



ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

XỬ LÝ DỮ LIỆU MẠNG XÃ HỘI DỰA TRÊN CÔNG NGHỆ GOOGLE CLOUD PLATFORM

GVHD: ThS. Nguyễn Thị Anh Thư

SVTH: 21522019 Âu Trường Giang
21521806 Nguyễn Nguyễn Thành An
21522039 Trần Lê Khánh Hân



1. Giới thiệu
2. Đánh giá
3. Cách thức triển khai
4. Kết luận và hướng phát triển



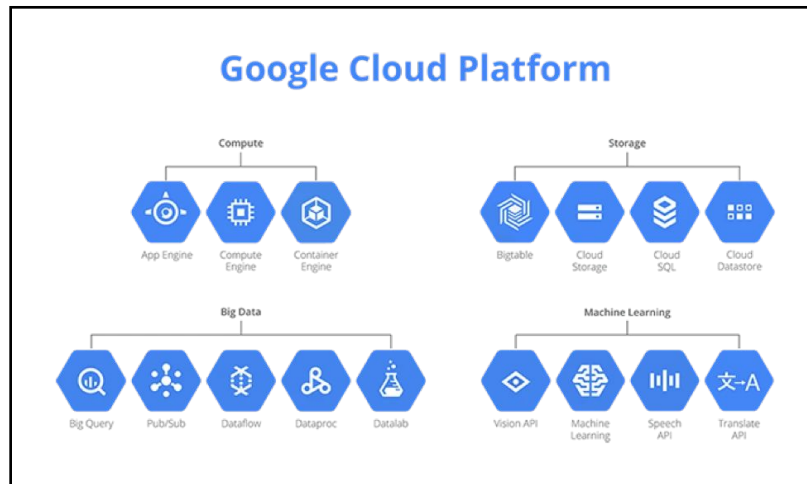
Giới thiệu

- Các công nghệ triển khai dữ liệu lớn trên thế giới: Apache Hadoop, Apache Spark, Tableau,....
- Trong đó, Google Cloud Platform là một trong những sự lựa chọn cho việc xử lý dữ liệu lớn mạng xã hội.
- Google Cloud Platform (GCP) là nền tảng điện toán đám mây do Google cung cấp.



Giới thiệu

Gồm một loạt các dịch vụ được lưu trữ để tính toán, lưu trữ và phát triển ứng dụng.



Hình 1: Các dịch vụ phổ biến của GCP



Đánh giá

a. Ưu điểm

- Tính linh hoạt và mở rộng
- Hiệu suất cao
- Dịch vụ trí tuệ nhân tạo và Machine Learning
- Tính bảo mật
- Hỗ trợ đa nền tảng



b. Nhược điểm

- Chi phí
- Tính phức tạp
- Tương thích
- Khả năng di chuyển dữ liệu



c. Ứng dụng

- Phân tích dữ liệu và kho dữ liệu lớn
- Trí tuệ nhân tạo và Machine Learning
- An ninh
- Internet of Things (IoT)



Cách thức triển khai

I. Data House

1. Lakehouse

- Mô hình kết hợp giữa data lake và data warehouse.
- Trong GCP, có thể triển khai bằng cách kết hợp GCS và BigQuery.



Cách thức triển khai

a. Lựa chọn Dịch vụ GCP phù hợp

- Cloud Storage: lưu trữ dữ liệu gốc.
- BigQuery: nơi truy vấn và phân tích dữ liệu.



Cách thức triển khai

b. Xử lý và Biến đổi Dữ liệu

Cloud Dataflow: xử lý và biến đổi dữ liệu trước khi nạp vào BigQuery hoặc lưu trữ trong Cloud Storage



Cách thức triển khai

c. Xây dựng Lakehouse

Tạo cấu trúc thư mục hợp lý trong Cloud Storage để lưu trữ dữ liệu gốc và dữ liệu đã được tiền xử lý.



Cách thức triển khai

d. Phân tích và Trực quan hóa Dữ liệu

- Sử dụng các công cụ BigQuery, Data Studio,...
- Triển khai kiến trúc lakehouse phức tạp hơn và đòi hỏi kế hoạch và quản lý dữ liệu kỹ thuật.



