THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/PSsWDRzeA58*
- Link slides (dạng .pdf đặt trên Github): *https://github.com/tainhuit/CS2205.MAR2024/blob/main/DuBaoThoiTietTheoThoiGianThucBangHocMay_Kafka_Spark.pdf*
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Ngô Huy Tài • MSSV: 230201028 	• Lớp: CS2205.MAR2024 • Tự đánh giá (điểm tổng kết môn): 8.5/10 • Số buổi vắng: 0 • Số câu hỏi QT cá nhân: 3 • Link Github: https://github.com/tainhuit/CS2205.MAR2024

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

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

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

REAL-TIME WEATHER FORECAST USING MACHINE LEARNING ON CONFLUENT KAFKA AND PYSPARK DATABRICKS PLATFORMS

TÓM TẮT *(Tối đa 400 từ)*

Dự báo thời tiết đóng vai trò rất quan trọng đối đời sống con người, tác động đến nhiều lĩnh vực như nông nghiệp, giao thông, du lịch, năng lượng... Việc dự đoán chính xác diễn biến thời tiết giúp chúng ta chủ động ứng phó với các hiện tượng bất thường, giảm thiểu thiệt hại và tối ưu hóa hoạt động sản xuất, kinh doanh.

Vấn đề then chốt trong dự báo thời tiết hiện đại nằm ở việc phân tích và xử lý lượng dữ liệu rất lớn được được truyền tải về từ cảm biến, từ các trạm quan trắc, vệ tinh, radar... Dữ liệu thời tiết có đặc điểm đa dạng, phức tạp, thay đổi và cập nhật liên tục, đòi hỏi các giải pháp và công nghệ tiên tiến để xử lý một cách hiệu quả.

Để xử lý dữ liệu thời tiết trên quy mô lớn, Hadoop MapReduce [2] là một công cụ được sử dụng rất phổ biến trước đây. Tuy nhiên, nền tảng này vẫn còn một số hạn chế nhất định, trong khi đó Spark [3] là một công nghệ mới, có những ưu điểm nổi trội và đang dần thay thế Hadoop MapReduce nhờ khả năng tính toán trong bộ nhớ (In-memory computing), cho hiệu suất xử lý cao hơn.

Đề tài này đề xuất dùng hệ thống tính toán song song, trong đó Confluent Platform [1] được dùng để xây dựng ứng dụng dữ liệu dạng luồng (stream). Dữ liệu từ nhiều nguồn sẽ được tích hợp và truyền tới một nơi tập trung đó là Kafka [1]. Đồng thời, sử dụng Spark để xử lý và phân tích dữ liệu thời tiết một cách nhanh chóng và hiệu quả. Để dự báo thời tiết, đề tài đề xuất 3 mô hình máy học là Decision Tree [4], Random Forest [5] và Ensemble Gradient-boosted Tree [6]. Huấn luyện chúng với dữ liệu mẫu và chọn ra mô hình tốt nhất sử dụng cho đề tài.

Việc triển khai Spark trong dự báo thời tiết hứa hẹn mang lại nhiều lợi ích, bao gồm: độ chính xác cao, xử lý nhanh, mở rộng hay thu hẹp dễ dàng, giảm chi phí vận hành.

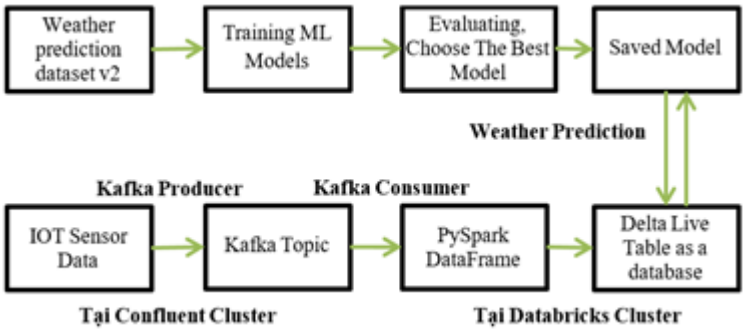
GIỚI THIỆU *(Tối đa 1 trang A4)*

Trong cuộc sống hiện đại ngày nay, 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. Đóng góp rất lớn vào sự phát triển kinh tế, xã hội.

