TRƯỜNG ĐẠI HỌC VĂN LANG

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**CÁC NỀN TẢNG PHÁT TRIỂN PHẦN MỀM**

NGÀNH: CÔNG NGHỆ THÔNG TIN

***Đề tài:***

**TÌM HIỂU**

**SVTH: Phạm Khánh Tài**

**MSSV: 2274802010776**

**GVHD: ThS. Nguyễn Minh Tân**

TP. Hồ Chí Minh – năm 2023

**MỤC LỤC**

1. **Giới thiệu**

**2. Các định dạng dữ liệu**

**3. Lưu trữ tập tin**

**4. Cơ sở dữ liệu**

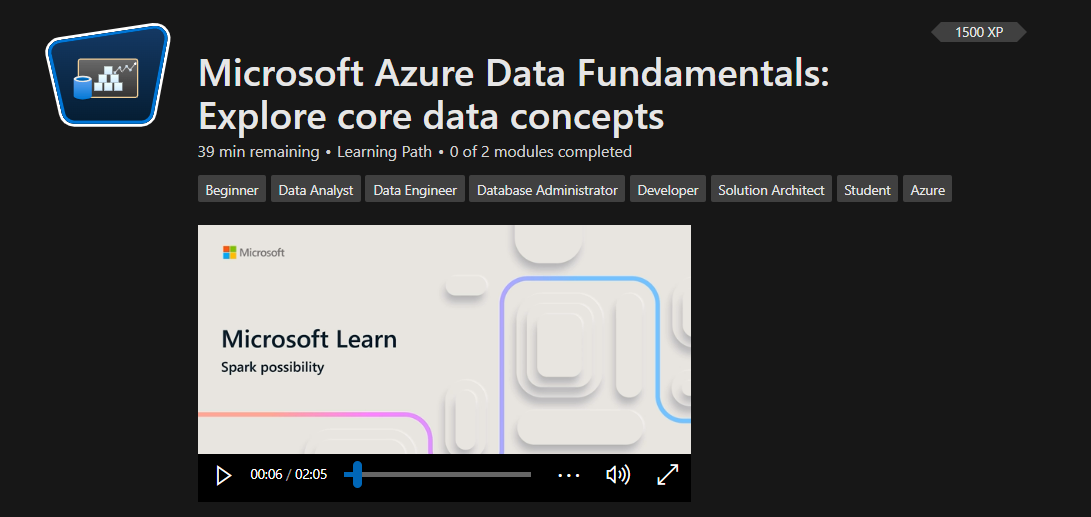
**5. Xử lý dữ liệu giao dịch**

**6. Xử lý dữ liệu phân tích**

**7. Kiểm tra**

**8. Tổng kết**

**NỘI DUNG**



1. **Giới thiệu**

Trong vài thập kỷ qua, lượng dữ liệu được tạo ra bởi các hệ thống, ứng dụng và thiết bị đã tăng lên đáng kể. Dữ liệu có ở khắp mọi nơi, ở vô số cấu trúc và định dạng.

Dữ liệu giờ đây dễ thu thập hơn và lưu trữ rẻ hơn, khiến hầu hết mọi doanh nghiệp đều có thể truy cập được. Giải pháp dữ liệu bao gồm các công nghệ phần mềm và nền tảng có thể giúp tạo điều kiện thuận lợi cho việc thu thập, phân tích và lưu trữ thông tin có giá trị. Trong mô-đun này, bạn sẽ tìm hiểu về các tùy chọn để biểu diễn và lưu trữ dữ liệu cũng như về khối lượng công việc dữ liệu điển hình. Khi hoàn thành mô-đun này, bạn sẽ xây dựng nền tảng để tìm hiểu về các kỹ thuật và dịch vụ được sử dụng để làm việc với dữ liệu.

**Mục tiêu học tập**

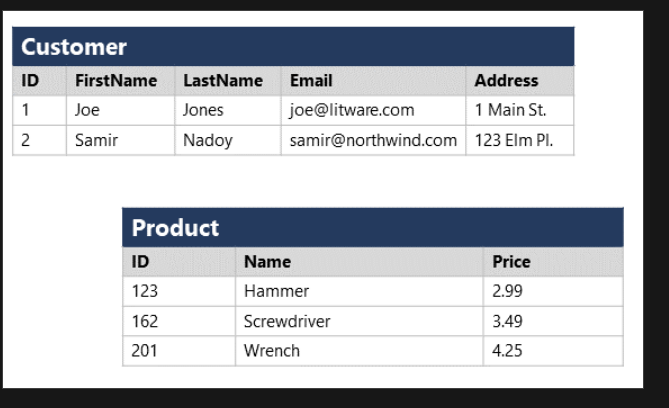
* Xác định các định dạng dữ liệu phổ biến
* Mô tả các tùy chọn để lưu trữ dữ liệu trong tập tin
* Mô tả các tùy chọn để lưu trữ dữ liệu trong cơ sở dữ liệu
* Mô tả đặc điểm của giải pháp xử lý dữ liệu giao dịch
* Mô tả đặc điểm của giải pháp xử lý dữ liệu phân tích

1. **Cách định dạng dữ liệu**

Dữ liệu là tập hợp các sự kiện như con số, mô tả và quan sát được sử dụng để ghi lại thông tin. Cấu trúc dữ liệu trong đó dữ liệu này được tổ chức thường đại diện cho *các thực thể* quan trọng đối với một tổ chức. Mỗi thực thể thường có một hoặc nhiều *thuộc tính* hoặc đặc điểm và có thể phân loại dữ liệu thành *có cấu trúc* , *bán cấu trúc* hoặc *không cấu trúc* .