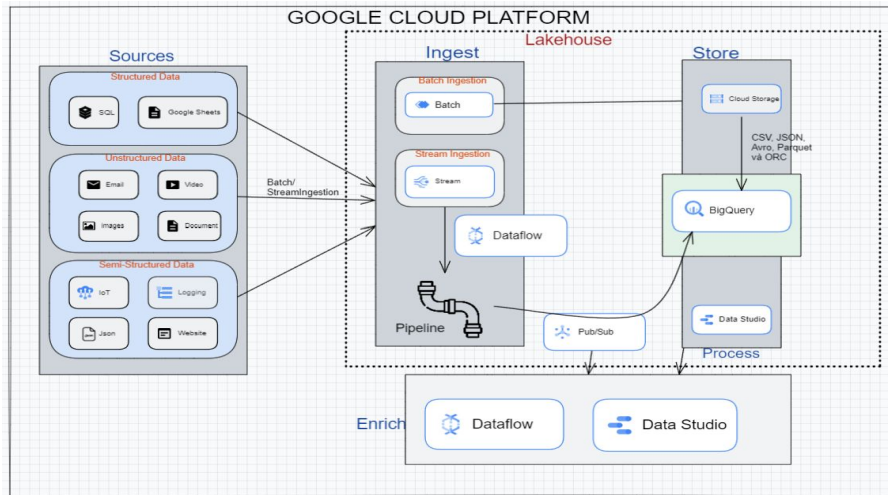
Cách thức triển khai

1. Data House

2. Kiến trúc component

Chia thành ba lớp chính:

