



(12) BẢN MÔ TẢ SÁNG CHẾ THUỘC BẰNG ĐỘC QUYỀN SÁNG CHẾ

(19) Cộng hòa xã hội chủ nghĩa Việt Nam (VN)
CỤC SỞ HỮU TRÍ TUỆ

(11)



1-0052734

(51)^{2020.01} H04L 12/00

(13) B

(21) 1-2021-04085

(22) 02/07/2021

(45) 27/10/2025 451

(43) 27/09/2021 402A

(73) TẬP ĐOÀN CÔNG NGHIỆP - VIỆN THÔNG QUÂN ĐỘI (VN)

Lô D26 Khu đô thị mới Cầu Giấy, phường Yên Hoà, quận Cầu Giấy, thành phố Hà Nội

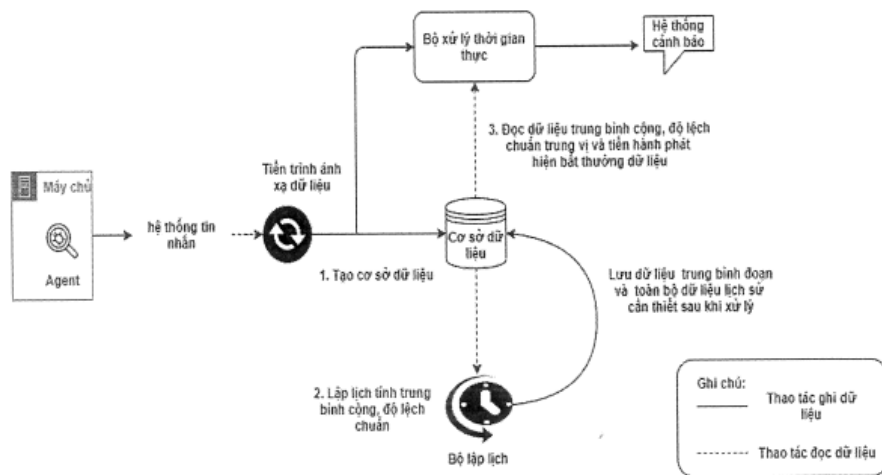
(72) Cao Đăng Sao (VN); Trần Văn Thuyết (VN); Nguyễn Đức Hiếu (VN); Nguyễn Đình Tâm (VN).

(74) Công ty TNHH Tư vấn Quốc Dân (NACILAW)

(54) PHƯƠNG PHÁP XỬ LÝ VÀ LƯU TRỮ DỮ LIỆU ĐỂ PHÁT HIỆN DỮ LIỆU
BẤT THƯỜNG THỜI GIAN THỰC

(21) 1-2021-04085

(57) Phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường theo thời gian thực giúp cảnh báo bất thường theo thời gian thực, bao gồm các bước: bước 1: xây dựng cơ sở dữ liệu lịch sử theo thời gian, cơ sở dữ liệu trung bình và độ lệch chuẩn; bước 2: tiến hành lựa chọn, phân chia dữ liệu lịch sử thành các khối kích thước bằng nhau và xây dựng công thức tính trung bình cộng, độ lệch chuẩn của từng khối dữ liệu và toàn bộ dữ liệu; bước 3: tạo tiến trình ánh xạ dữ liệu chạy độc lập thực hiện đọc dữ liệu thu thập được, chuẩn hóa dữ liệu và tương tác với hệ quản trị cơ sở dữ liệu trên bộ nhớ (in-memory database) để ghi dữ liệu lịch sử theo thời gian; bước 4: thực hiện phát hiện dữ liệu bất thường dữ liệu mới đến với trung bình cộng, độ lệch chuẩn của dữ liệu lịch sử đã được lưu trữ sẵn trong cơ sở dữ liệu trên bộ nhớ chỉ đọc (RAM).



Hình 1

Lĩnh vực được đề cập

Sáng chế đề cập đến phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực. Phương pháp được đề xuất trong sáng chế được sử dụng trên nền tảng công nghệ phát hiện dữ liệu bất thường (anomaly detection) và được ứng dụng trong lĩnh vực tính toán thời gian thực.

Tình trạng kỹ thuật của sáng chế

Thông thường, phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực được thể hiện bởi các bước giản lược như sau:

Bước 1: dữ liệu mới đến sẽ luôn được lưu trữ trong cơ sở dữ liệu.

Bước 2: đồng thời thực hiện so sánh dữ liệu mới đến với các điểm dữ liệu trong quá khứ để đưa ra kết luận dữ liệu mới đến có bất thường hay không và đưa ra cảnh báo.

Tuy nhiên, khi số lượng các điểm dữ liệu lịch sử cần sử dụng để so sánh tăng lên thì xảy ra ba vấn đề sau:

Một là yêu cầu máy tính cần lưu trữ dung lượng lớn dữ liệu lịch sử trên bộ nhớ truy cập ngẫu nhiên hay bộ nhớ chỉ đọc (Random Access Memory – RAM) trong khi dung lượng RAM là có giới hạn.

Hai là yêu cầu việc truy xuất dữ liệu lịch sử từ cơ sở dữ liệu liên tục gây tốn kém và dẫn đến lỗi cơ sở dữ liệu về lâu dài.

Ba là thời gian tính toán tăng lên, trong khi đối với bài toán phát hiện dữ liệu bất thường thời gian thực (bài toán ràng buộc về thời gian từ khi xảy ra sự kiện cho tới khi hệ thống phản ứng với sự kiện đó), thời gian tính toán các phép toán cơ bản cần đạt được tốc độ hay giới hạn thời gian nhất định.

Phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực giải quyết tốt ba vấn đề trên. Đáp ứng cho bài toán phát hiện dữ liệu bất thường thời gian thực và đưa ra hướng xử lý cho các bài toán tương tự có thể áp dụng để tăng tốc độ tính toán.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực. Phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực giúp tính toán tăng lên nhiều lần (phụ thuộc vào cách phân chia việc lưu trữ dữ liệu và tính toán trên bộ nhớ chỉ đọc RAM).

