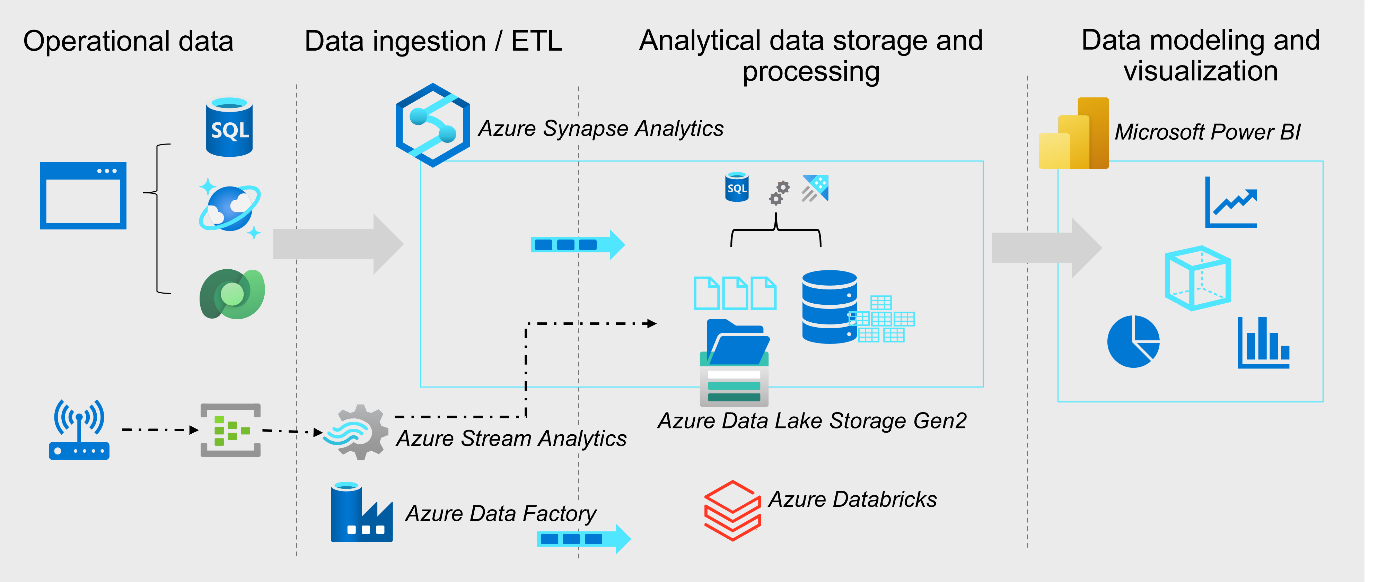
**Data engineering in Microsoft Azure**

Completed100 XP

* 6 minutes



Microsoft Azure includes many services that can be used to implement and manage data engineering workloads.

The diagram displays the flow from left to right of a typical enterprise data analytics solution, including some of the key Azure services that may be used. Operational data is generated by applications and devices and stored in Azure data storage services such as Azure SQL Database, Azure Cosmos DB, and Microsoft Dataverse. Streaming data is captured in event broker services such as Azure Event Hubs.

This operational data must be captured, ingested, and consolidated into analytical stores; from where it can be modeled and visualized in reports and dashboards. These tasks represent the core area of responsibility for the data engineer. The core Azure technologies used to implement data engineering workloads include:

* Azure Synapse Analytics
* Azure Data Lake Storage Gen2
* Azure Stream Analytics
* Azure Data Factory
* Azure Databricks

The analytical data stores that are populated with data produced by data engineering workloads support data modeling and visualization for reporting and analysis, often using sophisticated visualization tools such as Microsoft Power BI.

Microsoft Azure cung cấp nhiều dịch vụ có thể được sử dụng để triển khai và quản lý các khối lượng công việc về công trình dữ liệu (data engineering).

Biểu đồ cho thấy luồng dữ liệu từ trái sang phải của một giải pháp phân tích dữ liệu doanh nghiệp điển hình, bao gồm một số dịch vụ Azure chính có thể được sử dụng. Dữ liệu vận hành được tạo bởi các ứng dụng và thiết bị và được lưu trữ trong các dịch vụ lưu trữ dữ liệu Azure như Azure SQL Database, Azure Cosmos DB và Microsoft Dataverse. Dữ liệu streaming (luồng dữ liệu) được thu thập trong các dịch vụ môi giới sự kiện như Azure Event Hubs.

