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

Ngô Huy Tài - 230201028

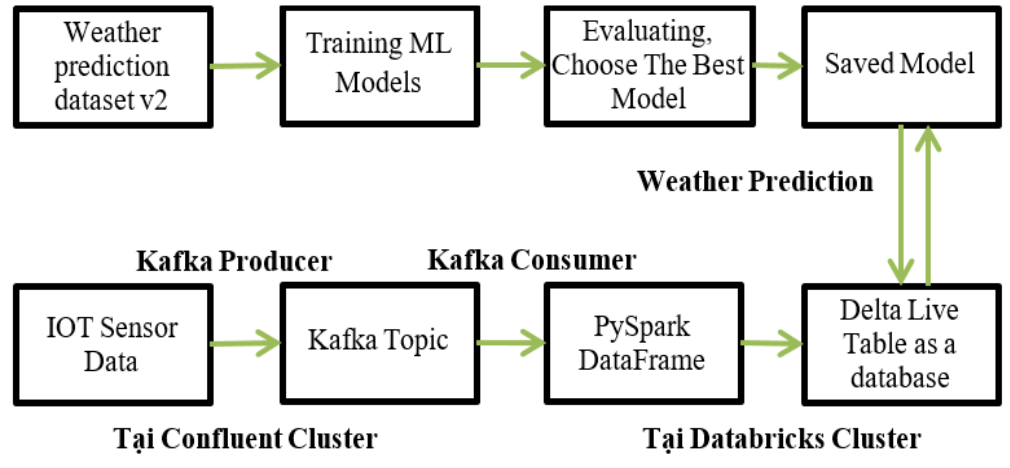
Tóm tắt

- Lớp: CS2205.MAR2024
- Link Github: https://github.com/tainhuit/CS2205.MAR2024/blob/main/DuBaoThoiTietTheoThoiGianThucBangHocMay_Kafka_Spark.pdf
- Link YouTube video: https://youtu.be/PSsWDRzeA58
- Ảnh + Họ và Tên:

Ngô Huy Tài

Giới thiệu

- Dự báo thời tiết đóng vai trò vô cùng quan trọng trong đời sống con người.
- Dùng hệ thống tính toán song song và mô hình máy học để dự báo có độ chính xác cao, xử lý nhanh.

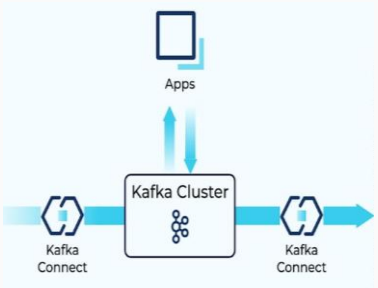


Mục tiêu

- **Xây dựng hệ thống tính toán song song**: Dữ liệu được truyền dạng luồng, được quản lý, xử lý trên nền tảng Confluent Kafka và PySpark trên Azure Databricks.
- **Chọn mô hình dự báo tối ưu nhất:** Huấn luyện 3 mô hình máy học trên bộ dữ liệu “Weather prediction dataset v2”. Chọn ra mô hình tối ưu nhất.
- **Thực hiện dự báo trên dữ liệu thời gian thực:** Dùng mô hình vừa được chọn, triển khai trên hệ thống dữ liệu đang cập nhật theo thời gian thực.

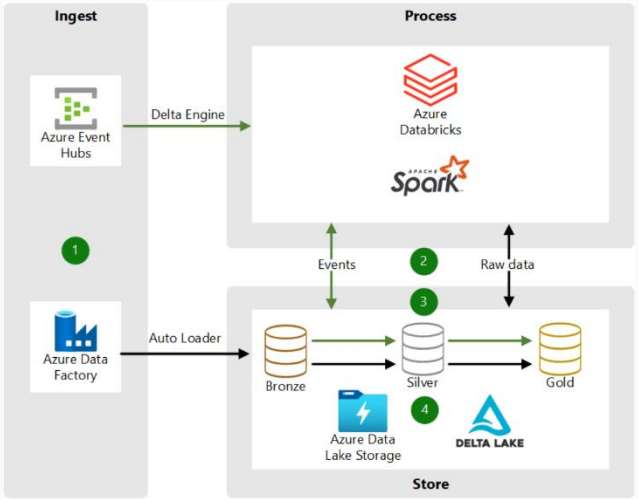
Nội dung và Phương pháp

Nội dung 1: Xây dựng hệ thống tính toán song song:



