

GIẢI QUYẾT BÀI TOÁN DỮ LIỆU LỚN TRONG GIAO THÔNG BẰNG NGÔN NGỮ XỬ LÝ LUỒNG

Đào Xuân Phúc, Vũ Thị Phương Hoa*, Nguyễn Thống Nhất**
Email: phucdx@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 01/11/2023

Ngày phản biện đánh giá: 17/06/2024

Ngày bài báo được duyệt đăng: 26/06/2024

DOI: 10.59266/houjs.2024.415

Tóm tắt: Giao thông vận tải tại các trung tâm đô thị lớn, ở hầu hết các quốc gia, được coi là một trong những hệ quả của quá trình đô thị hóa. Để góp phần giảm thiểu tác động của vấn đề này thì việc áp dụng các tiến bộ khoa học kỹ thuật trong quản lý là cần thiết. Những phát triển công nghệ mới như ngôn ngữ xử lý luồng cho phép xử lý khối lượng lớn dữ liệu từ các thiết bị giám sát hành trình. Trong nghiên cứu này, tác giả mô phỏng gửi dữ liệu từ thiết bị giám sát hành trình đến trung tâm điều khiển sử dụng ngôn ngữ xử lý luồng. Dữ liệu sau đó được xử lý và người ra quyết định căn cứ trên đó để xuất giải pháp cho từng tình huống giao thông cụ thể. Ngôn ngữ xử lý luồng có cấu trúc dễ hiểu cho phép chúng ta tập trung vào nghiệp vụ hơn là logic đằng sau thuật toán và dữ liệu bên dưới. Trong điều kiện thực tế, tất cả những gì đơn vị quản lý cần làm là thay đổi các tham số vào/ra và chọn nền tảng phù hợp để triển khai.

Từ khóa: Giao thông vận tải, ngôn ngữ xử lý luồng, dữ liệu lớn, giám sát hành trình.

I. Đặt vấn đề

Giao thông vận tải là một trong những điểm nghẽn phát triển của Việt Nam. Mới đây nhất, Nghị quyết số 48/NQ-CP ngày 05/04/2022 của Chính phủ về tăng cường bảo đảm trật tự an toàn giao thông và chống ùn tắc giao thông giai đoạn 2022-2025 đã được Bộ GTVT hiện thực hóa bằng kế hoạch hành động [1]. Các giải pháp mang tính chính sách, cơ học và vật lý như hạn chế đăng ký mới

phương tiện; quy hoạch mở rộng đường; xây dựng cầu vượt cho phương tiện giao thông qua các nút giao cắt phức tạp; phát triển thêm các tuyến xe bus công cộng, bus nhanh, tàu điện trên cao, tàu điện ngầm... đòi hỏi nhiều thời gian và chi phí cao để thực hiện.

Trong khi đó, những giải pháp mang tính chất kỹ thuật như phát triển các hệ thống thu phí không dừng ETC; TBGSHT; camera giám sát giao thông... nếu được

đầu tư đồng bộ, khai thác tốt sẽ giúp giảm thiểu ùn tắc và tai nạn giao thông và nâng cao ý thức chấp hành của người tham gia giao thông.

Lộ trình bắt buộc phải lắp thiết bị giám sát hành trình (TBGSHT) được quy định rõ đối với mỗi loại phương tiện trong Nghị định số 86/2014/NĐ-CP ngày 10 tháng 09 năm 2014 như sau [2]:

