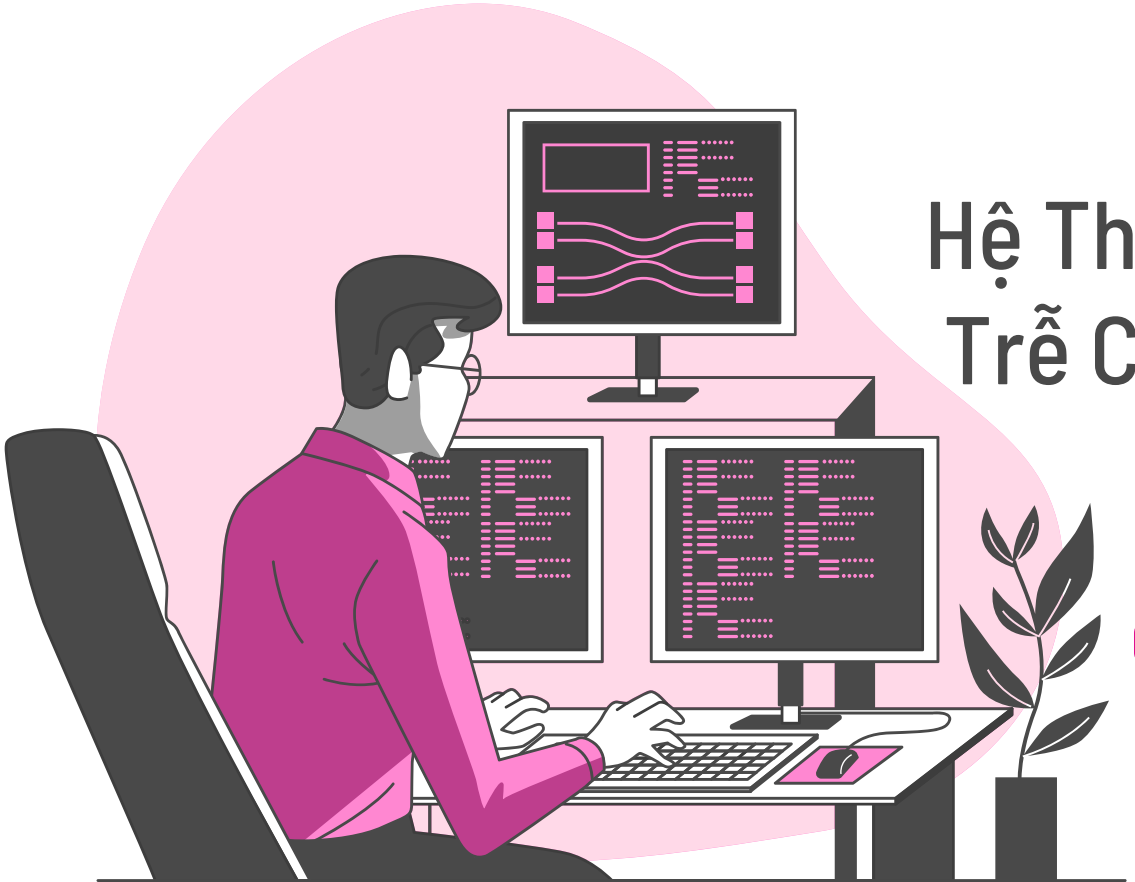
Phân Tích Dữ Liệu Lớn



Hệ Thống Dự Đoán Độ Trễ Chuyến Bay Theo Thời Gian Thực

GVHD: TS. Đỗ Trọng Hợp

Võ Minh Trí
19522396

Trần Triệu Vũ
19522539

Phạm Đức Thể
19522253

Our Solutions

01 Giới Thiệu

02 Dataset

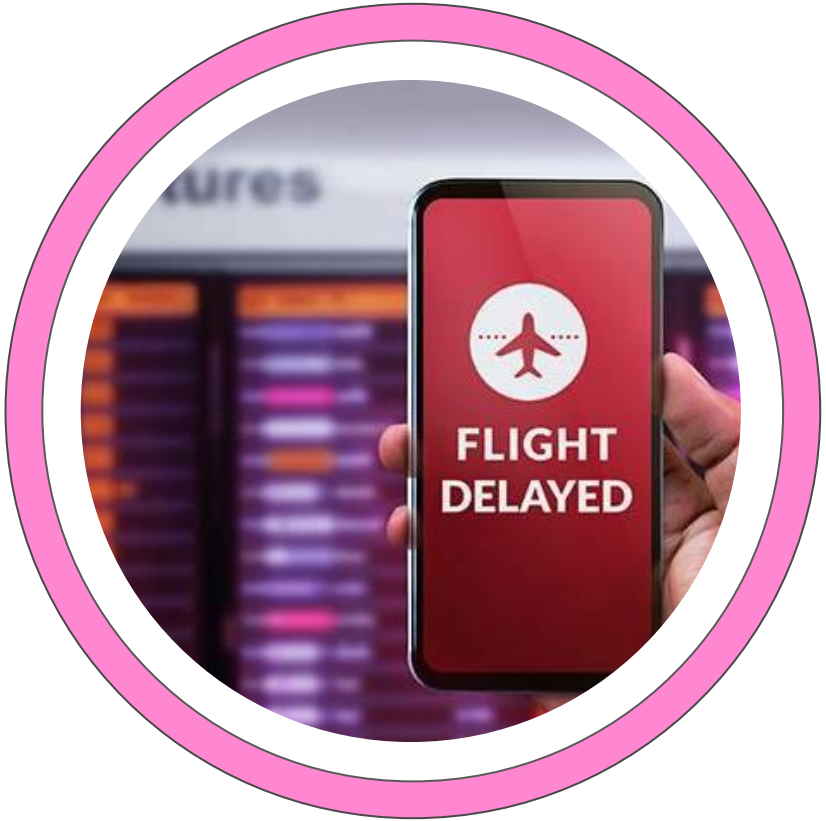
03 Kiến Trúc Hệ Thống

04 Demo

01

Giới Thiệu

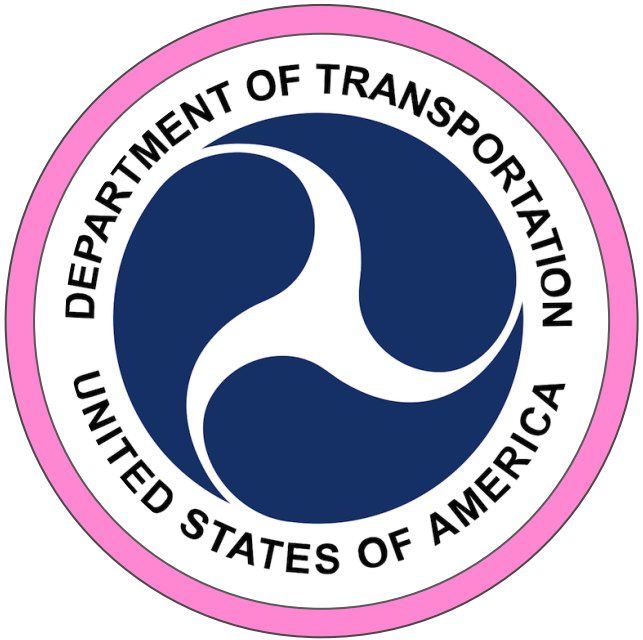
Dự đoán độ trễ chuyến bay là một bài toán dựa trên nền tảng phân tích các thông tin được cung cấp trước