**Dữ liệu có cấu trúc**

Dữ liệu có cấu trúc là dữ liệu tuân theo một *lược đồ* cố định , vì vậy tất cả dữ liệu đều có cùng trường hoặc thuộc tính. Thông thường nhất, lược đồ cho các thực thể dữ liệu có cấu trúc là *dạng bảng.*



**Dữ liệu bán cấu trúc**

*Dữ liệu bán cấu trúc* là thông tin có một số cấu trúc nhưng cho phép có một số biến thể giữa các phiên bản thực thể. Một định dạng phổ biến cho dữ liệu bán cấu trúc là *Ký hiệu đối tượng JavaScript* (JSON).

**Dữ liệu phi cấu trúc**

Không phải tất cả dữ liệu đều có cấu trúc hoặc thậm chí bán cấu trúc. Ví dụ: tài liệu, hình ảnh, dữ liệu âm thanh và video và tệp nhị phân có thể không có cấu trúc cụ thể. Loại dữ liệu này được gọi là dữ liệu *phi cấu trúc* .



**Kho dữ liệu**

Các tổ chức thường lưu trữ dữ liệu ở định dạng có cấu trúc, bán cấu trúc hoặc không cấu trúc để ghi lại chi tiết về các thực thể, các sự kiện cụ thể (chẳng hạn như giao dịch bán hàng) hoặc thông tin khác dưới dạng tài liệu, hình ảnh và các định dạng khác. Dữ liệu được lưu trữ sau đó có thể được lấy ra để phân tích và báo cáo sau này.

Có hai loại lưu trữ dữ liệu được sử dụng phổ biến:

* Lưu trữ tập tin
* Cơ sở dữ liệu

1. **Lưu trữ tập tin**

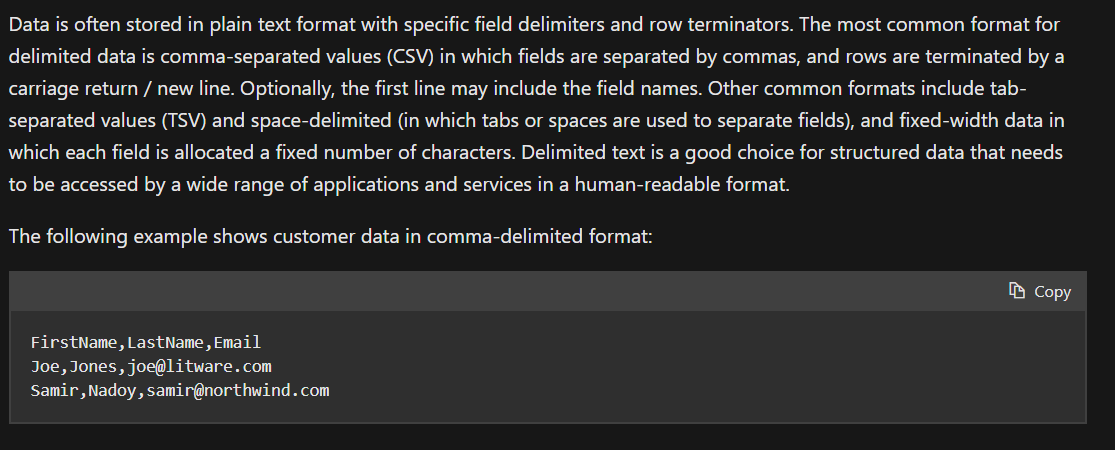
Khả năng lưu trữ dữ liệu dưới dạng tệp là yếu tố cốt lõi của bất kỳ hệ thống máy tính nào. Các tệp có thể được lưu trữ trong hệ thống tệp cục bộ trên đĩa cứng của máy tính cá nhân của bạn và trên phương tiện di động như ổ USB; nhưng ở hầu hết các tổ chức, các tệp dữ liệu quan trọng được lưu trữ tập trung trong một số loại hệ thống lưu trữ tệp dùng chung.

Định dạng tệp cụ thể được sử dụng để lưu trữ dữ liệu phụ thuộc vào một số yếu tố, bao gồm:

* Loại dữ liệu đang được lưu trữ (có cấu trúc, bán cấu trúc hoặc không cấu trúc).
* Các ứng dụng và dịch vụ sẽ cần đọc, ghi và xử lý dữ liệu.
* Nhu cầu về các tệp dữ liệu mà con người có thể đọc được hoặc được tối ưu hóa để lưu trữ và xử lý hiệu quả.