- Lớp thu thập dữ liệu
- Lớp lưu trữ dữ liệu
- Lớp xử lý dữ liệu

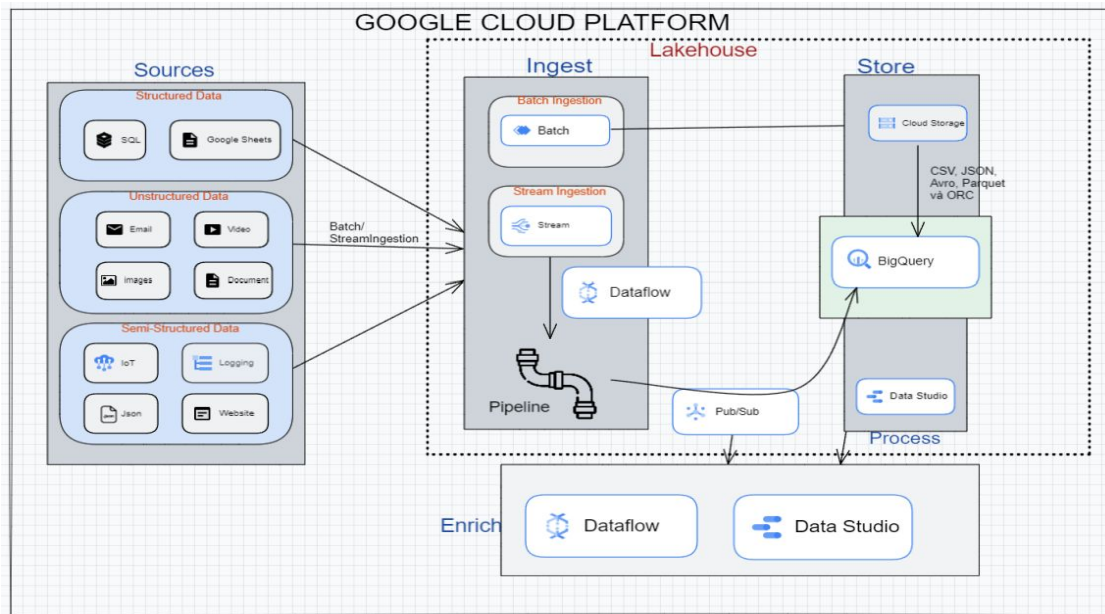


Hình 2: Kiến trúc components



Cách thức triển khai

II. Các components



Hình 2: Kiến trúc components

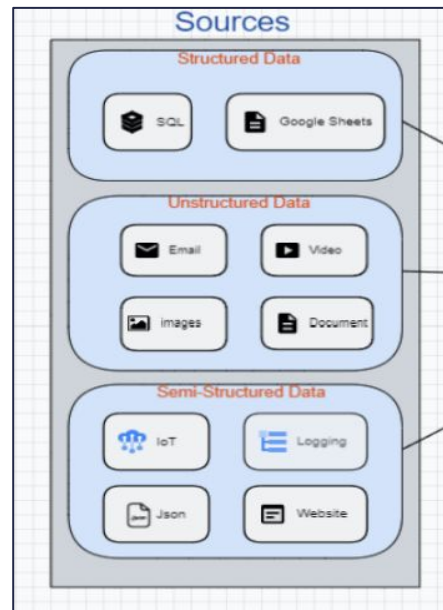


Cách thức triển khai

1. Sources

a. Các loại dữ liệu

- Dữ liệu có cấu trúc (Excel, csdl SQL,...)
- Dữ liệu phi cấu trúc (tệp văn bản, ảnh, video,...)
- Dữ liệu bán cấu trúc (JSON, XML,...)





Cách thức triển khai

1. Sources

b. Thu thập dữ liệu thô

- **Raw data**: dữ liệu thô hoặc không xử lý.
- Chưa qua quá trình biến đổi, làm sạch, phân tích.
- Thường được thu thập tự động từ các hệ thống, thiết bị, hoặc nguồn dữ liệu khác.





Cách thức triển khai

2. Ingest

a. Batch Ingestion

Nhập dữ liệu theo loạt liên quan đến nạp các tập dữ liệu lớn, có giới hạn, không cần xử lý trong thời gian thực.



Cách thức triển khai

2. Ingest

b. Stream Ingestion

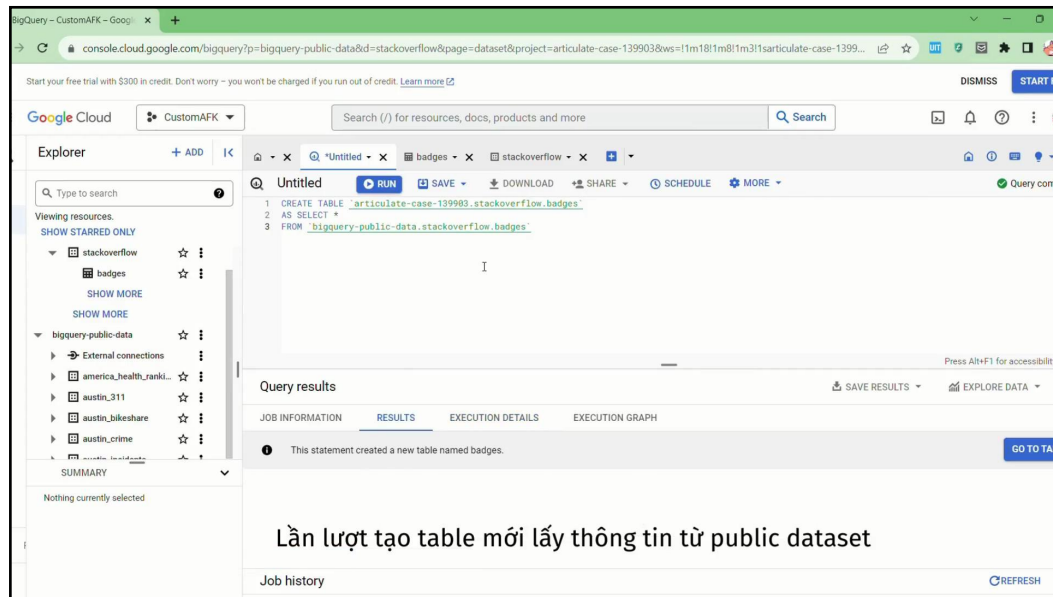
Nhập dữ liệu theo luồng dùng cho yêu cầu phân tích lượng dữ liệu liên tục.