Để đạt được mục đích nêu trên, sáng chế đề xuất phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực với các bước thực hiện cụ thể như sau:

Bước 1: xây dựng cơ sở dữ liệu lịch sử theo thời gian, cơ sở dữ liệu trung bình và độ lệch chuẩn. Cụ thể hơn: dữ liệu sau khi đến với hệ thống sẽ được lưu vào cơ sở dữ liệu theo dấu thời gian, sau các chu kì thời gian xác định thì tiến hành tính trung bình dữ liệu và lưu vào cơ sở dữ liệu.

Bước 2: tiến hành lựa chọn, phân chia dữ liệu lịch sử thành các khối kích thước bằng nhau và xây dựng công thức tính trung bình cộng, độ lệch chuẩn của từng khối dữ liệu và toàn bộ dữ liệu:

Thực tế, việc phát hiện dữ liệu bất thường dữ liệu sử dụng các thuật toán khác nhau thì yêu cầu về việc xử lý và lưu trữ dữ liệu là khác nhau. Đối với các thuật toán yêu cầu sử dụng trung bình cộng, độ lệch chuẩn của dữ liệu lịch sử để đưa ra đánh giá bất thường thì áp dụng các bước sau:

Bước 2.1: phân chia dữ liệu lịch sử thành các khối bằng nhau, cụ thể: giả sử dữ liệu lịch sử cần tính trung bình, độ lệch chuẩn là $n \times m$ điểm dữ liệu, ta chia thành m khối dữ liệu, mỗi khối chứa n điểm dữ liệu.

Bước 2.2: xác định số lượng các điểm dữ liệu lịch sử cần sử dụng

Bước 2.3: xây dựng công thức tính trung bình cộng, độ lệch chuẩn của dữ liệu theo khối và toàn bộ dữ liệu.

Bước 3: tạo tiến trình ánh xạ dữ liệu chạy độc lập thực hiện đọc dữ liệu thu thập được, chuẩn hóa dữ liệu và tương tác với hệ quản trị cơ sở dữ liệu trên bộ nhớ (in-memory database) để ghi dữ liệu lịch sử theo thời gian.

Bước 4: thực hiện phát hiện dữ liệu bất thường dữ liệu mới đến với trung bình cộng, độ lệch chuẩn của dữ liệu lịch sử đã được lưu trữ sẵn trong cơ sở dữ liệu trên bộ nhớ chỉ đọc (RAM).

Để thực hiện phát hiện dữ liệu bất thường theo sự phân chia dữ liệu ở bước 2. Ta sử dụng hai tiến trình độc lập: tiến trình tính toán trung bình cộng, độ lệch chuẩn thực hiện tính toán khi đã thu thập được n điểm dữ liệu cho khối đó và cho toàn bộ dữ liệu lịch sử được thể hiện ở bước 4.1; tiến trình thời gian thực phát hiện dữ liệu bất thường đọc dữ liệu theo thời gian thực và tiến hành kiểm tra xem điểm dữ liệu đó có phải là bất thường hay không thực hiện ở bước 4.2.

Bước 4.1: xử lý tính toán trung bình cộng, độ lệch chuẩn của các khối dữ liệu, toàn bộ dữ liệu và lưu vào cơ sở dữ liệu cho giá trị trung bình, giá trị độ lệch chuẩn có cấu trúc dữ liệu như Bảng 2, và được lưu trữ trực tiếp trên RAM:

Tiến trình tính toán trung bình cộng, độ lệch chuẩn được lập lịch thực thi sau $n \times t$ thời gian bởi vì dữ liệu được ghi vào cơ sở dữ liệu theo chu kỳ t thời gian, nên sau $n \times t$ thời gian ta tiến hành các bước tiếp theo đây.

Bước 4.1.1: Đọc dữ liệu lịch sử n điểm gần nhất trong cơ sở dữ liệu lưu trữ tại Bước 3.

Bước 4.1.2: Tính trung bình cộng, độ lệch chuẩn của n điểm vừa thu được.

Bước 4.1.3: Tiến hành tính toán trung bình cộng của toàn bộ các khối dữ liệu lịch sử đã được lưu trên cơ sở dữ liệu: dựa trên các trung bình cộng, độ lệch chuẩn của tối đa $m - 1$ khối dữ liệu đã tính trước đó và trung bình cộng, độ lệch chuẩn của n điểm gần nhất, tính trung bình cộng, độ lệch chuẩn của toàn bộ dữ liệu lịch sử sử dụng các công thức được thành lập ở Bước 2.3.

Bước 4.1.4: Lưu trung bình cộng n điểm gần nhất, độ lệch chuẩn n điểm gần nhất, trung bình cộng toàn bộ dữ liệu lịch sử và trung vị độ lệch chuẩn toàn bộ dữ liệu lịch sử vào cơ sở dữ liệu cho giá trị trung bình, giá trị độ lệch chuẩn có cấu trúc dữ liệu Bảng 2 để truy vấn.

Bước 4.2: Tiến trình thời gian thực phát hiện dữ liệu bất thường đọc dữ liệu thời gian thực từ cơ sở dữ liệu và tiến hành phát hiện dữ liệu bất thường

Khi đó, bởi vì ở Bước 4.1 đã tính toán trung bình cộng, độ lệch chuẩn dữ liệu lịch sử nên không cần thiết phải tính toán lại trung bình cộng, độ lệch chuẩn của dữ liệu lịch sử mỗi lần có dữ liệu mới, dẫn đến tăng tốc độ tính toán phát hiện dữ liệu bất thường và đáp ứng được thời gian thực cho bài toán.

Giải pháp này giúp giải quyết bài toán về tính toán thời gian thực của cả quá trình phát hiện dữ liệu bất thường, tránh việc tiến hành quét ổ cứng và thao tác đóng mở tập tin cơ sở dữ liệu nhiều lần.