Thời điểm	Loại xe áp dụng
Trước ngày 01/7/2015	Xe taxi, xe đầu kéo kéo rơ moóc, sơ mi rơ moóc kinh doanh vận tải phải lắp thiết bị giám sát hành trình
Trước ngày 01/1/2016	Xe ô tô kinh doanh vận tải hàng hóa có trọng tải thiết kế từ 10 tấn trở lên phải lắp thiết bị giám sát hành trình
Trước ngày 04/7/2016	Xe ô tô kinh doanh vận tải hàng hóa có trọng tải thiết kế từ 07 tấn đến dưới 10 tấn phải lắp thiết bị giám sát hành trình
Trước ngày 01/1/2017	Xe ô tô kinh doanh vận tải hàng hóa có trọng tải thiết kế từ 3,5 tấn đến dưới 07 tấn phải lắp thiết bị giám sát hành trình
Trước ngày 01/7/2018	Xe ô tô kinh doanh vận tải hàng hóa có trọng tải thiết kế dưới 3,5 tấn phải lắp thiết bị giám sát hành trình

Bảng 1. Lộ trình áp dụng lắp TBGSHT

Cuối năm 2016, các Sở GTVT địa phương đã nhận được văn bản của Tổng cục đường bộ về tăng cường xử phạt các xe trong diện quy định không lắp TBGSHT. Trong văn bản cũng nêu rõ các Sở tổ chức công tác chuyên đề về truyền dữ liệu từ TBGSHT. Trong trường hợp các phương tiện cần giám sát không có dữ liệu truyền về do xe hỏng, không hoạt động, thiết bị gặp trục trặc...phải giải trình [3]. Đối với loại hình dịch vụ vận tải (DVVT) gọi xe

công nghệ, thủ đô Hà Nội cũng có yêu cầu đối với Grab phải chia sẻ cơ sở dữ liệu [4]. Như vậy, theo ước tính, số lượng ô tô bắt buộc phải lắp đặt TBGSHT có thể lên đến hàng triệu xe.

Số liệu trên TBGSHT gửi về Trung tâm dữ liệu của Tổng cục đường bộ theo Thông tư 73/2014/TT-BGTVT mặc dù có cấu trúc đơn giản nhưng với việc đều đặn gửi về các thông tin với tần suất 30 giây / lần (đối với xe chạy) và 15 phút / lần (đối với xe đỗ) thì lượng thông tin phải xử lý cũng là rất lớn khi nhân với số lượng đầu xe tham gia kinh doanh DVVT [5].

Ngày 17/01/2020 Chính phủ đã ban hành Nghị định 10/2020/NĐ-CP quy định về kinh doanh và điều kiện kinh doanh bằng xe ô tô đã có hiệu lực (thay thế Nghị định 86/2014/NĐ-CP) đã cập nhật được những thay đổi trước tình hình mới. Riêng điều 12 trong Nghị định 10 quy định rất rõ về việc bắt buộc phải lắp TBGSHT [6].

Mặc dù cơ sở dữ liệu lưu trữ từ TBGSHT trên trung tâm điều khiển giao thông đã được lưu trữ từ lâu, tuy nhiên phần lớn là các hệ thống đã được viết từ cách đây nhiều năm và bằng các ngôn ngữ như Java, C++, ... Các nâng cấp hệ thống sẽ tốn nhiều chi phí, do vậy, các giải pháp tích hợp có tính tương thích ngược cần được lựa chọn triển khai.

Để tận dụng hết sức mạnh của một trong những ngôn ngữ tiên tiến nhất trong các bài toán về xử lý dữ liệu giao thông, nhóm tác giả thiết kế ra một chương trình bằng ngôn ngữ SPL để mô phỏng dữ liệu gửi từ các thiết bị giám sát hành trình về trung tâm điều hành. Thông tin có được từ dữ liệu nêu trên cho phép phát hiện ùn tắc thông qua mật độ xe tại một tọa độ địa lý

cụ thể hoặc phát hiện xe chạy quá tốc độ. Những thông tin này đạt được nhờ việc lưu trữ / phân tích / xử lý theo các luồng, tại các hệ thống của các đơn vị chức năng.