Ví dụ: Theo dõi sự kiện trên ứng dụng di động.



Cách thức triển khai

2. Ingest



Hình 3: Minh họa Ingest



Cách thức triển khai

3. Process

- Xử lý dữ liệu (xử lý trùng lặp, thiếu, chuẩn hóa)
- Thống kê data
- Trực quan hóa dữ liệu



Cách thức triển khai

3. Process

a. Xử lý dữ liệu

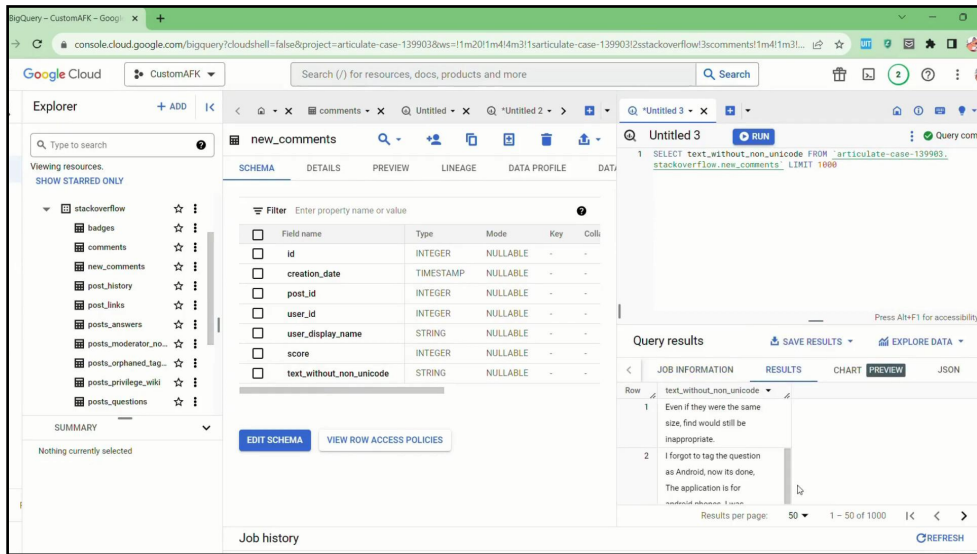
- Xử lý dữ liệu trùng lặp.
- Xử lý dữ liệu bị thiếu.
- Chuẩn hóa dữ liệu.



Cách thức triển khai

3. Process

a. Xử lý dữ liệu



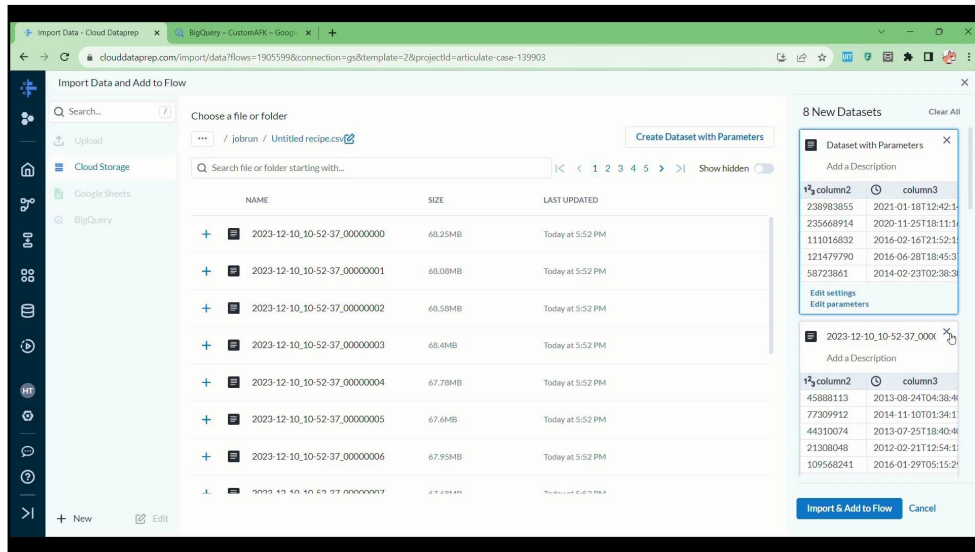
Hình 4: Minh họa cho chuẩn hóa dữ liệu ở comments



