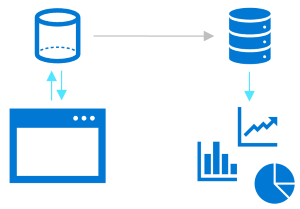
**Important data engineering concepts**

Completed100 XP

* 6 minutes

There are some core concepts with which data engineers should be familiar. These concepts underpin many of the workloads that data engineers must implement and support.

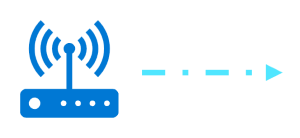
**Operational and analytical data**



*Operational* data is usually transactional data that is generated and stored by applications, often in a relational or non-relational database. *Analytical* data is data that has been optimized for analysis and reporting, often in a data warehouse.

One of the core responsibilities of a data engineer is to design, implement, and manage solutions that integrate operational and analytical data sources or extract operational data from multiple systems, transform it into appropriate structures for analytics, and load it into an analytical data store (usually referred to as ETL solutions).

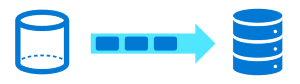
**Streaming data**



Streaming data refers to perpetual sources of data that generate data values in real-time, often relating to specific events. Common sources of streaming data include internet-of-things (IoT) devices and social media feeds.

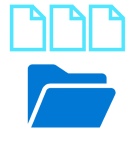
Data engineers often need to implement solutions that capture real-time stream of data and ingest them into analytical data systems, often combining the real-time data with other application data that is processed in batches.

**Data pipelines**



Data pipelines are used to orchestrate activities that transfer and transform data. Pipelines are the primary way in which data engineers implement repeatable extract, transform, and load (ETL) solutions that can be triggered based on a schedule or in response to events.

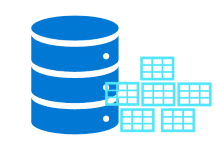
**Data lakes**



A data lake is a storage repository that holds large amounts of data in native, raw formats. Data lake stores are optimized for scaling to massive volumes (terabytes or petabytes) of data. The data typically comes from multiple heterogeneous sources, and may be structured, semi-structured, or unstructured.

The idea with a data lake is to store everything in its original, untransformed state. This approach differs from a traditional data warehouse, which transforms and processes the data at the time of ingestion.

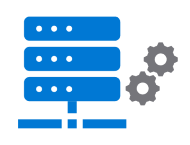
**Data warehouses**



A data warehouse is a centralized repository of integrated data from one or more disparate sources. Data warehouses store current and historical data in relational tables that are organized into a schema that optimizes performance for analytical queries.

Data engineers are responsible for designing and implementing relational data warehouses, and managing regular data loads into tables.

**Apache Spark**



Apache Spark is a parallel processing framework that takes advantage of in-memory processing and a distributed file storage. It's a common open-source software (OSS) tool for big data scenarios.

Data engineers need to be proficient with Spark, using notebooks and other code artifacts to process data in a data lake and prepare it for modeling and analysis.

**Dữ liệu vận hành và dữ liệu phân tích**

* **Dữ liệu vận hành (Operational data):** Thường là dữ liệu giao dịch được tạo và lưu trữ bởi các ứng dụng, thường nằm trong cơ sở dữ liệu quan hệ hoặc phi quan hệ.
* **Dữ liệu phân tích (Analytical data):** Là dữ liệu đã được tối ưu hóa để phân tích và báo cáo, thường nằm trong kho dữ liệu (data warehouse).

Một trong những trách nhiệm cốt lõi của kỹ sư dữ liệu là thiết kế, triển khai và quản lý các giải pháp tích hợp các nguồn dữ liệu vận hành và phân tích, hoặc trích xuất dữ liệu vận hành từ nhiều hệ thống khác nhau, chuyển đổi dữ liệu thành các cấu trúc phù hợp để phân tích và tải dữ liệu vào kho lưu trữ dữ liệu phân tích (thường được gọi là các giải pháp ETL - Extract, Transform, Load - Trích xuất, Chuyển đổi, Tải).

**Dữ liệu Streaming (Streaming data)**

Dữ liệu Streaming là các nguồn dữ liệu liên tục tạo ra các giá trị dữ liệu theo thời gian thực, thường liên quan đến các sự kiện cụ thể. Các nguồn dữ liệu streaming phổ biến bao gồm thiết bị Internet vạn vật (IoT) và nguồn cấp dữ liệu mạng xã hội.

Kỹ sư dữ liệu thường cần triển khai các giải pháp để thu thập luồng dữ liệu theo thời gian thực và đưa chúng vào các hệ thống dữ liệu phân tích, thường kết hợp dữ liệu thời gian thực với các dữ liệu ứng dụng khác được xử lý theo từng đợt (batch).

**Các luồng dữ liệu (Data pipelines)**

Các luồng dữ liệu được sử dụng để điều phối các hoạt động truyền và chuyển đổi dữ liệu. Luồng dữ liệu là cách chính mà kỹ sư dữ liệu triển khai các giải pháp ETL (Extract, Transform, Load) có thể lặp lại, được kích hoạt theo lịch trình hoặc để phản hồi các sự kiện.

**Hồ dữ liệu (Data lakes)**

Hồ dữ liệu là một kho lưu trữ chứa một lượng lớn dữ liệu ở định dạng gốc, chưa qua xử lý. Lưu trữ hồ dữ liệu được tối ưu hóa để mở rộng quy mô tới khối lượng dữ liệu khổng lồ (terabyte hoặc petabyte). Dữ liệu thường đến từ nhiều nguồn không đồng nhất và có thể được cấu trúc, bán cấu trúc hoặc không cấu trúc.

Ý tưởng của hồ dữ liệu là lưu trữ mọi thứ ở trạng thái gốc, chưa được chuyển đổi. Cách tiếp cận này khác với kho dữ liệu truyền thống, nơi chuyển đổi và xử lý dữ liệu tại thời điểm đưa vào.

**Kho dữ liệu (Data warehouses)**

Kho dữ liệu là một kho lưu trữ tập trung của dữ liệu tích hợp từ một hoặc nhiều nguồn khác nhau. Kho dữ liệu lưu trữ dữ liệu hiện tại và lịch sử trong các bảng quan hệ được tổ chức thành một lược đồ (schema) để tối ưu hóa hiệu suất cho các truy vấn phân tích.

Kỹ sư dữ liệu chịu trách nhiệm thiết kế và triển khai các kho dữ liệu quan hệ, đồng thời quản lý việc tải dữ liệu thường xuyên vào các bảng.

**Apache Spark**

Apache Spark là một khung xử lý song song tận dụng xử lý trong bộ nhớ và lưu trữ tệp phân tán. Đây là một công cụ phần mềm mã nguồn mở (OSS) phổ biến cho các tình huống dữ liệu lớn (big data).

Kỹ sư dữ liệu cần thành thạo Spark, sử dụng sổ ghi chép (notebooks) và các sản phẩm mã khác để xử lý dữ liệu trong hồ dữ liệu và chuẩn bị dữ liệu cho mô hình hóa và phân tích.