II. Đối tượng và phương pháp nghiên cứu

2.1 Đối tượng nghiên cứu

Tác giả sử dụng bộ dữ liệu giả lập (theo đúng định dạng mẫu được công bố) để mô phỏng việc gửi dữ liệu từ TBSGHT về trung tâm điều hành (có định dạng dữ liệu JSON - GPS đã được công bố và sử dụng rộng rãi trong nhiều hệ thống). Trong trường hợp kết nối về hệ thống thực tế, chỉ cần thay đổi các tham số hệ thống đầu vào để cấu hình và đưa dữ liệu thực vào phân tích cũng như kết xuất ra các kho chứa dữ liệu phục vụ các nhu cầu khác nhau cho những nhóm đối tượng người dùng khác nhau. Mỗi một bản tin trong bộ dữ liệu giả lập có các trường như sau:

Tên	Kiểu	Giải thích
id	rstring	Định danh phương tiện (chữ cái kết hợp chữ số)
Time	int64	Thời điểm quan sát (tính bằng milisecond kể từ 00:00:00 01-01-1970)
latitude	float64	Vĩ độ
longitude	float64	Kinh độ
speed	float64	Vận tốc (km/h)
heading	float64	Hướng di chuyển (độ, theo chiều kim đồng hồ từ phía Bắc)

Bảng 2. Định dạng một bản tin vị trí phương tiện (LocationType)

Chương trình giả lập sẽ tạo một bộ dữ liệu mô phỏng có định dạng LocationType và gửi vào luồng chính để xử lý. Bộ dữ liệu giả lập tuân thủ miền giá trị của các trường time, latitude, longitude, speed và heading như trong định nghĩa của dữ liệu

JSON-GPS. Riêng đối với trường id thì để cho bài toán đơn giản, chương trình giả lập sẽ sinh ra các chuỗi có định dạng Cxxx để thay thế cho định danh của phương tiện (giả lập biển số xe - trong đó, C là chữ cái viết hoa, là viết tắt của car, xxx là 3 chữ số bất kỳ). Như vậy trong hệ thống sẽ có dữ liệu giả lập của 1000 phương tiện có định danh từ C000 đến C999. Bộ sinh dữ liệu giả lập cũng để một xác suất nhỏ dưới 0.1% sẽ sinh ra một chuỗi id khác định dạng nêu trên (để sử dụng trong trường hợp ngoại lệ và chuyển sang luồng báo cáo lỗi).

2.2 Phương pháp nghiên cứu

- Phương pháp phân tích dữ liệu lớn dựa trên các cú pháp SQL.

- Phương pháp tạo chuỗi số ngẫu nhiên.

- Phương pháp thiết kế luồng thông tin theo định dạng SPL.

2.3 Thiết kế hệ thống

Luồng dữ liệu chạy tự động và xuyên suốt kể từ lúc chương trình tạo ngẫu nhiên các bản ghi LocationType, thông qua các toán tử xử lý trung gian, các luồng sẽ đến các điểm xử lý riêng biệt tùy vào các dữ liệu được cấu hình đường đi trong mạng. Hệ thống được chia làm 2 phần FileIngest (Nhập liệu) và MainCompostie (Tổng hợp):

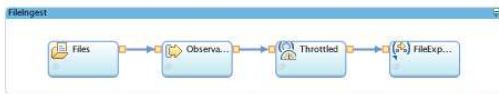
FileIngest gồm 4 toán tử:

DirectoryScan tên là Files (quét toàn bộ tệp tin trong thư mục chứa các dữ liệu được khởi tạo); luồng ra từ DirectoryScan là các file.

1. FileSource tên là Observation (đọc dữ liệu từ các file có trong thư mục - được tạo từ thuật toán tạo dữ liệu ngẫu nhiên - với định dạng csv); luồng ra từ Observation là các bản ghi có định dạng LocationType.

2. Throttle tên là Throttled (xác định ngưỡng nhận bản ghi trong 1 đơn vị thời gian, trong trường hợp này xét là 50 bản ghi /1 giây); luồng ra cũng vẫn là các bản ghi có định dạng LocationType với tốc độ là 50 bản ghi /1 giây.