Cách thức triển khai

3. Process

a. Xử lý dữ liệu



Hình 5: Minh họa cho chuẩn hóa dữ liệu ở comments với Dataprep



Cách thức triển khai

3. Process

b. Thống kê dữ liệu

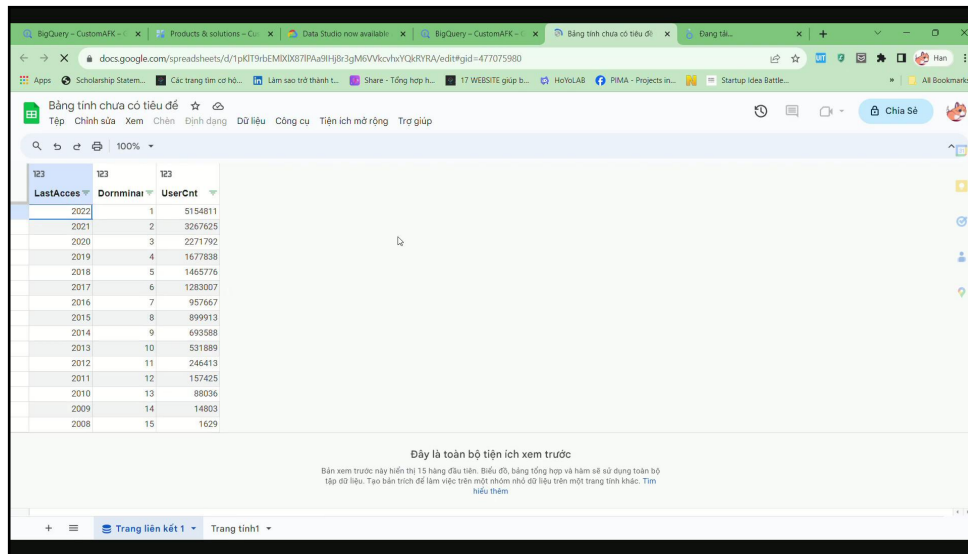
- Tạo ra thông tin hữu ích từ các dữ liệu.
- Dùng SQL trong BigQuery để truy vấn.
- Một số hàm: COUNT(), AVG(), STDDEV(), VAR(),...



Cách thức triển khai

3. Process

b. Thống kê dữ liệu



LastAccess	Dominant	UserCnt
2022	1	5154811
2021	2	3267625
2020	3	2271792
2019	4	1677838
2018	5	1465776
2017	6	1283007
2016	7	957567
2015	8	899913
2014	9	693588
2013	10	531889
2012	11	246413
2011	12	157425
2010	13	88036
2009	14	14803
2008	15	1629

