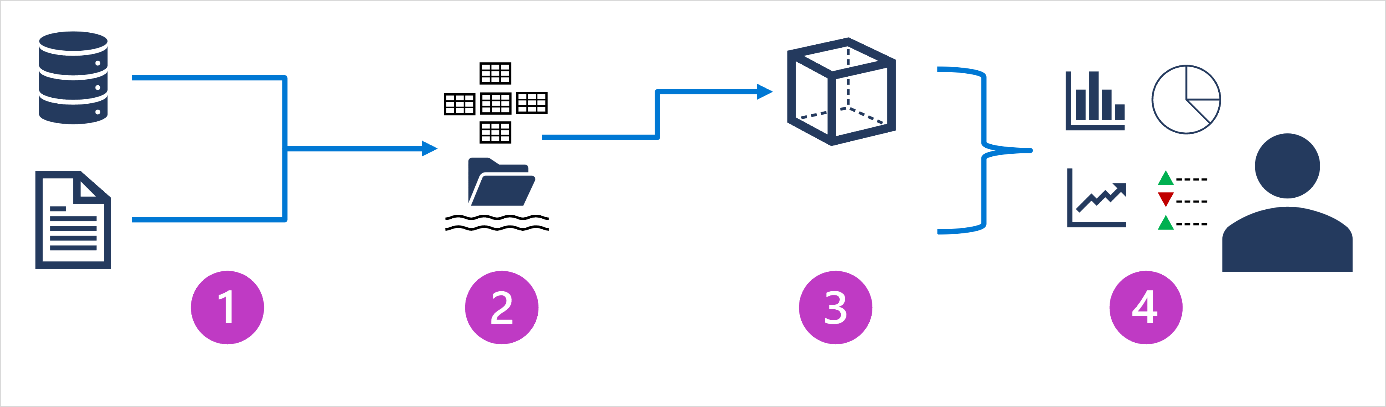
**Describe data warehousing architecture**

Completed100 XP

* 3 minutes

Large-scale data analytics architecture can vary, as can the specific technologies used to implement it; but in general, the following elements are included:



1. **Data ingestion and processing** – data from one or more transactional data stores, files, real-time streams, or other sources is loaded into a data lake or a relational data warehouse. The load operation usually involves an *extract, transform, and load* (ETL) or *extract, load, and transform* (ELT) process in which the data is cleaned, filtered, and restructured for analysis. In ETL processes, the data is transformed before being loaded into an analytical store, while in an ELT process the data is copied to the store and then transformed. Either way, the resulting data structure is optimized for analytical queries. The data processing is often performed by distributed systems that can process high volumes of data in parallel using multi-node clusters. Data ingestion includes both batch processing of static data and real-time processing of streaming data.
2. **Analytical data store** – data stores for large scale analytics include relational *data warehouses*, file-system based *data lakes*, and hybrid architectures that combine features of data warehouses and data lakes (sometimes called *data lakehouses* or *lake databases*). We'll discuss these in more depth later.
3. **Analytical data model** – while data analysts and data scientists can work with the data directly in the analytical data store, it’s common to create one or more data models that pre-aggregate the data to make it easier to produce reports, dashboards, and interactive visualizations. Often these data models are described as *cubes*, in which numeric data values are aggregated across one or more dimensions (for example, to determine total sales by product and region). The model encapsulates the relationships between data values and dimensional entities to support "drill-up/drill-down" analysis.
4. **Data visualization** – data analysts consume data from analytical models, and directly from analytical stores to create reports, dashboards, and other visualizations. Additionally, users in an organization who may not be technology professionals might perform self-service data analysis and reporting. The visualizations from the data show trends, comparisons, and key performance indicators (KPIs) for a business or other organization, and can take the form of printed reports, graphs and charts in documents or PowerPoint presentations, web-based dashboards, and interactive environments in which users can explore data visually.