Mô tả vắn tắt các hình vẽ

Hình 1: là hình vẽ mô tả luồng dữ liệu được xử lý trong hệ thống phát hiện dữ liệu bất thường thời gian thực;

Hình 2: là hình vẽ mô tả luồng xử lý ánh xạ dữ liệu từ nguồn vào cơ sở dữ liệu;

Hình 3: là hình vẽ mô tả sự phân chia dữ liệu lịch sử thành các khối nhỏ hơn để xử lý tính trung bình cộng trên từng khối và trên toàn bộ dữ liệu;

Hình 4: là hình vẽ mô tả tiến trình thời gian thực phát hiện dữ liệu bất thường vào thời điểm cụ thể;

Hình 5: là hình vẽ mô tả luồng xử lý dữ liệu thời gian thực của hệ thống phát hiện dữ liệu bất thường bằng thuật toán sử dụng trung bình dữ liệu.

Mô tả chi tiết sáng chế

Trong hệ thống phát hiện dữ liệu bất thường (Anomaly Detection System) là việc phát hiện dữ liệu bất thường xảy ra trong hệ thống, yêu cầu đặt ra là bất thường cần được phát hiện sớm nhất có thể để giảm thiểu rủi ro do nó tác động đến một phần hay toàn bộ hệ thống hay nói cách khác là phát hiện theo thời gian thực.

Phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực được đề xuất trong sáng chế bao gồm các bước thực hiện tuần tự được nêu chi tiết dưới đây:

Bước 1: xây dựng cơ sở dữ liệu lịch sử theo thời gian, cơ sở dữ liệu trung bình và độ lệch chuẩn.

Tham chiếu Hình 1, mô tả luồng dữ liệu được xử lý trong hệ thống phát hiện dữ liệu bất thường thời gian thực và Hình 2, mô tả luồng xử lý ánh xạ dữ liệu từ nguồn vào cơ sở dữ liệu.

Dữ liệu của hệ thống được thu thập bởi các tác nhân (agent) cài đặt trên máy chủ bao gồm thông tin như phần trăm bộ xử lý trung tâm sử dụng, phần trăm bộ nhớ trong sử dụng, độ trễ mạng, Dữ liệu đó sẽ được lưu trữ trên một hệ thống tin nhắn tập trung nhằm nhiệm vụ cho các hệ thống khác nhau sử dụng chung nguồn dữ liệu. Nhờ một tiến trình ánh xạ dữ liệu (Process Mapping Data) chạy độc lập thực hiện đọc dữ liệu từ hệ thống tin nhắn tập trung, chuẩn hóa dữ liệu và tương tác với hệ quản trị cơ sở dữ liệu trên bộ nhớ (in-memory database) để ghi dữ liệu theo thời gian.

Nội dung của dữ liệu bao gồm: thời gian dữ liệu được ghi, nguồn dữ liệu được ghi, giá trị dữ liệu được ghi.

Khi có một bản ghi được gửi đến hệ thống tin nhắn, tiến trình ánh xạ dữ liệu sẽ ghi dữ liệu vào cơ sở dữ liệu có cấu trúc như sau:

Tên trường	Kiểu dữ liệu	Ý nghĩa
Mã định danh	Số nguyên	Khóa chính của bảng, có kiểu dữ liệu số nguyên, mã định danh duy nhất của dữ liệu
Thời gian dữ liệu được ghi	Thời gian mili giây	Thời gian dữ liệu được ghi, có kiểu dữ liệu thời gian
Nguồn dữ liệu được ghi	Chuỗi	Thông tin dữ liệu được ghi, có kiểu dữ liệu chuỗi
Giá trị dữ liệu được ghi	Số thực	Giá trị dữ liệu nhận được, có kiểu dữ liệu số thực

Bảng 1. Bảng dữ liệu lịch sử theo thời gian.

Ngoài ra, cần xây dựng cấu trúc lưu trữ cơ sở dữ liệu cho giá trị trung bình, giá trị độ lệch chuẩn của các điểm dữ liệu lịch sử như sau:

Tên trường	Kiểu dữ liệu	Ý nghĩa
Mã định danh	Số nguyên	Khóa chính của bảng, có kiểu dữ liệu số nguyên, mã định danh duy nhất
Thời gian tạo	Thời gian mili giây	Thời gian ghi dữ liệu trung bình lịch sử, có kiểu dữ liệu thời gian
Giá trị trung bình toàn bộ dữ liệu	Số thực	Giá trị trung bình toàn bộ dữ liệu lịch sử cần thiết, có kiểu dữ liệu số thực
Giá trị trung vị độ lệch chuẩn của toàn bộ dữ liệu	Số thực	Giá trị trung vị của các độ lệch chuẩn các khối của toàn bộ dữ liệu lịch sử sử dụng, có kiểu dữ liệu số thực
Giá trị trung bình khối dữ liệu gần nhất	Số thực	Giá trị trung bình khối dữ liệu nhận được gần nhất, có kiểu dữ liệu số thực
Giá trị độ lệch chuẩn khối dữ liệu gần nhất	Số thực	Giá trị độ lệch chuẩn khối dữ liệu nhận được gần nhất, có kiểu dữ liệu số thực

Bảng 2. Bảng giá trị trung bình và độ lệch chuẩn.

Bước 2: tiến hành lựa chọn, phân chia dữ liệu lịch sử thành các khối kích thước bằng nhau và xây dựng công thức tính trung bình cộng, độ lệch chuẩn của từng khối dữ liệu và toàn bộ dữ liệu:

Từ ý tưởng xuất phát là chia dữ liệu lịch sử thành các khối nhỏ hơn thuận tiện tính toán phát hiện dữ liệu bất thường thời gian thực, việc tính trung bình cộng, độ lệch chuẩn được thực hiện như sau: với trung bình cộng, trung bình cộng của $n \times m$ điểm dữ liệu bằng với trung bình cộng của các trung bình cộng của m khối, trong đó mỗi khối có n điểm dữ liệu; với độ lệch chuẩn, lấy trung vị độ lệch chuẩn của m khối, trong đó mỗi khối có n điểm dữ liệu sẽ tính được độ lệch chuẩn khối đó. Cụ thể phương pháp phân chia khối dữ liệu và tính trung bình cộng, độ lệch chuẩn từng khối và toàn bộ dữ liệu bao gồm các bước dưới đây:

Bước 2.1: Phân chia dữ liệu lịch sử thành các khối bằng nhau: giả sử dữ liệu lịch sử cần tính trung bình là $n \times m$ điểm dữ liệu, ta chia thành m khối dữ liệu, mỗi khối chứa n điểm dữ liệu. Việc lựa chọn hệ số n, m phụ thuộc vào đặc trưng của từng loại dữ liệu khác nhau, dựa trên yêu cầu giữa tốc độ xử lý dữ liệu và trung bình dữ liệu được sử dụng để phát hiện dữ liệu bất thường là gần nhất. Ví dụ, khi ta chia càng nhiều khối (m

lớn) và mỗi khối có số điểm lớn (n lớn) thì tốc độ xử lý dữ liệu sẽ chậm hơn, đồng thời việc so sánh dữ liệu mới đến với trung bình dữ liệu sẽ kém chính xác hơn.