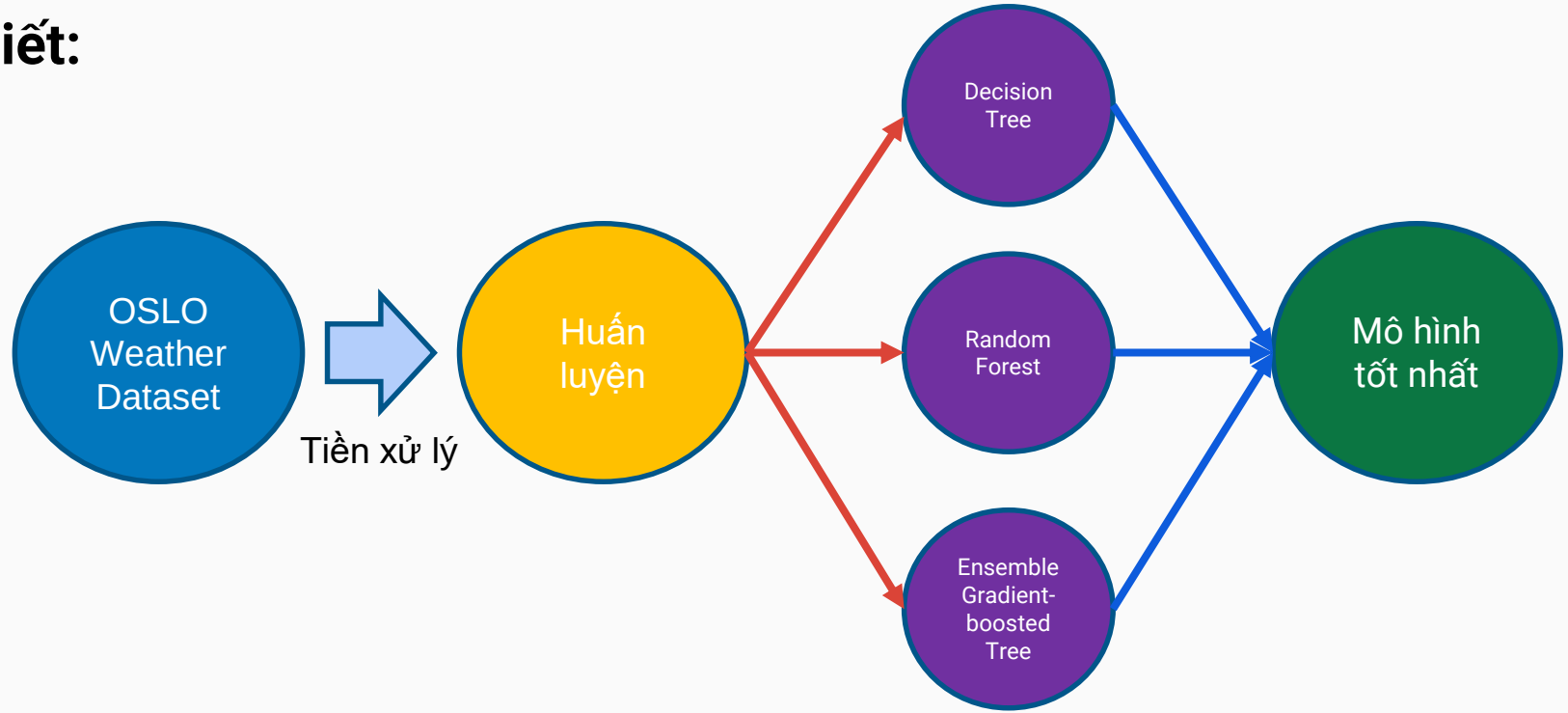
- Dùng Kafka tạo hệ thống dữ liệu luồng.

