commands are run:

```
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


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

3.1. Data Transformation (Chuyển đổi dữ liệu)

Metric Normalization (Chuẩn hóa số liệu):

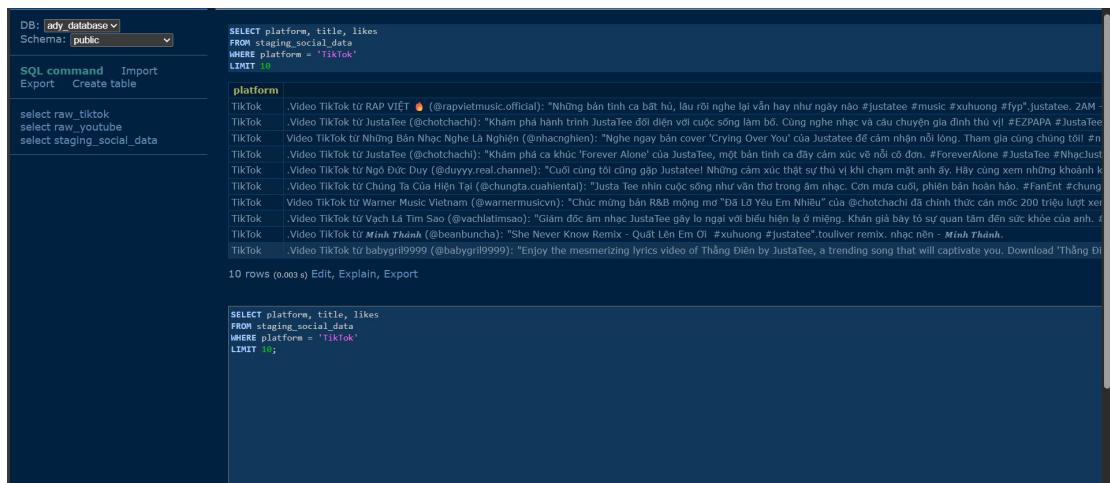
Vấn đề: TikTok trả về dữ liệu dạng chuỗi ("1.5K").

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.



The screenshot shows the Adminer 5.4.1 interface connected to the 'ady_database' schema. A query is run to select data from staging_social_data where platform is 'TikTok' and limit it to 10 rows:

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

The results show various TikTok video titles and like counts, such as:

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 #fyp".Justatee. 2AM	
TikTok	.Video TikTok từ Justatee (@chotchach): "Khám phá hành trình Justatee đối diện với cuộc sống lầm bẩm. Cùng nghe nhạc và câu chuyện gia đình thú vị #EZPAPA #Justatee	
TikTok	.Video TikTok từ Những Bản Nhạc Nghe Lá Nghiện (@nhacnghe): "Nghe ngay bản cover 'Crying Over You' của Justatee để cảm nhận nỗi lòng. Tham gia cộng đồng tại #EZPAPA #Justatee	
TikTok	.Video TikTok từ Justatee (@chotchach): "Khám phá ca khúc 'Forever Alone' của Justatee, một bản tình ca đầy cảm xúc và nội cõi đơn. #ForeverAlone #Justatee #NhacJust	
TikTok	.Video TikTok từ Ngô Đức Duy (@duyyreal.channel): "Cuối cùng tôi cũng gặp Justatee! Những cảm xúc thật sự thú vị khi chạm mặt anh ấy. Hãy cùng xem những khoảnh khắc này #EZPAPA #Justatee	
TikTok	.Video TikTok từ Chung Ta Của Hiện Tại (@chungta.cuahtientai): "Just Tee nhí nhảnh sống như vẫn thở trong âm nhạc. Cơn mưa cuối, phiên bản hoàn hảo. #FanEnt #chungta	
TikTok	.Video TikTok từ Warner Music Vietnam (@warnermusicvn): "Chúc mừng bản R&B mông má "Đã Lỡ Yêu Em Nhìu" của @chotchach đã chính thức cán mốc 200 triệu lượt xem #EZPAPA #Justatee	
TikTok	.Video TikTok từ Vạch Lá Tim Sao (@vachlatimsao): "Giảm 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. #EZPAPA #Justatee	
TikTok	.Video TikTok từ Minh Thành (@bebanbuncha): "She Never Know Remix - Quất Lèn Em Oi #xuhuong #Justatee".touliver remix, nhạc nền - <i>Minh Thành</i> .	
TikTok	.Video TikTok từ babygril9999 (@babycat9999): "Enjoy the mesmerizing lyrics video of Tháng Đen by Justatee, a trending song that will captivate you. Download 'Tháng Đen' #EZPAPA #Justatee	

likes
181900
425400
836
12900
815500
217800
166
3524
35000
34400

ng lòng khán giả. Ca khúc là lời thư nhận về một tình cảm chân thành dành cho đối phương, đến mức dù có ra sao đi nữa thì cũng bắt chấp yêu người. ❤ #JustaTee #WarnerMusicVN". Đà Lò Yêu Em Nhiều - JustaTee.