khi máy bay khởi hành nhằm trích xuất các thông tin có ích sau đó sử dụng các mô hình máy học để dự đoán độ trễ máy bay.



02
Dataset

Airline Service Quality Performance 234 (On-Time performance data)

Nguồn: Bureau of Transportation Statistics (BTS), một bộ phận thuộc The Department of Transportation (DOT).
Số thuộc tính sử dụng: 10.
Data training: Hơn 1,7 triệu data point



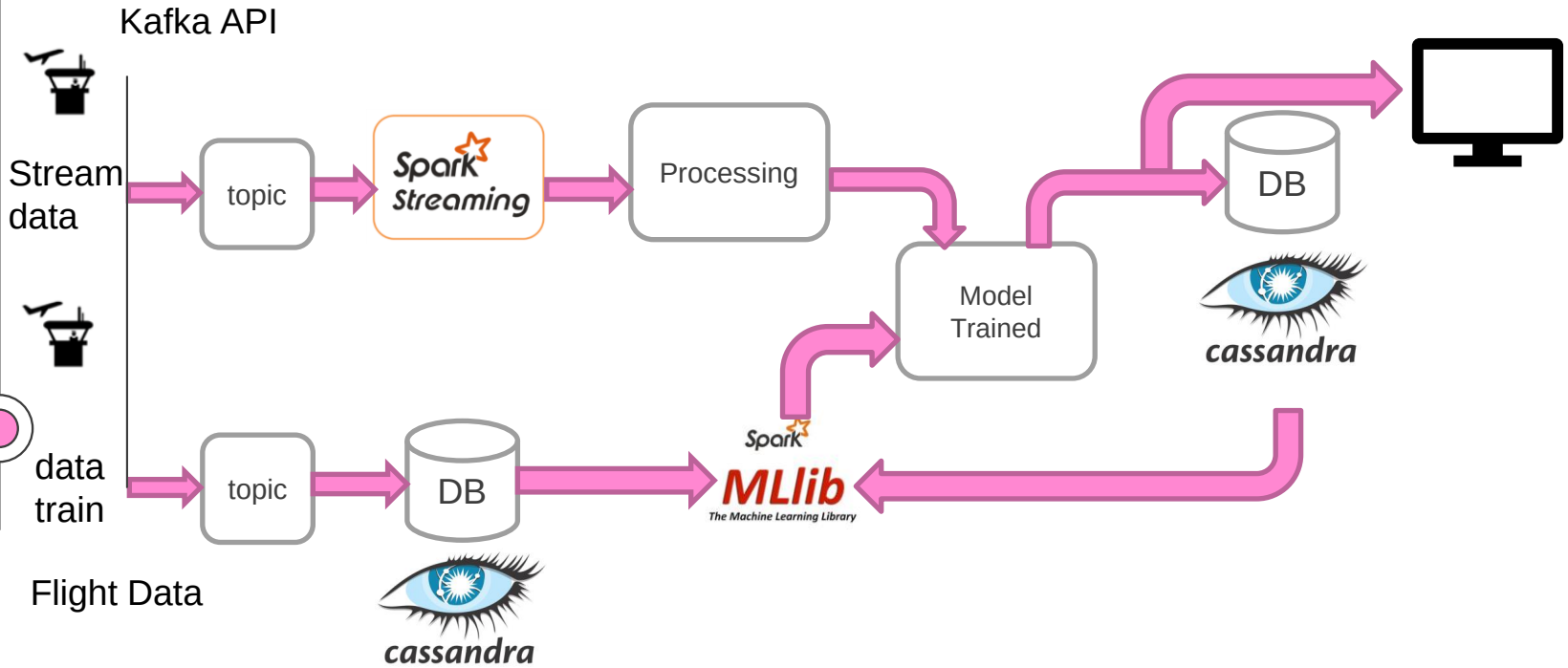
Airline Service Quality Performance 234 (On-Time performance data)