Bước 2.2: xác định các điểm dữ liệu lịch sử cần sử dụng, các điểm này là các dữ liệu quá khứ tính từ thời điểm hiện tại, giả sử các điểm đó được ký hiệu là $a_{11}, a_{21}, \dots, a_{n1}, a_{12}, a_{22}, \dots, a_{n2}, \dots, a_{1m}, a_{2m}, \dots, a_{nm}$ tương ứng là điểm dữ liệu thứ nhất, điểm dữ liệu thứ hai, ..., điểm dữ liệu thứ $n \times m$

Bước 2.3: Trung bình cộng được tính bằng các cộng tất cả các điểm dữ liệu và chia kết quả cho số lượng các điểm dữ liệu, và độ lệch chuẩn trung vị được tính bằng trung vị độ lệch chuẩn của các khối nhỏ hơn lần lượt theo các công thức dưới đây:

$$\begin{aligned} \text{Trung_bình_cộng} &= \frac{a_{11} + a_{21} + \dots + a_{nm}}{n \times m} \\ &= \frac{\frac{a_{11} + a_{21} + \dots + a_{n1}}{n} + \frac{a_{12} + a_{22} + \dots + a_{n2}}{n} + \dots + \frac{a_{1m} + a_{2m} + \dots + a_{nm}}{n}}{m} \\ &= \frac{\text{trung_bình_cộng_khối_1} + \text{trung_bình_cộng_khối_2} + \dots + \text{trung_bình_cộng_khối_m}}{m} \end{aligned}$$

Độ_lệch_chuẩn_trung_vị

$$= \text{trung_vị}(\text{độ_lệch_chuẩn_khối_1}, \text{độ_lệch_chuẩn_khối_2}, \dots, \text{độ_lệch_chuẩn_khối_m})$$

Trong đó, độ lệch chuẩn khối từng khối được tính theo công thức sau:

$$\text{độ_lệch_chuẩn_khối_i} = \sqrt{\frac{\sum_{k=1}^n |a_{ki} - \text{trung_bình_cộng_khối_i}|^2}{n - 1}}$$

Tham chiếu Hình 3, mô tả sự phân chia dữ liệu lịch sử thành các khối nhỏ hơn để xử lý tính trung bình cộng, độ lệch chuẩn trên từng khối và trung bình cộng, độ lệch chuẩn trung vị trên toàn bộ dữ liệu.

Bước 3: tạo tiến trình ánh xạ dữ liệu chạy độc lập thực hiện đọc dữ liệu thu thập được, chuẩn hóa dữ liệu và tương tác với hệ quản trị cơ sở dữ liệu trên bộ nhớ (in-memory database) để ghi dữ liệu lịch sử theo thời gian. Bởi vì dữ liệu thu thập được do các tác nhân (agent) đẩy ra thường dưới dạng thô (thường có định dạng json – javascript object notation) bao gồm nhiều trường khác nhau nên ta cần phân tách dữ liệu ra thành các trường cần thực hiện phát hiện bất thường và chuẩn hóa dữ liệu về định dạng số thực. Dữ liệu sau chuẩn hóa được tiến trình ghi vào cơ sở dữ liệu trên bộ nhớ (in-memory database) theo thời gian. Dữ liệu trong cơ sở dữ liệu có cấu trúc dữ liệu như Bảng 1.

Bước 4: thực hiện phát hiện dữ liệu bất thường dữ liệu mới đến với trung bình cộng, độ lệch chuẩn của dữ liệu lịch sử đã được lưu trữ sẵn trong cơ sở dữ liệu trên bộ nhớ chỉ đọc (RAM).

Để thực hiện phát hiện dữ liệu bất thường theo sự phân chia dữ liệu ở bước 2. Ta sử dụng hai tiến trình độc lập: Tiến trình tính toán trung bình cộng, độ lệch chuẩn thực hiện tính toán khi đã thu thập được n điểm dữ liệu cho khối đó và cho toàn bộ dữ liệu lịch sử được thể hiện ở bước 4.1; Tiến trình thời gian thực phát hiện dữ liệu bất thường đọc dữ liệu theo thời gian thực và tiến hành kiểm tra xem điểm dữ liệu đó có phải là bất thường hay không thực hiện ở bước 4.2. Cụ thể như sau:

Bước 4.1: xử lý tính toán trung bình cộng, độ lệch chuẩn của các khối dữ liệu, toàn bộ dữ liệu và lưu vào cơ sở dữ liệu cho giá trị trung bình, giá trị độ lệch chuẩn có cấu trúc dữ liệu như Bảng 2, và được lưu trữ trực tiếp trên RAM:

Tiến trình tính toán trung bình cộng, độ lệch chuẩn được lập lịch thực thi sau $n \times t$ thời gian bởi vì dữ liệu được ghi vào cơ sở dữ liệu theo chu kỳ t thời gian, nên sau $n \times t$ thời gian ta tiến hành các bước tiếp theo đây.

Bước 4.1.1: đọc dữ liệu lịch sử n điểm gần nhất trong cơ sở dữ liệu lưu trữ tại bước 3.

Bước 4.1.2: tính trung bình cộng, độ lệch chuẩn của n điểm vừa thu được.