3. Export tên là FileExporter để xác định luồng từ FileIngest ra MainComposite vẫn là LocationType với tốc độ là 50 bản ghi /1 giây.



Hình 1. Thiết kế thành phần FileIngest

MainCompostie gồm 7 toán tử:

4. Import tên là Observation nhận luồng vào MainComposite ra từ FileInvest là các bản ghi có định dạng Location Type với tốc độ là 50 bản ghi dữ liệu/1 giây

5. Split tên là IDCheck dùng để kiểm tra định dạng phương tiện (biển số) với đầu vào vẫn là 50 bản ghi dữ liệu /1 giây có định dạng LocationType.

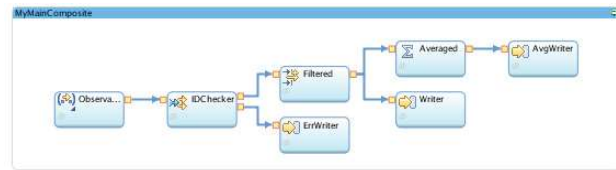
+ Trong trường hợp biển số nằm ngoài định dạng Cxxx (0.1% của phần tạo biển số ngẫu nhiên) luồng sẽ được điều hướng xuống FileSink có tên là ErrWriter (file error.observations).

+ Trong trường hợp biển số đúng định dạng Cxxx, luồng được chuyển tới bộ lọc Filter có tên là Filtered (cụ thể chỉ tìm xe có biển số là C111 và C999)

6. Luồng từ bộ lọc Filtered ra sẽ đưa vào:

+ Toán tử Aggregate để tính trung bình cộng tốc độ của 5 bản ghi theo id và đưa vào FileSink có tên là AvgWriter (file average.speeds).

+ Toán tử FileSink có tên là Writer (file filtered.cars).



Hình 2. Thiết kế thành phần MainComposite

2.4. Môi trường chạy hệ thống:

- Môi trường chạy: Cấu hình trên localhost của máy chủ Centos 6.0.

- Môi trường phát triển: Stream Studio (Eclipse).

- Ngôn ngữ lập trình: Stream Processing Language.

III. Kết quả nghiên cứu

Sau khi chạy, chương trình sẽ tạo ra 3 file có định dạng csv, sẵn sàng để phân tích, xử lý sâu hơn:

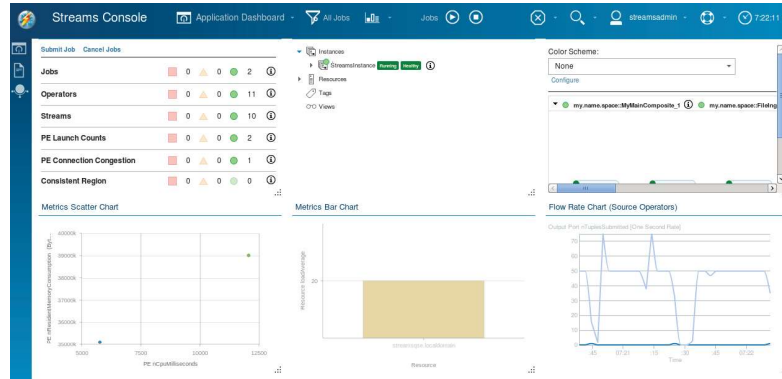
- File filtered.cars là bộ lọc các phương tiện theo danh sách các id (cụ thể là C111 và C999 trong ví dụ này). Trong tình huống thực tế, có thể cấu hình các bộ lọc thỏa mãn các điều kiện cho trước bằng các cú pháp SQL đa dạng hơn.

- File average_speed để tính vận tốc bình quân của các phương tiện được lọc (trung bình cộng vận tốc của phương tiện sau 5 lần nhận được dữ liệu liên tiếp). Trong trường hợp thực tế, khi đặt ngưỡng vận tốc theo cú pháp SQL hoàn toàn có thể phát hiện được xe vượt quá tốc độ và gửi cảnh báo về lực lượng chức năng.