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



Nội dung và Phương pháp

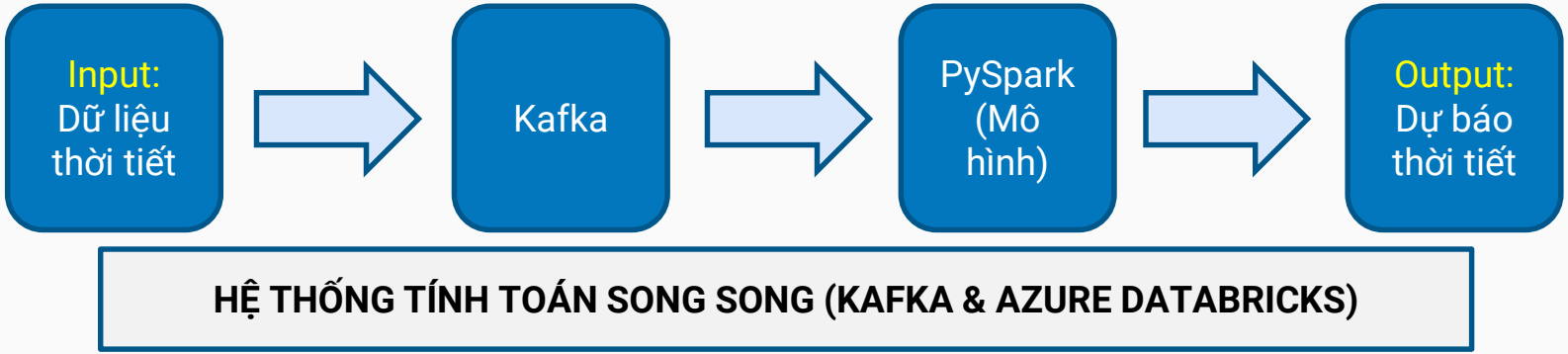
Nội dung 2: Huấn luyện và chọn mô hình tối ưu cho dự báo thời tiết:



Nội dung và Phương pháp

Nội dung 3: Thực hiện dự báo trên dữ liệu thực:

- Triển khai hệ thống, dữ liệu thời tiết mới từ các cảm biến sẽ được gởi 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.



Kết quả dự kiến

- Báo cáo tổng hợp về các nền tảng Kafka, Spark.
- Xây dựng được hệ thống tính toán song song phục vụ dự báo thời tiết.
- Xây dựng bộ dữ liệu thời tiết mẫu của thành phố OSLO trong 10 năm.
- Chọn được mô hình tối ưu phục vụ cho dự báo thời tiết.
- Hoàn thiện hệ thống dự báo thời tiết theo thời gian thực bằng học máy.

Tài liệu tham khảo

[1]. Shapira, G., Palino, T., Sivaram, R., & Petty, K.: *Kafka: the definitive guide*. O'Reilly Media, Inc. 2021

[2]. Padhy, Rabi Prasad: Big data processing with Hadoop-MapReduce in cloud systems. International Journal of Cloud Computing and Services Science 2.1. 2013: 16-27.

[3]. Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, Ameet Talwalkar: Mllib: Machine learning in apache spark. Journal of Machine Learning Research 17.34. 2016: 1-7.

[4]. Rokach, Lior, and Oded Maimon: Decision trees. Data mining and knowledge discovery handbook. 2005: 165-192.

[5]. Breiman, Leo: Random forests. Machine learning 45. 2001: 5-32.

Tài liệu tham khảo

[6]. Friedman, Jerome H: Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001: 1189-1232.
[7]. Ashofteh, Afshin: Big data for credit risk analysis: Efficient machine learning models using PySpark. International Workshop on Simulation and Statistics. Cham: Springer International Publishing. 2019.
[8]. Jayanthi, D., and G. Sumathi: Weather data analysis using spark—an in-memory computing framework. Innovations in Power and Advanced Computing Technologies i-PACT. IEEE 2017.
[9]. Huber, F: Weather Prediction Dataset, v2. Zenodo. 2021.
[10]. Etaati, Leila, and Leila Etaati: Azure databricks. Machine Learning with Microsoft Technologies: Selecting the Right Architecture and Tools for Your Project. 2019: 159-171.