

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA MẠNG MÁY TÍNH VÀ TRUYỀN THÔNG



Nguyễn Trần Thành Tâm - 21521404

Lê Xuân Sơn - 21521386

BÁO CÁO

Môn học: Đồ án chuyên ngành

Lớp: NT114.P11.ATCL

TÌM HIỂU VÀ PHÁT TRIỂN HỆ THỐNG PHÁT HIỆN RÒ RỈ
DỮ LIỆU TRONG DOANH NGHIỆP

GVHD: ThS. Nguyễn Duy

TP. HỒ CHÍ MINH, 2024

MỤC LỤC

I. GIỚI THIỆU CHUNG	4
1. Ngữ cảnh	4
2. Tổng quát	4
2.1. Vì sao DLP quan trọng ?	4
2.2. Các phần của DLP:	4
a.DLP in Technology	4
b.DLP in Business Processes	5
c.DLP in Organizational culture	5
II. CƠ SỞ LÝ THUYẾT	5
1. Tổng quan	5
2. DSPM	5
3. OpenMetadata	6
3.1. Giới thiệu	6
3.2. Các tính năng chính	6
3.3. Cách thức hoạt động	7
4. Airbyte	8
4.1. Giới thiệu	8
4.2. Các tính năng chính	8
4.3. Cách thức hoạt động	9
5. Airflow	10
5.1. Giới thiệu	10
5.2. Các tính năng chính	10
5.3. Hoạt động	11
6. MinIO	12
6.1. Giới thiệu	12
6.2. Các tính năng chính	12
6.3. Lợi ích khi sử dụng MinIO	13
7. PostgreSQL	13
7.1. Giới thiệu	13
7.2. Các tính năng chính	14
8. Varonis	14
8.1. Giới thiệu	14
8.2. DSPM trong Varonis	14
III. HƯỚNG GIẢI QUYẾT	16
1. Tổng quan giải pháp	16
2. Mô hình chung	16

3. Triển khai hệ thống.....	17
a. Database	17
b. Ingestion and Managing Dataflow.....	18
c. Data Governance.....	20
IV. DEMO.....	22

I. GIỚI THIỆU CHUNG

1. Ngữ cảnh

Công ty X lưu trữ các thông tin, dữ liệu về nhân viên trong công ty như địa chỉ, số điện thoại, vị trí làm việc,... Ngoài ra, công ty còn lưu trữ các kết quả thử nghiệm, danh sách tài khoản,... đều là những dữ liệu nhạy cảm. Để bảo vệ những thông tin nhạy cảm này không bị rò rỉ ra ngoài và gây thiệt hại cho ngân hàng cũng như khách hàng, cần có một biện pháp để ngăn chặn việc rò rỉ dữ liệu.

2. Tổng quát

Data Loss Prevention là tập hợp các công cụ và quy trình được thiết kế để ngăn chặn rò rỉ hoặc mất dữ liệu.

2.1. Vì sao DLP quan trọng ?

- Nhiều ngành nghề có các quy định nghiêm ngặt về việc bảo mật dữ liệu, nếu không tuân thủ sẽ phải chịu các hình phạt pháp lí nghiêm trọng.
- Việc dữ liệu bị rò rỉ có thể dẫn đến sự tổn hại về danh tiếng, uy tín của doanh nghiệp.
- Cùng với việc dữ liệu bị rò rỉ, việc mất dữ liệu có thể xảy ra và dẫn đến các chi phí đáng kể để khắc phục hậu quả.

2.2. Các phần của DLP:

- DLP giám sát và bảo vệ các mục nhạy cảm của người dùng cũng như tổ chức khỏi các hoạt động rủi ro, ngay cả khi người dùng chưa được nâng cao nhận thức về việc bảo vệ dữ liệu.

a.DLP in Technology

~ DLP là công nghệ có thể giám sát dữ liệu ở trạng thái nghỉ, trạng thái dữ liệu đang sử dụng và trạng thái dữ liệu đang chuyển động trên các dịch vụ như Microsoft 365, Windows 10/11, macOS,...

b.DLP in Business Processes

~ DLP có thể chặn người dùng thực hiện các hành động cấm, như chia sẻ thông tin nhạy cảm qua email,... Chủ sở hữu của sẽ cung cấp cách xác định các hành vi thích hợp được cho phép và các hành vi không thích hợp với người dùng.

c.DLP in Organizational culture

~ Sử dụng các mẹo chính sách để nâng cao nhận thức cho người dùng, giúp người dùng làm quen với các biện pháp ngăn ngừa mất dữ liệu bằng những kế hoạch đào tạo cụ thể.

II. CƠ SỞ LÝ THUYẾT

1. Tổng quan

Để thực hiện DLP (Data Loss Prevention) trong môi trường doanh nghiệp, ta cần giám sát, phân loại, phân tích dữ liệu nhằm phát hiện những bất thường trong xử lý, trao đổi dữ liệu. Từ đó, ngăn chặn được việc thất thoát dữ liệu từ những tác nhân bên trong và bên ngoài.

2. DSPM

DSPM (Data Security Posture Management) là một giải pháp quản lý và bảo mật dữ liệu hiện đại, giúp doanh nghiệp giám sát, phân loại và bảo vệ dữ liệu nhạy cảm trên nhiều môi trường khác nhau, bao gồm hệ thống on-premises, đám mây, và hybrid (lai). DSPM không chỉ tập trung vào dữ liệu tĩnh mà còn theo dõi trạng thái bảo mật tổng thể của dữ liệu trong thời gian thực, giúp

doanh nghiệp phát hiện và khắc phục các rủi ro trước khi chúng trở thành vấn đề nghiêm trọng.