**Kiến trúc data warehouse - Tổng quan dễ hiểu**

Kiến trúc phân tích dữ liệu quy mô lớn có thể thay đổi tùy theo từng doanh nghiệp, cũng như các công nghệ cụ thể được sử dụng để triển khai nó. Tuy nhiên, nhìn chung, kiến trúc data warehouse thường bao gồm các yếu tố sau:

1. **Thu thập và xử lý dữ liệu (Data ingestion and processing):** Dữ liệu từ một hoặc nhiều kho lưu trữ dữ liệu giao dịch, tệp tin, luồng dữ liệu thời gian thực hoặc các nguồn khác được đưa vào data lake (hồ dữ liệu) hoặc data warehouse quan hệ. Quá trình tải dữ liệu thường liên quan đến việc trích xuất, chuyển đổi và tải (ETL) hoặc trích xuất, tải và chuyển đổi (ELT). Trong quá trình này, dữ liệu được làm sạch, lọc và sắp xếp lại để phục vụ cho việc phân tích.
   * **ETL (Extract, Transform, Load):** Dữ liệu được chuyển đổi trước khi được nạp vào kho lưu trữ phân tích.
   * **ELT (Extract, Load, Transform):** Dữ liệu được sao chép vào kho lưu trữ và sau đó mới được chuyển đổi.
   * Bất kể phương pháp nào được sử dụng, cấu trúc dữ liệu kết quả đều được tối ưu hóa cho các truy vấn phân tích. Việc xử lý dữ liệu thường được thực hiện bởi các hệ thống phân tán có khả năng xử lý khối lượng lớn dữ liệu song song bằng cách sử dụng các cụm nhiều nút. Thu thập dữ liệu bao gồm cả xử lý theo lô đối với dữ liệu tĩnh và xử lý theo thời gian thực đối với dữ liệu streaming.
2. **Kho lưu trữ dữ liệu phân tích (Analytical data store):** Các kho lưu trữ dữ liệu để phân tích quy mô lớn bao gồm data warehouse quan hệ, data lake dựa trên hệ thống tập tin và kiến trúc lai kết hợp các tính năng của data warehouse và data lake (đôi khi được gọi là data lakehouse or lake database). Chúng ta sẽ thảo luận chi tiết hơn về những điều này sau.
3. **Mô hình dữ liệu phân tích (Analytical data model):** Mặc dù các nhà phân tích dữ liệu và nhà khoa học dữ liệu có thể trực tiếp làm việc với dữ liệu trong kho lưu trữ phân tích, nhưng việc tạo ra một hoặc nhiều mô hình dữ liệu để tổng hợp sẵn dữ liệu trước sẽ giúp việc tạo báo cáo, bảng điều khiển và hình ảnh trực quan tương tác dễ dàng hơn. Thông thường, các mô hình dữ liệu này được mô tả dưới dạng khối (cube), trong đó các giá trị dữ liệu số được tổng hợp trên một hoặc nhiều chiều (ví dụ: để xác định tổng doanh số theo sản phẩm và khu vực). Mô hình này đóng gói các mối quan hệ giữa các giá trị dữ liệu và các thực thể theo chiều để hỗ trợ phân tích "cuộn lên/cuộn xuống" (drill-up/drill-down).
4. **Trực quan hóa dữ liệu (Data visualization):** Các nhà phân tích dữ liệu sử dụng dữ liệu từ các mô hình phân tích và trực tiếp từ các kho lưu trữ phân tích để tạo báo cáo, bảng điều khiển và các hình ảnh trực quan khác. Ngoài ra, những người dùng trong tổ chức không phải là chuyên gia công nghệ có thể tự thực hiện phân tích và báo cáo dữ liệu. Các hình ảnh trực quan từ dữ liệu hiển thị xu hướng, so sánh và các chỉ số hiệu suất chính (KPIs) cho một doanh nghiệp hoặc tổ chức khác và có thể dưới dạng báo cáo in, biểu đồ và hình minh họa trong tài liệu hoặc bản trình bày PowerPoint, bảng điều khiển trên web và môi trường tương tác nơi người dùng có thể khám phá dữ liệu trực quan.