Id	Quarter	Month	Day_of_ month	Day_of_ week	OP_ unique_ carrier	Origin	Dest	Distance	Crs_ dep_ time	Dep_ delay
1	1	1	12	2	AA	MDT	CLT	413	1108	-10
2	1	4	14	2	AA	MDT	CLT	413	1108	0
3	2	6	2	3	OH	MKE	PHL	690	942	12

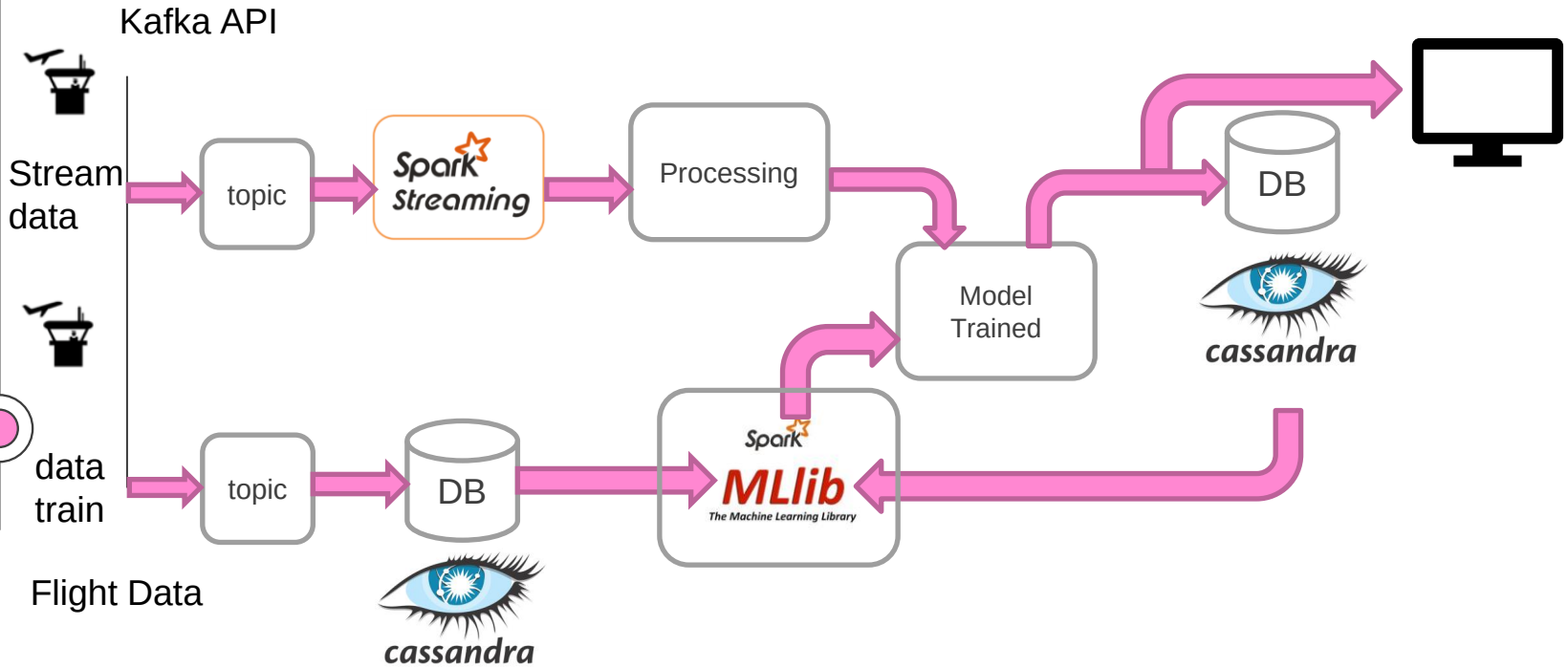
03

Kiến Trúc Hệ Thống

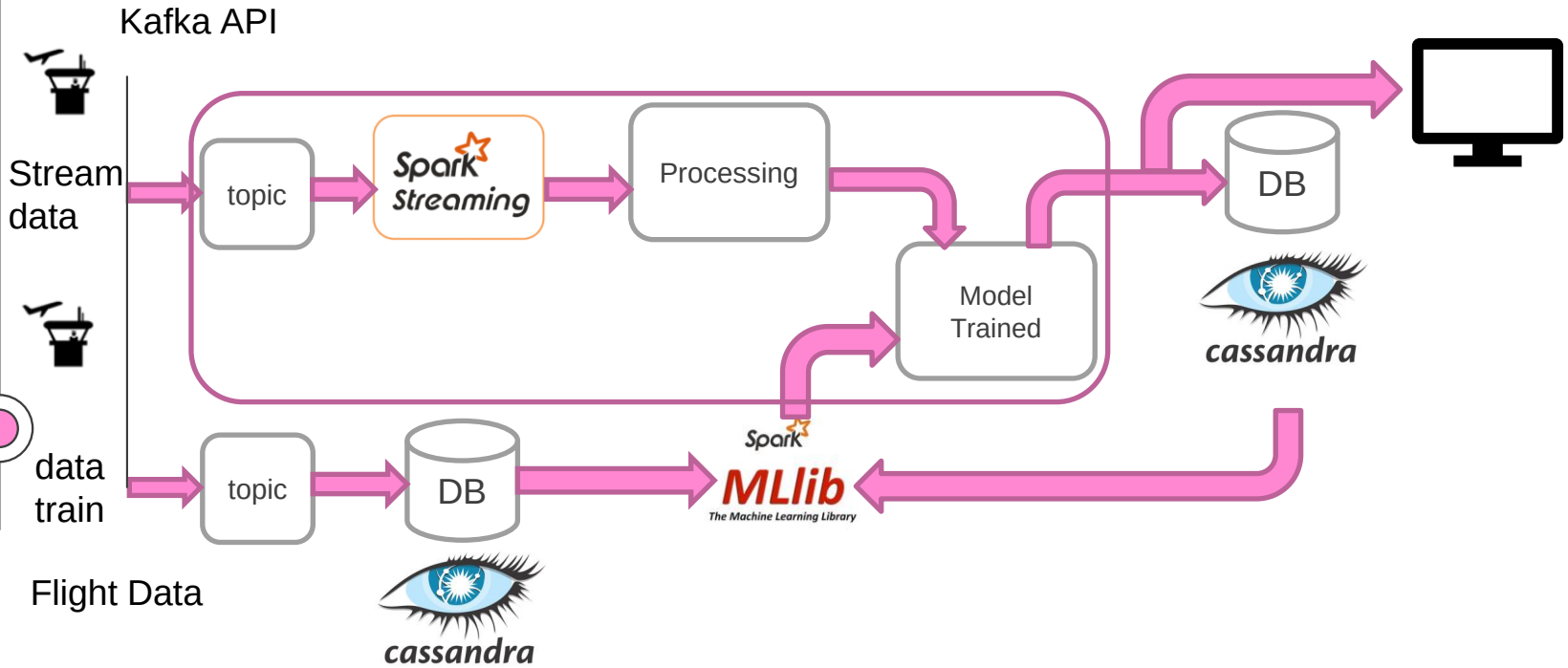
Tổng Quan Hệ Thống



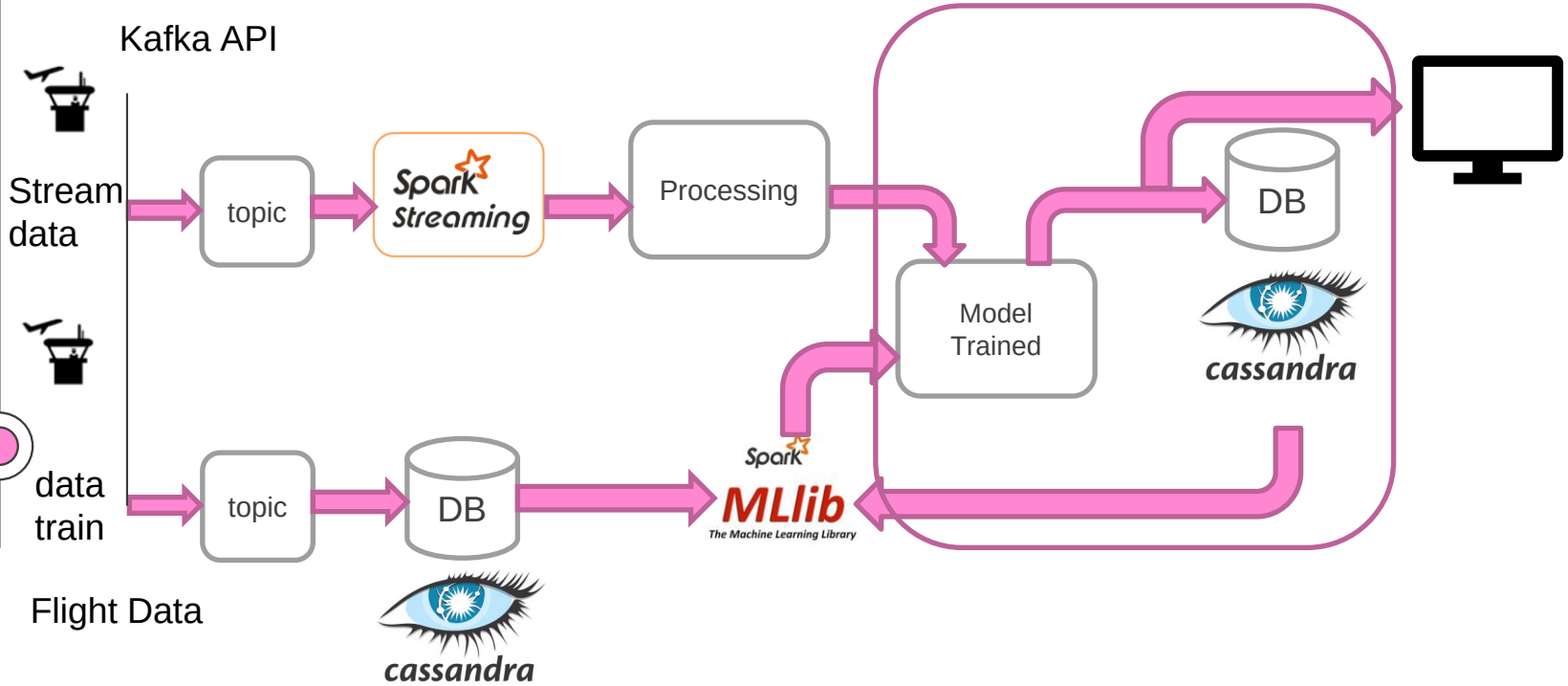
Tổng Quan Hệ Thống



Tổng Quan Hệ Thống



Tổng Quan Hệ Thống



Training Model

Mã hoá dữ liệu

Huấn luyện mô hình

Kết quả