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

SQL command		mytam_admin Logout
<pre>SELECT content_id, platform, title, likes, upload_date FROM staging_social_data ORDER BY upload_date DESC LIMIT 15</pre>		
content_id	platform	
wanghedi	TikTok	Video TikTok từ 💕BerryMcDidi💕 (@berrymciddi): "婀娜多姿の舞を披露してますよ♪ #wanghedi #dylanwang #妩媚多姿 #theinn #theinn2026".som original - aveerplayer-Đô.
wanghedi	TikTok	Video TikTok từ 💕DaDa.Love.DIDI💕 (@napada2523): "dyylanwang王鹤棣 #wanghedi #dyylanwang #didi".som original - Đô.
wanghedi	TikTok	Video TikTok từ 💕..สาวน้อยที่รักมัน - Rathawit.
wanghedi	TikTok	Video TikTok từ S.editzz15 (@shivalingbhradar24): "My chaotic babies #dylanwang #ballu #dliu #wanghedi #viral".bản lu xin su. original sound - S.editzz15.
wanghedi	TikTok	Video TikTok từ ✨愛 Sonja ✨ (@sheleavesalittlesparkle): ""Love Letter of Devotion" 🌸 A little story I came up with ❤️ As he writes his love down on a piece of paper, he knows, with his wife Alex Warren "Ordinary (Wedding Version)" Ordinary (Wedding Version) - Alex Warren.
vittthon	TikTok	Video TikTok từ T_ppt (@tp.tpt): "Ra vlog lẻ đi ac oii #siroofficial #vittthon #rovavit".nhạc nền - Aris Music.
wanghedi	TikTok	Video TikTok từ Võ tri thich gi đang này (@mae_mel09): "Dùng số trưởng của ánh #wanghedi #vuonghadde #vuonghadde_ #wanghedi #dylanwang #quantrothanhtuong #王鹤棣".nhạc nền - Florence wanghedi TikTok từ HmeeDidi (@hmeedidi): "Nhó nhau thật nhiều #wanghedi王鹤棣 #vuonghadde#wanghedi #wanghedi王鹤棣 ❤️.nhạc nền - Huệ bông 's.
wanghedi	TikTok	Video TikTok từ tn (@ruanshizhuer): "#vuonghadde #dylanwang #ruanshizhuer #wanghedi".Ruanshizhuer 海边探戈 (完整版) - 王鹤棣/王齐铭/朴董.
wanghedi	TikTok	Video TikTok từ 🌟 (@junhee009): "One bà chủ quán chúc chúc mại dâm đó #ballu #vuonghadde #wanghedi #vuonghadde_xth #viral #tiktok" nhạc nền - NYC CAPUT - NYC CAPUT 🌟.
wanghedi	TikTok	Video TikTok từ ARJA♪ ♪ ~BMY~ (@aria_sv26): "Anh trân trọng cái bằng của chí vẻ hòn cát kho báu☀ Sợ nó bị ngã theo bá vỉ bá lờ vấp xiu.. lấy tay giữ báng ... sợ hứa nên vác vào trong đai".
wanghedi	TikTok	Video TikTok từ dylan_wang_fr (@dylan_wang_fr): "在那个地方相遇 #dyylanwang王鹤棣 #wanghedi #dyylanwang #parisfashionweek".11 Binaural Beats 1 Hz - August Son Productions.
wanghedi	TikTok	Video TikTok từ DidI ❤️ (@di_did18): "#wanghedi #dyylanwang #vuonghadde #vuonghadde #virolikotk".DidI ❤️ (2005) - Yu Isobe.
vittthon	TikTok	Video TikTok từ Huỳnh Thiên Tân (@moenhandn): "Không còn demo nữa, cãi dỗi mò em @VittHon. ngày nào giờ đã chính thức cưới em #siroofficial #vittthon".sura - Bibin.
vittthon	TikTok	Video TikTok từ selina5 🌸 (@selina_520...): "dẫn lu dمن the là cung 😊 #selina_250 🌸 #siroofficial #vittthon #rovavit #xuhuong".nhạc nền - Boy Thủ Môn? - Klett?.
wanghedi	TikTok	Video TikTok từ P h u o n g 🌸 (@chutheo1812): "Những ngày đó body🔥 #vuonghadde #dyylanwang #wanghedi #王鹤棣 #dylan".original sound - Ruido Negro 🌸.

	likes	upload_date
y of her love to move forward 🖤 #dylanwang #王鹤棣 #wanghedidi #dylanwangedit #wanghediedit".Lyric Edit 47	44	2026-01-26 23:49:12
	38	2026-01-26 23:49:07
	129	2026-01-26 20:29:26
	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.

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
23	244

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

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

The screenshot shows the Adminer 5.4.1 web-based PostgreSQL client. The top navigation bar includes 'Language: English', 'PostgreSQL » db » ady_database » public > SQL command', and user information 'mytam_admin' and 'Logout'. The main area is titled 'SQL command' and contains a query window with the following SQL code:

```
SELECT
    platform,
    ROUND(AVG(likes), 0) AS likes_trung_binh,
    SUM(likes) AS tong_luot_like
FROM staging_social_data
GROUP BY platform;
```

Below the query window is a table displaying the results:

platform	likes_trung_binh	tong_luot_like
TikTok	42551	181139021
YouTube	56097	333160828

A status message at the bottom indicates '2 rows (0.006s)' followed by links for 'Edit', 'Explain', and 'Export'. At the bottom of the interface are buttons for 'Execute', 'Limit rows:', 'Stop on error', and 'Show only errors'.