Lợi ích của DSPM:

- Với xu hướng chuyển đổi số, dữ liệu ngày càng phân tán trên nhiều môi trường. DSPM giúp quản lý và bảo vệ dữ liệu toàn diện bất kể nơi lưu trữ.
- Nhiều quy định pháp lý đòi hỏi doanh nghiệp phải biết chính xác dữ liệu của mình ở đâu và ai có quyền truy cập. DSPM cung cấp khả năng giám sát và báo cáo chi tiết.
- Thay vì chỉ phản ứng sau sự cố, DSPM giúp doanh nghiệp phát hiện và xử lý các mối đe dọa trước khi chúng gây thiệt hại.
- Đảm bảo rằng chỉ những người có thẩm quyền mới có quyền truy cập vào dữ liệu nhạy cảm, giúp giảm nguy cơ rò rỉ từ bên trong.
- Vi phạm dữ liệu có thể dẫn đến tổn thất tài chính lớn và làm tổn hại danh tiếng. DSPM giúp ngăn ngừa điều này thông qua việc quản lý rủi ro hiệu quả.

3. OpenMetadata

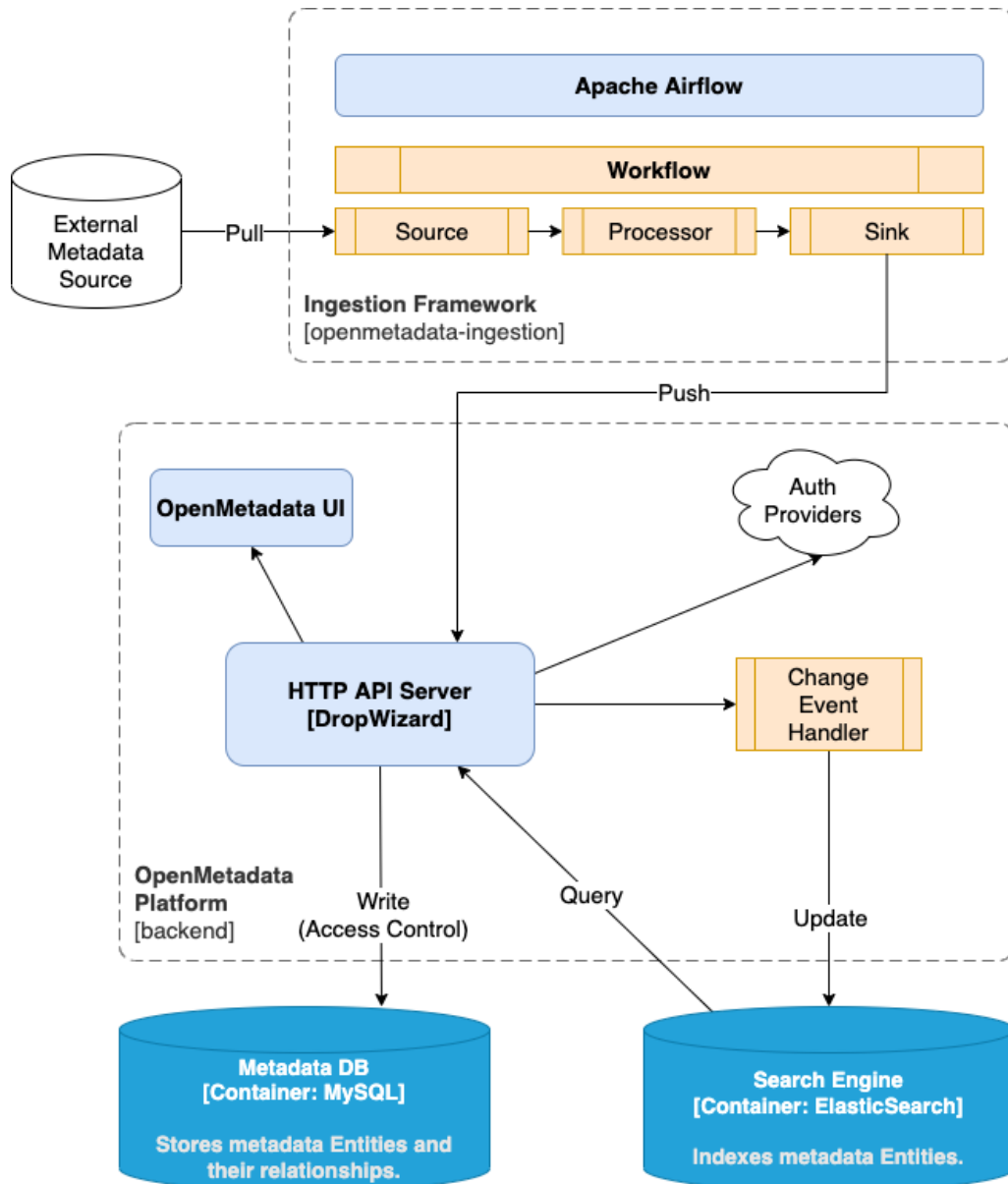
3.1. Giới thiệu

OpenMetadata là một nền tảng quản lý siêu dữ liệu mã nguồn mở giúp các doanh nghiệp quản lý dữ liệu một cách hiệu quả, từ việc thu thập, phân tích, cho đến chia sẻ thông tin siêu dữ liệu trong toàn tổ chức. OpenMetadata được thiết kế để đơn giản hóa việc quản lý dữ liệu trong các hệ sinh thái phức tạp, hỗ trợ nhiều nguồn dữ liệu và công cụ BI (Business Intelligence), ETL, máy học (ML), và lưu trữ dữ liệu.