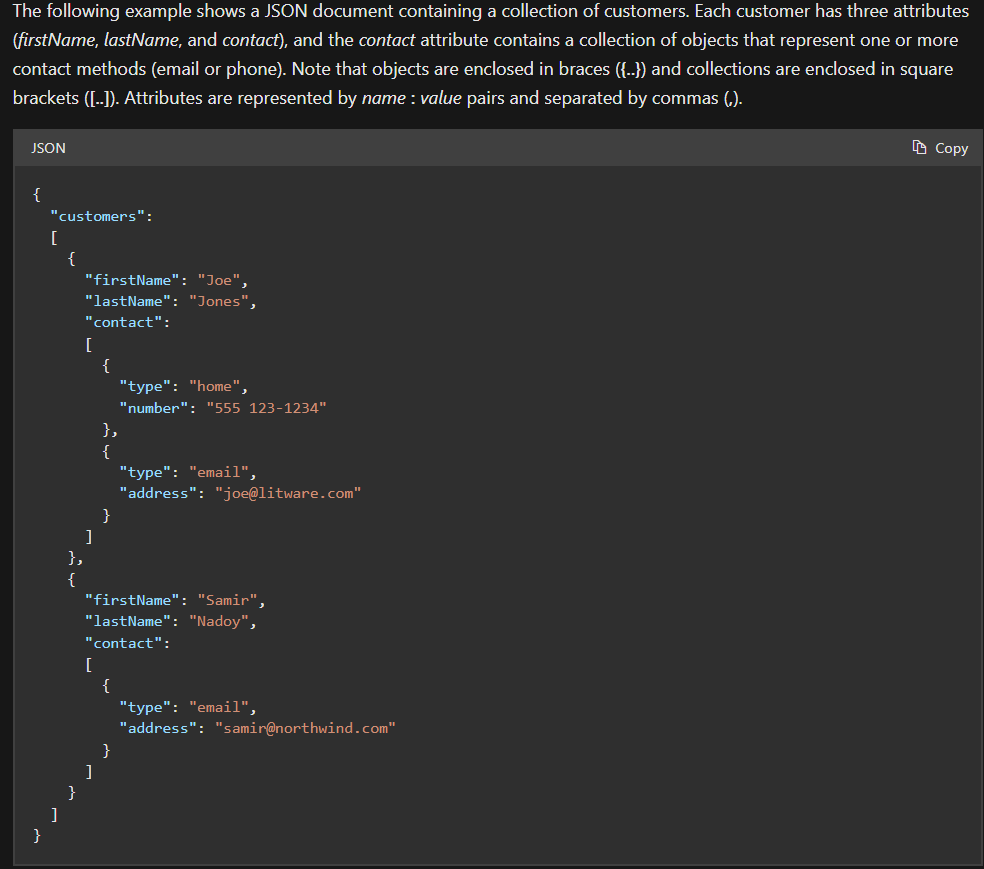
**Tệp văn bản được phân tách**

Dữ liệu thường được lưu trữ ở định dạng văn bản thuần túy với các dấu phân cách trường và dấu kết thúc hàng cụ thể. Định dạng phổ biến nhất cho dữ liệu được phân tách là các giá trị được phân tách bằng dấu phẩy (CSV), trong đó các trường được phân tách bằng dấu phẩy và các hàng được kết thúc bằng dấu xuống dòng/dòng mới.