Hình 6: Minh họa cho bảng xử lý thống kê users

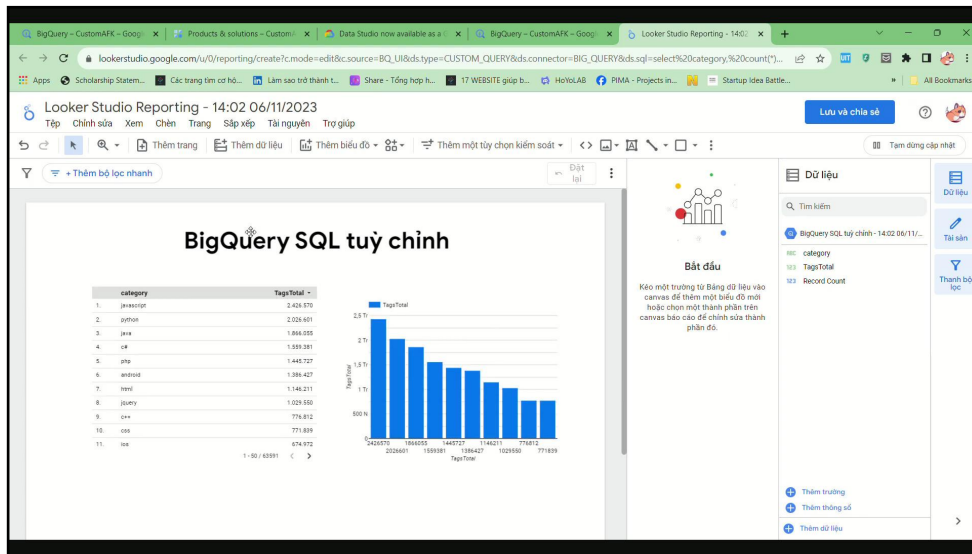


Cách thức triển khai

3. Process

b. Thống kê dữ liệu

Hình 7: Minh họa cho trực quan hóa dạng cột xử lý thống kê posts_questions





Cách thức triển khai

3. Process

c. Trực quan hóa dữ liệu.

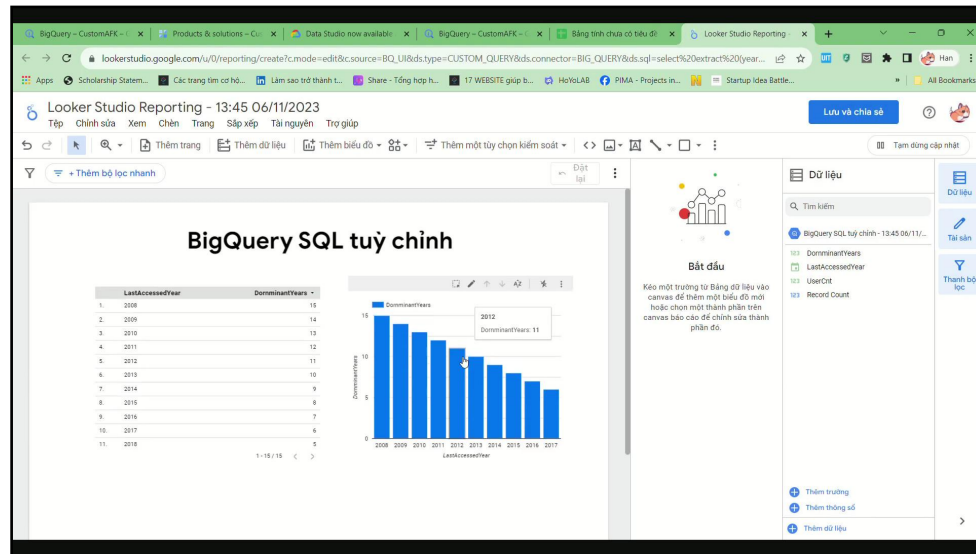
- Nhằm dễ hiểu những dữ liệu phức tạp, phát hiện xu hướng và tạo biểu đồ,...
- Google Data Studio cho phép bạn tạo báo cáo và biểu đồ tương tác.



Cách thức triển khai

3. Process

c. Trực quan hóa dữ liệu.