Dữ liệu vận hành này cần được thu thập, đưa vào và hợp nhất thành các kho lưu trữ phân tích; từ đó nó có thể được mô hình hóa và trực quan hóa trong các báo cáo và bảng điều khiển. Các nhiệm vụ này đại diện cho lĩnh vực trách nhiệm cốt lõi của kỹ sư dữ liệu. Các công nghệ cốt lõi của Azure được sử dụng để triển khai các khối lượng công việc về công trình dữ liệu bao gồm:

* **Azure Synapse Analytics:** Là một dịch vụ phân tích dữ liệu doanh nghiệp kết hợp kho dữ liệu doanh nghiệp và các tính năng phân tích dữ liệu lớn. Nó cho phép bạn lưu trữ, quản lý và phân tích một lượng lớn dữ liệu có cấu trúc, bán cấu trúc và không cấu trúc từ nhiều nguồn khác nhau.
* **Azure Data Lake Storage Gen2:** Là một giải pháp lưu trữ hồ dữ liệu highly scalable (có khả năng mở rộng cao) và an toàn cho dữ liệu lớn của bạn. Bạn có thể lưu trữ bất kỳ loại dữ liệu nào ở bất kỳ quy mô nào trong hồ dữ liệu của mình, bao gồm dữ liệu thô từ cảm biến, log ứng dụng, phương tiện xã hội, tệp văn bản, hình ảnh, video và các loại dữ liệu khác.
* **Azure Stream Analytics:** Là một dịch vụ thực hiện phân tích thời gian thực trên các luồng dữ liệu liên tục. Bạn có thể sử dụng Azure Stream Analytics để xử lý và phân tích dữ liệu theo thời gian thực khi nó được tạo ra. Điều này cho phép bạn phản ứng với các sự kiện theo thời gian thực và thực hiện các hành động ngay lập tức.
* **Azure Data Factory:** Là dịch vụ tích hợp dữ liệu đám mây cho phép bạn tạo, lập lịch và quản lý các quy trình di chuyển dữ liệu giữa các kho lưu trữ dữ liệu khác nhau. Bạn có thể sử dụng Azure Data Factory để tự động hóa việc di chuyển dữ liệu giữa các nguồn dữ liệu khác nhau, bao gồm các nguồn dữ liệu cục bộ, nguồn dữ liệu đám mây và các nguồn dữ liệu SaaS (phần mềm dưới dạng dịch vụ).
* **Azure Databricks:** Là một nền tảng phân tích dữ liệu mở rộng (open-source analytics platform) cung cấp các cụm Apache Spark được quản lý hoàn toàn. Bạn có thể sử dụng Azure Databricks để xử lý và phân tích lượng lớn dữ liệu có cấu trúc, bán cấu trúc và không cấu trúc.

Các kho lưu trữ dữ liệu phân tích được lấp đầy bằng dữ liệu do các khối lượng công việc về công trình dữ liệu tạo ra hỗ trợ mô hình hóa dữ liệu và trực quan hóa cho báo cáo và phân tích, thường sử dụng các công cụ trực quan hóa tinh vi như Microsoft Power BI.

**Tóm tắt:**

* Microsoft Azure cung cấp nhiều dịch vụ để hỗ trợ các công việc về công trình dữ liệu.
* Dữ liệu vận hành được tạo ra bởi các ứng dụng và thiết bị được lưu trữ trong các dịch vụ lưu trữ dữ liệu của Azure.
* Dữ liệu streaming được thu thập trong các dịch vụ môi giới sự kiện.
* Công việc chính của kỹ sư dữ liệu là thu thập, đưa vào và hợp nhất dữ liệu vận hành thành các kho lưu trữ phân tích để phân tích và báo cáo.
* Microsoft Azure cung cấp nhiều dịch vụ để thực hiện các tác vụ này, chẳng hạn như Azure Synapse Analytics, Azure Data Lake Storage Gen2, Azure Stream Analytics, Azure Data Factory và Azure Databricks.
* Dữ liệu phân tích sau đó được sử dụng để tạo báo cáo và bảng điều khiển bằng các công cụ trực quan hóa như Microsoft Power BI.