Bước 4.1.3: tiến hành tính toán trung bình cộng của toàn bộ các khối dữ liệu lịch sử đã được lưu trên cơ sở dữ liệu: dựa trên các trung bình cộng, độ lệch chuẩn của tối đa $m - 1$ khối dữ liệu đã tính trước đó và trung bình cộng, độ lệch chuẩn của n điểm gần nhất, tính trung bình cộng, độ lệch chuẩn của toàn bộ dữ liệu lịch sử sử dụng các công thức được thành lập ở bước 2.3.

Bước 4.1.4: lưu trung bình cộng n điểm gần nhất, độ lệch chuẩn n điểm gần nhất, trung bình cộng toàn bộ dữ liệu lịch sử và trung vị độ lệch chuẩn toàn bộ dữ liệu lịch sử vào cơ sở dữ liệu cho giá trị trung bình, giá trị độ lệch chuẩn có cấu trúc dữ liệu Bảng 2 để truy vấn.

Bước 4.2: tiến trình thời gian thực phát hiện dữ liệu bất thường đọc dữ liệu thời gian thực từ cơ sở dữ liệu và tiến hành phát hiện dữ liệu bất thường:

Dữ liệu hiện tại sẽ được kiểm tra điều kiện bất thường hay không theo phương pháp tham số dựa trên thống kê, cụ thể là thuật toán dựa trên trung bình cộng và độ lệch chuẩn dữ liệu lịch sử như sau:

Gọi $x_{hiện_tại}$ là giá trị hiện tại của dữ liệu thu được, $trung_bình_cộng$ và $độ_lệch_chuẩn$ lần lượt là trung bình cộng, độ lệch chuẩn của dữ liệu lịch sử gần nhất so với điểm dữ liệu hiện tại đã tính ở bước 4.1.3. Khi đó:

$x_{hiện_tại}$ là bất thường nếu

$$x_{hiện_tại} < trung_bình_cộng - hệ_số \times độ_lệch_chuẩn_trung_vị$$

hoặc $x_{hiện_tại} > trung_bình_cộng + hệ_số \times độ_lệch_chuẩn_trung_vị$

Trong đó, $hệ_số$ sẽ được xác định bằng quy tắc thực nghiệm (empirical rule), thường sẽ lấy bằng 3.

Nếu $x_{hiện_tại}$ là điểm dữ liệu bất thường thì sẽ được lưu vào cơ sở dữ liệu đồng thời sẽ gửi trực tiếp sang hệ thống cảnh báo để người vận hành hệ thống mạng sẽ tiến hành kiểm tra và sửa lỗi.

Tham chiếu Hình 4, Mô tả tiến trình thời gian thực phát hiện dữ liệu bất thường vào thời điểm cụ thể và Hình 5, mô tả luồng xử lý dữ liệu thời gian thực của hệ thống phát hiện dữ liệu bất thường. Khi có dữ liệu mới đến hệ thống phát hiện dữ liệu bất thường, tiến trình bộ xử lý thời gian thực sẽ thực hiện việc đọc trung bình cộng, độ lệch chuẩn của dữ liệu lịch sử từ cơ sở dữ liệu trên bộ nhớ (in-memory database), tiếp đó so sánh dữ liệu mới đến với trung bình cộng dữ liệu lịch sử đó và cuối cùng đưa ra kết luận điểm dữ liệu đó có phải bất thường hay không, nếu có thì phát ra cảnh báo cho hệ thống cảnh báo.

Hiệu quả đạt được của sáng chế

Giải quyết được bài toán phát hiện bất thường thời gian thực, thời gian phản hồi bất thường < 1 phút (từ thời gian xuất hiện bất thường đến khi đưa ra cảnh báo).

Tiết kiệm chi phí lưu trữ trên RAM và không phải tiến hành quét ổ cứng nhiều lần.

Yêu cầu bảo hộ

1. Phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực với các bước thực hiện cụ thể như sau:

bước 1: xây dựng cơ sở dữ liệu lịch sử theo thời gian, cơ sở dữ liệu trung bình và độ lệch chuẩn;

tại bước này, dữ liệu của hệ thống được thu thập bởi các tác nhân (agent) cài đặt trên máy chủ bao gồm thông tin như phần trăm bộ xử lý trung tâm sử dụng, phần trăm bộ nhớ trong sử dụng, độ trễ mạng; dữ liệu đó sẽ được lưu trữ trên một hệ thống tin nhắn tập trung nhằm nhiệm vụ cho các hệ thống khác nhau sử dụng chung nguồn dữ liệu; nhờ một tiến trình ánh xạ dữ liệu (process mapping data) chạy độc lập thực hiện đọc dữ liệu từ hệ thống tin nhắn tập trung, chuẩn hóa dữ liệu và tương tác với hệ quản trị cơ sở dữ liệu trên bộ nhớ (in-memory database) để ghi dữ liệu theo thời gian; nội dung của dữ liệu bao gồm: thời gian dữ liệu được ghi, nguồn dữ liệu được ghi, giá trị dữ liệu được ghi;

bước 2: tiến hành lựa chọn, phân chia dữ liệu lịch sử thành các khối kích thước bằng nhau và xây dựng công thức tính trung bình cộng, độ lệch chuẩn của từng khối dữ liệu và toàn bộ dữ liệu; cụ thể bao gồm:

bước 2.1: phân chia dữ liệu lịch sử thành các khối bằng nhau, cụ thể: giả sử dữ liệu lịch sử cần tính trung bình, độ lệch chuẩn là $n \times m$ điểm dữ liệu, ta chia thành m khối dữ liệu, mỗi khối chứa n điểm dữ liệu;

bước 2.2: xác định số lượng các điểm dữ liệu lịch sử cần sử dụng, giả sử các điểm đó được ký hiệu là

