

DỰ BÁO THỜI TIẾT THỜI GIAN THỰC BẰNG HỌC MÁY TRÊN NỀN TẢNG CONFLUENT KAFKA VÀ PYSPARK DATABRICKS

Ngô Huy Tài

Trường Đại học Công nghệ Thông tin - Đại học Quốc gia Tp.HCM

What ?

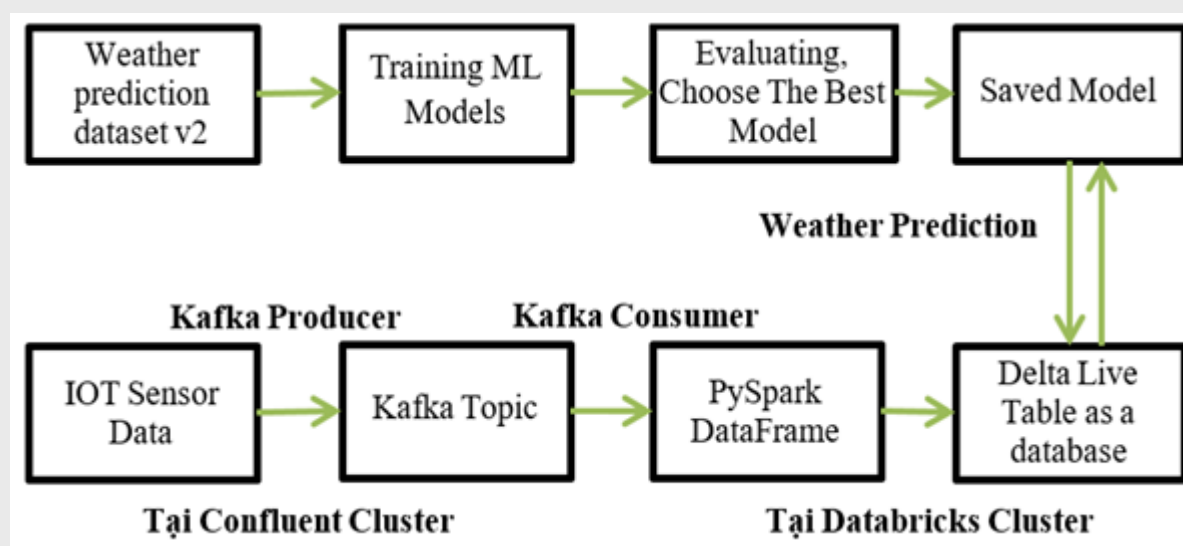
Chúng tôi nghiên cứu và thiết kế hoàn chỉnh một hệ thống dự báo thời tiết thời gian thực bằng học máy:

- Xây dựng hệ thống tính toán song song, quản lý và xử lý dữ liệu thời tiết.
- Huấn luyện các mô hình trên dữ liệu mẫu, chọn ra mô hình dự báo tốt nhất.
- Hoàn thiện hệ thống dự báo với mô hình học máy trên dữ liệu thực.

Why ?

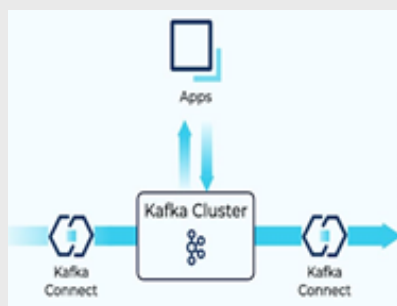
- Dự báo thời tiết trong vai trò vô cùng quan trọng trong cuộc sống con người. Dự báo thời tiết sớm, chính xác, giúp chúng ta chủ động trong sinh hoạt hằng ngày, tối ưu trong hoạt động sản xuất, kinh doanh và có biện pháp, kế hoạch ứng phó sớm, giảm thiểu tác hại do thiên tai gây ra.
- Hệ thống tính toán song song **xử lý hiệu quả dữ liệu lớn** and dự báo thời tiết bằng mô hình học máy **chính xác, nhanh chóng** theo thời gian thực.

Overview



Description

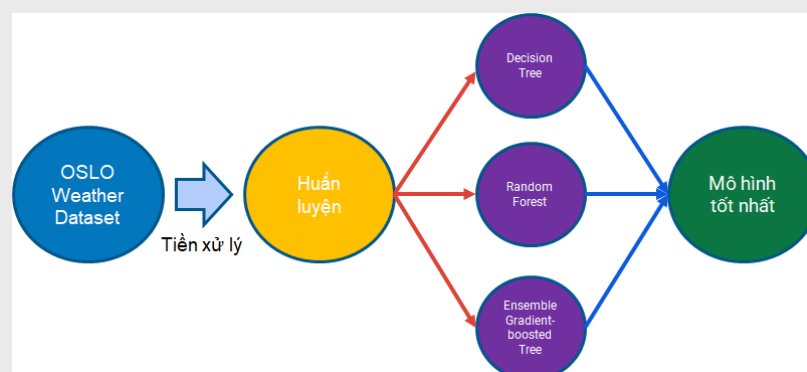
1. Xây dựng hệ thống tính toán song song:



Hình 1. Sơ đồ thiết lập Kafka Cluster.

- Dùng Kafka tạo hệ thống dữ liệu luồng.

2. Huấn luyện và chọn mô hình tối ưu cho dự báo thời tiết:



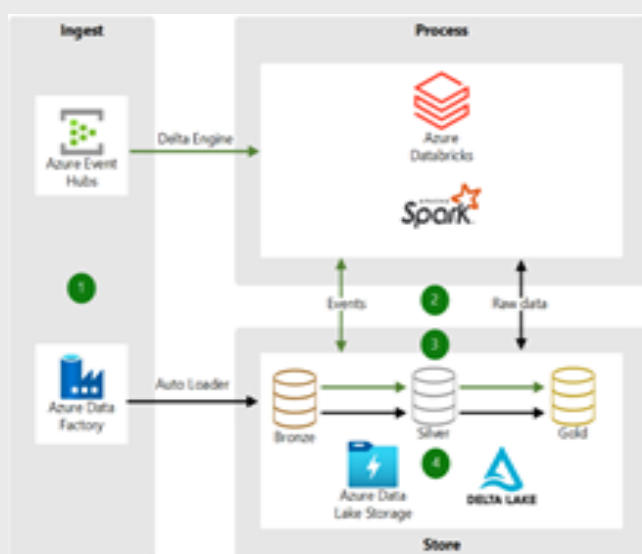
Hình 3. Sơ đồ xử lý dữ liệu mẫu, huấn luyện và chọn ra mô hình dự báo tối ưu nhất

3. Thực hiện dự báo trên dữ liệu thực:

- Triển khai hệ thống, dữ liệu thời tiết mới từ các cảm biến sẽ được gửi về và cập nhật liên tục trên Kafka.
- Sử dụng mô hình tốt nhất vừa được chọn để dự báo với dữ liệu mới từ hệ thống vừa cập nhật.



Hình 4. Sơ đồ hệ thống dự báo thời tiết thời gian thực.



Hình 2. Sơ đồ thiết lập Azure Databricks Cluster.

- Trên Azure Databricks tạo Spark Cluster gồm 3 workers.
- Thiết lập các Jupyter Notebooks phục vụ cho hệ thống.