3.2. Các tính năng chính

- Quản lý và phát hiện siêu dữ liệu (Metadata Management & Discovery)
- Hồ sơ dữ liệu (Data Lineage)
- Hợp tác dữ liệu (Collaboration)
- Tích hợp linh hoạt (Flexible Integration)
- Kiểm soát truy cập và bảo mật (Access Control & Security)
- Tuân thủ và quản trị dữ liệu (Compliance & Data Governance)

3.3. Cách thức hoạt động



- Thu thập siêu dữ liệu (Metadata Ingestion): Dữ liệu từ các nguồn bên ngoài được thu thập thông qua Ingestion Framework tích hợp với Apache Airflow.
- Xử lý siêu dữ liệu: Dữ liệu sau khi được thu thập sẽ được đẩy vào HTTP API Server (được xây dựng trên nền tảng DropWizard). API Server này đóng vai trò trung tâm, quản lý toàn bộ giao tiếp giữa các thành phần khác nhau.
- Quản lý và hiển thị: Người dùng tương tác với siêu dữ liệu thông qua OpenMetadata UI, giúp tìm kiếm, khám phá và quản lý dễ dàng. Quyền truy cập và bảo mật được quản lý thông qua Auth Providers.
- Quản lý sự kiện: Hệ thống có Change Event Handler để theo dõi và xử lý các sự kiện thay đổi, đảm bảo dữ liệu luôn được cập nhật theo thời gian thực.

4. Airbyte

4.1. Giới thiệu

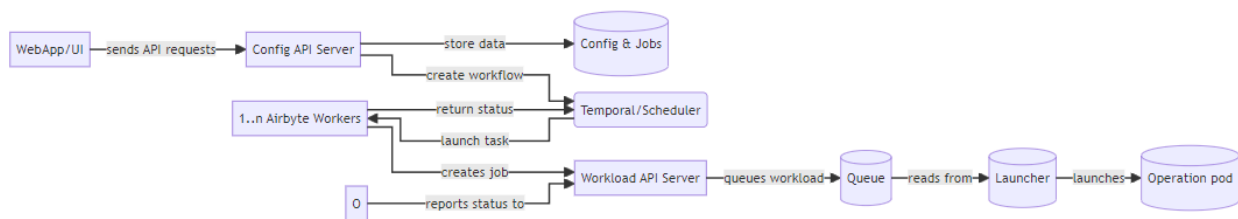
Airbyte là một nền tảng mã nguồn mở giúp tích hợp dữ liệu từ nhiều nguồn khác nhau, cho phép người dùng trích xuất, chuyển đổi và tải dữ liệu (ETL) vào kho dữ liệu hoặc hệ thống phân tích.

4.2. Các tính năng chính

- Hỗ trợ đa dạng nguồn dữ liệu: Airbyte cung cấp hơn 300+ connector có sẵn, cho phép kết nối đến cơ sở dữ liệu như PostgreSQL, MySQL, MongoDB, Oracle, v.v., ứng dụng SaaS như Salesforce, Shopify, HubSpot, Google Analytics, v.v., hay các dịch vụ đám mây như AWS S3, Google Cloud Storage, Azure Blob Storage, v.v.

- Airbyte có mã nguồn mở, giúp dễ dàng tùy chỉnh và mở rộng, đồng thời được cộng đồng phát triển và duy trì, cập nhật liên tục các tính năng mới.
- Airbyte hỗ trợ triển khai trên nhiều môi trường như Docker, Kubernetes hay Airbyte Cloud.
- Airbyte tích hợp dữ liệu thời gian thực, cho phép phát hiện và đồng bộ những thay đổi từ nguồn dữ liệu mà không cần phải tải lại toàn bộ dữ liệu.
- Airbyte có khả năng mã hóa dữ liệu, đồng thời quản lý quyền truy cập.
- Airbyte có thể đẩy dữ liệu đến các kho dữ liệu phổ biến như Snowflake, BigQuery, Redshift,...

4.3. Cách thức hoạt động



- Giao diện Web/UI: Dễ dàng sử dụng để tương tác với máy chủ Airbyte.
- Cấu hình máy chủ API: Đây là bộ điều khiển chính của Airbyte. Mọi hoạt động trong Airbyte như tạo nguồn, đích, kết nối, quản lý cấu hình, v.v. đều được cấu hình và gọi từ API.
- Cấu hình cơ sở dữ liệu và công việc: Lưu trữ tất cả cấu hình (thông tin xác thực, tần suất,...) và lịch sử công việc.
- Dịch vụ tạm thời: Quản lý việc lập kế hoạch và sắp xếp hàng đợi nhiệm vụ cũng như quy trình công việc.
- Worker: Đọc từ hàng đợi tác vụ và thực thi logic lập lịch kết nối cũng như trình tự, thực hiện lệnh gọi đến API khối lượng công việc.
- Workload API: Giao diện HTTP để sắp xếp khối lượng công việc - các nhóm riêng biệt chạy các hoạt động của trình kết nối.

- Trình khởi chạy: Sử dụng các sự kiện từ API khối lượng công việc và giao diện với k8s để khởi chạy khối lượng công việc.

5. Airflow

5.1. Giới thiệu

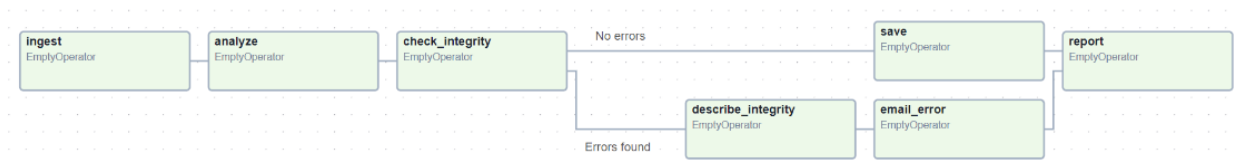
Được phát triển bởi Airbnb và sau đó trở thành một dự án mã nguồn mở trong cộng đồng Apache Software Foundation, Airflow là một nền tảng mã nguồn mở dùng để lập kế hoạch, giám sát và quản lý các quy trình công việc (workflow). Nó cho phép người dùng thiết kế các quy trình công việc phức tạp dưới dạng đồ thị có hướng (DAG).

