

# Dự Đoán Rủi Ro Tín Dụng

## Ứng dụng Big Data trong Fintech

Nguyễn Công Cường    Hoàng Tiên Đạt

Nhóm 8 - K68 UET  
Đại học Công nghệ - ĐHQGHN

Ngày 27 tháng 11 năm 2025

# Nội dung trình bày

- 1 Giới thiệu & mục tiêu
- 2 Bài toán & Dữ liệu
- 3 Kiến trúc hệ thống
- 4 Kết quả
- 5 Kết luận & hướng phát triển

# Giới thiệu đề tài & mục tiêu

- Bài toán: dự đoán khả năng vỡ nợ của khách hàng.
- Sử dụng công nghệ (Spark, Kafka, Hadoop HDFS, RestAPI, Zookeeper) trong chấm điểm tín dụng.
- Kết hợp xử lý batch (offline) và realtime (online).

## RỦI RO TÍN DỤNG & KHÁCH HÀNG UNBANKED



# Bài toán dự đoán rủi ro tín dụng

- Phân loại nhị phân:

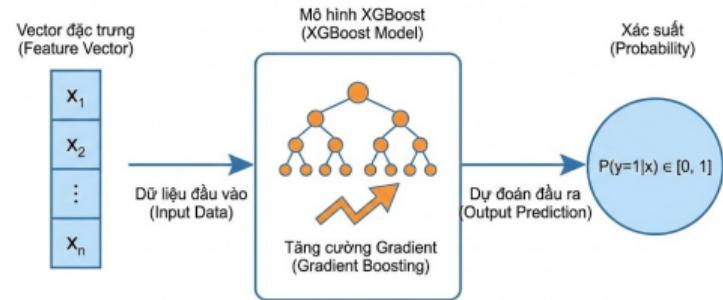
- Target = 1: có nguy cơ vỡ nợ.
- Target = 0: trả nợ bình thường.

- Đầu vào:

- Nhân khẩu học, thu nhập, nghề nghiệp.
- Thông tin khoản vay, lịch sử tín dụng.

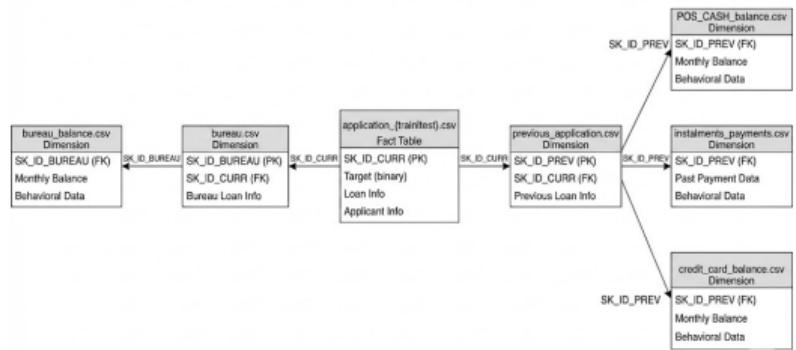
- Model sử dụng:

- Thư viện SparkML: XGBoost
- Một số model khác: Logistic Regression, Random Forest.



