

# **ỨNG DỤNG GOOGLE BIGQUERY VÀO DỮ ĐOÁN BỆNH TIỂU ĐƯỜNG Ở NỮ GIỚI**





# GIỚI THIỆU BÀI TOÁN

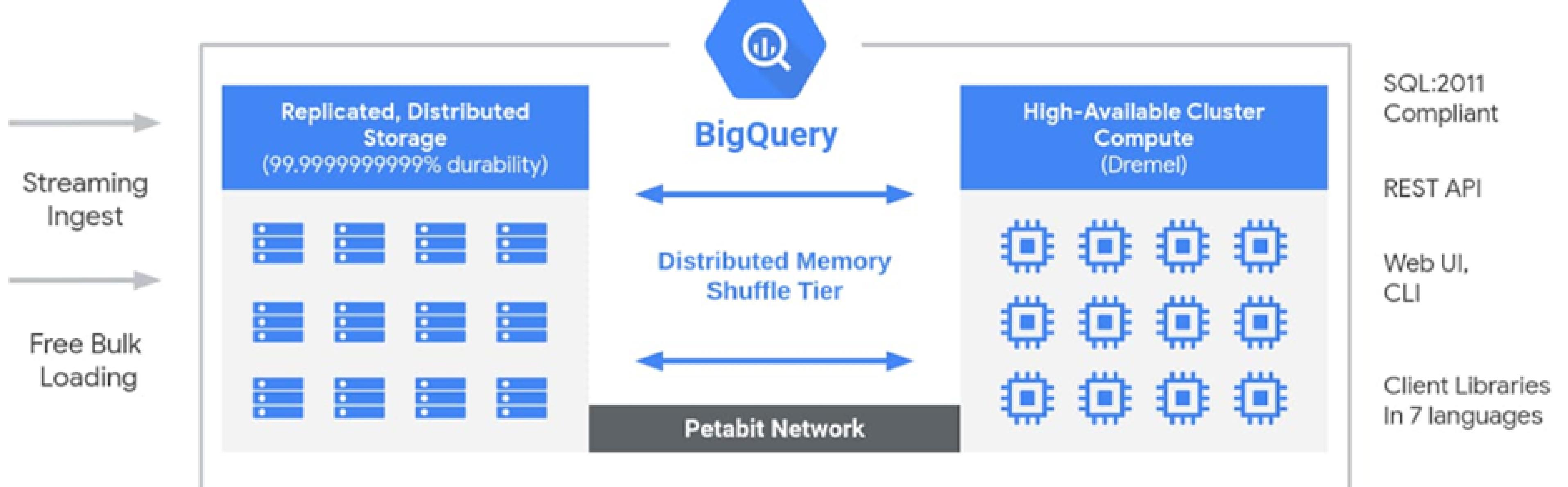
Google BigQuery là một dịch vụ đám mây của Google cho phép lưu trữ và xử lý dữ liệu lớn bằng cách sử dụng SQL và công nghệ xử lý phân tán. Nó tích hợp với các công cụ thống kê và khai phá dữ liệu của Google và có thể được tích hợp với các công cụ như Tableau, QlikView và Excel thông qua kết nối JDBC/ODBC.

BigQuery được sử dụng rộng rãi trong các lĩnh vực như phân tích dữ liệu, khai thác dữ liệu, kho dữ liệu và phát triển ứng dụng. Với việc áp dụng BigQuery trong đề tài dự đoán mắc bệnh tiểu đường ở nữ giới, giúp phát hiện sớm các trường hợp và quản lý và điều trị bệnh hiệu quả hơn.

# GOOGLE BIGQUERY LÀ GÌ?

Google BigQuery là một kho dữ liệu trên nền tảng điện toán đám mây của Google. Nó cho phép chạy các truy vấn siêu nhanh trên các tập dữ liệu lớn. Bạn có thể quản lý và thực hiện các truy xuất dữ liệu từ nhiều nguồn khác nhau và sử dụng Machine Learning để xây dựng các module bằng cách sử dụng cú pháp SQL đơn giản.





# CẤU TRÚC CỦA BIGQUERY

Dữ liệu đưa vào được chia làm 2 loại: dữ liệu được đưa vào liên tục trong một luồng dữ liệu (streaming ingest) và dữ liệu được đưa vào theo một khối (free bulk loading).

# CẤU TRÚC CỦA BIGQUERY

01

Dremel là một cụm đa nhiệm lớn được sử dụng để thực thi các truy vấn SQL. Nó biến các truy vấn thành cây thực thi và tự động sắp xếp các truy vấn theo mức độ ưu tiên tương đương cho người dùng.