5.2. Các tính năng chính

- Airflow cho phép lên lịch chạy các công việc theo định kỳ hoặc theo yêu cầu. Bạn có thể sử dụng cú pháp cron hoặc các lịch trình đặc biệt như @daily, @hourly, v.v.
- Airflow cho phép xác định các mối quan hệ phụ thuộc giữa các task trong DAG, giúp đảm bảo rằng các công việc chỉ được thực hiện khi các công việc phụ thuộc đã hoàn thành.
- Airflow có giao diện web giúp dễ dàng theo dõi trạng thái của các DAG và task. Bạn có thể kiểm tra lịch sử thực thi, xem báo cáo lỗi, và chỉnh sửa DAG nếu cần thiết.
- Airflow hỗ trợ khả năng mở rộng với kiến trúc phân tán. Bạn có thể triển khai Airflow trên nhiều máy chủ để xử lý hàng nghìn task cùng một lúc, đảm bảo hệ thống hoạt động hiệu quả dù quy mô công việc lớn.

- Airflow cung cấp nhiều Operator để thực hiện các loại công việc khác nhau như BashOperator, PythonOperator, EmailOperator, PostgresOperator, v.v.
- Airflow có thể tích hợp với nhiều công cụ và dịch vụ bên ngoài như:
 - Hệ thống lưu trữ: AWS S3, Google Cloud Storage.
 - Cơ sở dữ liệu: MySQL, PostgreSQL, MongoDB.
 - Dịch vụ đám mây: AWS, Google Cloud, Azure.
 - Hệ thống container: Kubernetes, Docker.
- Airflow có thể giới hạn tài nguyên cho các task, giúp tránh tình trạng quá tải hệ thống. Bạn có thể cấu hình Airflow để chỉ sử dụng một lượng tài nguyên nhất định cho mỗi DAG hoặc task.

5.3. Hoạt động



Airflow quản lý luồng Data và các Dependencies hoạt động cùng với nó dưới dạng các Tasks. Airflow phân công và tạo các Task, cũng như quản lý hoạt động của Task, từ đó điều khiển Data Flow theo kiến trúc được thiết lập. Các thành phần của Airflow và chức năng của nó bao gồm:

- Bộ lập lịch xử lý việc kích hoạt quy trình công việc đã lên lịch và gửi task cho người thực thi để chạy. Trình thực thi là thuộc tính cấu hình của bộ lập lịch, không phải là thành phần riêng biệt và chạy trong quy trình lập lịch.
- Airflow có máy chủ web, nơi có một giao diện người dùng tiện dụng để kiểm tra, kích hoạt và gỡ lỗi hành vi của DAG và tác vụ.
- Một thư mục chứa các tệp DAG, được bộ lập lịch đọc để tìm ra tác vụ nào cần chạy và thời điểm chạy chúng.
- Cơ sở siêu dữ liệu mà các thành phần của Airflow sử dụng để lưu trữ trạng thái của quy trình công việc và tasks. Việc thiết lập cơ sở siêu dữ liệu là bắt buộc để Airflow hoạt động.

6. MinIO

6.1. Giới thiệu

MinIO là một hệ thống lưu trữ đối tượng mã nguồn mở, được thiết kế để lưu trữ dữ liệu dạng đối tượng với hiệu suất cao. MinIO tương thích với giao thức Amazon S3.

6.2. Các tính năng chính

- Hỗ trợ giao thức S3 (Amazon S3 API): Giúp MinIO dễ dàng tích hợp với các hệ thống đã sử dụng Amazon S3. Thêm vào đó, có thể di chuyển ứng dụng từ S3 sang MinIO mà không cần thay đổi mã nguồn.
- Lưu trữ phân tán (Distributed Object Storage):
 - Hỗ trợ mở rộng dễ dàng bằng cách thêm node vào cluster.
 - Nếu một hoặc nhiều node gặp sự cố, dữ liệu vẫn được truy cập nhờ cơ chế nhân bản (replication).
 - Bảo vệ dữ liệu chống mất mát thông qua kỹ thuật phân tách và tái tạo dữ liệu.
- Hiệu suất cao: Tối ưu hóa cho dữ liệu không cấu trúc, xử lý khối lượng lớn các file như video, hình ảnh, và log, cùng với đó là khả năng đạt hiệu suất đọc/ghi lên tới **183 GB/s** trên phần cứng tiêu chuẩn.
- Bảo mật mạnh mẽ:
 - Dữ liệu được mã hóa khi truyền và lưu trữ thông qua mã hóa End-to End.
 - Đảm bảo giao tiếp an toàn giữa máy khách và máy chủ bằng TLS/SSL

- Cung cấp cơ chế kiểm soát truy cập chi tiết theo vai trò, người dùng, và chính sách bằng IAM (Identity and Access Management)
- Khả năng tích hợp và tương thích cao: Tích hợp dễ dàng với các công cụ và nền tảng như Hadoop, Spark, Presto, TensorFlow,... Hỗ trợ giao thức Kubernetes, phù hợp với các hệ thống container hóa và ứng dụng cloud-native. Ngoài ra, còn hỗ trợ client SDKs đa ngôn ngữ: Python, Go, Java, .NET, v.v.
- Giao diện quản lý trực quan: MinIO Console là giao diện web thân thiện, dễ sử dụng để quản lý bucket, kiểm tra log và giám sát hệ thống. MinIO Console hỗ trợ các công cụ dòng lệnh (CLI) như mc (MinIO Client) để thao tác nhanh chóng.
- Triển khai đa nền tảng: Chạy được trên bất kỳ nền tảng nào: từ máy chủ vật lý, máy ảo, container (Docker) đến môi trường Kubernetes. Hỗ trợ các hệ điều hành phổ biến như Linux, Windows và macOS.