Hình 8: Minh họa cho trực quan hóa của xử lý thống kê users

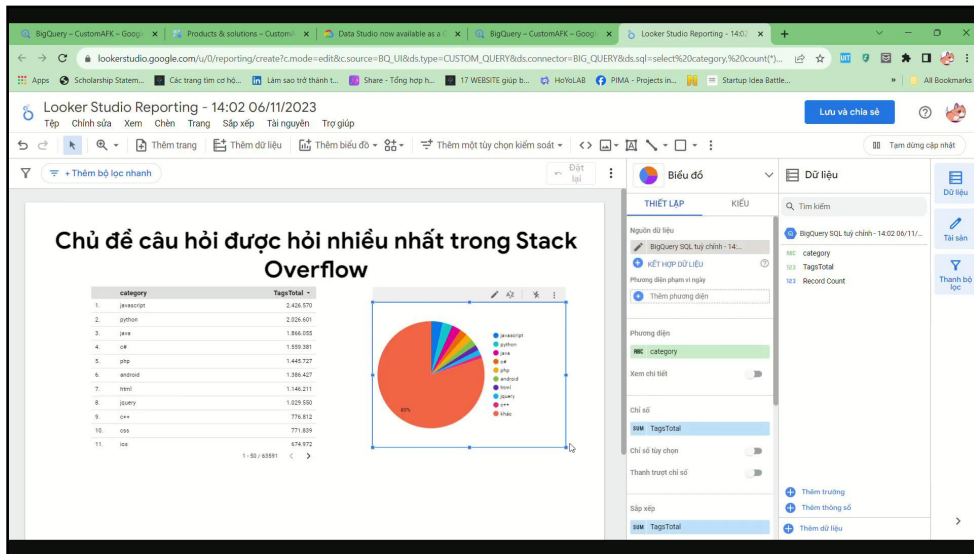


Cách thức triển khai

3. Process

c. Trực quan hóa dữ liệu.

Hình 9: Minh họa cho trực quan hóa dạng tròn xử lý thống kê posts_questions

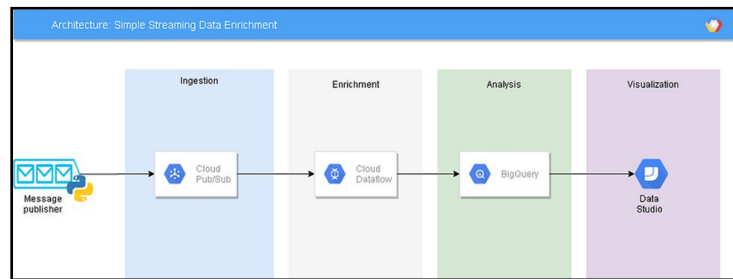




Cách thức triển khai

4. Enrich

- Còn được gọi là làm giàu dữ liệu.
- Thêm thông tin, làm chi tiết vào dữ liệu để làm dữ liệu cơ động và hữu ích hơn.



Hình 10: Kiến trúc thành phần enrich



Cách thức triển khai

4. Enrich

Hình 11: Minh họa tạo bảng ở phần Enrich

Đã tạo xong table mới

Row	location	reputation	up_votes	down_votes	views	profile_image_url	website_url	total_questions
46	null		1	0	0	https://www.gravatar.com/avatar/118556c7edc12c9f30357780e5a18477	null	0
47	null		1	0	0	https://www.gravatar.com/avatar/c972c2f549c2946930ec0584a3652f5f	null	0
48	null		1	0	0	https://www.gravatar.com/avatar/6b84f054e4f96c08c5c3284e6d4b17	null	0
49	null		1	0	0	https://lh4.googleusercontent.com/-zuK1bEmu0u/AAAAAAAAAAI/AAAAAAAAACH3j/IEKJn3H32811b5xHWw3g	null	0
50	null		1	0	0	null	null	0



Cách thức triển khai

4. Enrich

Hình 12: Minh họa kiểm tra các user có hơn 1000 câu hỏi

The screenshot shows the Google Cloud BigQuery console. The query executed is: `select * from articulate-case-139903.stackoverflow.enriched_users where total_questions > 1000`. The results table displays user information and their total questions.

Row	reputation	up_votes	down_votes	views	profile_image_url	website_url	total_questions
1	43934	1860	43	4569	https://i.stack.imgur.com/Yw9...	https://medium.com/@ViaCog...	1162
2	52314	1590	300	3662	https://www.gravatar.com/avatar/ea8f12a687ab83fec70d259fe2c081c?...	https://www.google.com/...	1465

Kiểm tra các user có hơn 1000 câu hỏi > có ích cho việc phân tích sau này



Kết luận và hướng phát triển

- Tìm hiểu được các thành phần của kiến trúc dữ liệu lớn.
- Triển khai được một số dịch vụ của Google Cloud.
- Hoàn thành ở mức đọc hiểu và triển khai theo tài liệu của công nghệ.
- Nghiên cứu kỹ hơn khi có cơ hội tiếp cận.



Cảm ơn cô đã theo dõi
phần trình bày nhóm em!
