**What is data engineering**

Completed100 XP

* 5 minutes

The data engineer will often work with multiple types of data to perform many operations using many scripting or coding languages that are appropriate to their individual organization.

**Types of data**

There are three primary types of data that a data engineer will work with.

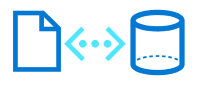
Expand table

| **Structured** | **Semi-structured** | **Unstructured** |
| --- | --- | --- |
| Diagram of Structured data type. | Diagram of Semi-structured data type. | Diagram of Unstructured data type. |
| Structured data primarily comes from table-based source systems such as a relational database or from a flat file such as a comma separated (CSV) file. The primary element of a structured file is that the rows and columns are aligned consistently throughout the file. | Semi-structured data is data such as JavaScript object notation (JSON) files, which may require flattening prior to loading into your source system. When flattened, this data doesn't have to fit neatly into a table structure. | Unstructured data includes data stored as key-value pairs that don't adhere to standard relational models and Other types of unstructured data that are commonly used include portable data format (PDF), word processor documents, and images. |

**Data operations**

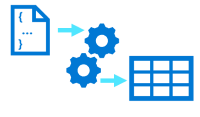
As a data engineer some of the main tasks that you'll perform in Azure include *data integration*, *data transformation*, and *data consolidation*.

**Data integration**



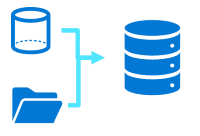
Data Integration involves establishing links between operational and analytical services and data sources to enable secure, reliable access to data across multiple systems. For example, a business process might rely on data that is spread across multiple systems, and a data engineer is required to establish links so that the required data can be extracted from all of these systems.

**Data transformation**



Operational data usually needs to be *transformed* into suitable structure and format for analysis, often as part of an *extract, transform, and load* (ETL) process; though increasingly a variation in which you *extract, load, and transform* (ELT) the data is used to quickly ingest the data into a data lake and then apply "big data" processing techniques to transform it. Regardless of the approach used, the data is prepared to support downstream analytical needs.

**Data consolidation**



Data consolidation is the process of combining data that has been extracted from multiple data sources into a consistent structure - usually to support analytics and reporting. Commonly, data from operational systems is extracted, transformed, and loaded into analytical stores such as a data lake or data warehouse.

**Common languages**

Data Engineers must be proficient with a range of tools and scripting languages - in particular SQL and Python, and potentially others.

* **SQL** - One of the most common languages data engineers use is SQL, or Structured Query Language, which is a relatively easy language to learn. SQL uses queries that include SELECT, INSERT, UPDATE, and DELETE statements to directly work with the data stored in tables.
* **Python** - Python is one of the most popular and fastest growing programming languages in the world. It's used for all sorts of tasks, including web programming and data analysis. It has emerged as the language to learn for machine learning, and is increasing in popularity in data engineering with the use of notebooks.
* **Others** - Depending upon the needs of the organization and your individual skill set, you may also use other popular languages within or outside of notebooks including R, Java, Scala, .NET, and more. The use of notebooks is growing in popularity, and allows collaboration using different languages within the same notebook.

Công trình dữ liệu (data engineering) là gì?

Trong lĩnh vực công trình dữ liệu, kỹ sư dữ liệu sẽ thường xuyên làm việc với nhiều loại dữ liệu khác nhau để thực hiện nhiều tác vụ sử dụng nhiều ngôn ngữ kịch bản hoặc lập trình phù hợp với từng tổ chức của họ.

Các loại dữ liệu

Có ba loại dữ liệu chính mà kỹ sư dữ liệu sẽ làm việc:

* Dữ liệu có cấu trúc (Structured data): dữ liệu có cấu trúc chủ yếu đến từ các hệ thống nguồn dựa trên bảng như cơ sở dữ liệu quan hệ hoặc từ một tệp phẳng như tệp dấu phẩy ngăn cách (CSV). Đặc điểm chính của một tệp có cấu trúc là các hàng và cột được căn chỉnh thống nhất trong toàn bộ tệp.
* Dữ liệu bán cấu trúc (Semi-structured data): là dữ liệu như các tệp JavaScript Object Notation (JSON), có thể cần làm phẳng trước khi tải vào hệ thống nguồn của bạn. Khi được làm phẳng, dữ liệu này không cần phải khớp hoàn toàn với cấu trúc bảng.
* Dữ liệu không cấu trúc (Unstructured data): bao gồm dữ liệu được lưu trữ dưới dạng các cặp key-value không tuân theo các mô hình quan hệ tiêu chuẩn và các loại dữ liệu không cấu trúc khác thường được sử dụng bao gồm định dạng dữ liệu di động (PDF), tài liệu của bộ xử lý văn bản và hình ảnh.