02

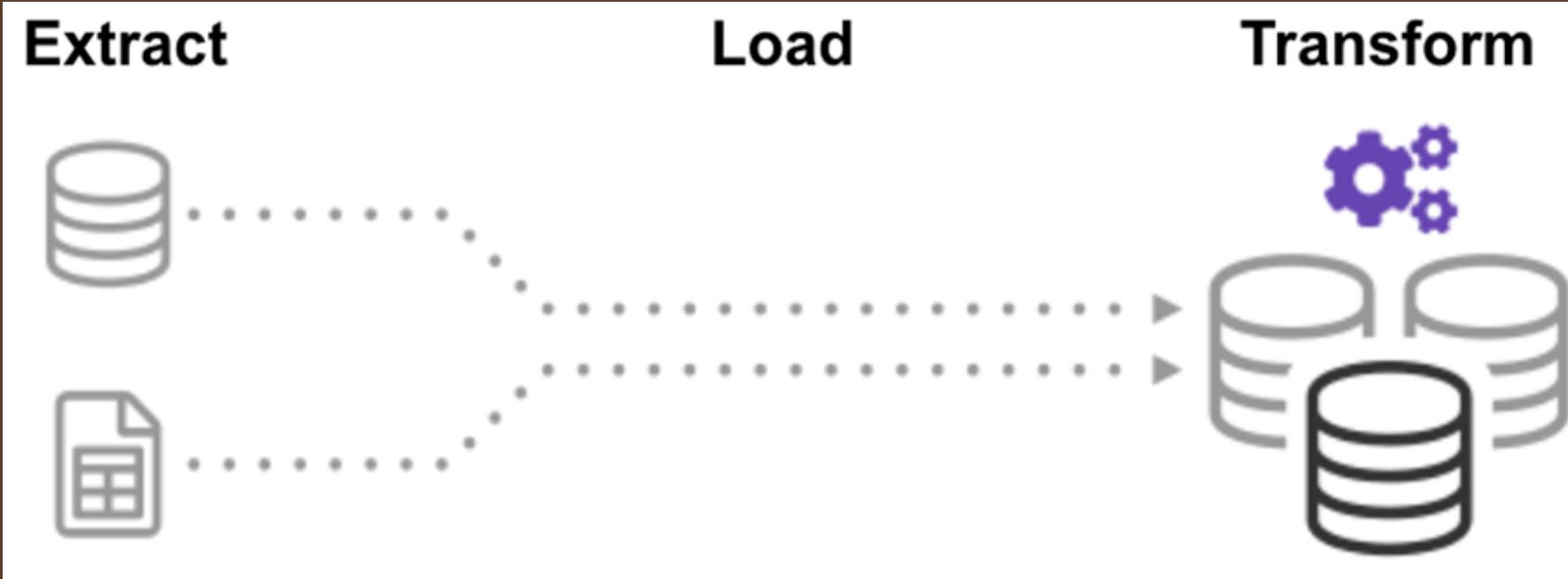
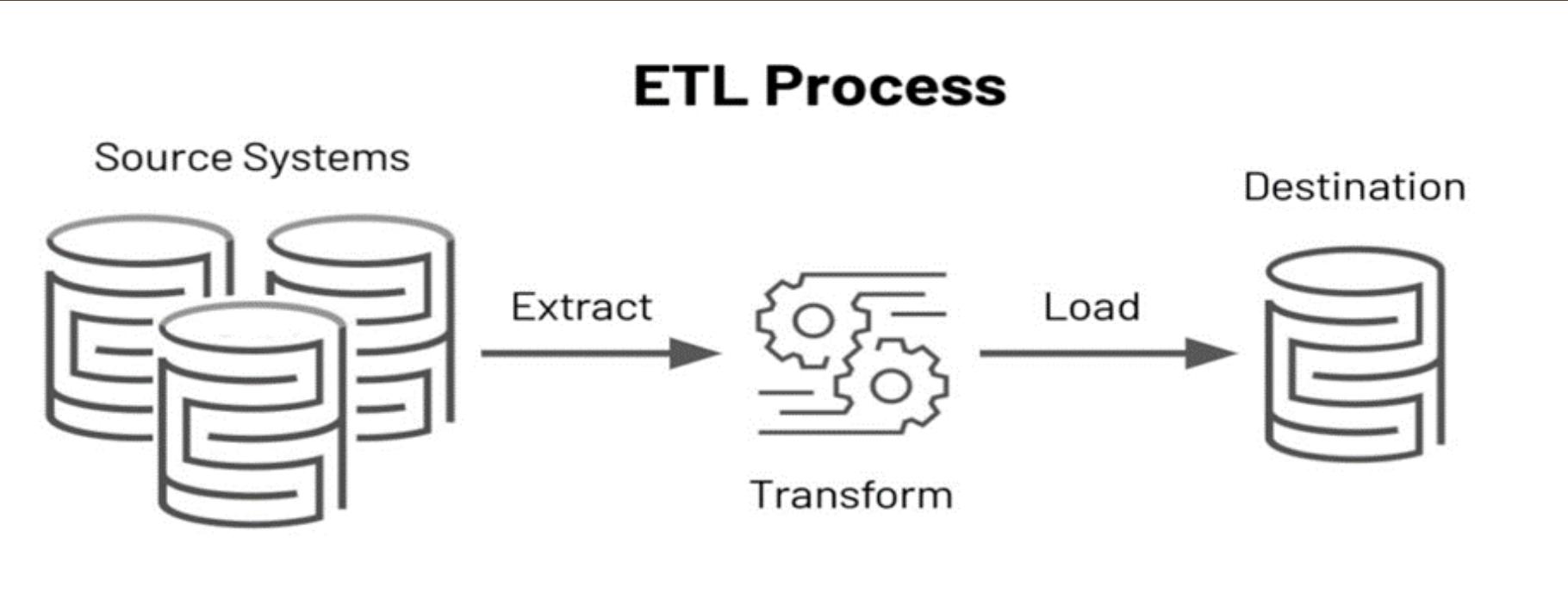
Colossus là hệ thống lưu trữ toàn cầu của Google sử dụng bởi BigQuery để lưu trữ dữ liệu trong định dạng cột và tối ưu hóa thuật toán nén. Nó có khả năng xử lý sao chép, phục hồi và quản lý phân tán mà không có điểm lỗi nào.

