BỘ GIÁO DỤC & ĐÀO TẠO
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM
KHOA CÔNG NGHỆ THÔNG TIN

NGHIÊN CỨU CÔNG CỤ CUNG CẤP TÍNH NĂNG ORCHESTRATION VÀ TỰ ĐỘNG HÓA WORKFLOWS CHO CÁC PIPELINE DỮ LIỆU GIAI ĐOẠN INGESTION

Môn học: Xử lý dữ liệu trực tuyến
Bộ môn: Hệ thống thông tin
Mã lớp: CQ2022/1
GVHD: TS. Nguyễn Trần Minh Thư
Tên nhóm: Nhóm 4
Thành viên:

Lê Đức Cường	MSSV: 21120213
Nguyễn Dương Trường Sinh	MSSV: 21120322
Phạm Trần Trung Hậu	MSSV: 22120100
Bùi Đoàn Thuý Vy	MSSV: 22120448

Thành phố Hồ Chí Minh, ngày 8 tháng 4 năm 2025

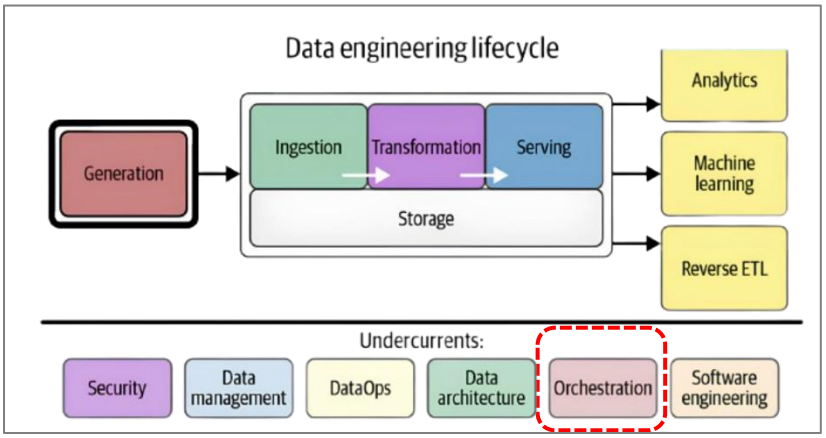
Mục lục

1 Giới thiệu

1.1 Khái niệm:

- Data Ingestion là quá trình thu thập và di chuyển dữ liệu (thô hoặc đã xử lý sơ) từ nhiều nguồn khác nhau (database, file, API, stream…) vào hệ thống lưu trữ trung tâm (data lake, warehouse…) để phục vụ cho các bước xử lý và phân tích tiếp theo.

- Orchestration là quá trình tự động hóa và quản lý luồng dữ liệu giữa nhiều hệ thống, ứng dụng và kho lưu trữ khác nhau. Quá trình này bao gồm việc điều phối việc di chuyển, chuyển đổi và tích hợp dữ liệu từ nhiều nguồn khác nhau vào một môi trường thống nhất để phục vụ phân tích.



Ảnh 1: Data engineering lifecycle

1.2 Mục tiêu của Data Orchestration

- Tự động hóa các pipeline dữ liệu phức tạp

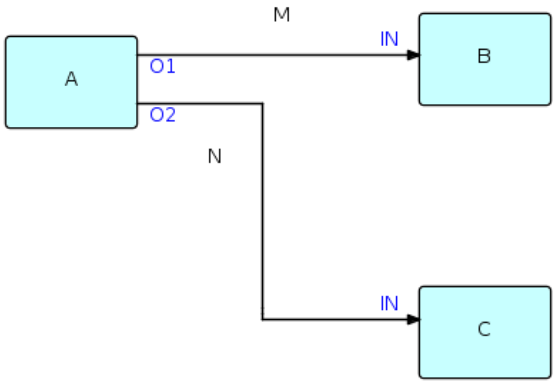
- Đảm bảo tính chính xác và nhất quán của dữ liệu

- Tối ưu hiệu suất khi xử lý dữ liệu ở quy mô lớn

- Giảm thiểu sự can thiệp thủ công

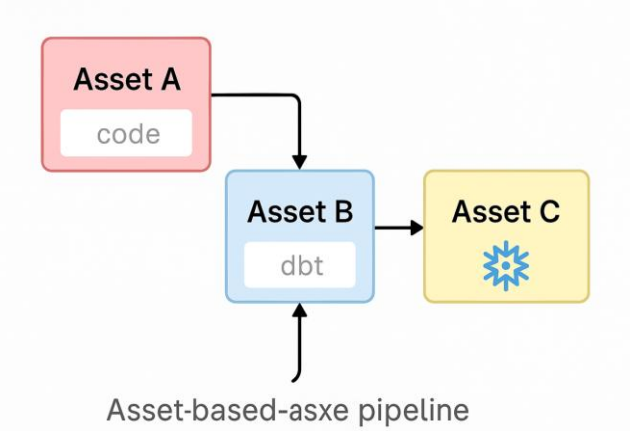
1.3 Lý do chọn công cụ

- Apache NiFi: mạnh về ingestion real-time và xử lý dữ liệu theo mô hình flow-based.