Mã hoá dữ liệu

StringIndexer

Dest	DestIndex
ABY	245
ATL	0
AGS	161

OnehotEncoder

DestIndex	DestOneHot
245	(371,[0],[1.0])
0	(371,[245],[1.0])
161	(371,[161],[1.0])

Huấn luyện mô hình

Logistic Regression

Decision Tree

Random Forest

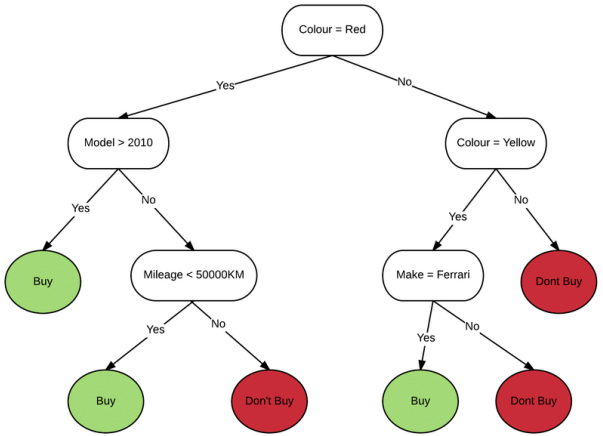
Navie Bayes

Huấn luyện mô hình

Logistic Regression

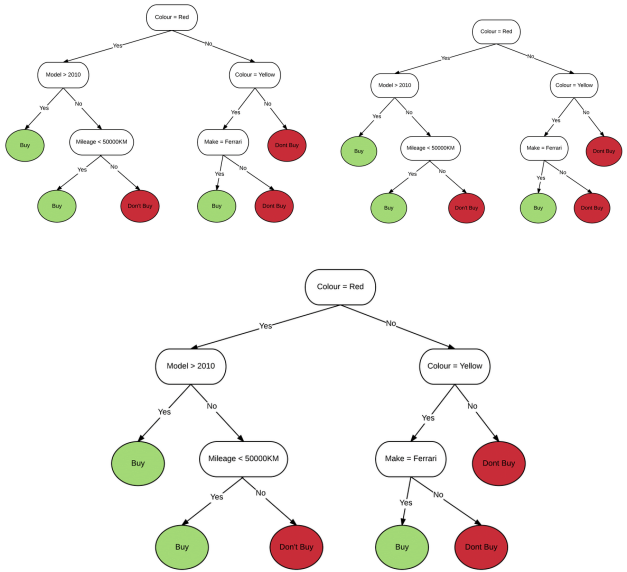
Xác định mối quan hệ giữa dữ liệu và phân loại dữ liệu theo mối quan hệ giữa chúng. Đồng thời dự đoán xác suất của biến phụ thuộc.