$a_{11}, a_{21}, \dots, a_{n1}, a_{12}, a_{22}, \dots, a_{n2}, \dots, a_{1m}, a_{2m}, \dots, a_{nm}$ tương ứng là điểm dữ liệu thứ nhất, điểm dữ liệu thứ hai, ..., điểm dữ liệu thứ $n \times m$;

bước 2.3: xây dựng công thức tính trung bình cộng, độ lệch chuẩn của dữ liệu theo khối và toàn bộ dữ liệu;

bước 3: tạo tiến trình ánh xạ dữ liệu chạy độc lập thực hiện đọc dữ liệu thu thập được, chuẩn hóa dữ liệu và tương tác với hệ quản trị cơ sở dữ liệu trên bộ nhớ (in-memory database) để ghi dữ liệu lịch sử theo thời gian; bởi vì dữ liệu thu thập được do các tác nhân (agent) đẩy ra thường dưới dạng thô (thường có định dạng json – javascript object notation) bao gồm nhiều trường khác nhau nên ta cần phân tách dữ liệu ra thành các trường cần thực hiện phát hiện bất thường và chuẩn hóa dữ liệu về định dạng số thực; dữ liệu sau chuẩn hóa được tiến trình ghi vào cơ sở dữ liệu trên bộ nhớ (in-memory database) theo thời gian;

bước 4: thực hiện phát hiện dữ liệu bất thường dữ liệu mới đến với trung bình cộng, độ lệch chuẩn của dữ liệu lịch sử đã được lưu trữ sẵn trong cơ sở dữ liệu trên bộ nhớ chỉ đọc (RAM);

tại bước này, sử dụng hai tiến trình độc lập: tiến trình tính toán trung bình cộng, độ lệch chuẩn thực hiện tính toán khi đã thu thập được n điểm dữ liệu cho khối đó và cho toàn bộ dữ liệu lịch sử được thể hiện ở bước 4.1; tiến trình thời gian thực phát hiện dữ liệu bất thường đọc dữ liệu theo thời gian thực và tiến hành kiểm tra xem điểm dữ liệu đó có phải là bất thường hay không thực hiện ở bước 4.2; cụ thể:

bước 4.1: xử lý tính toán trung bình cộng, độ lệch chuẩn của các khối dữ liệu, toàn bộ dữ liệu và lưu vào cơ sở dữ liệu cho giá trị trung bình, giá trị độ lệch chuẩn có cấu trúc dữ liệu như bảng dưới, và được lưu trữ trực tiếp trên RAM:

tên trường	kiểu dữ liệu	ý nghĩa
mã định danh	số nguyên	khóa chính của bảng, có kiểu dữ liệu số nguyên, mã định danh duy nhất
thời gian tạo	thời gian mili giây	Thời gian ghi dữ liệu trung bình lịch sử, có kiểu dữ liệu thời gian
giá trị trung bình toàn bộ dữ liệu	số thực	Giá trị trung bình toàn bộ dữ liệu lịch sử cần thiết, có kiểu dữ liệu số thực
giá trị trung vị độ lệch chuẩn của toàn bộ dữ liệu	số thực	giá trị trung vị của các độ lệch chuẩn các khối của toàn bộ dữ liệu lịch sử sử dụng, có kiểu dữ liệu số thực
giá trị trung bình khối dữ liệu gần nhất	số thực	giá trị trung bình khối dữ liệu nhận được gần nhất, có kiểu dữ liệu số thực
giá trị độ lệch chuẩn khối dữ liệu gần nhất	số thực	giá trị độ lệch chuẩn khối dữ liệu nhận được gần nhất, có kiểu dữ liệu số thực

tiến trình tính toán trung bình cộng, độ lệch chuẩn được lập lịch thực thi sau $n \times t$ thời gian bởi vì dữ liệu được ghi vào cơ sở dữ liệu theo chu kì t thời gian, nên sau $n \times t$ thời gian ta tiến hành các bước tiếp theo đây:

bước 4.1.1: đọc dữ liệu lịch sử n điểm gần nhất trong cơ sở dữ liệu lưu trữ tại bước 3;

bước 4.1.2: tính trung bình cộng, độ lệch chuẩn của n điểm vừa thu được;

bước 4.1.3: tiến hành tính toán trung bình cộng của toàn bộ các khối dữ liệu lịch sử đã được lưu trên cơ sở dữ liệu: dựa trên các trung bình cộng, độ lệch chuẩn của tối đa

$m - 1$ khối dữ liệu đã tính trước đó và trung bình cộng, độ lệch chuẩn của n điểm gần nhất, tính trung bình cộng, độ lệch chuẩn của toàn bộ dữ liệu lịch sử sử dụng các công thức được thành lập ở bước 2.3;

bước 4.1.4: lưu trung bình cộng n điểm gần nhất, độ lệch chuẩn n điểm gần nhất, trung bình cộng toàn bộ dữ liệu lịch sử và trung vị độ lệch chuẩn toàn bộ dữ liệu lịch sử vào cơ sở dữ liệu cho giá trị trung bình, giá trị độ lệch chuẩn có cấu trúc dữ liệu bảng ở trên để truy vấn;

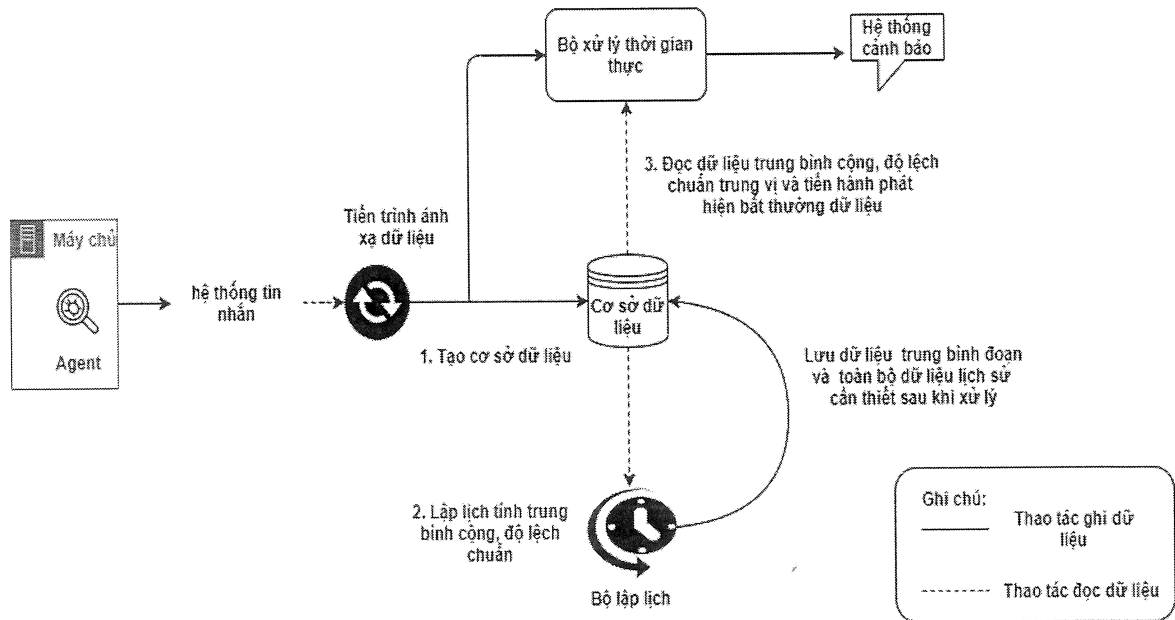
bước 4.2: tiến trình thời gian thực phát hiện dữ liệu bất thường đọc dữ liệu thời gian thực từ cơ sở dữ liệu và tiến hành phát hiện dữ liệu bất thường.

