**Ex1: Online Extraction vs Offline Extraction. Example?**

Trích xuất trực tuyến là quá trình lấy dữ liệu từ nguồn trực tuyến như trang web, API hoặc cơ sở dữ liệu đám mây. Quá trình này thường được thực hiện bằng các công cụ hoặc ứng dụng web, cho phép người dùng truy cập và trích xuất dữ liệu mà không cần cài đặt phần mềm.

Ví dụ về trích xuất trực tuyến:

Sử dụng công cụ trích xuất dữ liệu web: Có rất nhiều công cụ trích xuất dữ liệu web cho phép bạn lấy dữ liệu từ các trang web. Ví dụ, bạn có thể sử dụng Import.io để trích xuất dữ liệu từ các bảng HTML, hoặc sử dụng Octoparse để trích xuất dữ liệu từ các trang web phức tạp hơn.

Sử dụng API: Nhiều dịch vụ trực tuyến cung cấp API cho phép bạn truy cập và trích xuất dữ liệu của họ. Ví dụ, bạn có thể sử dụng API của Google Search để trích xuất kết quả tìm kiếm, hoặc sử dụng API của Twitter để trích xuất tweet.

Sử dụng dịch vụ trích xuất dữ liệu đám mây: Một số công ty cung cấp dịch vụ trích xuất dữ liệu đám mây cho phép bạn trích xuất dữ liệu từ nhiều nguồn trực tuyến khác nhau. Ví dụ, bạn có thể sử dụng Cloud Data Fusion của Google để trích xuất dữ liệu từ các nguồn như Salesforce, Marketo và Zendesk.

Trích xuất ngoại tuyến là quá trình lấy dữ liệu từ nguồn ngoại tuyến như tệp tin, CD hoặc DVD. Quá trình này thường được thực hiện bằng phần mềm hoặc thiết bị chuyên dụng.

Ví dụ về trích xuất ngoại tuyến:

Sử dụng phần mềm quét ảnh: Bạn có thể sử dụng phần mềm quét ảnh để trích xuất dữ liệu từ ảnh hoặc tài liệu giấy. Ví dụ, bạn có thể sử dụng ABBYY FineReader để trích xuất văn bản từ ảnh quét.

Sử dụng phần mềm trích xuất dữ liệu từ CD/DVD: Bạn có thể sử dụng phần mềm trích xuất dữ liệu từ CD/DVD để trích xuất dữ liệu từ các đĩa quang. Ví dụ, bạn có thể sử dụng PowerISO để trích xuất dữ liệu từ CD/DVD.

Sử dụng máy sao chép dữ liệu: Bạn có thể sử dụng máy sao chép dữ liệu để trích xuất dữ liệu từ ổ cứng hoặc thiết bị lưu trữ khác.

Ex2: Motivation of ELT model. What's the difference between ETL and ELT?

Động lực của mô hình ELT:

1. Khả năng mở rộng: ELT cho phép xử lý dữ liệu một cách linh hoạt và hiệu quả hơn so với mô hình ETL truyền thống. Việc tách rời giai đoạn biến đổi dữ liệu khỏi giai đoạn tải dữ liệu cho phép các hệ thống ELT xử lý lượng dữ liệu lớn hơn mà không gặp phải tình trạng tắc nghẽn.

2. Tăng tốc độ: Bằng cách trì hoãn giai đoạn biến đổi dữ liệu, ELT có thể tải dữ liệu vào kho dữ liệu nhanh hơn, giúp giảm thời gian cần thiết để cung cấp dữ liệu cho phân tích.

3. Tính linh hoạt: ELT cung cấp nhiều sự linh hoạt hơn trong việc xử lý dữ liệu. Do dữ liệu thô được lưu trữ trong kho dữ liệu, người dùng có thể dễ dàng thay đổi cách thức biến đổi dữ liệu mà không cần tải lại toàn bộ dữ liệu.

4. Tiết kiệm chi phí: ELT có thể tiết kiệm chi phí bằng cách giảm thiểu nhu cầu lưu trữ dữ liệu đã được biến đổi. Dữ liệu thô thường chiếm ít dung lượng lưu trữ hơn dữ liệu đã được biến đổi.

5. Hợp tác tốt hơn: ELT hỗ trợ cộng tác tốt hơn giữa các nhà phân tích dữ liệu và các nhà khoa học dữ liệu. Do dữ liệu thô được lưu trữ trong kho dữ liệu, các nhà khoa học dữ liệu có thể dễ dàng truy cập và khám phá dữ liệu mà không cần phụ thuộc vào các nhà phân tích dữ liệu để chuẩn bị dữ liệu.

Sự khác biệt giữa ETL và ELT:

|  |  |  |
| --- | --- | --- |
| Đặc điểm | ETL | ELT |
| Giai đoạn biến đổi | Trước khi tải | Sau khi tải |
| Lưu trữ dữ liệu | Dữ liệu đã được biến đổi | Dữ liệu thô |
| Khả năng mở rộng | Hạn chế | Tốt |
| Tốc độ | Chậm hơn | Nhanh hơn |
| Tính linh hoạt | Thấp hơn | Cao hơn |
| Chi phí | Cao hơn | Thấp hơn |
| Hợp tác | Khó khăn hơn | Dễ dàng hơn |

**Ex3: Investigate the overall architecture of Docker; The differences between Docker Container and Docker Image; How to build a docker image; How to run a container from a self-built docker image; How to run a container from docker hub image.**

Kiến trúc tổng thể của Docker bao gồm các thành phần chính sau:

1. Docker Engine: Đây là thành phần cốt lõi của Docker, chịu trách nhiệm quản lý các container và image. Docker Engine bao gồm các công cụ để xây dựng, chạy, quản lý và kết nối với các container.

2. Docker Image: Image là một bản mẫu chỉ đọc chứa tất cả các tệp và phụ thuộc cần thiết để chạy một ứng dụng trong một container. Image được tạo từ Dockerfile, một tập lệnh văn bản mô tả cách cấu hình image.

3. Docker Container: Container là một instance chạy của một image. Nó bao gồm tất cả các tệp và phụ thuộc được chỉ định trong image, cũng như một hệ thống tệp riêng biệt và một môi trường mạng riêng biệt.

4. Docker Hub: Đây là một kho lưu trữ public cho các image Docker. Docker Hub cho phép bạn chia sẻ và tải xuống image với người khác.

5. Docker Registry: Registry là một kho lưu trữ private cho các image Docker. Registry có thể được lưu trữ trên máy chủ riêng của bạn hoặc trên dịch vụ registry đám mây như Amazon Elastic Container Registry (ECR).

Sự khác biệt giữa Docker Container và Docker Image:

Docker Image:

Là một bản mẫu chỉ đọc.

Chứa tất cả các tệp và phụ thuộc cần thiết để chạy một ứng dụng.

Được sử dụng để tạo các container.

Không thể thay đổi khi đã được tạo.

Docker Container:

Là một instance chạy của một image.

Bao gồm tất cả các tệp và phụ thuộc được chỉ định trong image, cũng như một hệ thống tệp riêng biệt và một môi trường mạng riêng biệt.

Có thể được khởi động, dừng, di chuyển và xóa.

Có thể được thay đổi khi đang chạy.