Decision Tree



Huấn luyện mô hình

Random Forest



Navie Bayes

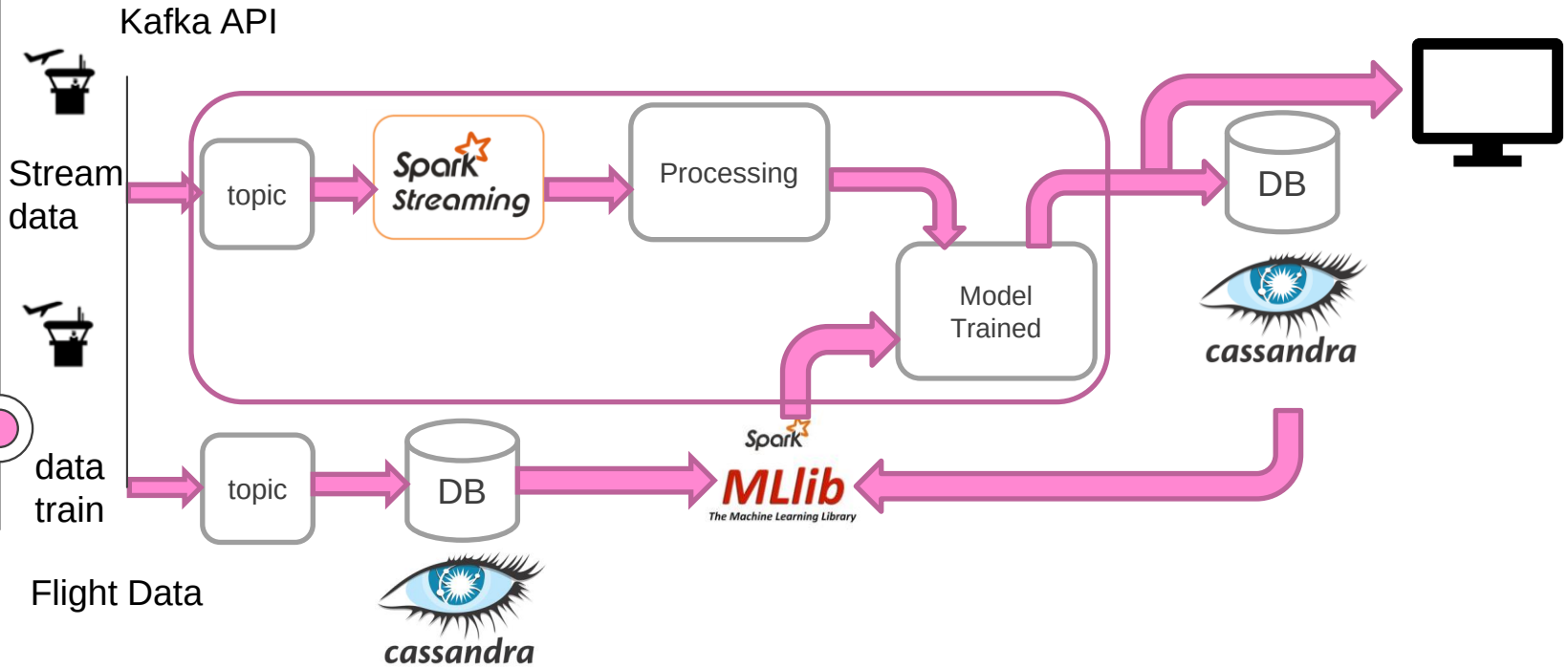
Naıve Bayes là một trong họ "bộ phân loại theo xác suất" dựa trên việc áp dụng định lý Bayes trong xác suất thống kê