2. Phương pháp xử lý và lưu trữ dữ liệu để phát hiện dữ liệu bất thường thời gian thực theo điểm 1, trong đó tại bước 2, tiến hành:

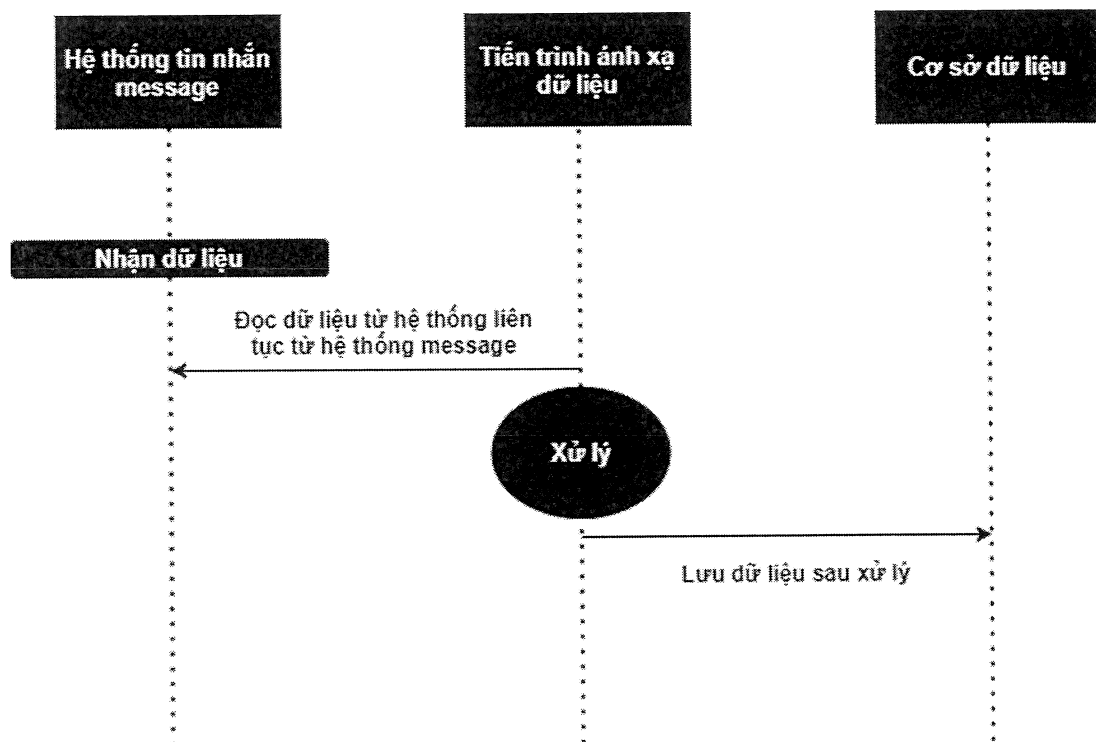
phân chia dữ liệu lịch sử thành các khối bằng nhau (m khối, mỗi khối n điểm dữ liệu);

sau khi thu được n điểm dữ liệu, tiến hành tính trung bình cộng, độ lệch chuẩn của n điểm đó và lưu vào cơ sở dữ liệu;

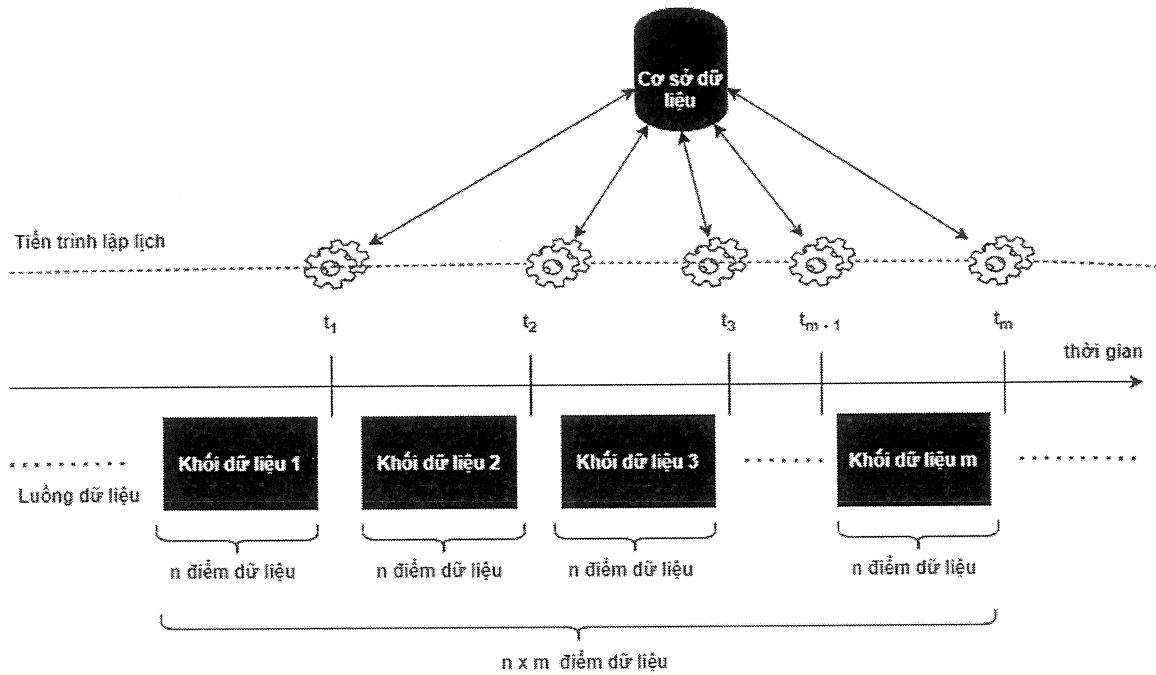
tiến hành tính trung bình cộng, độ lệch chuẩn trung vị của toàn bộ $n \times m$ điểm dữ liệu bằng cách lấy trung bình cộng của trung bình cộng của mỗi n điểm dữ liệu của m khối.