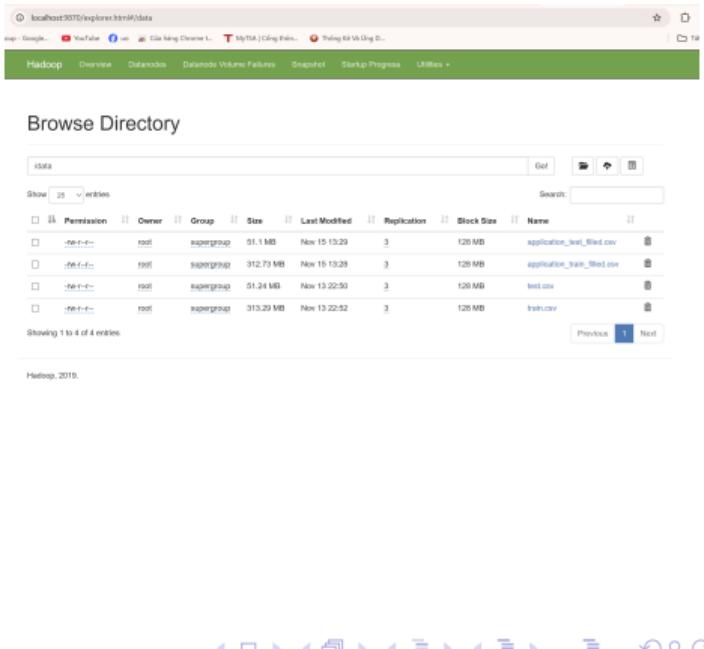
# Bộ dữ liệu & các bảng quan hệ

- **Bảng quan hệ:**
  - Bảng ứng dụng chính (application).
  - Các bảng lịch sử vay, thẻ tín dụng, trả góp, v.v.
- Dữ liệu từ file CSV, nạp vào HDFS & Spark.
- Biến mục tiêu: TARGET, cùng các biến nhân khẩu học, tài chính, lịch sử.



# Tiền xử lý dữ liệu

- Nạp data từ các file csv vào Spark.
- Chuẩn hóa kiểu dữ liệu, join theo SK\_ID\_CURR.
- Làm sạch:
  - Xử lý giá trị thiếu, ngoại lệ (ví dụ 365243 ngày).
  - Loại bỏ bản ghi kém chất lượng.
  - Lưu data trên HDFS Namenode.

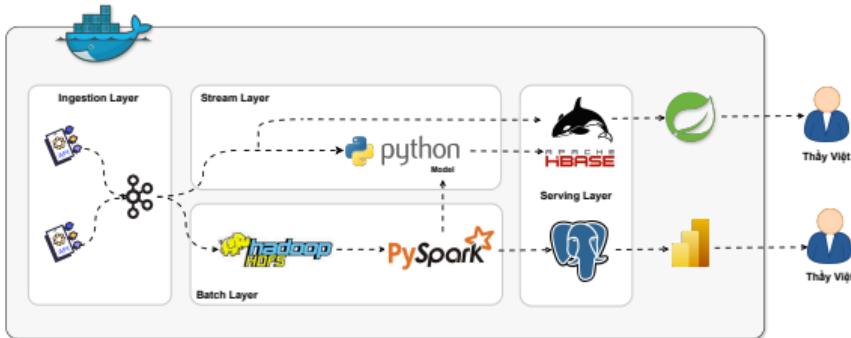


The screenshot shows a web-based Hadoop File Explorer interface. The URL in the address bar is "localhost:9070/explorer.html#/data". The page title is "Browse Directory". The main content area displays a table of files in the "/data" directory. The columns in the table are: Name, Block Size, Replication, Last Modified, Size, Group, Owner, and Permission. There are four entries listed:

Name	Block Size	Replication	Last Modified	Size	Group	Owner	Permission
application_test_file.csv	128 MB	3	Nov 15 13:29	51.1 MB	supergroup	root	-rwt----
application_train_file.csv	128 MB	3	Nov 15 13:28	312.73 MB	supergroup	root	-rwt----
test.csv	128 MB	3	Nov 15 13:29	51.34 MB	supergroup	root	-rwt----
train.csv	128 MB	3	Nov 15 13:29	313.29 MB	supergroup	root	-rwt----

At the bottom of the table, it says "Showing 1 to 4 of 4 entries". Below the table, there is a footer note: "Hadoop, 2019." The browser tabs at the top include "Gmail", "YouTube", "Facebook", "GiaiNhungChrome L.", "TinTuc | Công Nghệ...", "Thống Kê Vũ Ông D.", and "Tin Tuc". The status bar at the bottom right shows navigation icons.