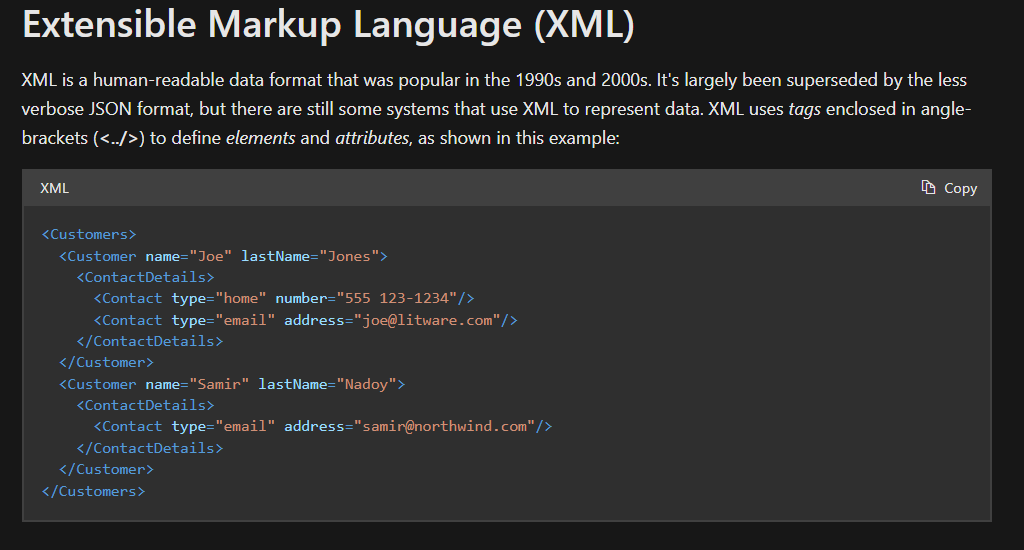


**Ký hiệu đối tượng Javascript (JSON)**

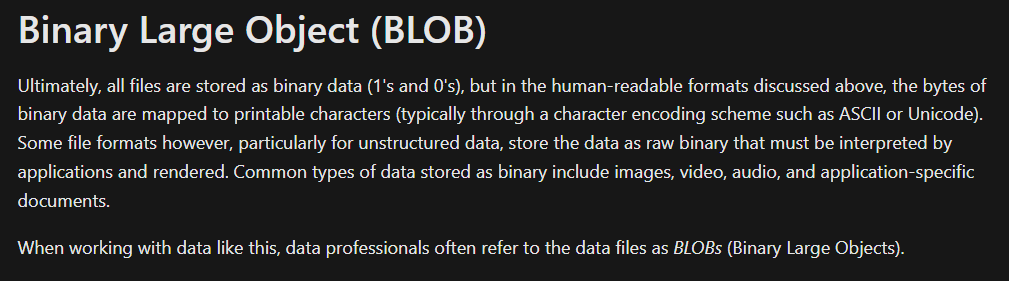
JSON là một định dạng phổ biến trong đó lược đồ tài liệu phân cấp được sử dụng để xác định các thực thể dữ liệu (đối tượng) có nhiều thuộc tính. Mỗi thuộc tính có thể là một đối tượng (hoặc một tập hợp các đối tượng); biến JSON thành định dạng linh hoạt phù hợp với cả dữ liệu có cấu trúc và bán cấu trúc.



**Ngôn ngữ đánh dấu mở rộng**

****

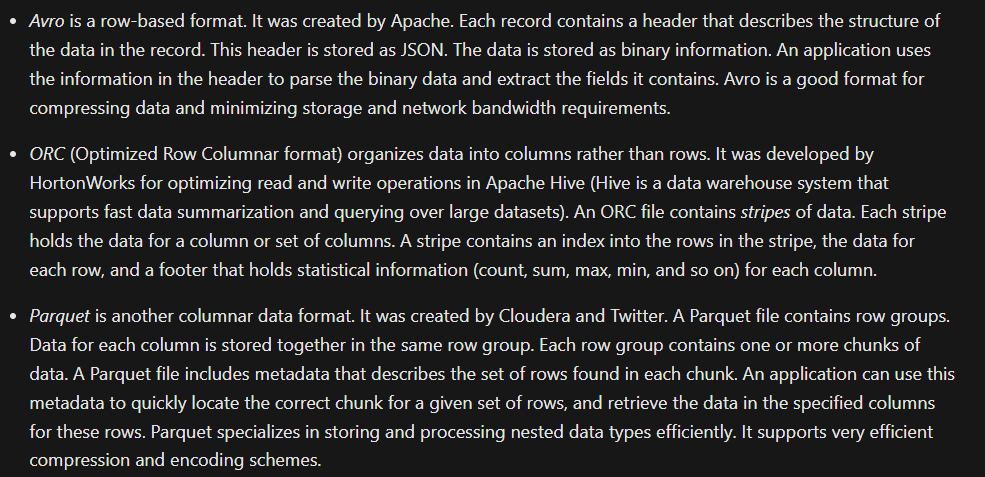
**Đối tượng lớn nhị phân**

****

**Định dạng tập tin được tối ưu hóa**

Mặc dù các định dạng mà con người có thể đọc được đối với dữ liệu có cấu trúc và bán cấu trúc có thể hữu ích nhưng chúng thường không được tối ưu hóa cho không gian lưu trữ hoặc xử lý. Theo thời gian, một số định dạng tệp chuyên dụng cho phép nén, lập chỉ mục cũng như lưu trữ và xử lý hiệu quả đã được phát triển.

Một số định dạng tệp được tối ưu hóa phổ biến mà bạn có thể thấy bao gồm *Avro* , *ORC* và *Parquet*

