

BÁO CÁO CUỘC THI

AUTOSCALING ANALYSIS
DỰ ĐOÁN LƯU LƯỢNG HTTP VÀ TỰ ĐỘNG MỞ RỘNG MÁY CHỦ

CUỘC THI: CUỘC THI DATAFLOW 2026 - NĂM: 2026

[Tên thành viên 1] [MSSV 1] [Email 1]
[Tên thành viên 2] [MSSV 2] [Email 2]
[Tên thành viên 3] [MSSV 3] [Email 3]

BAN TỔ CHỨC: DATAFLOW 2026

MỤC LỤC

Mục lục	iii
1 Giới thiệu bài toán	1
1.1 Giới thiệu	1
1.2 Bài toán hồi quy: Dự đoán lưu lượng	1
1.3 Bài toán tối ưu: Autoscaling	2
1.4 Mục tiêu dự án	2

Chương 1

GIỚI THIỆU BÀI TOÁN

1.1 GIỚI THIỆU

Trong bối cảnh điện toán đám mây phát triển mạnh mẽ, việc tối ưu hóa tài nguyên thông qua Autoscaling (tự động mở rộng quy mô) đã trở thành một yêu cầu cấp thiết. Do lưu lượng truy cập web thường xuyên biến động, các hệ thống luôn phải đổi mới với bài toán cân bằng giữa chi phí và hiệu năng: cấp phát thừa tài nguyên (over-provisioning) sẽ gây lãng phí ngân sách, trong khi cấp phát thiếu (under-provisioning) lại dẫn đến quá tải và gián đoạn dịch vụ. Để giải quyết vấn đề này, các chiến lược Autoscaling thường được chia thành hai nhóm chính là Reactive Scaling và Predictive Scaling. Trong khi Reactive Scaling chỉ phản ứng dựa trên các chỉ số hiện tại như CPU hay RAM, thì Predictive Scaling lại tận dụng kỹ thuật dự báo chuỗi thời gian (Time series forecasting) để phân tích dữ liệu lịch sử, nhận diện các chu kỳ (mùa vụ) và chủ động điều chỉnh tài nguyên trước khi lưu lượng thực tế tăng lên, giúp giảm thiểu độ trễ phản ứng của hệ thống.

1.2 BÀI TOÁN HỒI QUY: DỰ ĐOÁN LUU LƯỢNG

Trọng tâm của bài toán hồi quy là xây dựng các mô hình máy học có khả năng dự báo chính xác hai chỉ số quan trọng: số lượng HTTP request và lượng dữ liệu (bytes) được chuyển tải trong tương lai. Để đảm bảo tính thực tiễn và linh hoạt, các mô hình này sẽ được huấn luyện và đánh giá trên ba khung thời gian khác nhau: cửa sổ 1 phút giúp hệ thống phản ứng nhanh với các biến động tức thời, cửa sổ 5 phút để cân bằng giữa độ nhạy và độ ổn định, và cửa sổ 15 phút phục vụ cho các chiến lược dài hạn. Dữ liệu huấn luyện được trích xuất từ logs của máy chủ NASA (giai đoạn 1/7 - 31/8/1995), bao gồm các thông tin chi tiết về host, thời gian, request và trạng thái phản hồi. Cần lưu ý rằng dữ liệu được phân chia theo trình tự thời gian nghiêm ngặt với tập Training từ 1/7 đến 22/8 và tập Test cho giai đoạn còn lại, đồng thời đã xử lý các khoảng downtime do sự cố thiên tai bằng dữ liệu giả lập.

Về phương pháp luận, dự án sẽ triển khai và so sánh hiệu quả giữa các nhóm kỹ thuật khác nhau, bao gồm các phương pháp thống kê truyền thống như ARIMA/SARIMA, các mô hình Deep Learning như LSTM/RNN và các thuật toán Machine Learning hiện đại như XGBoost hay LightGBM. Mô hình tối ưu nhất cho mỗi khung thời gian sẽ được lựa chọn dựa trên bộ chỉ số đánh giá toàn diện. Cụ thể, RMSE sẽ được dùng để phạt nặng các sai số lớn, trong khi MSE và MAE cung cấp cái nhìn tổng quát ít nhạy cảm với ngoại lai hơn, và MAPE sẽ giúp

đánh giá mức độ sai số dưới dạng phần trăm trực quan.

1.3 BÀI TOÁN TỐI UU: AUTOSCALING

Bài toán tối ưu hóa đóng vai trò chuyển đổi các kết quả dự báo thành hành động cụ thể nhằm tìm ra chính sách autoscaling cân bằng nhất giữa việc giảm thiểu chi phí vận hành (tính theo server-hours) và đảm bảo cam kết chất lượng dịch vụ (SLA). Hệ thống sẽ triển khai ba chiến lược để so sánh: Scaling dựa trên CPU (điều chỉnh khi tải vượt ngưỡng 70% hoặc giảm dưới 30%), Scaling dựa trên số lượng request thực tế, và Predictive Scaling sử dụng kết quả từ mô hình học máy để đón đầu xu hướng. Để đảm bảo hệ thống hoạt động ổn định và tránh hiện tượng dao động liên tục (oscillation), các cơ chế như thời gian chờ (Cooldown Period) và ngưỡng trễ (Hysteresis) cũng được áp dụng chặt chẽ trong quá trình vận hành. Hiệu quả cuối cùng của các chính sách này sẽ được đo lường dựa trên tổng chi phí, tỷ lệ thời gian hệ thống không bị quá tải, tần suất thay đổi trạng thái máy chủ và độ trễ trong phản ứng.

1.4 MỤC TIÊU DỰ ÁN

Dự án hướng tới việc xây dựng một quy trình khép kín từ phát triển mô hình đến triển khai hệ thống thực nghiệm.

Ở giai đoạn phát triển, mục tiêu là triển khai tối thiểu hai mô hình thuộc các nhóm thuật toán khác nhau cho