Ảnh 2 Ví dụ mô hình flow-based đơn giản

- Dagster: công cụ orchestration hiện đại, định nghĩa pipeline dưới dạng assets, tích hợp tốt với DevOps và CI/CD.



Ảnh 3 Ví dụ sơ đồ pipeline asset-based

2 Tổng quan về Apache NiFi

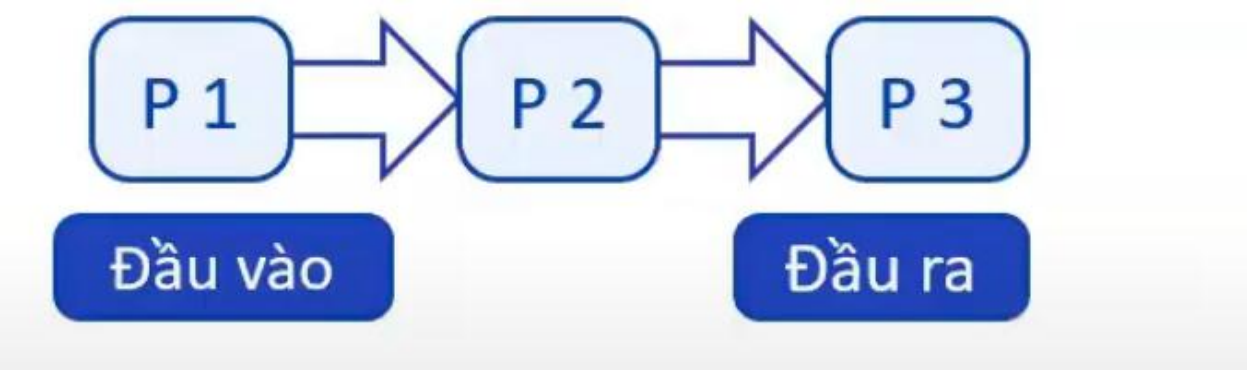
2.1 Giới thiệu và lịch sử phát triển

- Apache NiFi bắt nguồn từ dự án "NiagaraFiles" do NSA phát triển, sau đó được open-source và gia nhập Apache Foundation năm 2014.

- Được thiết kế để xử lý dòng dữ liệu (data flows) một cách trực quan, dễ cấu hình.

2.2 Một số khái niệm

- Nhóm qui trình: Là tập hợp các qui trình và kết nối giữa chúng. Các qui trình có thể nhận, gửi dữ liệu qua các cổng đầu vào và đầu ra.