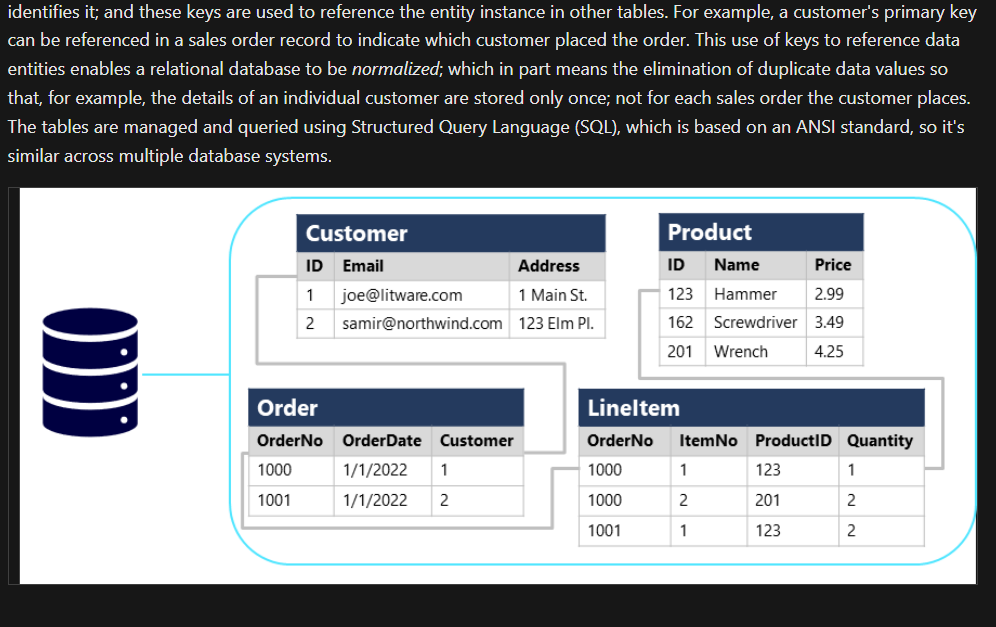


1. **Cơ sở dữ liệu**

Cơ sở dữ liệu được sử dụng để xác định một hệ thống trung tâm trong đó dữ liệu có thể được lưu trữ và truy vấn. Theo nghĩa đơn giản, hệ thống tệp nơi các tệp được lưu trữ là một loại cơ sở dữ liệu; nhưng khi chúng tôi sử dụng thuật ngữ này trong ngữ cảnh dữ liệu chuyên nghiệp, chúng tôi thường muốn nói đến một hệ thống chuyên dụng để quản lý các bản ghi dữ liệu thay vì các tệp.

**Cơ sở dữ liệu quan hệ**

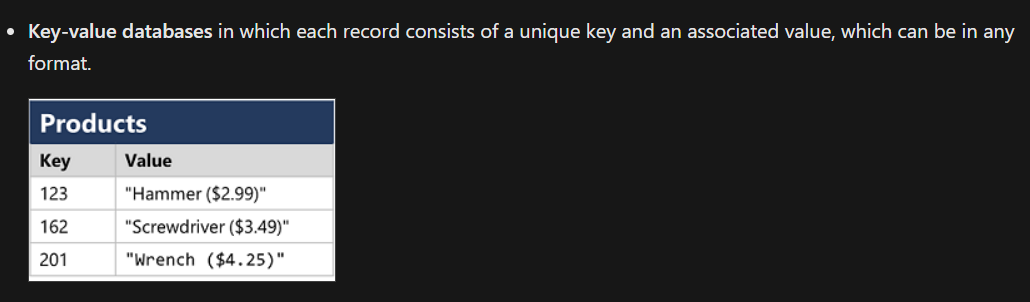
Cơ sở dữ liệu quan hệ thường được sử dụng để lưu trữ và truy vấn dữ liệu có cấu trúc. Dữ liệu được lưu trữ trong các bảng đại diện cho các thực thể, chẳng hạn như khách hàng, sản phẩm hoặc đơn đặt hàng. Mỗi phiên bản của một thực thể được gán một *khóa chính* xác định duy nhất nó; và các khóa này được sử dụng để tham chiếu thể hiện thực thể trong các bảng khác.