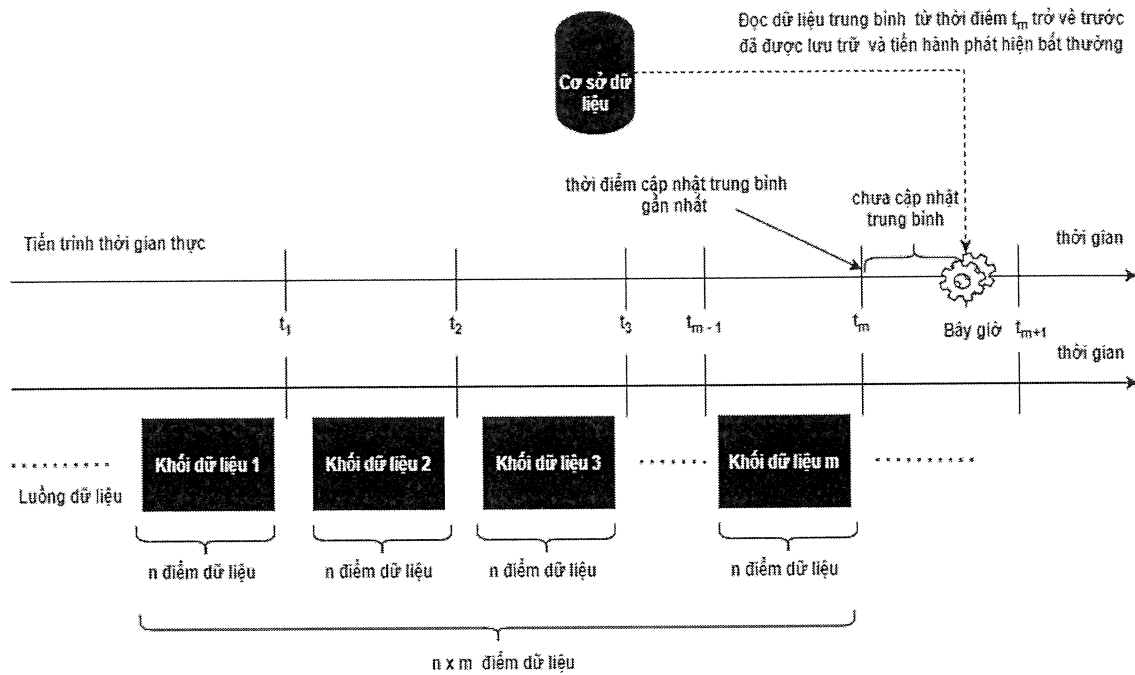
Hình 1



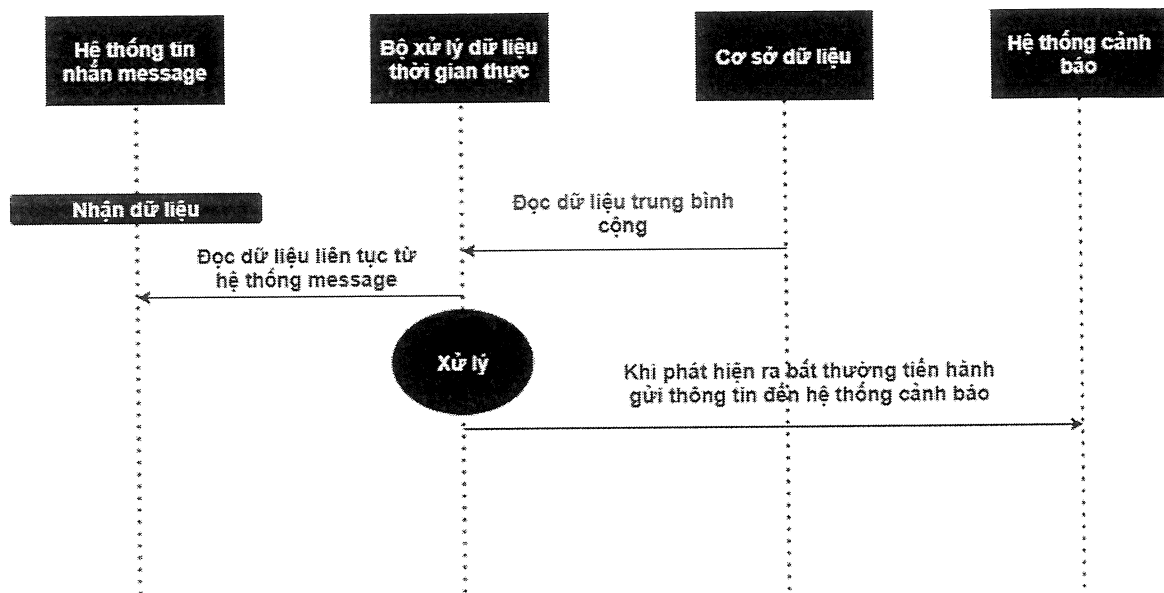
Hình 2



Hình 3



Hình 4



Hình 5