6.3. Lợi ích khi sử dụng MinIO

- MinIO có mã nguồn mở miễn phí, giúp cho việc tùy chỉnh trở nên dễ dàng.
- MinIO có hiệu suất cao, là công cụ tối ưu hóa cho các ứng dụng cần xử lý dữ liệu lớn.
- MinIO có khả năng mở rộng theo chiều ngang.

7. PostgreSQL

7.1. Giới thiệu

PostgreSQL là hệ quản trị cơ sở dữ liệu quan hệ mã nguồn mở, được phát triển từ năm 1986 tại Đại học California, Berkeley. Nó hỗ trợ cả dữ liệu quan hệ và phi quan hệ, cho phép lưu trữ và truy vấn dữ liệu sử dụng SQL và JSON.

7.2. Các tính năng chính

- Khả năng mở rộng: Xử lý lượng dữ liệu lớn và truy vấn phức tạp, đồng thời hỗ trợ hệ thống plugin và module mở rộng như PostGIS (xử lý dữ liệu không gian), TimescaleDB (dữ liệu chuỗi thời gian).
- Hỗ trợ truy vấn song song (parallel queries).
- Hỗ trợ ACID (Atomicity, Consistency, Isolation, Durability), đảm bảo dữ liệu chính xác và đáng tin cậy.
- Kết nối mã hóa bằng SSL.
- Hỗ trợ đa dạng loại dữ liệu.
- Có thể sử dụng với nhiều ngôn ngữ lập trình như Python, Java, C#, Ruby, PHP, Go.
- Tuân thủ đầy đủ SQL chuẩn, đồng thời hỗ trợ mở rộng cú pháp.
- Cung cấp các công cụ như pgAdmin, DBeaver, hoặc dòng lệnh psql, đồng thời tích hợp logging và monitoring chi tiết.

8. Varonis

8.1. Giới thiệu

Varonis là một nền tảng quản lý và bảo mật dữ liệu hàng đầu, được thiết kế để bảo vệ dữ liệu nhạy cảm, phát hiện các mối đe dọa bên trong và bên ngoài, và đảm bảo tuân thủ các quy định về bảo mật dữ liệu. Varonis tập trung vào việc giám sát quyền truy cập và hành vi sử dụng dữ liệu, đặc biệt là dữ liệu phi cấu trúc (unstructured data) như tài liệu, email, bảng tính và các tệp lưu trữ khác trong hệ thống doanh nghiệp.

8.2. DSPM trong Varonis

DSPM trong Varonis có những sự khác biệt so với các nền tảng khác. Có thể kể đến:

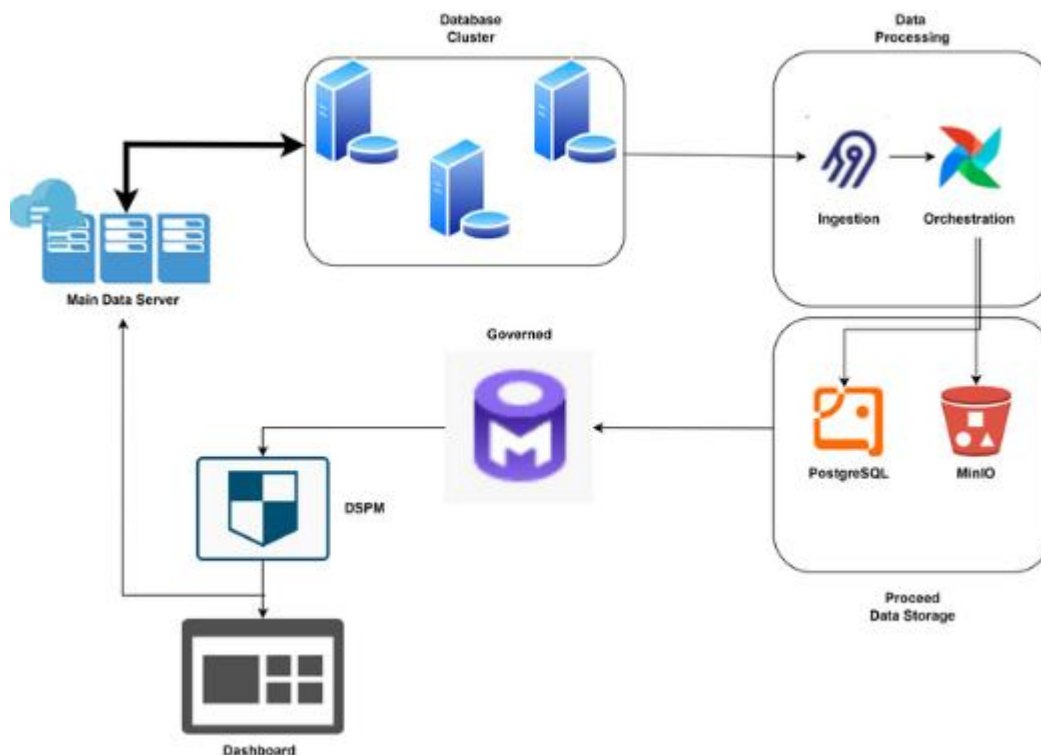
- **Chuyên Sâu Về Dữ Liệu Phi Cấu Trúc:** Varonis nổi bật với khả năng xử lý dữ liệu phi cấu trúc như tệp, thư mục, email, và nội dung trên các nền tảng chia sẻ tệp như SharePoint, OneDrive, Google Drive.
- **Phân Tích Hành Vi Người Dùng (UEBA) Tích Hợp Sâu:** Varonis sử dụng User and Entity Behavior Analytics (UEBA) để phát hiện hành vi bất thường của người dùng và thiết bị truy cập dữ liệu, giúp ngăn chặn các mối đe dọa nội bộ và tấn công từ bên ngoài.
- **Quản Lý Quyền Truy Cập Chi Tiết và Tự Động Hóa:** Varonis cung cấp khả năng quản lý quyền truy cập chi tiết đến từng cấp độ tệp hoặc thư mục, giúp giảm thiểu quyền truy cập không cần thiết và tự động hóa việc điều chỉnh quyền.
- **Hỗ Trợ Tích Hợp Rộng Rãi Với Các Hệ Thống On-Premises:** Varonis tích hợp mạnh với các hệ thống On-Premises như Active Directory, NAS, SharePoint On-Prem, và hệ thống tệp Windows, điều này rất hữu ích cho các tổ chức có hạ tầng hybrid.
- **Phát Hiện Ransomware và Phản Ứng Tự Động:** Varonis có khả năng phát hiện và phản ứng tự động với các cuộc tấn công ransomware dựa trên phân tích hành vi, đưa ra cảnh báo kịp thời và tự động khóa quyền truy cập để ngăn chặn lây lan.
- **Độ Chính Xác Cao Trong Phân Loại Dữ Liệu Nhạy Cảm:** Varonis sử dụng công nghệ Machine Learning để phân loại dữ liệu nhạy cảm với độ chính xác cao, đồng thời nhận diện các mẫu dữ liệu cụ thể như PII, PHI, PCI một cách nhanh chóng.
- **Báo Cáo và Khả Năng Kiểm Toán Mạnh Mẽ:** Varonis cung cấp báo cáo chi tiết về quyền truy cập, hành vi người dùng và các lỗ hổng bảo mật, đồng thời hỗ trợ kiểm toán đầy đủ để đáp ứng các yêu cầu tuân thủ nghiêm ngặt.
- **Trải Nghiệm Người Dùng (UX) Tốt và Tích Hợp Tối Ưu:** Varonis cung cấp giao diện trực quan, dễ sử dụng, cùng khả năng tích hợp mạnh mẽ với các công cụ bảo mật khác như SIEM, DLP, IAM, giúp dễ dàng triển khai trong hạ tầng hiện có.