03

Máy tính và bộ nhớ của Google giao tiếp với nhau thông qua mạng petabit Jupiter, sử dụng shuffle để di chuyển dữ liệu từ bộ nhớ sang bộ xử lý tính toán.

04

BigQuery được điều phối thông qua Borg - tiền thân của Kubernetes và cải tiến liên tục để mang lại hiệu suất, độ bền, hiệu quả và khả năng mở rộng cho người dùng.



# PHƯƠNG THỨC VẬN CHUYỂN DATA ETL VÀ ELT

ETL (Extract, Transform, Load) và ELT (Extract, Load, Transform) là các phương pháp tích hợp dữ liệu giúp chuyển đổi dữ liệu từ một nguồn bên ngoài vào data warehouse.

ETL trích xuất dữ liệu từ các hệ thống nguồn sau đó chuyển đổi dữ liệu và tải vào data warehouse.

ELT cho phép lưu trực tiếp dữ liệu thô vào kho. BigQuery chuộng sử dụng ELT vì nó dễ sử dụng hơn và giảm khối lượng công việc cho Data Engineer.

# ĐẶC TRƯNG CỦA BIGQUERY

- Tính linh hoạt, giá cả có thể dự đoán và hiệu suất giá tốt nhất:
- Học máy tích hợp
- Phân tích và chia sẻ dữ liệu trên các đám mây
- Phân tích thời gian thực với các đường ống truyền dữ liệu
- Thống nhất, quản lý và chi phối tất cả các loại dữ liệu
- Chia sẻ thông tin chuyên sâu với thông tin kinh doanh tích hợp sẵn
- Quản trị và bảo mật dữ liệu
- Phân tích không gian địa lý với BigQuery
- Thu thập và sao chép dữ liệu thay đổi theo thời gian thực
- SQL chuẩn

# PRICING

"Định giá phân tích" là chi phí để xử lý các truy vấn dữ liệu, bao gồm SQL, hàm người dùng, tập lệnh và câu lệnh DML và DDL. "Định giá bộ nhớ" là chi phí để lưu trữ dữ liệu vào BigQuery.



## Phân Tích Mô Hình Định Giá

"Định giá dựa trên nhu cầu" là mô hình định giá theo số lượng bytes được xử lý cho mỗi query, với 1 TB dữ liệu query miễn phí đầu tiên vào mỗi tháng. "Định giá thống nhất" là mô hình định giá khi mua slots, tương đương với việc mua khả năng xử lý chuyên dụng để sử dụng trong việc chạy query.

## Định Giá Lưu Trữ

"Định giá lưu trữ" là chi phí để lưu trữ dữ liệu load vào BigQuery. "Lưu trữ ngắn hạn" bao gồm table và partition chỉnh sửa trong vòng 90 ngày, trong khi "lưu trữ lâu dài" bao gồm table và partition không được chỉnh sửa trong vòng 90 ngày liên tiếp. Miễn phí cho 10GB lưu trữ đầu tiên mỗi tháng.

## Tài Khoản Dịch Vụ (Service Account)

Để tích hợp BigQuery vào một công cụ thứ ba, cần sử dụng tài khoản Dịch vụ để xác thực kết nối. Tài khoản này được sử dụng bởi ứng dụng hoặc phần mềm, không phải bởi cá nhân. Tài khoản dịch vụ có thể được ủy quyền để thực hiện các lệnh gọi API với tư cách là tài khoản dịch vụ hoặc người dùng Google Workspace, Cloud Identity thông qua ủy quyền trên toàn miền.



## Ưu điểm của BigQuery

khả năng xử lý dữ liệu lớn, tích hợp dữ liệu dễ dàng, tốc độ xử lý nhanh, hỗ trợ nhiều ngôn ngữ lập trình và tính bảo mật.