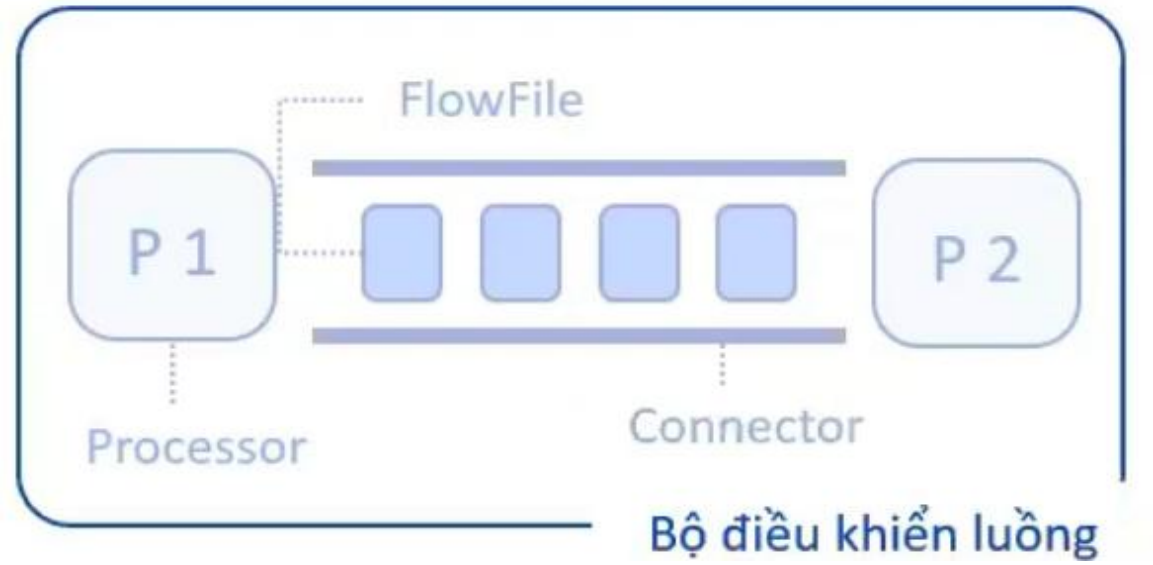
Ảnh 4 Nhóm qui trình của Apache Nifi

- Bộ xử lí Flow-file (Processor): nhận dữ liệu, xử lý dữ liệu (chuyển đổi, lọc, parse, nén, giải nén...) và gửi dữ liệu đi.

- FlowFile: đại diện cho một phần tử dữ liệu đang được xử lý trong hệ thống.

- Connector: hàng đợi trung gian giữa 2 Processor khi P1 xử lí xong và chờ P2 xử lí.

- Bộ điều khiển luồng (Flow Controller): quản lí Processor, Flowfile, Connector; quản lí luồng dữ liệu trong hệ thống; quản lí lịch trình chạy của các Processor và ghi lại lịch sử.



Ảnh 5 Cấu trúc Flow Controller

2.3 Kiến trúc và tính năng chính

- Flow-based Programming: Giao diện thiết kế kéo-thả, thể hiện các bước xử lý dưới dạng khối và kết nối.

- Ingestion mạnh mẽ: Kết nối được với file system, API, Kafka, MQTT, FTP, v.v.

- Orchestration & Automation: Các processor có thể được cấu hình để chạy tự động, với khả năng branching, routing, retry logic, scheduling.

- Clustering: Hỗ trợ mở rộng theo cụm để xử lý lượng dữ liệu lớn.

2.4 Ưu điểm

+ Giao diện trực quan, dễ theo dõi.

+ Xử lý tốt dữ liệu streaming.

+ Dễ tích hợp nhiều nguồn dữ liệu.

2.5 Nhược điểm

+ Không lý tưởng cho pipeline phức tạp có nhiều logic điều kiện.

+ Tài liệu nâng cao còn thiếu.

3 Tổng quan về Dagster

3.1 Giới thiệu và lịch sử phát triển

- Dagster là một open-source data orchestrator do Elementl phát triển.

- Hướng tới lập trình theo asset-based pipeline: tập trung vào dữ liệu được tạo ra chứ không chỉ là job.

3.2 Kiến trúc và tính năng chính

- Asset-based pipeline: xác định các "asset" (bảng dữ liệu, file...) và mối quan hệ giữa chúng.

- Ingestion qua integration: không thu thập dữ liệu trực tiếp như NiFi, nhưng có thể tích hợp với nguồn dữ liệu qua code (SQL, API, dbt...).

- Orchestration mạnh mẽ: Có scheduler, retry logic, conditional branching, logging, tích hợp CI/CD.

- Tích hợp mạnh mẽ: Docker, Kubernetes, dbt, Airbyte, Snowflake…

3.3 Ưu điểm

- Dễ kiểm thử, dễ tích hợp DevOps.

- Hệ sinh thái lập trình hiện đại.

- Thích hợp cho team Data Engineer chuyên sâu.

3.4 Nhược điểm

- Yêu cầu viết code – không có giao diện trực quan.

- Cần thời gian làm quen với khái niệm assets và IO manager.

4 So sánh giữa Apache NiFi và Dagster

4.1 Giống nhau

- Điều phối dữ liệu: Cả hai đều hỗ trợ quản lý và điều phối các bước xử lý dữ liệu từ đầu vào đến đầu ra.
- Xử lý theo pipeline: Cả hai đều tổ chức các tác vụ xử lý dữ liệu theo pipeline (luồng), với nhiều bước nối tiếp nhau.
- Khả năng mở rộng cao: Hỗ trợ xử lý dữ liệu trên quy mô lớn, có thể mở rộng theo chiều ngang (scale-out).
- Hỗ trợ đa nguồn dữ liệu: Có thể tích hợp với nhiều nguồn dữ liệu khác nhau như databases, APIs, file systems, cloud storage...

