**Understand the stages for processing big data**

Completed100 XP

* 5 minutes

Data lakes have a fundamental role in a wide range of big data architectures. These architectures can involve the creation of:

* An enterprise data warehouse.
* Advanced analytics against big data.
* A real-time analytical solution.

There are four stages for processing big data solutions that are common to all architectures:

* **Ingest** - The ingestion phase identifies the technology and processes that are used to acquire the source data. This data can come from files, logs, and other types of unstructured data that must be put into the data lake. The technology that is used will vary depending on the frequency that the data is transferred. For example, for batch movement of data, pipelines in Azure Synapse Analytics or Azure Data Factory may be the most appropriate technology to use. For real-time ingestion of data, Apache Kafka for HDInsight or Stream Analytics may be an appropriate choice.
* **Store** - The store phase identifies where the ingested data should be placed. Azure Data Lake Storage Gen2 provides a secure and scalable storage solution that is compatible with commonly used big data processing technologies.
* **Prep and train** - The prep and train phase identifies the technologies that are used to perform data preparation and model training and scoring for machine learning solutions. Common technologies that are used in this phase are Azure Synapse Analytics, Azure Databricks, Azure HDInsight, and Azure Machine Learning.
* **Model and serve** - Finally, the model and serve phase involves the technologies that will present the data to users. These technologies can include visualization tools such as Microsoft Power BI, or analytical data stores such as Azure Synapse Analytics. Often, a combination of multiple technologies will be used depending on the business requirements.

**Các giai đoạn xử lý Big Data**

Hồ dữ liệu đóng vai trò nền tảng cho nhiều kiến trúc Big Data khác nhau. Kiến trúc này có thể bao gồm việc xây dựng:

* Kho dữ liệu doanh nghiệp (Enterprise Data Warehouse)
* Phân tích nâng cao trên dữ liệu lớn (Advanced Analytics against Big Data)
* Giải pháp phân tích thời gian thực (Real-time analytical solution)

Có bốn giai đoạn xử lý dữ liệu lớn phổ biến áp dụng cho tất cả các kiến trúc:

**1. Thu thập dữ liệu (Ingest)**

Giai đoạn thu thập xác định công nghệ và quy trình được sử dụng để lấy dữ liệu nguồn. Dữ liệu này có thể đến từ các file, log và các loại dữ liệu phi cấu trúc khác cần được đưa vào hồ dữ liệu. Công nghệ được sử dụng sẽ thay đổi tùy thuộc vào tần suất truyền dữ liệu. Ví dụ:

* Đối với việc di chuyển dữ liệu theo batch (xử lý hàng loạt), các pipeline trong Azure Synapse Analytics hoặc Azure Data Factory có thể là công nghệ phù hợp nhất.
* Đối với việc thu thập dữ liệu theo thời gian thực, Apache Kafka for HDInsight hoặc Stream Analytics có thể là lựa chọn phù hợp.

**2. Lưu trữ (Store)**

Giai đoạn lưu trữ xác định vị trí lưu trữ dữ liệu đã được thu thập. Azure Data Lake Storage Gen2 cung cấp giải pháp lưu trữ an toàn, có khả năng mở rộng và tương thích với các công nghệ xử lý dữ liệu lớn thông dụng.

**3. Chuẩn bị và huấn luyện (Prep and Train)**

Giai đoạn chuẩn bị và huấn luyện xác định các công nghệ được sử dụng để thực hiện chuẩn bị dữ liệu, đào tạo và chấm điểm mô hình cho các giải pháp học máy. Các công nghệ thông dụng được sử dụng trong giai đoạn này là Azure Synapse Analytics, Azure Databricks, Azure HDInsight và Azure Machine Learning.

**4. Cung cấp mô hình (Model and Serve)**

Cuối cùng, giai đoạn cung cấp mô hình liên quan đến các công nghệ sẽ hiển thị dữ liệu cho người dùng. Các công nghệ này có thể bao gồm các công cụ trực quan hóa như Microsoft Power BI hoặc kho dữ liệu phân tích như Azure Synapse Analytics. Thông thường, tùy thuộc vào yêu cầu kinh doanh, có thể sử dụng kết hợp nhiều công nghệ khác nhau.