# Kiến trúc tổng thể batch + realtime



## • Luồng batch:

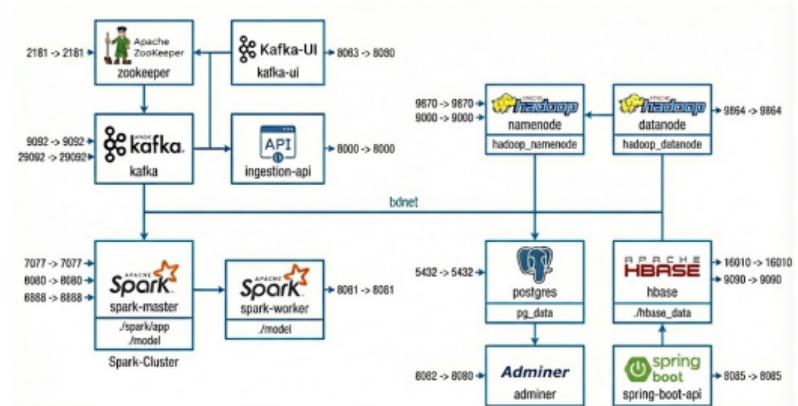
- Spark xử lý dữ liệu.
- Xây dựng đặc trưng & huấn luyện XGBoost.
- Kết quả được lưu trong PostgreSQL
- Power Bi lấy data từ PostgreSQL xây dựng dashboard

## • Luồng realtime:

- REST API → Kafka topic credit\_application.
- Spark Streaming chấm điểm, ghi kết quả ra PostgreSQL/HBase, kafka topic credit\_score lưu kết quả
- Spring Boot đóng vai trò là cầu nối giữa user và Hbase

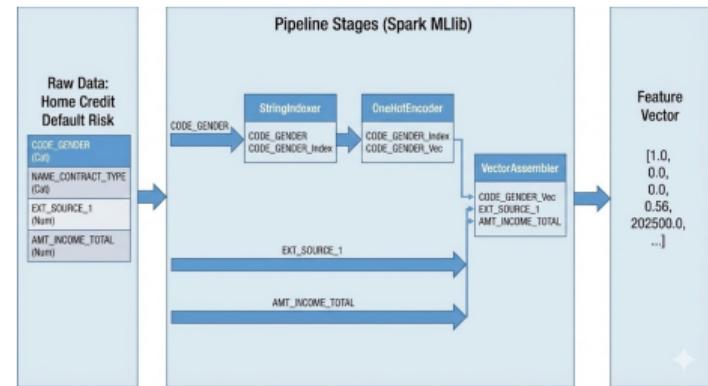
# Môi trường triển khai

- Các dịch vụ chạy trong Docker:
  - Kafka, Zookeeper, Spark Master, Worker.
  - HDFS (NameNode, DataNode), PostgreSQL, HBase
  - Spring boot, Power Bi
- Cấu hình tài nguyên: 4 cores, 6GB RAM cho Spark worker.



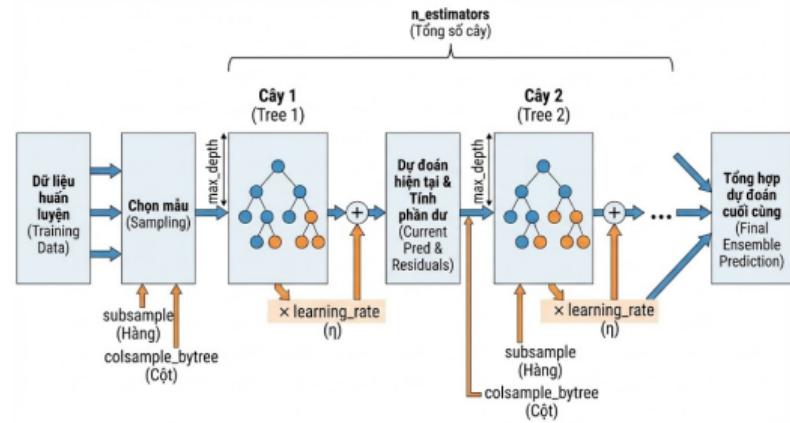
# Xây dựng pipeline đặc trưng

- Mã hóa biến phân loại:
  - StringIndexer, OneHotEncoder.
- Gộp đặc trưng bằng VectorAssembler.
- Lưu **pipeline đặc trưng** để dùng chung cho:
  - Huấn luyện batch.
  - Chấm điểm realtime.