Huấn luyện mô hình

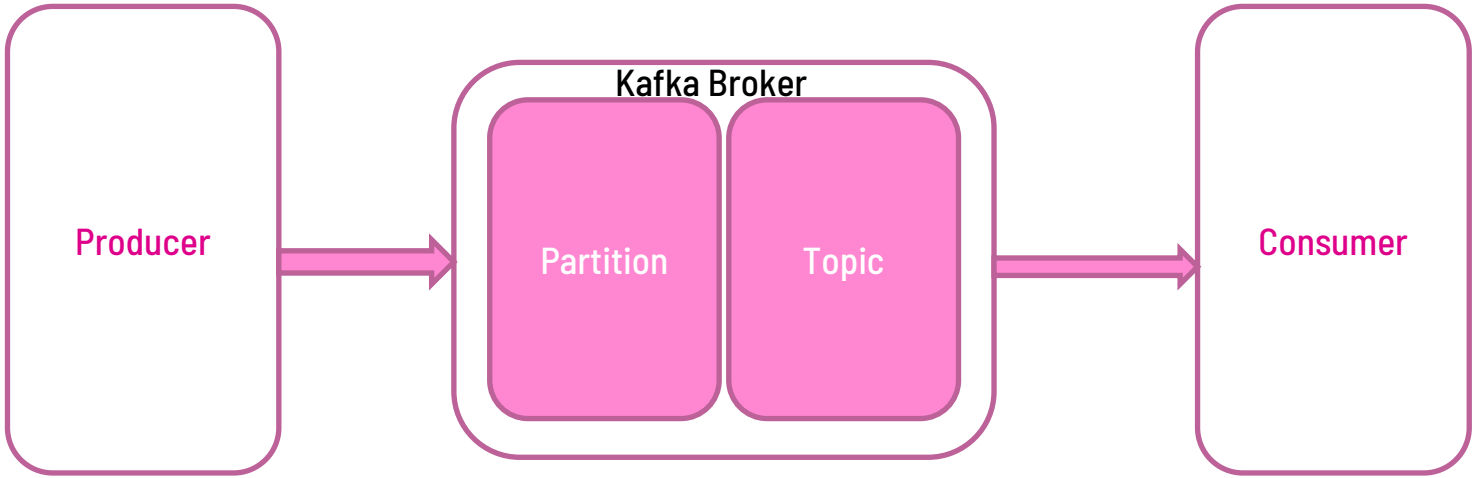
Kết quả

	Accuracy	F1-macro
Logistic Regression	0.63	0.36
Decision Tree	**0.65**	**0.45**
Random Forest	0.62	0.26
Navie Bayes	0.60	0.40