III. HƯỚNG GIẢI QUYẾT

1. Tổng quan giải pháp

Triển khai giám sát luồng dữ liệu sử dụng OpenMetadata và DSPM trong Varonis kết hợp các khả năng phát hiện, quản lý và bảo vệ dữ liệu nhạy cảm trong tổ chức. Triển khai giám sát luồng dữ liệu thô đã được ingest qua airbyte, đi qua OpenMetadata để govern và đánh dấu trước khi đến DSPM của Varonis. Varonis sẽ thực hiện kiểm soát quyền, giám sát hoạt động của data và người dùng.. Kết hợp các công nghệ này, chúng em sẽ thực hiện tạo ra 1 demo bao gồm 1 database cluster được quản lý bởi OpenMetadata và sử dụng DSPM của Varonis để kiểm soát và theo dõi data flow nhằm demo việc phát hiện và ngăn chặn data leak

2. Mô hình chung



Mô hình trên miêu tả quá trình Data flow đi từ Database qua vùng Giám sát được thiết lập bởi OpenMetaData để govern data từ nhiều database khác

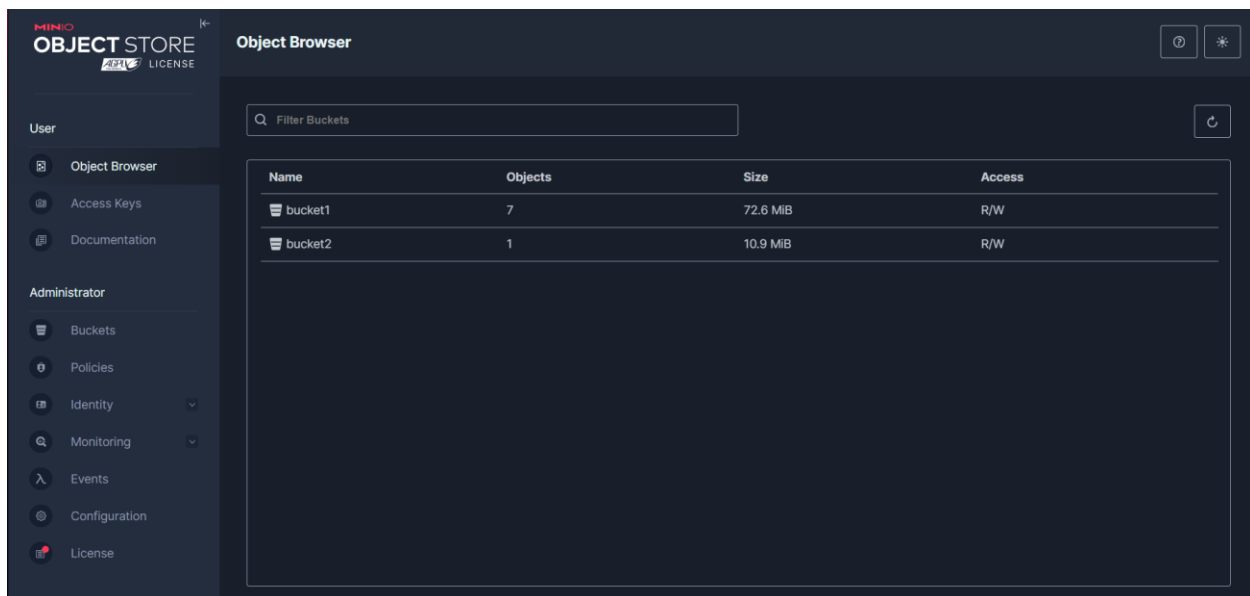
nhau, từ đó tạo ra 1 thể thống nhất để đưa vào DSPM để phát hiện các bất thường. Từ đó có thể phát hiện các rò rỉ, lỗ hổng trong các Database trong môi trường doanh nghiệp để ngăn chặn và vá kịp thời.