# Huấn luyện & triển khai mô hình XGBoost

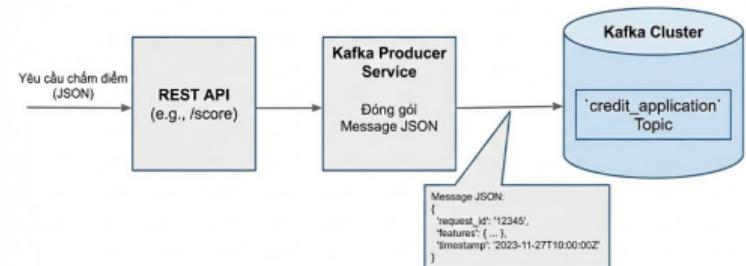
- Mục tiêu: output , đánh giá bằng AUC-ROC.
- Một số siêu tham số:
  - max\_depth, learning\_rate, n\_estimators.
  - subsample, colsample\_bytree.
- Lưu mô hình & pipeline trên HDFS, hỗ trợ versioning.



# Luồng dữ liệu realtime qua Kafka

- REST API nhận yêu cầu chấm điểm.
- Producer đẩy JSON vào topic `credit_application`.
- Message chứa: `request_id`, đặc trưng đầu vào, timestamp.

Sơ đồ Luồng Dữ liệu Realtime: REST API -> Kafka Topic



Sơ đồ Luồng Dữ liệu Realtime: REST API -> Kafka Topic

# Kafka topic credit\_application

```
Key Value Headers
{
    "SK_ID_CURR": 676962,
    "NAME_CONTRACT_TYPE": "Cash loans",
    "CODE_GENDER": "F",
    "FLAG_OWN_CAR": "Y",
    "FLAG_OWN_REALTY": "Y",
    "AMT_INCOME_TOTAL": 174591.0,
    "AMT_CREDIT": 976932.0,
    "AMT_ANNUITY": 70429.22284554469,
    "AMT_GOODS_PRICE": 797008.6104726905,
    "DAYS_BIRTH": -25550.0,
    "DAYS_EMPLOYED": -8030.0,
    "NAME_EDUCATION_TYPE": "Higher education",
    "NAME_FAMILY_STATUS": "Single / not married",
    "NAME_HOUSING_TYPE": "Office apartment",
    "ORGANIZATION_TYPE": "Business Entity Type 1",
    "EXT_SOURCE_1": 0.645434911544178,
    "EXT_SOURCE_2": 0.8087240489415076,
    "EXT_SOURCE_3": 0.7044959144446498,
    "TARGET": 0,
    "event_time": "2025-11-28T01:00:26.662051"
}
```