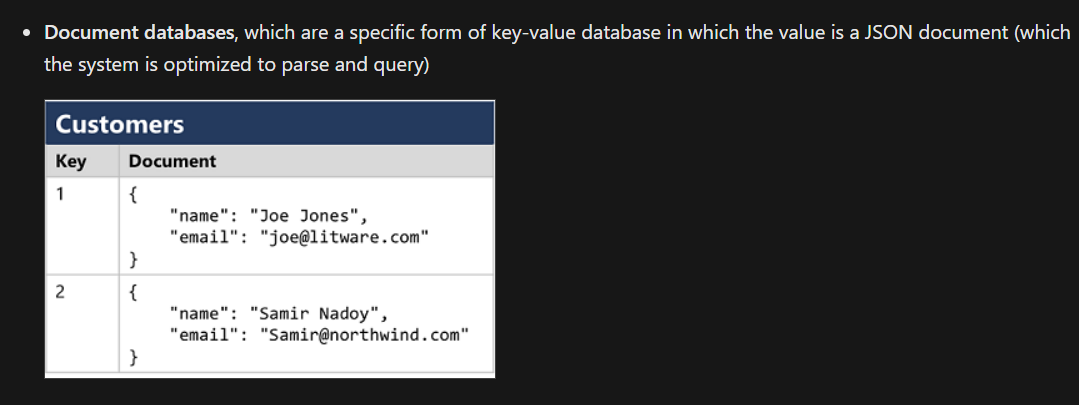


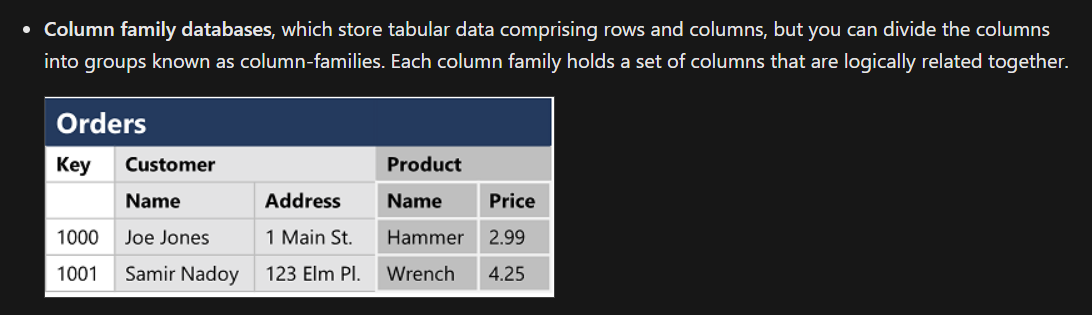
**Cơ sở dữ liệu phi quan hệ**

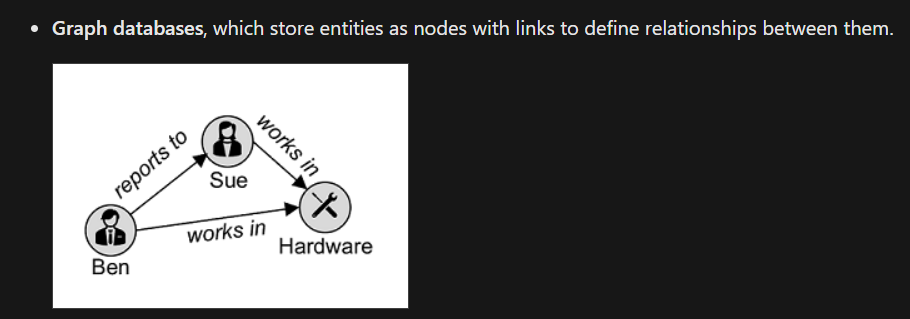
Cơ sở dữ liệu phi quan hệ là hệ thống quản lý dữ liệu không áp dụng lược đồ quan hệ cho dữ liệu. Cơ sở dữ liệu phi quan hệ thường được gọi là cơ sở dữ liệu NoSQL, mặc dù một số hỗ trợ một biến thể của ngôn ngữ SQL.

Có bốn loại cơ sở dữ liệu phi quan hệ thường được sử dụng.