4.2 Khác nhau

Tiêu chí		**Apache NiFi**	**Dagster**
tổng quan về tính năng và đặc điểm	Ingestion	Kéo dữ liệu từ nhiều nguồn với giao diện kéo-thả	Tích hợp ingestion qua asset/script
	Orchestration	Có scheduling và routing, nhưng giới hạn	Mạnh mẽ, hỗ trợ retry, conditional logic
	Khả năng tích hợp	Clustering, REST API, Kafka, Hadoop	Tốt trong môi trường DevOps: Docker, dbt, Spark...
	Khả năng mở rộng	Hỗ trợ clustering, scale-out tốt	Tốt, có thể triển khai trên Kubernetes, nhưng phức tạp hơn

	Độ phức tạp	Dễ dùng cho người mới	Phù hợp với lập trình viên, phức tạp hơn
	CI/CD Support	Có thể tích hợp nhưng không mạnh, thiếu tooling chuyên dụng	Có hỗ trợ
Về hiệu suất	Xử lý real-time	Mạnh, xử lý dòng dữ liệu theo thời gian thực	Chủ yếu xử lý batch, không tối ưu cho real-time
	Latency (độ trễ)	Thấp với dữ liệu stream, tối ưu cho thời gian thực	Phụ thuộc scheduler, latency cao hơn trong batch
	Conditional branching	Có, nhưng giới hạn và chủ yếu qua cấu hình UI	Có hỗ trợ
	Plugin/extension hỗ trợ	Processor phong phú	Qua Python lib
	Độ ổn định khi chạy lâu dài	Ổn định cao trong môi trường production stream	Ổn định nếu cấu hình đúng, nhưng dễ gặp lỗi nếu CI/CD không tốt
	Quản lý lỗi (error handling)	Hỗ trợ retry logic ở từng processor, log chi tiết	Retry logic mạnh, dễ kiểm thử nhưng yêu cầu viết code

4.3 Phân tích theo từng trường hợp sử dụng

- Apache NiFi:

+ Thích hợp với những doanh nghiệp cần kết nối nhanh nhiều nguồn dữ liệu, yêu cầu real-time, không muốn viết nhiều code.

+ NiFi dùng tốt với Hadoop ecosystem

- Dagster:

+ Thích hợp với các hệ thống cần kiểm soát pipeline bằng code, yêu cầu kiểm thử và dễ dàng tích hợp CI/CD.

+ Dagster phù hợp với modern stack (Snowflake, dbt, Airbyte...).

5 Kết luận

- Trong bối cảnh dữ liệu ngày càng đa dạng và khối lượng lớn, việc lựa chọn công cụ phù hợp để thực hiện Ingestion và Orchestration cho các pipeline dữ liệu là rất quan trọng.

- Apache NiFi và Dagster là hai công cụ đại diện cho hai cách tiếp cận khác nhau:

- NiFi phù hợp với những hệ thống cần xử lý dữ liệu real-time, yêu cầu kết nối nhiều nguồn một cách linh hoạt mà không cần nhiều code. Với giao diện trực quan và khả năng kéo-thả, NiFi đặc biệt hữu ích cho các tổ chức cần triển khai nhanh các flow ingestion và xử lý streaming.

- Dagster lại phù hợp với môi trường phát triển theo hướng lập trình hiện đại, đặc biệt khi pipeline cần được kiểm thử, kiểm soát chặt chẽ và tích hợp tốt vào quy trình CI/CD. Dagster hỗ trợ orchestration mạnh mẽ và cung cấp khả năng quản lý tài nguyên dữ liệu (assets) theo cách có cấu trúc và dễ bảo trì.

- Tùy theo nhu cầu và kiến trúc hệ thống, mỗi công cụ sẽ có điểm mạnh riêng:

+ Nếu mục tiêu là xử lý luồng dữ liệu thời gian thực và dễ dàng tích hợp với nhiều nguồn: Nên chọn Apache NiFi.

+ Nếu pipeline hướng tới batch processing, kiểm thử tự động, DevOps-friendly: Dagster là lựa chọn phù hợp hơn.

-Cả hai công cụ đều đang được phát triển mạnh mẽ, có cộng đồng hỗ trợ tốt và có thể mở rộng linh hoạt theo nhu cầu doanh nghiệp

Tài liệu tham khảo

[1] Apache Software Foundation. (n.d.). *Apache NiFi*. https://nifi.apache.org
[2] Apache. (n.d.). *apache/nifi* [GitHub repository]. GitHub. https://github.com/apache/nifi
[3] Elementl. (n.d.). *Dagster*. https://dagster.io
[4] Elementl. (n.d.). *Dagster Documentation*. https://docs.dagster.io

Danh mục hình ảnh