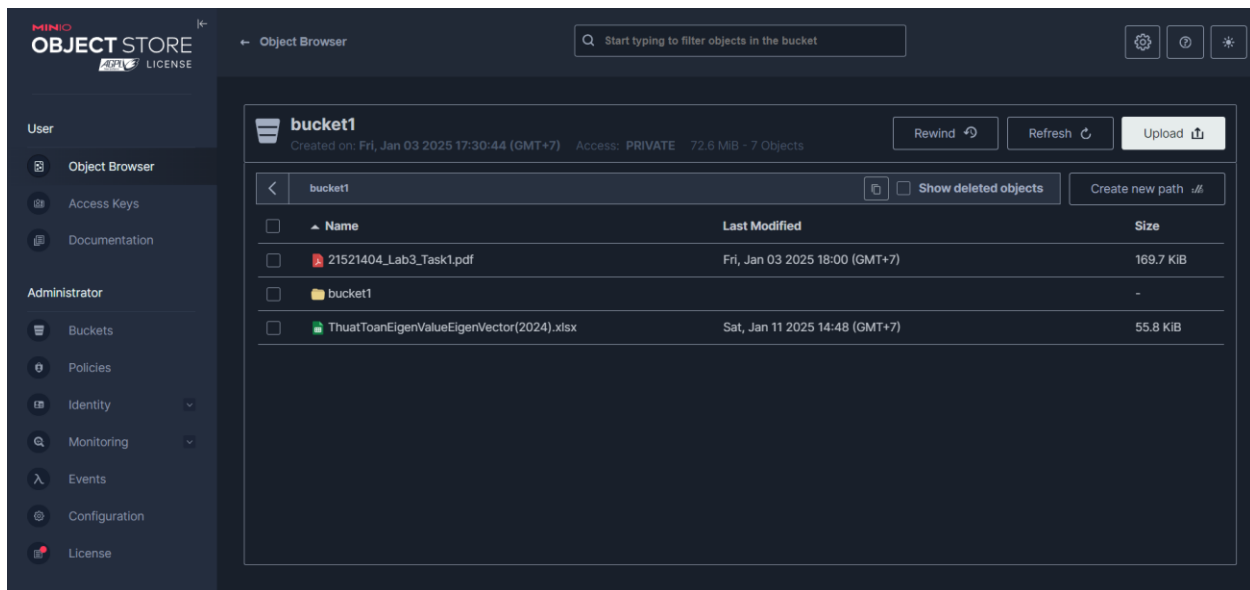
3. Triển khai hệ thống

a. Database

Khi dữ liệu đã được ingest, chúng ta cần một nơi để lưu trữ và quản lý nó. Đây là lúc MinIO phát huy vai trò của mình:

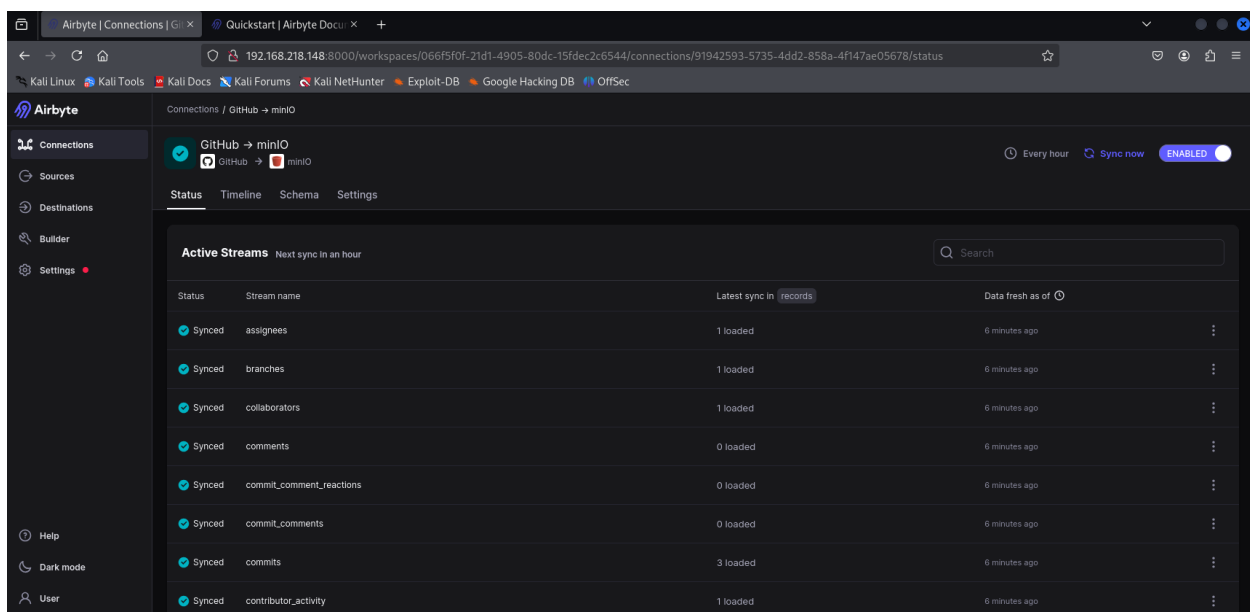
- **Lưu trữ dữ liệu:** MinIO là một hệ thống object storage, cung cấp nơi lưu trữ an toàn cho dữ liệu đã được ingest.
- **Quản lý metadata:** Sau khi dữ liệu được lưu trữ, MinIO sẽ được gắn các metadata để phục vụ cho quá trình phân tích và tìm kiếm sau này.