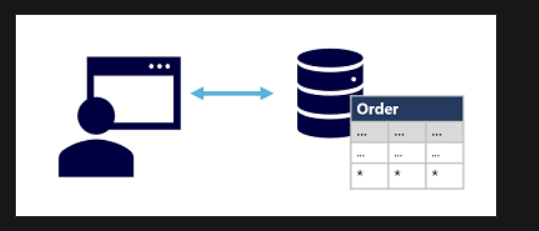




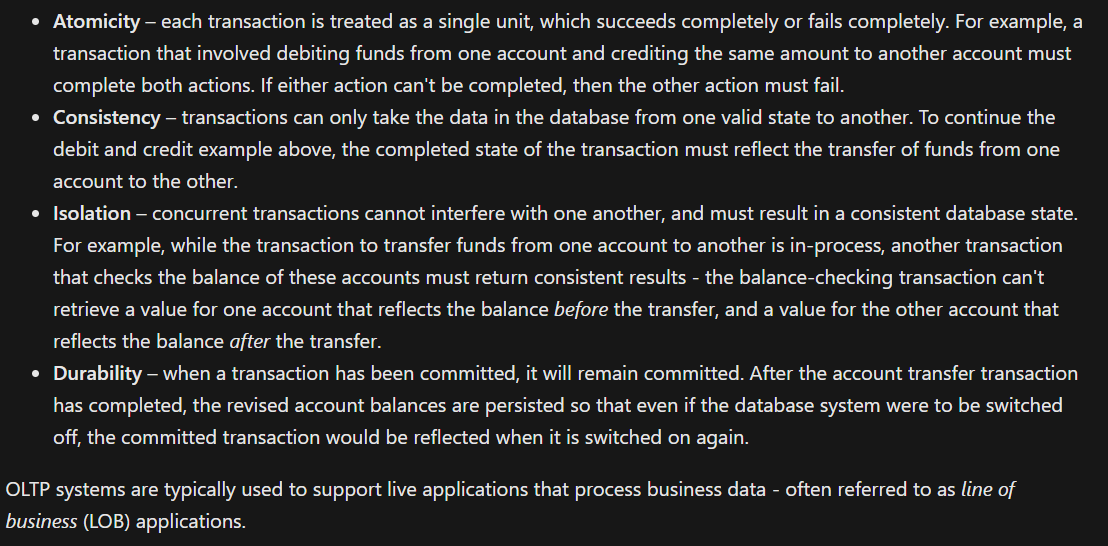
1. **Xử Lý Dữ Liệu Giao Dịch**

Hệ thống xử lý dữ liệu giao dịch là thứ mà hầu hết mọi người coi là chức năng chính của điện toán doanh nghiệp. Hệ thống giao dịch ghi lại *các giao dịch* gói gọn các sự kiện cụ thể mà tổ chức muốn theo dõi. Một giao dịch có thể mang tính chất tài chính, chẳng hạn như chuyển tiền giữa các tài khoản trong hệ thống ngân hàng hoặc có thể là một phần của hệ thống bán lẻ, theo dõi các khoản thanh toán hàng hóa và dịch vụ từ khách hàng.

Hệ thống giao dịch thường có khối lượng lớn, đôi khi xử lý hàng triệu giao dịch trong một ngày. Dữ liệu đang được xử lý phải được truy cập rất nhanh. Công việc được thực hiện bởi các hệ thống giao dịch thường được gọi là Xử lý giao dịch trực tuyến (OLTP).

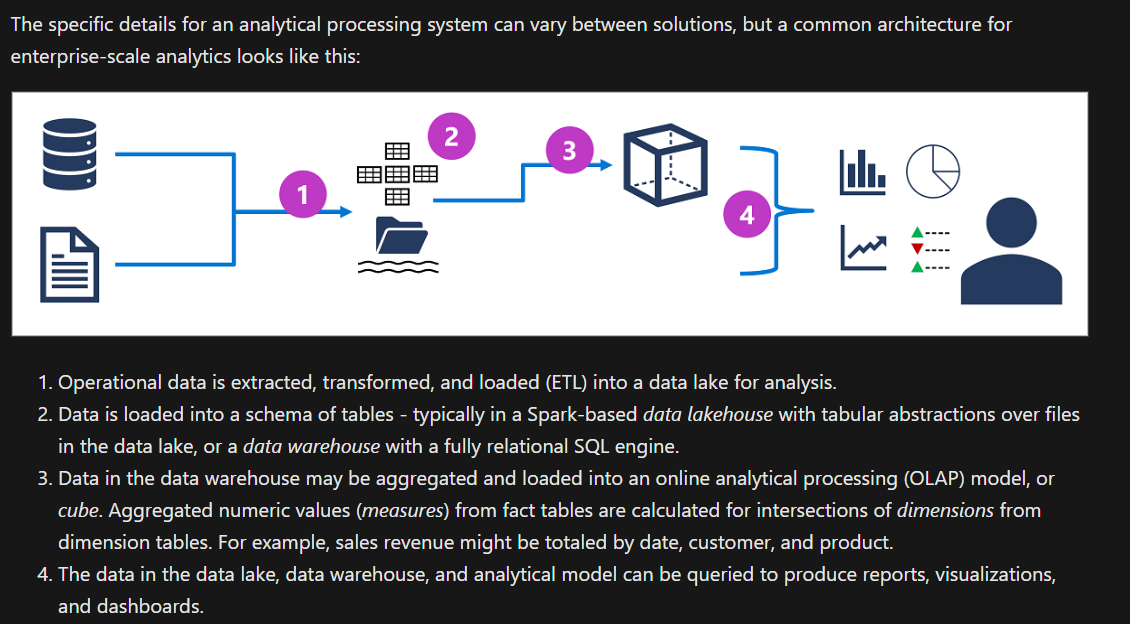


Các giải pháp OLTP dựa trên hệ thống cơ sở dữ liệu trong đó việc lưu trữ dữ liệu được tối ưu hóa cho cả hoạt động đọc và ghi nhằm hỗ trợ khối lượng công việc giao dịch trong đó các bản ghi dữ liệu được tạo, truy xuất, cập nhật và xóa (thường được gọi là hoạt động *CRUD* ). Các hoạt động này được áp dụng mang tính giao dịch, theo cách đảm bảo tính toàn vẹn của dữ liệu được lưu trữ trong cơ sở dữ liệu. Để thực hiện điều này, hệ thống OLTP thực thi các giao dịch hỗ trợ cái gọi là ngữ nghĩa ACID:



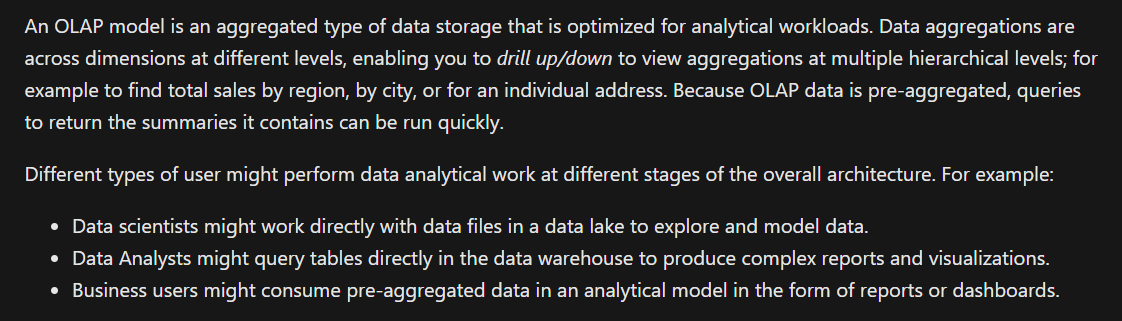
1. **Xử lý dữ liệu phân tích**

Xử lý dữ liệu phân tích thường sử dụng các hệ thống chỉ đọc (hoặc *hầu hết* đọc ) để lưu trữ khối lượng lớn dữ liệu lịch sử hoặc số liệu kinh doanh. Phân tích có thể dựa trên ảnh chụp nhanh dữ liệu tại một thời điểm nhất định hoặc một loạt ảnh chụp nhanh.

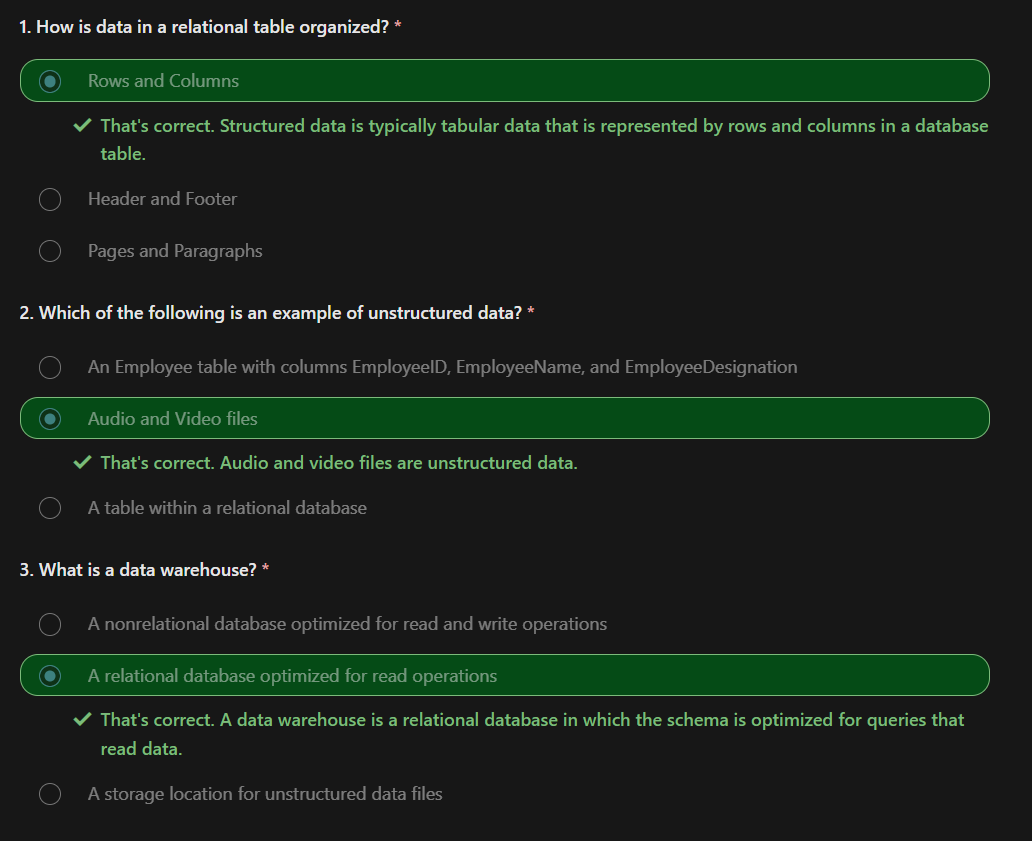


*Hồ dữ liệu* rất phổ biến trong các tình huống xử lý phân tích dữ liệu quy mô lớn, trong đó phải thu thập và phân tích một khối lượng lớn dữ liệu dựa trên tệp.

*Kho dữ liệu* là một cách được thiết lập để lưu trữ dữ liệu trong một lược đồ quan hệ được tối ưu hóa cho các hoạt động đọc – chủ yếu là các truy vấn để hỗ trợ báo cáo và trực quan hóa dữ liệu.



1. **Kiểm Tra Kiến Thức**

****

1. **Tổng Kết**

Dữ liệu là cốt lõi của hầu hết các ứng dụng và giải pháp phần mềm. Nó có thể được thể hiện dưới nhiều định dạng, được lưu trữ trong các tệp và cơ sở dữ liệu và được sử dụng để ghi lại các giao dịch hoặc hỗ trợ phân tích và báo cáo.

