# Tổng quan về Datahub

## **Datahub là gì ?**

* **DataHub** là một nền tảng siêu dữ liệu (metadata) mã nguồn mở cho doanh nghiệp dữ liệu hiện đại. Hiểu đơn giản, nó là một hệ thống giúp bạn quản lý, tổ chức, khám phá và hiểu rõ tất cả thông tin về dữ liệu của bạn (chứ không phải bản thân dữ liệu đó).
* **DataHub đóng vai trò như hệ thống mục lụ**c cho kho dữ liệu. Nó thu thập thông tin về dữ liệu (siêu dữ liệu) từ nhiều nguồn khác nhau và cung cấp một nơi tập trung để bạn có thể dễ dàng tìm kiếm, truy cập và hiểu rõ về dữ liệu của mình.

## **Tại sao cần dùng Datahub**

* DataHub ra đời để giải quyết những thách thức, cụ thể là:
  + **Khó khăn trong việc tìm kiếm dữ liệu**: Dữ liệu thường nằm rải rác ở nhiều nơi, gây khó khăn cho việc tìm kiếm và truy cập. DataHub tập hợp siêu dữ liệu từ các nguồn khác nhau vào một nơi, giúp bạn dễ dàng tìm kiếm và khám phá dữ liệu.
  + **Không hiểu rõ về dữ liệu**: Thiếu thông tin về nguồn gốc, chất lượng, ý nghĩa của dữ liệu dẫn đến việc sử dụng dữ liệu không hiệu quả. DataHub cung cấp thông tin chi tiết về dữ liệu, bao gồm nguồn gốc, lineage, chất lượng, và các thông tin ngữ nghĩa, giúp bạn hiểu rõ hơn về dữ liệu.
  + **Khó khăn trong việc quản trị dữ liệu**: Việc thiếu một hệ thống quản trị dữ liệu tập trung dẫn đến các vấn đề về bảo mật, tuân thủ và chất lượng dữ liệu. DataHub hỗ trợ quản trị dữ liệu bằng cách cho phép định nghĩa và thực thi các chính sách, quản lý quyền truy cập và theo dõi việc sử dụng dữ liệu.
  + **Thiếu sự cộng tác**: Các nhóm dữ liệu thường làm việc độc lập, dẫn đến việc trùng lặp công việc và thiếu sự chia sẻ kiến thức. DataHub cung cấp nền tảng để các nhóm dữ liệu cộng tác và chia sẻ thông tin về dữ liệu.
  + **Không nắm bắt được giá trị từ dữ liệu**: Do những khó khăn trên, các tổ chức thường không khai thác được tối đa giá trị từ dữ liệu của mình. DataHub giúp giải quyết những thách thức này, từ đó giúp các tổ chức khai thác tối đa giá trị từ dữ liệu.

## **Các khái niệm cốt lõi trong Datahub**

* **Metadata Model**: DataHub sử dụng một mô hình dữ liệu linh hoạt gọi là PDL (Pegasus Definition Language) để mô tả các thực thể và mối quan hệ giữa chúng. Mô hình này bao gồm các khái niệm chính sau:
  + **Entities (Thực thể):** Đại diện cho các đối tượng trong hệ sinh thái dữ liệu, ví dụ: Dataset, Dashboard, User, Pipeline,...
  + **Aspects (Khía cạnh):** Các thuộc tính của một thực thể, ví dụ: DatasetSchema (mô tả schema của dataset), DatasetProfile (thông tin thống kê của dataset), Ownership (thông tin về người sở hữu),...
  + **Relationships (Mối quan hệ):** Biểu diễn mối liên kết giữa các thực thể, ví dụ: một Dataset có thể được sử dụng bởi một Dashboard, hoặc một User là chủ sở hữu của một Dataset,...
* **Metadata Ingestion**: Quá trình thu thập siêu dữ liệu từ các nguồn dữ liệu khác nhau. DataHub cung cấp nhiều phương thức ingestion khác nhau, bao gồm:
  + **Push-based**: Đẩy siêu dữ liệu trực tiếp đến DataHub thông qua API.
  + **Pull-based**: DataHub chủ động kết nối và lấy siêu dữ liệu từ các nguồn dữ liệu.
* **Metadata Storage**: DataHub lưu trữ siêu dữ liệu trong một kho lưu trữ có thể mở rộng, hỗ trợ nhiều loại cơ sở dữ liệu như MySQL, Elasticsearch, Kafka.
* **Metadata Serving**: Cung cấp giao diện để truy cập và sử dụng siêu dữ liệu, bao gồm:
  + **GraphQL API**: Cho phép truy vấn siêu dữ liệu một cách linh hoạt.
  + **React-based UI**: Giao diện người dùng trực quan để tìm kiếm, duyệt và khám phá siêu dữ liệu.

## **Ưu điểm và nhược điểm của Datahub**

* **Ưu điểm:**
  + **Mã nguồn mở**: Miễn phí sử dụng và có thể tùy chỉnh theo nhu cầu.
  + **Khả năng mở rộng**: Được thiết kế để xử lý lượng lớn siêu dữ liệu và hỗ trợ nhiều người dùng đồng thời.
  + **Linh hoạt**: Hỗ trợ nhiều nguồn dữ liệu và có thể tích hợp với các công cụ khác trong hệ sinh thái dữ liệu.
  + **Khám phá dữ liệu mạnh mẽ**: Cung cấp giao diện tìm kiếm và duyệt trực quan, giúp người dùng dễ dàng tìm thấy dữ liệu họ cần.
  + **Quản trị dữ liệu hiệu quả**: Hỗ trợ định nghĩa và thực thi các chính sách quản trị dữ liệu, đảm bảo tính tuân thủ và bảo mật.
  + **Cộng đồng hỗ trợ tích cực**: Được hỗ trợ bởi một cộng đồng nguồn mở lớn và năng động.
  + **Data Lineage**: Hỗ trợ theo dõi luồng dữ liệu (Data Lineage) tốt.
* **Nhược điểm:**
  + **Độ phức tạp**: Việc triển khai và vận hành DataHub có thể phức tạp, đòi hỏi kiến thức chuyên môn về kỹ thuật.
  + **Tài nguyên**: Cần đầu tư về hạ tầng và nhân lực để triển khai và duy trì DataHub.
  + **Tính năng đang phát triển**: Một số tính năng vẫn đang trong quá trình phát triển và hoàn thiện.
  + **Tài liệu**: Mặc dù tài liệu ngày càng được cải thiện, nhưng vẫn có thể gặp khó khăn trong việc tìm kiếm thông tin cho một số trường hợp sử dụng cụ thể.

# Kiến trúc thành phần, Sơ đồ các component và vai trò của các component