- File error.observations lưu trữ các phương tiện gửi dữ liệu định danh về bị lỗi (do lỗi phương tiện, lỗi thiết bị, đường truyền v.v.)

Ngoài ra, Stream Studio (Eclipse) còn cung cấp hệ thống theo dõi luồng bằng đồ thị. Thông qua đồ họa trực quan, sinh động, quản trị viên hệ thống có thể

theo dõi và giám sát trạng thái của toàn bộ hệ thống: các luồng ra /vào hệ thống, nghẽn mạch ở đâu, tại sao lại nghẽn mạch, các tài nguyên được phân bổ, ...



Hình 3. Giám sát hệ thống thông qua Stream Console của Stream Studio (Eclipse)

IV. Kết luận

Mặc dù chỉ là hệ thống giả lập, nhưng công cụ được viết bằng SPL cũng đã mô phỏng được một hệ thống xử lý dữ liệu lớn như hệ thống thu thập dữ liệu giao thông từ TBGSHT. Bản chất của các bài toán xử lý dữ liệu lớn theo thời gian thực trong trường hợp này không phải là sự phức tạp, ràng buộc cũng như số lượng trường của mỗi bản ghi mà là số lượng bản ghi được gửi về trong thời gian thực (real time). Mô hình đơn giản nhưng cũng đã khái quát được những vấn đề của xử lý dữ liệu giao thông lớn bằng SPL. Để triển khai và nâng cấp lên thành hệ thống thật, hệ thống cần phải tích hợp thêm các mô hình nghiệp vụ của các đơn vị chức năng có liên quan. Ứng dụng SPL vào các bài toán xử lý dữ liệu lớn dạng luồng thời gian thực như các dịch vụ nghe nhạc trực tuyến, video, thị trường hàng hóa, chứng khoán v.v. là rất lớn [7].

Tài liệu tham khảo:

- [1]. Nghị quyết về đảm bảo trật tự an toàn giao thông & chống ùn tắc giao thông (48/NQ-CP 2022)
- [2]. Nghị định về kinh doanh và điều kiện kinh doanh vận tải bằng xe ô tô (86/2014/NĐ-CP)
- [3]. http://drvn.gov.vn/tin-tuc-su-kien/-/view_content/content/402696/ket-qua-tong-hop-phan-tich-tinh-hinh-vi-pham-tu-du-lieu-thiet-bi-giam-sat-hanh-trinh-thang-08-2016-
- [4]. <http://hanoitv.vn/ha-noi-de-nghi-uber chia-se-co-so-du-lieu-d69611.html>
- [5]. Quy chuẩn quốc gia về thiết bị giám sát hành trình của xe ô tô (73/2014/TT-BGTVT)
- [6]. Nghị định về kinh doanh và điều kiện kinh doanh vận tải bằng xe ô tô (10/2020/NĐ-CP)

SOLVING THE BIG DATA PROBLEM IN TRAFFIC BY STREAMINGS PROCESSING LANGUAGE

Dao Xuan Phuc[†], Vu Thi Phuong Hoa[†], Nguyen Thong Nhat[†]

Abstract: *In most countries, Transportation in large cities is considered an urbanization challenge. To minimize the impact of this problem, it is necessary to apply science and technology in management. New technological developments such as stream processing languages can process large volumes of data from cruise monitoring devices. In this report, the author simulates sending data from the cruise monitoring device to the control center using stream processing language. The data is then processed, and decision-makers are based on the proposed solutions for each specific traffic situation. The easily understood structured stream processing language allows us to focus on the business rather than the logic behind the algorithm and underlying data. In real situations, the operation unit only needs to change the I/O parameters and choose the right platform for deployment.*

Keywords: *Transportation, streamings processing language, big data, cruise monitoring.*

[†] Hanoi Open University