# Spark Structured Streaming chấm điểm realtime

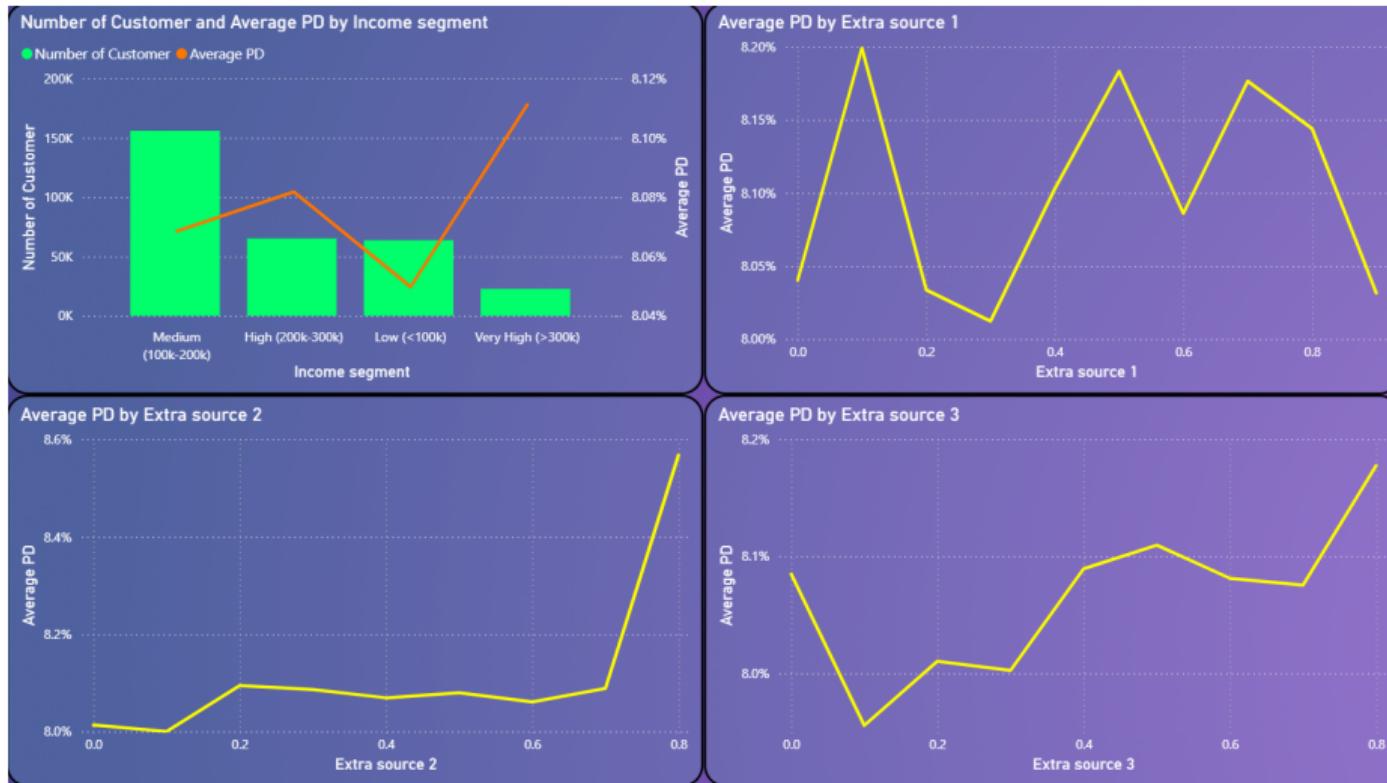
- Áp dụng pipeline đặc trưng đã lưu.
- Gọi mô hình XGBoost để ra:
  - Xác suất vỡ nợ.
  - Nhãn dự đoán.
- Ghi kết quả ra PostgreSQL/HBase (view batch & realtime).
- Qua spring boot hiện thị UI cho user

The screenshot shows the Apache Kafka UI interface. At the top, there is a navigation bar with tabs: Topics (selected), Messages, Consumers, Settings, and Statistics. Below the navigation bar, there are dropdown menus for Seek Type (Offset), Partitions (All items are selected.), Key Serde (String), and Value Serde (String). A search bar and a 'Add Filters' button are also present. The main area displays two messages in the 'credit\_scores' topic. Message 0 has a key of 517677 and a value of { "sk\_id\_curr": 517677, "pd\_1": 0.133346644639969, "ts": "2025-11-27T23:21:32.705Z" }. Message 1 has a key of 676962 and a value of { "sk\_id\_curr": 676962, "pd\_1": 0.08626745641231537, "ts": "2025-11-28T01:00:30.026Z" }.