Ngoài việc đảm bảo khả năng truy xuất dữ liệu nhanh chóng, MinIO còn hỗ trợ bảo vệ dữ liệu thông qua các tính năng bảo mật tích hợp.

b. Ingestion and Managing Dataflow



Airbyte là công cụ tiếp theo trong hệ thống. Vai trò chính của Airbyte là thu thập và chuyển dữ liệu từ các nguồn khác nhau vào một cơ sở dữ liệu chung của doanh nghiệp. Các bước xử lý bao gồm:

- **Ingest dữ liệu:** Airbyte lấy dữ liệu từ các nguồn gốc như CRM, ERP hoặc các ứng dụng khác trong doanh nghiệp.
- **Đồng bộ:** Dữ liệu sau đó được chuyển vào một cơ sở dữ liệu riêng, nơi chúng ta có thể trích xuất metadata phục vụ phân tích.

Điểm mạnh của Airbyte là khả năng tích hợp linh hoạt với nhiều nguồn dữ liệu khác nhau, giúp tối ưu hóa quy trình thu thập và đồng bộ dữ liệu.

DAGs





All 0 Active 2 Paused 7

Running 0 Failed 0

Filter DAGs by tag

Search DAGs

Auto-refresh

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
00eca983-670f-47da-80f2-d29083bd8f79 Airbyte_metadata_rfcDYLNW OpenMetadata service:Airbyte type:metadata	admin	1	None	2025-01-15, 10:09:41		1	 	...
6d7b9e1a-43c6-4b1f-a8e1-c246a0c3b9b4 minio_metadata_ZLDDYelp OpenMetadata service:minio type:metadata	admin	2	30 * * * *	2025-01-15, 09:36:53	2025-01-15, 09:30:00	1	 	...

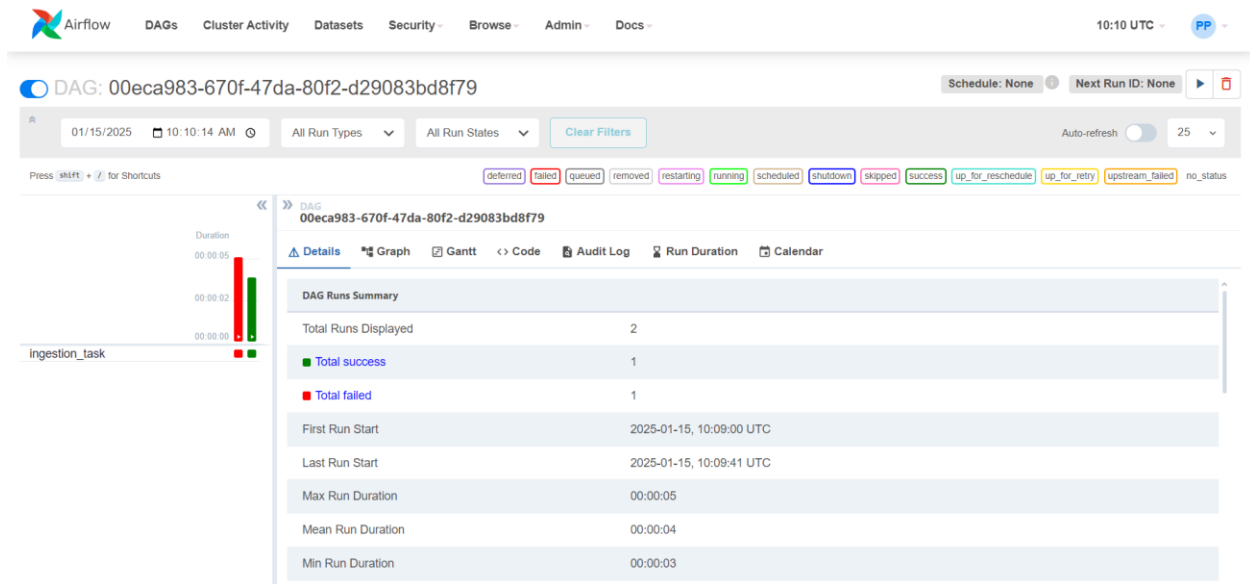
« 1 »

Showing 1-2 of 2 DAGs

AirFlow là một hệ thống mã nguồn mở dùng để quản lý dòng chảy dữ liệu trong pipeline của Airbyte. Trong hệ thống của chúng ta, AirFlow đảm nhận vai trò như sau:

- **Quản lý kết nối:** AirFlow chịu trách nhiệm giám sát và quản lý các kết nối giữa các thành phần trong pipeline của Airbyte.
- **Tự động hóa:** Với AirFlow, chúng ta có thể tự động hóa các tác vụ quan trọng, giúp giảm thiểu lỗi thủ công và tăng tính ổn định trong việc xử lý dữ liệu.

Ví dụ, khi có một thay đổi hoặc bổ sung nguồn dữ liệu, AirFlow sẽ đảm bảo rằng dòng chảy dữ liệu vẫn diễn ra một cách mượt mà mà không làm gián đoạn quy trình.



c. Data Governance

Cuối cùng, chúng ta sử dụng OpenMetadata để quản lý toàn diện dữ liệu và metadata:

- **Govern Data:** OpenMetadata cung cấp công cụ giám sát và quản lý dữ liệu theo các tiêu chuẩn bảo mật và tuân thủ.
- **Lưu trữ và tìm kiếm Metadata:** Metadata được lưu trữ trong OpenMetadata có thể dễ dàng được tìm kiếm và sử dụng trong quá trình phân tích bởi hệ thống DSPM của Varonis.

OpenMetadata là yếu tố then chốt giúp chúng ta kết nối dữ liệu và biến nó thành thông tin giá trị phục vụ cho các quyết định chiến lược.

IV. DEMO

https://drive.google.com/drive/folders/1YT7zWryfTTV_r7G4Yv1tKC0x4i9RJzee?usp=drive_link