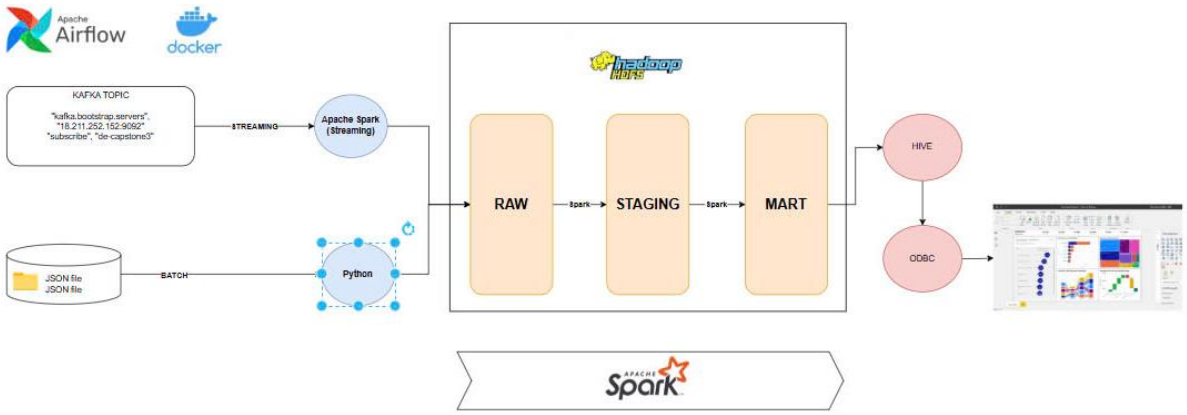
📌 **Tên dự án:**

Real-Time & Batch Data Ingestion Platform for JSON Data using Apache Kafka, Spark, Hadoop, and Hive

🎯 **Mục tiêu dự án:**

Thiết kế và triển khai một nền tảng thu thập, xử lý và phân tích dữ liệu JSON từ nhiều nguồn khác nhau, bao gồm luồng thời gian thực và batch, tích hợp với hệ sinh thái Big Data để phục vụ phân tích dữ liệu và trực quan hóa trên công cụ BI (như Power BI).



🧩 **Thành phần hệ thống:**

1. Nguồn dữ liệu đầu vào

- **Kafka Topic (Streaming)**:
    - Kafka Bootstrap Servers: 18.212.206.152:9092
    - Topic: "de-captions"
    - Consumer Group: "subscriber"
- **Batch JSON File**:
    - Định dạng: .json
    - Được ingest từ file system (local hoặc cloud).

2. Xử lý dữ liệu

a. Apache Spark Streaming (Dành cho dữ liệu từ Kafka):

- Đọc dữ liệu từ Kafka theo thời gian thực.
- Ghi dữ liệu vào zone RAW của HDFS.
- Chạy trên Docker hoặc cluster Hadoop (yarn client).

b. Python Batch Job:

- Chạy định kỳ (sử dụng Airflow).
- Đọc JSON files.
- Ghi dữ liệu vào RAW zone.

c. ETL Spark Jobs:

- **RAW → STAGING → MART**.
- Làm sạch dữ liệu, chuẩn hóa schema, enrich dữ liệu nếu cần.
- Tách xử lý cho từng layer:
    - RAW: dữ liệu thô chưa xử lý.
    - STAGING: dữ liệu được chuẩn hóa, tách cột, định dạng đúng kiểu.
    - MART: dữ liệu sẵn sàng phục vụ phân tích.

3. Hệ thống lưu trữ và phân tích

a. Hadoop HDFS:

- Lưu trữ toàn bộ dữ liệu RAW, STAGING, và MART theo định dạng phân tán (Parquet hoặc ORC).

b. Hive Metastore:

- Quản lý schema và bảng dữ liệu trên HDFS.
- Các bảng được định nghĩa cho từng zone.

c. ODBC/BI Tool:

- Dữ liệu được expose ra ngoài thông qua ODBC.
- Power BI kết nối với Hive thông qua ODBC để tạo báo cáo.

---

4. Orchestration

- **Apache Airflow**:
    - Quản lý job pipeline:
        - Batch job đọc JSON file.
        - Trigger Spark ETL.
        - Kiểm tra chất lượng dữ liệu.
    - Lên lịch chạy định kỳ (daily/hourly).

⚙️ **Yêu cầu kỹ thuật:**

Data Lake Structure:

bash

CopyEdit

/data_lake/

├— raw/

├— staging/

└— mart/

Spark Config:

- Mode: Yarn-client
- Format: Parquet
- Partitioning: theo ngày (partitioned by dt)

Hive:

- Tạo schema đồng bộ với cấu trúc dữ liệu đã transform.
- Dùng external table để mapping dữ liệu trên HDFS.

📊 **Tích hợp BI:**

- Kết nối Power BI với Hive thông qua ODBC.
- Truy xuất dữ liệu từ MART layer.
- Dữ liệu phục vụ trực quan hóa các chỉ số phân tích.

🔐 **Bảo mật:**

- Sử dụng ACL cho Kafka.
- Xác thực người dùng với Hive.
- Hạn chế truy cập vùng RAW.

🛠️ **Các công nghệ sử dụng:**

Công cụ	Mục đích
Apache Kafka	Thu thập dữ liệu thời gian thực
Apache Spark	Xử lý ETL cả streaming và batch
Hadoop HDFS	Lưu trữ dữ liệu phân tán
Hive	Metadata và phân tích SQL
Python	Xử lý batch file JSON
Airflow	Orchestration
Docker	Đóng gói và deploy môi trường
Power BI	BI Visualization

🗓️ **Lịch trình thực hiện (gợi ý)**

Giai đoạn	Thời gian
Phân tích yêu cầu	1 tuần
Setup hạ tầng	1 tuần
Dev ETL pipeline	2 tuần
Tích hợp & Test	1 tuần
BI & Báo cáo	1 tuần
Go-live	Sau khi kiểm thử