# Kết quả mô hình trên dữ liệu batch



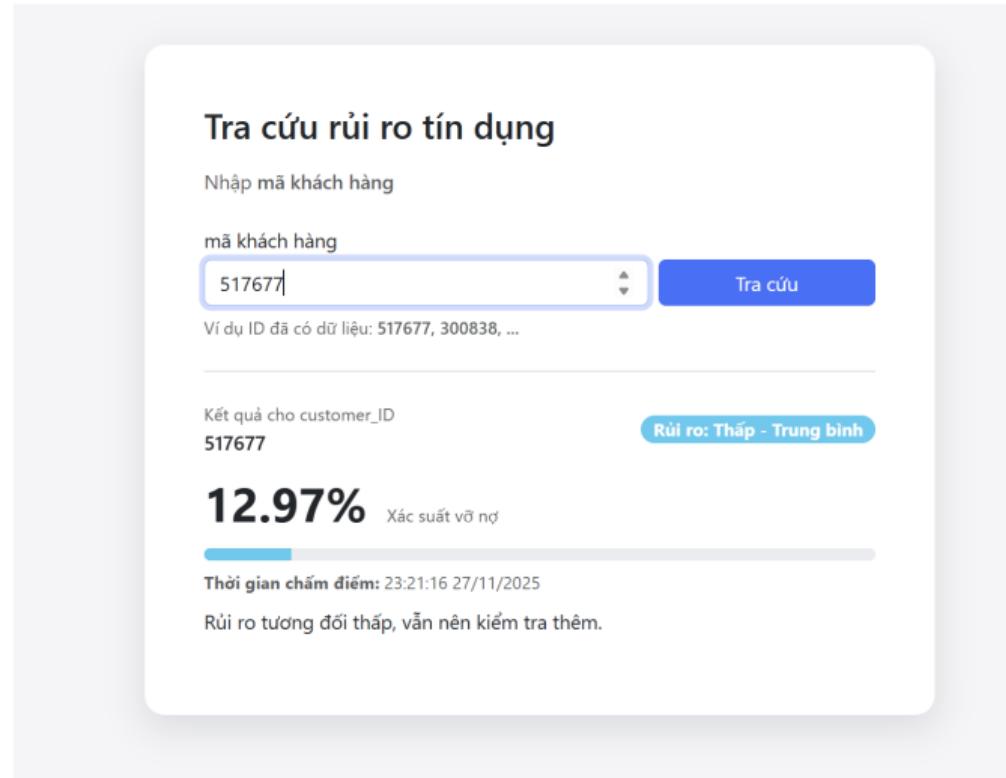
# Kết quả mô hình trên dữ liệu batch



# Kết quả pipeline realtime

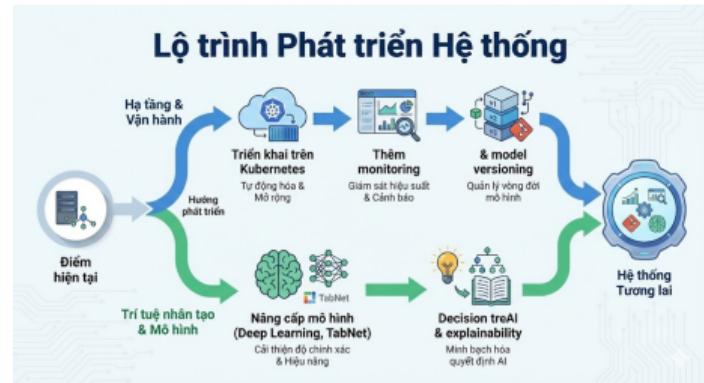
- Kết quả

- Nhập mã định danh của khách hàng
- Kết quả là mức độ rủi ro của tín dụng



# Kết luận & hướng phát triển

- Đã xây dựng hệ thống chấm điểm tín dụng:
  - Áp dụng các công nghệ Bigdata
  - Mang lại cho người dùng tổng quan về dữ liệu, tiện ích khi tập trung vào một flow duy nhất
  - Giải quyết được bài toán đã đề ra
- Hướng phát triển:
  - Triển khai trên Kubernetes, thêm monitoring & model versioning.
  - Nâng cấp mô hình (Deep Learning, TabNet) & explainability.
  - Sử dụng các mô hình gauss để sinh data



Cảm ơn Thầy và các bạn đã lắng nghe!