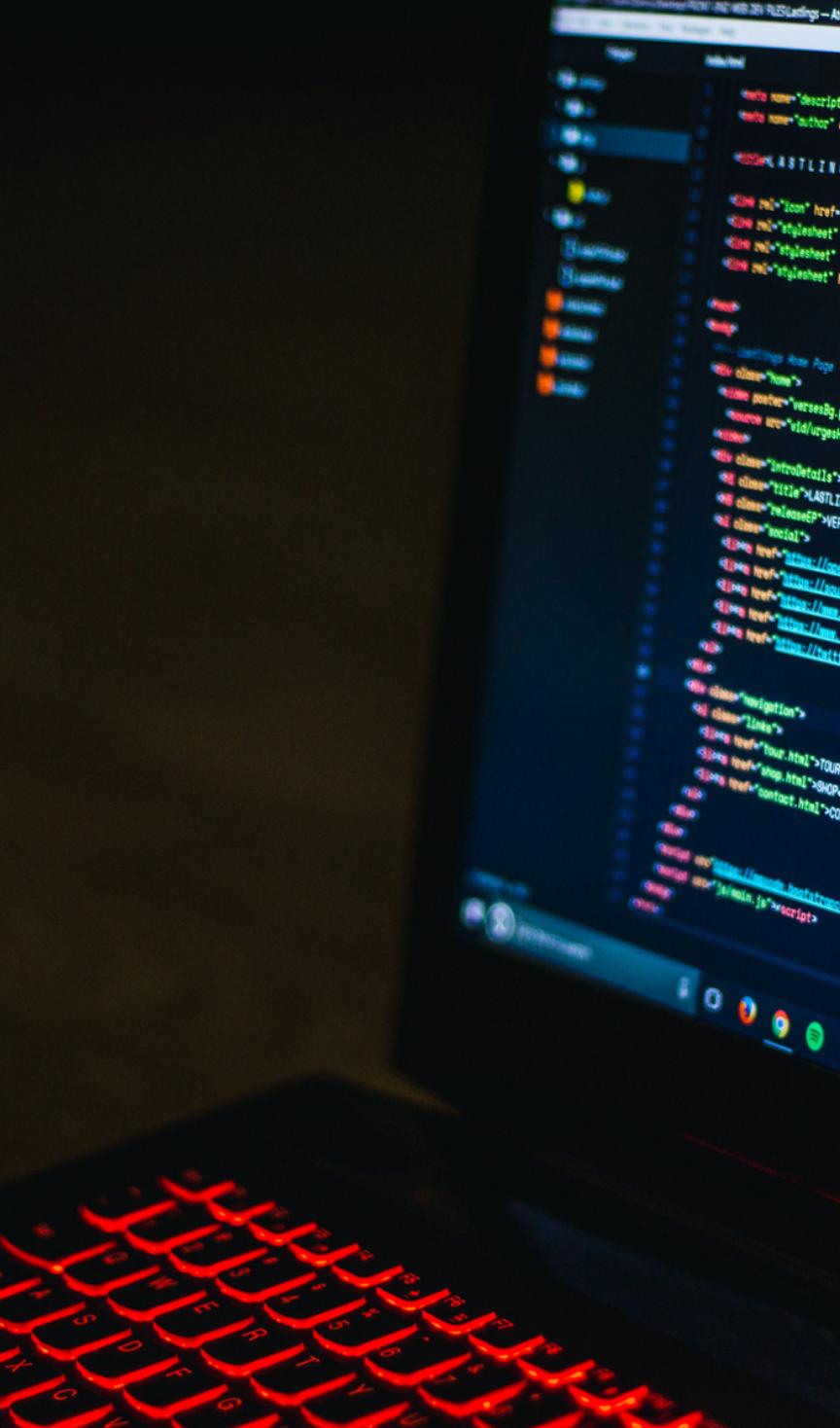


## Nhược điểm của BigQuery

sản phẩm có giá thành cao, không hỗ trợ các truy vấn phức tạp, phụ thuộc vào kết nối internet, không hỗ trợ cập nhật dữ liệu real-time và không phù hợp cho các ứng dụng với tần suất truy xuất dữ liệu cao.

# NHỮNG LỢI ÍCH MÀ BIGQUERY ĐEM LẠI:

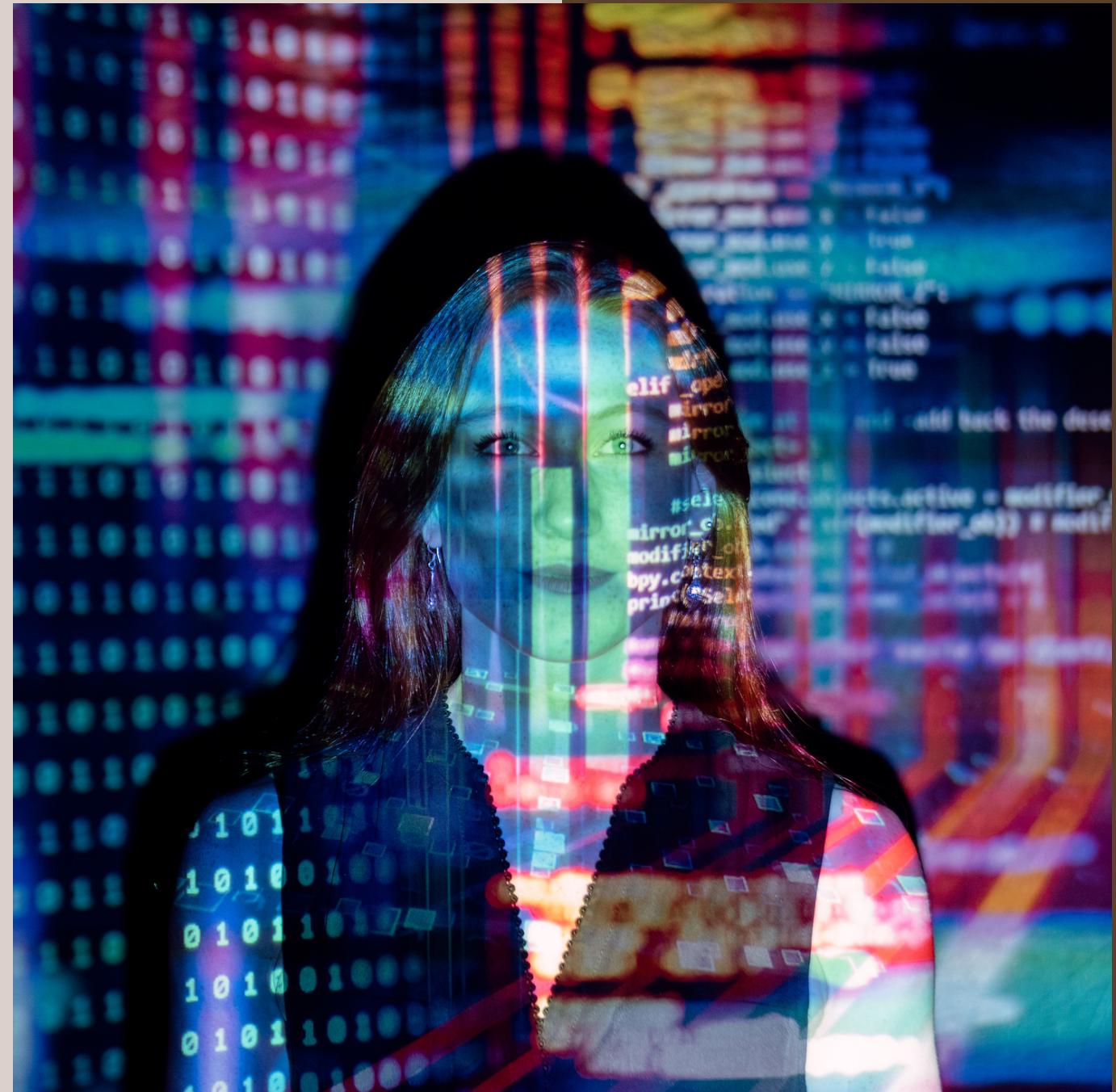
- Tự động phân phối dữ liệu
- Tăng khả năng tiếp cận các insights
- Xây dựng nền tảng cho trí tuệ nhân tạo (AI)
- Chuẩn bị các insights đúng thời điểm
- Bảo mật dữ liệu kinh doanh
- Đơn giản hóa quy trình dữ liệu hoạt động