Các hoạt động dữ liệu

Là một kỹ sư dữ liệu, một số nhiệm vụ chính mà bạn sẽ thực hiện trong Azure bao gồm tích hợp dữ liệu, chuyển đổi dữ liệu và hợp nhất dữ liệu.

* Tích hợp dữ liệu (Data integration): Tích hợp dữ liệu liên quan đến việc thiết lập các liên kết giữa các dịch vụ và nguồn dữ liệu hoạt động và phân tích để cho phép truy cập an toàn, đáng tin cậy vào dữ liệu trên nhiều hệ thống. Ví dụ, một quy trình kinh doanh có thể dựa trên dữ liệu được phân散 (phân tán) trên nhiều hệ thống và kỹ sư dữ liệu được yêu cầu thiết lập các liên kết để có thể trích xuất dữ liệu cần thiết từ tất cả các hệ thống này.
* Chuyển đổi dữ liệu (Data transformation): Dữ liệu hoạt động thường cần được chuyển đổi thành cấu trúc và định dạng phù hợp để phân tích, thường là một phần của quá trình trích xuất, chuyển đổi và tải (ETL); mặc dù ngày càng có nhiều biến thể trong đó bạn trích xuất, tải và chuyển đổi (ELT) dữ liệu được sử dụng để lấy dữ liệu nhanh chóng vào hồ dữ liệu (data lake) và sau đó áp dụng các kỹ thuật xử lý "dữ liệu lớn" để chuyển đổi nó. Bất kể phương pháp nào được sử dụng, dữ liệu đều được chuẩn bị để hỗ trợ các nhu cầu phân tích phía sau.
* Hợp nhất dữ liệu (Data consolidation): Hợp nhất dữ liệu là quá trình kết hợp dữ liệu đã được trích xuất từ nhiều nguồn dữ liệu khác nhau thành một cấu trúc nhất quán - thường để hỗ trợ phân tích và báo cáo. Thông thường, dữ liệu từ các hệ thống hoạt động được trích xuất, chuyển đổi và tải vào các kho lưu trữ phân tích như hồ dữ liệu (data lake) hoặc kho dữ liệu (data warehouse).

Các ngôn ngữ lập trình thông dụng

Kỹ sư dữ liệu phải thành thạo với một loạt các công cụ và ngôn ngữ kịch bản - đặc biệt là SQL và Python, và có khả năng là các ngôn ngữ khác.

* SQL (Structured Query Language): Một trong những ngôn ngữ phổ biến nhất mà kỹ sư dữ liệu sử dụng là SQL, đây là một ngôn ngữ tương đối dễ học. SQL sử dụng các truy vấn bao gồm các lệnh SELECT, INSERT, UPDATE và DELETE để làm việc trực tiếp với dữ liệu được lưu trữ trong các bảng.
* Python: Python là một trong những ngôn ngữ lập trình phổ biến và phát triển nhanh nhất trên thế giới. Nó được sử dụng cho tất cả các loại tác vụ, bao gồm lập trình web và phân tích dữ liệu. Nó đã nổi lên như ngôn ngữ để học cho học máy và ngày càng phổ biến trong lĩnh vực công trình dữ liệu với việc sử dụng sổ ghi chép (notebook).
* Các ngôn ngữ khác: Tùy thuộc vào nhu cầu của tổ chức và bộ kỹ năng cá nhân của bạn, bạn cũng có thể sử dụng các ngôn ngữ phổ biến khác trong hoặc ngoài sổ ghi chép bao gồm R, Java, Scala, .NET, v.v. Việc sử dụng sổ ghi chép ngày càng phổ biến và cho phép cộng tác bằng các ngôn ngữ khác nhau trong cùng một sổ ghi chép.