Dùng máy học để phân tích dữ liệu lớn là một giải pháp mới, như công trình của tác giả Afshin Ashofteh [7], trên PySpark dùng mô hình Logistic Regression để đánh giá dữ liệu tín dụng. Cho kết quả chính xác, đầy đủ và thận trọng đối với khách hàng tín dụng. Nhưng đề tài vẫn còn hạn chế về số liệu và chỉ thực nghiệm trên một mô hình.

Đối với dự báo thời tiết dùng máy học trên dữ liệu lớn thì nhóm tác giả D.Jayanth, G.Sumathi [8] đã tạo một Jupyter Notebook trên nền tảng Spark. Dữ liệu thời tiết được lấy từ các trang web thời tiết và đưa vào Jupyter Notebook này. Sau khi xử lý và phân tích cho ra báo cáo về lượng mưa và nhiệt độ của 10 trạm dự báo thời tiết. Tuy nhiên đề tài vẫn còn hạn chế thiếu nhiều dữ liệu thời tiết khác. Nên chưa thể thay thế cho một hệ thống dự báo thời tiết truyền thống.

Bằng việc kết hợp các nền tảng Kafka và PySpark, tham khảo và cải tiến hạn chế của những nghiên cứu trên. Khác với phương pháp dự báo truyền thống thực hiện bởi các dự báo viên. Đề tài này hoàn toàn không phụ thuộc vào con người, tất cả các tác vụ đều do máy tính thực hiện. Từ việc thu thập, quản lý, xử lý dữ liệu cho đến sử dụng các mô hình để đưa ra dự báo. Dữ liệu thời tiết như ảnh mây, tốc độ gió, độ ẩm, mực nước biển, bức xạ, lượng mưa/nắng, nhiệt độ… từ các cảm biến sẽ được xử lý theo dạng luồng và liên tục được chuyển đến Databricks Cluster thông qua Kafka Streaming. Bên cạnh đó, huấn luyện 3 mô hình máy học với dữ liệu mẫu "Weather prediction dataset v2" của Huber Florian [9]. Chúng tôi đánh giá và chọn ra mô hình tốt nhất dùng để dự báo thời tiết theo dữ liệu thời gian thực. Do đó đề tài mang tính ứng dụng cao, có thể thay thế cho hệ thống dự báo thời tiết truyền thống hiện tại.



Với quy trình như trên, mục tiêu của đề tài là giải quyết bài toán như sau:
Input: Dữ liệu thời tiết được cập nhật liên tục từ cảm biến.
Output: Dự báo thời tiết bằng mô hình học máy trên hệ thống tính toán song song.

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- **Xây dựng hệ thống tính toán song song:** Xây dựng hệ thống tính toán song song phục vụ dự báo thời tiết. Trong đó dữ liệu được truyền dạng luồng, được quản lý, xử lý trên nền tảng Confluent Kafka và PySpark trên Azure Databricks [10].
- **Chọn mô hình dự báo tối ưu nhất:** Sử dụng bộ dữ liệu công khai "Weather prediction dataset v2" của Huber Florian để huấn luyện các mô hình máy học Decision Tree, Random Forest và Ensemble Gradient-boosted Tree. Từ kết quả huấn luyện, thực hiện đánh giá và chọn ra được mô hình có độ chính xác cao nhất (kỳ vọng trên 90%) để phục vụ cho việc dự báo thời tiết theo thời gian thực của đề tài.
- **Thực hiện dự báo trên dữ liệu thời gian thực:** Thực hiện dự báo thời tiết trên hệ thống tính toán song song đã được triển khai bằng mô hình dự báo mới được chọn. Dữ liệu thời tiết mới từ các cảm biến sẽ được gởi về hệ thống và cập nhật liên tục. Sử dụng mô hình dự báo này, căn cứ trên dữ liệu thời tiết mới cập nhật đưa ra các dự báo một cách định kỳ, liên tục.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