# HỒI QUY LOGISTIC

Hồi quy logistic là một kỹ thuật phân tích dữ liệu để tìm mối quan hệ giữa hai yếu tố dữ liệu và sử dụng quan hệ đó để dự đoán giá trị của các yếu tố dựa trên yếu tố còn lại.

$$P(Y_i = 1) = \frac{e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})}}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})}}$$



# CÁC LOẠI HỒI QUY LOGISTIC

01

- Hồi quy logistic nhị phân

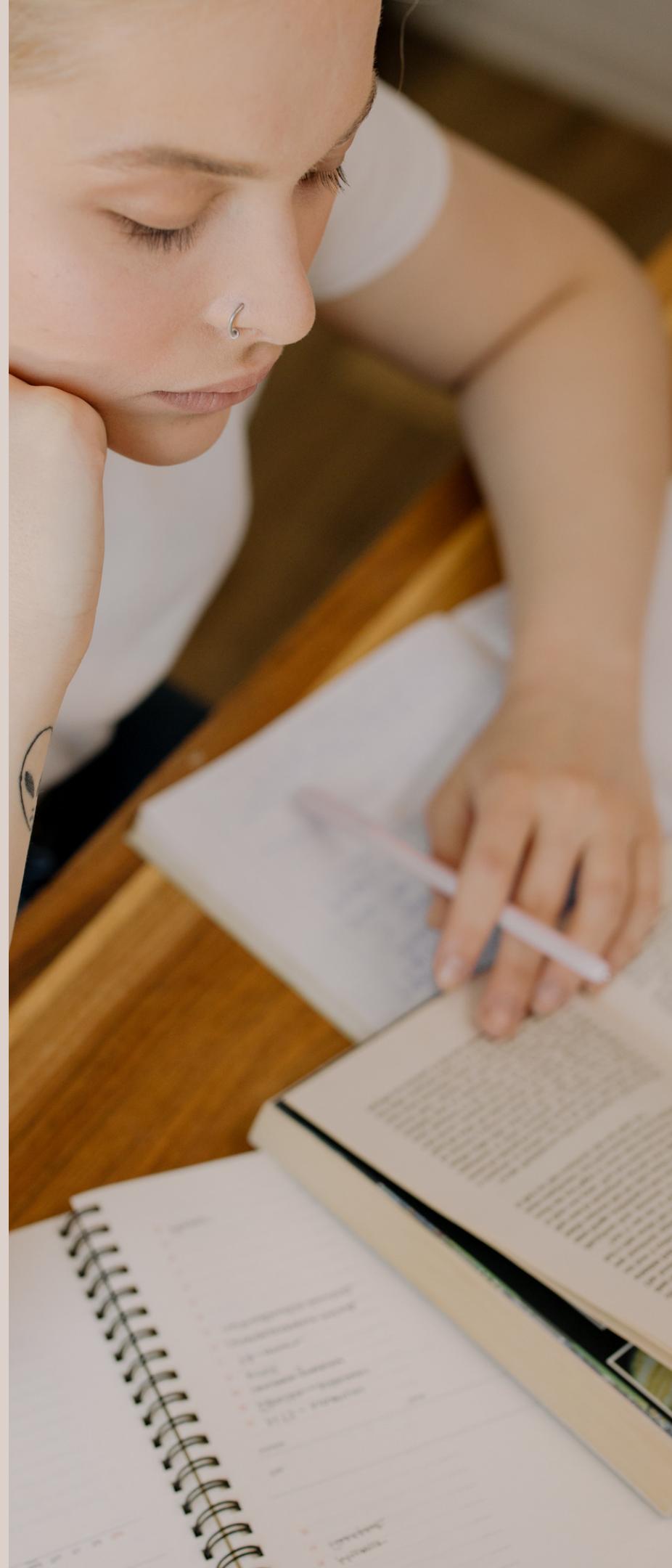
02

- Hồi quy logistic đa thức

03

- Hồi quy logistic thứ tự





# CÁC THÀNH PHẦN CỦA HỒI QUY LOGISTIC

**Hàm sigmoid**

**Biến phụ thuộc**

**Biến độc lập**

**Tham số số hồi quy**

**Hàm logistic**

**Hàm mất mát**



# CÁCH HOẠT ĐỘNG CỦA HỒI QUY LOGISTIC

Xác định câu hỏi

Thu thập dữ liệu lịch sử:

Đào tạo mô hình phân tích hồi quy:

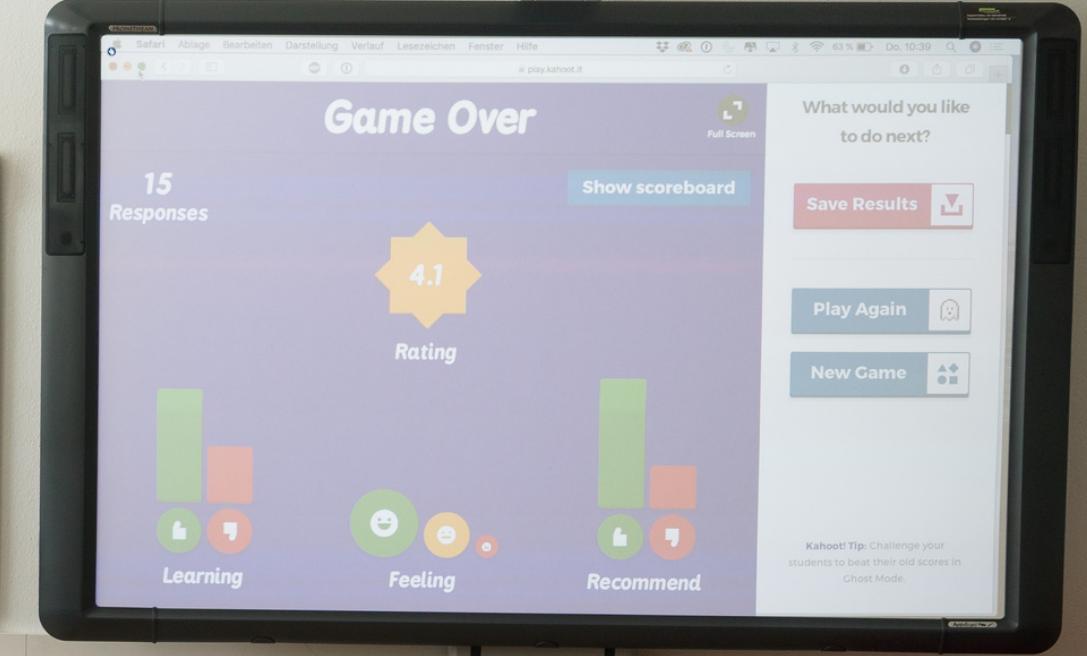
# MỤC ĐÍCH CỦA HỒI QUY LOGISTIC



Hồi quy logistic là một phương pháp trong học máy, được sử dụng để đo lường tác động của nhiều biến số trên một kết quả nhất định. Phương pháp này thường được áp dụng cho các bài toán phân loại nhị phân, tức là các bài toán có hai lớp giá trị.



- So với hồi quy tuyến tính: Hồi quy tuyến tính được sử dụng để dự đoán giá trị thực của biến phụ thuộc, trong khi đó hồi quy logistic là một thuật toán phân loại. Hồi quy logistic không thể dự đoán giá trị thực sự cho dữ liệu liên tục như hồi quy tuyến tính.



- So với học sâu: Hồi quy logistic ít phức tạp và có cường độ điện toán ít hơn so với học sâu. Các phép toán hồi quy logistic lại minh bạch và dễ khắc phục sự cố hơn so với các phép toán của học sâu.

# MỘT SỐ ƯU ĐIỂM VÀ NHƯỢC ĐIỂM CỦA VIỆC SỬ DỤNG HỒI QUY LOGISTIC SO VỚI CÁC KỸ THUẬT MACHINE LEARNING KHÁC

- **Ưu điểm**

Một số ưu điểm của hồi quy logistic là tính đơn giản, tốc độ xử lý nhanh, sự linh hoạt trong việc tìm đáp án cho các câu hỏi có hai hoặc nhiều kết quả hữu hạn và khả năng hiển thị dễ nhìn nhận.

- **Nhược điểm**

Thuật toán này cũng có một số nhược điểm như không thể xử lý được các tương tác non-linear giữa các đặc trưng, dễ bị overfitting khi số lượng đặc trưng quá lớn, khó khăn trong việc xử lý dữ liệu thiếu và phụ thuộc quá nhiều vào đặc trưng đầu vào. Việc chọn đúng đặc trưng đầu vào cũng ảnh hưởng rất lớn đến kết quả của thuật toán.





# KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

Nguồn dữ liệu

Tên: Tập dữ liệu dự đoán bệnh tiểu đường ở nữ giới ít nhất 21 tuổi của Ấn Độ Pima

Tập data bao gồm 9 đặc tính:

- Pregnancies
- Glucose
- BloodPressure
- SkinThickness
- BMI
- Diabetes
- Age
- Outcome

Tập data gồm 768 trường dữ liệu

# CÁC CÔNG CỤ, THƯ VIỆN DÙNG XÂY DỰNG

Google BigQuery ML cung cấp tính năng tích hợp dữ liệu trong BigQuery, hỗ trợ nhiều loại mô hình học máy, không yêu cầu kinh nghiệm về lập trình và tự động tối ưu siêu tham số của mô hình để đạt được hiệu suất tốt nhất

Các tính năng chính của Google BigQuery ML bao gồm:

- Tích hợp dữ liệu
- Hỗ trợ nhiều loại mô hình
- Không cần kinh nghiệm về lập trình
- Tự động tối ưu
- Thực hiện trên dữ liệu lớn
- Tích hợp với công cụ phân tích:

# XÂY DỰNG MÔ HÌNH SỬ DỤNG BIGQUERY UI

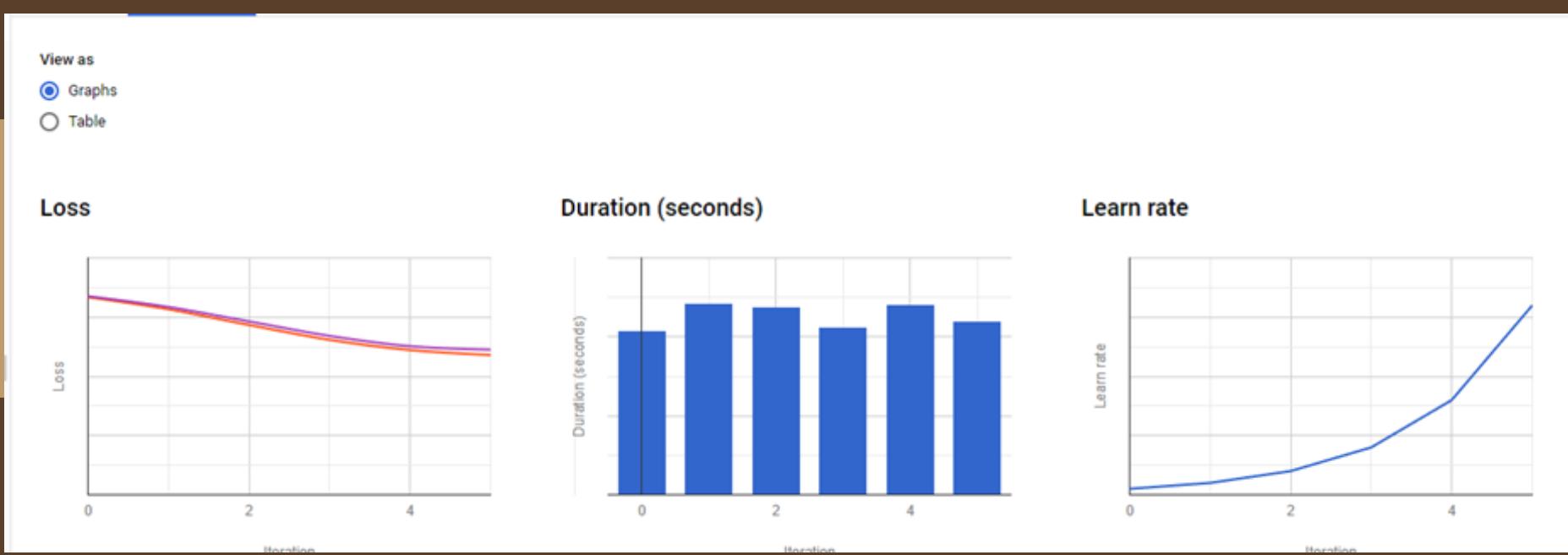
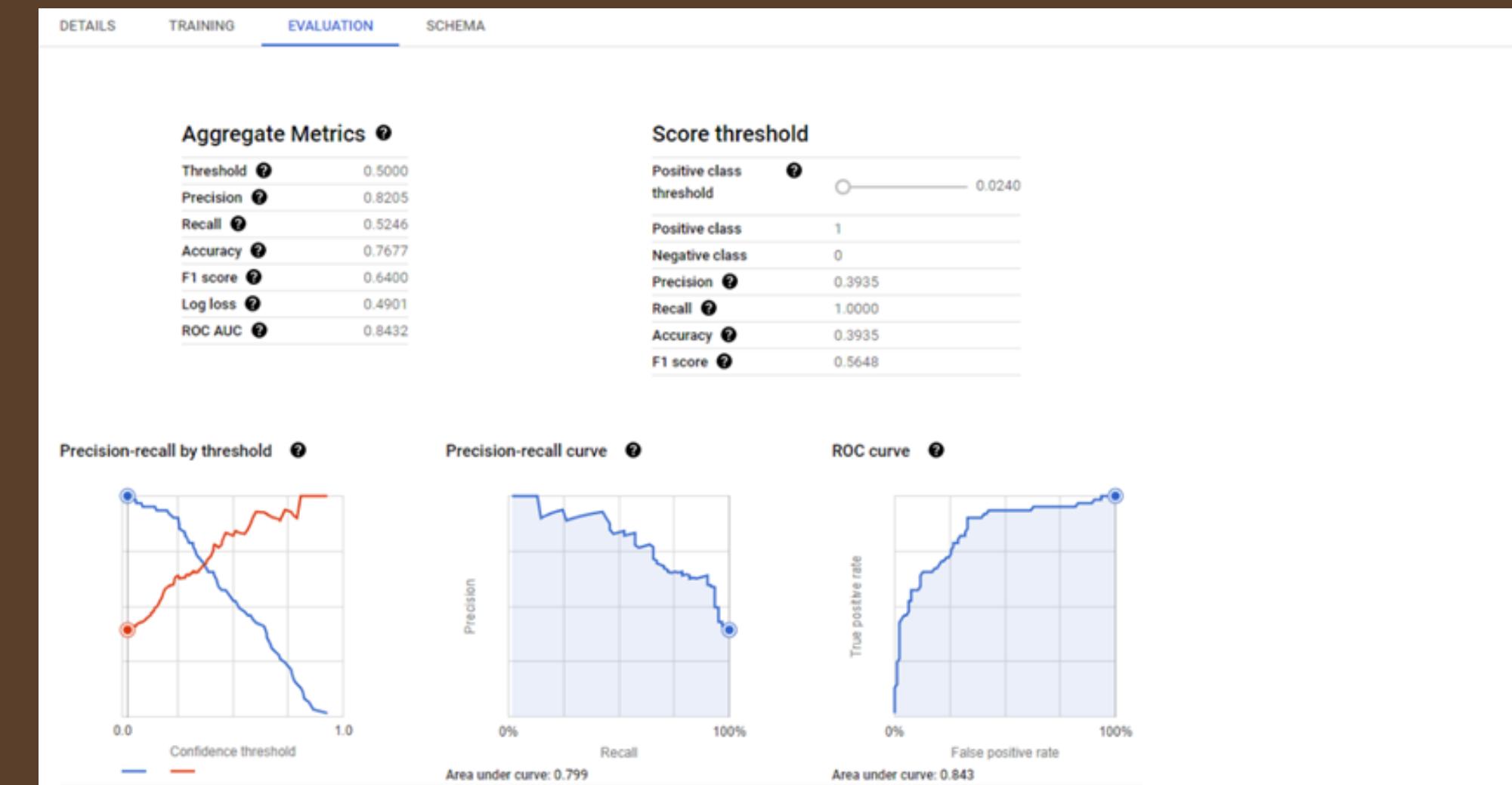
Đưa dữ liệu lên Google BigQuery