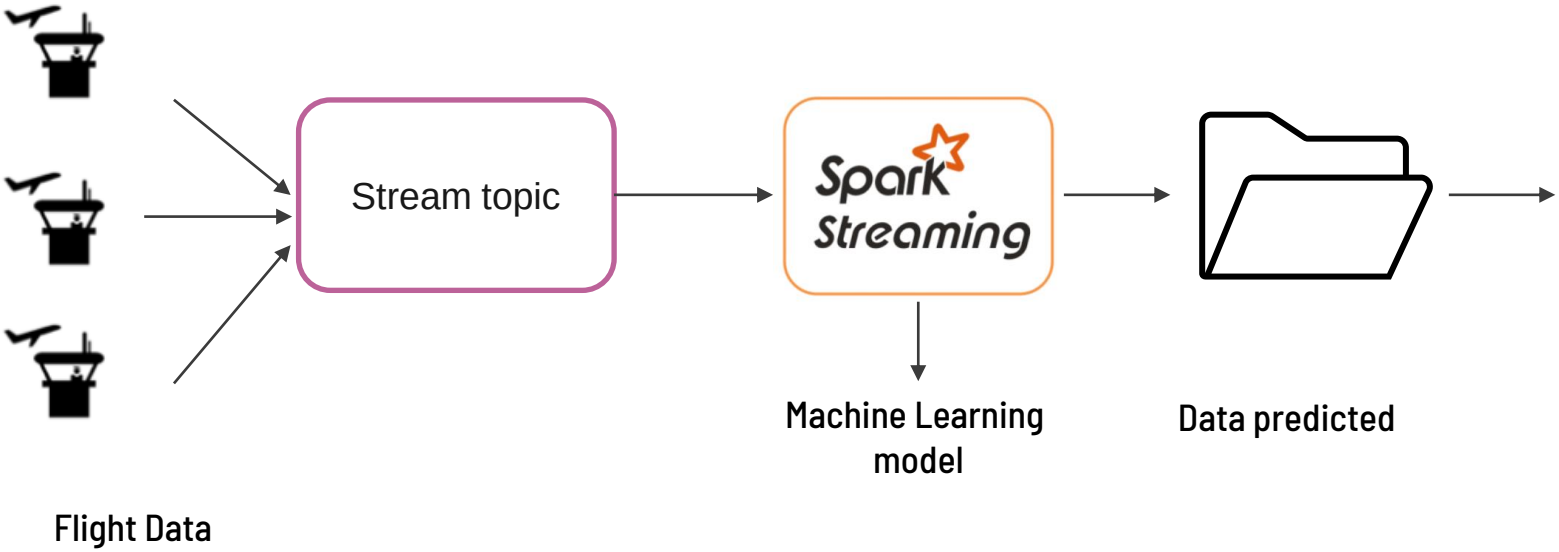
Tổng Quan Hệ Thống



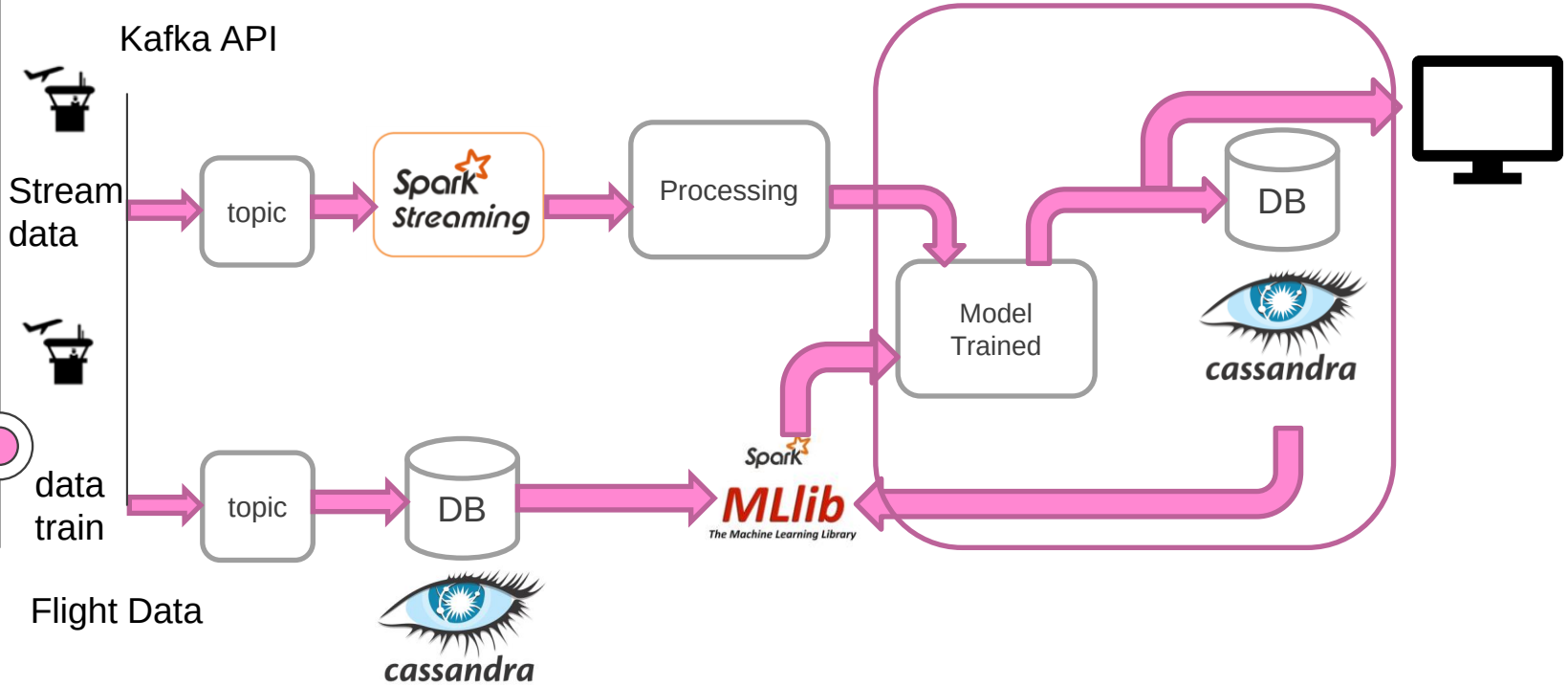
Apache Kafka



streaming


Stream topic
Spark Streaming
Machine Learning model
Data predicted
Flight Data

Tổng Quan Hệ Thống



Update Model

Data predicted

Update
Model

Apache Cassandra

Tốc độ đọc-ghi nhanh

Có thể xử lý các tập dữ liệu lớn

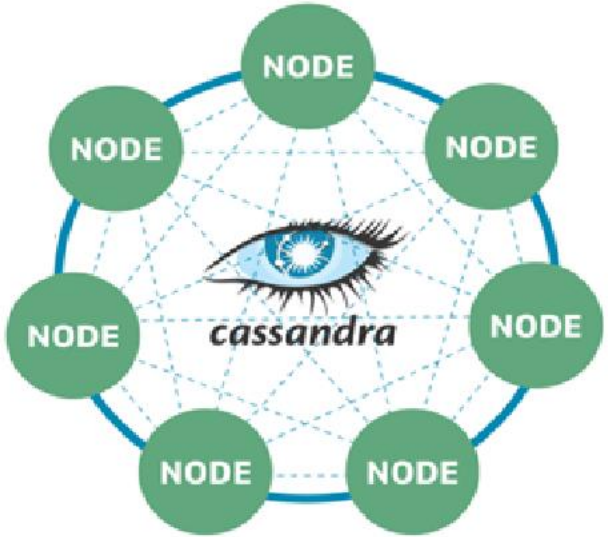
Khả năng chịu lỗi cao

Các ứng dụng cốt lõi dễ tích hợp

Dễ quản lý

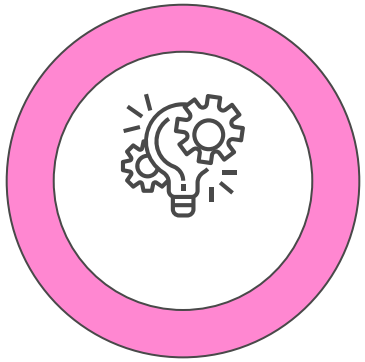
Apache Casandra

ApacheCassandra™= NoSQL Distributed Database



Demo


App
Chuyen Bay/Departures
Time: 07:13:19
ID
ORIGIN
DEST
CRS_DEP_TIME
prediction
4/13/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay
4/14/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay
4/15/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay
4/16/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay
4/17/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay
4/18/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay
4/19/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay
4/20/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay
4/21/2022-OH-CLT-CSG-5339
CLT
CSG
20.26
Not Delay



Thanks for watching