Nội dung 1: Xây dựng hệ thống tính toán song song: Nghiên cứu các nền tảng Apache Kafka, Apache Spark, Azure. Từ đó xây dựng hệ thống dự báo thời tiết với dữ liệu dạng luồng phục vụ cho việc truyền, quản lý, xử lý dữ liệu.

- Bước 1: Triển khai Confluent Kafka Cluster và PySpark trên Azure Databricks Cluster. Tạo Kafka Topic để chứa các Kafka Message, sử dụng Datagen của Confluent với vai trò như một Kafka Producer để kết nối và gửi dữ liệu đến Kafka Topic.
- Bước 2: Trên Azure Databricks tạo mới Spark Cluster gồm 3 Workers với cấu hình 12 vCPUs và 42 GB Memory.
- Bước 3: Trên Databricks thiết lập 3 tập tin Jupyter Notebook riêng biệt gồm:
    1.Oslo-ML-Model-Training.ipynb;
    2.Kafka-Consumer-Streaming.ipynb;
    3.Weather-Forecast-with-Kafka-Streaming-IoT-Data.ipynb;

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

- Sử dụng bộ dữ liệu "Weather prediction dataset v2" của Huber Florian. Đây là bộ dữ liệu dự báo thời tiết đa dạng, thu thập từ các cảm biến IoT, được ghi nhận hàng ngày, trong khoảng thời gian 3.654 ngày từ năm 2000 đến 2010 tại 18 thành phố của nhiều nước Châu Âu khác nhau. Trong bộ dữ liệu này, đề tài chỉ sử dụng dữ liệu thời tiết của thành phố Oslo (Na Uy). Sau khi trích lọc, dữ liệu này được tiền xử lý, bằng cách kết hợp 2 tập tin, trích xuất các cột dữ liệu liên quan đến thành phố Oslo

cùng cột DATE (ngày ghi nhận dữ liệu) sau đó lưu thành tập tin oslo.csv để sử dụng.
- Chạy Jupyter Notebook 1.Oslo-ML-Model-Training.ipynb để huấn luyện các mô hình máy học Decision Tree, Random Forest và Ensemble Gradient-boosted Tree.
- Từ kết quả huấn luyện, thực hiện đánh giá và chọn ra được mô hình có độ chính xác cao nhất để phục vụ cho việc dự báo thời tiết theo thời gian thực của đề tài.

Nội dung 3: Thực hiện dự báo trên dữ liệu thực:
- Chạy Jupyter Notebook 2.Kafka-Consumer-Streaming.ipynb để triển khai hệ thống, dữ liệu thời tiết mới từ các cảm biến sẽ được gửi về và cập nhật liên tục trên Kafka.
– Chạy Jupyter Notebook
3.Weather-Forecast-with-Kafka-Streaming-IoT-Data.ipynb
Trong đó, dùng mô hình tốt nhất vừa được chọn để dự báo với dữ liệu mới từ hệ thống vừa cập nhật.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Báo cáo tổng hợp về các nền tảng Kafka, Spark.
- Xây dựng được hệ thống tính toán song song 3 Workers với cấu hình 12 vCPUs và 42 GB Memory.
- Xây dựng bộ dữ liệu thời tiết mẫu của thành phố OSLO trong 10 năm (từ 2000 đến 2010).
- Chọn được mô hình tối ưu phục vụ cho dự báo thời tiết.
- Hoàn thiện hệ thống dự báo thời tiết bằng học máy theo thời gian thực.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[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.
[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.