The collage consists of five screenshots from the Google BigQuery UI:

- Screenshot 1:** Shows the BigQuery interface with two query tabs open. The left tab contains a CREATE OR REPLACE MODEL statement for a logistic regression model named "logistic\_reg\_model". The right tab shows the results of a SELECT query on the "diabetes" dataset.
- Screenshot 2:** Shows the "Add" dialog for connecting external data sources. It lists "Local file" and "Google Cloud Storage" as popular sources, and "Connections to external data sources" as additional sources.
- Screenshot 3:** Shows the "Source" dialog for creating a new table. It includes fields for "Source" (CSV file), "Destination" (Project, Dataset, Table), and "Schema" (Auto-detect).
- Screenshot 4:** Shows the "data" view for the "logistic\_reg\_model" dataset. It displays a preview of the data with columns: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigree, Age, and Outcome.
- Screenshot 5:** Shows the "data" view for the "logistic\_reg\_model" dataset, displaying the first 25 rows of the "data" table.

# XÂY DỰNG MÔ HÌNH

```
CREATE OR REPLACE MODEL `diabetes.logistic_reg_model`  
OPTIONS  
  (model_type='logistic_reg',  
   input_label_cols=['Outcome']) AS  
SELECT  
  *  
FROM  
  `diabetes.data`  
LIMIT 800 OFFSET 100
```



# CHẠY THỬ MÔ HÌNH

```
SELECT
    predicted_Outcome, raw.Outcome as real_OutCome
FROM
    ML.PREDICT(MODEL `graduationprojectms-
75dd9.diabetes.logistic_reg_model`,
(
    SELECT
        *
    FROM
        `diabetes.data`),
    `diabetes.data` as raw
LIMIT 100
```

Query results						
JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS	EXECUTION GRAPH	
Row	predicted_Outcome	real_OutCome				
1	0	1				
2	0	0				
3	0	1				
4	0	0				
5	0	0				
6	0	0				
7	0	0				
8	0	0				
9	0	0				
10	0	0				
11	0	0				
12	0	0				
13	0	0				
14	0	0				
15	0	0				

## ĐÁNH GIÁ ĐỘ CHÍNH XÁC CỦA MÔ HÌNH

```
SELECT
    *
FROM
    ML.EVALUATE(MODEL `graduationprojectms75dd9.diabetes.1
ogistic_reg_model`,
    (SELECT * FROM `diabetes.data` LIMIT 100))
```

Query results						
JOB INFORMATION		RESULTS	JSON	EXECUTION DETAILS		EXECUTION GRAPH
Row	precision	recall	accuracy	f1_score	log_loss	roc_auc
1	0.84210526...	0.45714285...	0.78	0.59259259...	0.45393689...	0.86568331...

# MA TRẬN NHẦM LÃN

```
SELECT
  *
FROM
  ML.CONFUSION_MATRIX
(MODEL `graduationprojectms75dd9.diabetes.logistic_regression`,
 (SELECT * FROM `diabetes.data` LIMIT 100))
```

Query results

JOB INFORMATION		RESULTS		JSON	EXECUTION DETAILS	EXECUTION GRAPH	PREVIEW
Row	expected_label	_0	_1				
1	0	62	3				
2	1	19	16				



# KẾT LUẬN

- Độ chính xác (Accuracy Score) cao hơn 70% có thể chấp nhận được
- Tỷ lệ recall khoảng 50% cho mô hình test có nghĩa là tỷ lệ dự đoán mắc bệnh đúng là 50%
- Tuy nhiên, một phần lý do dẫn đến độ chính xác không quá cao là bởi tập dữ liệu khá nhỏ (chỉ có 768 mẫu) nên khi chia tập train-test theo tỉ lệ 85-15, model có thể sẽ không train được tất cả các trường hợp đặc biệt.

# THANK YOU

