**Azure Data Factory và Azure Data Lake**

**Azure Data Factory**

* Là dịch vụ có khả năng thu thập khối lượng lớn dữ liệu thô, không tổ chức từ các hệ thống quan hệ và phi quan hệ.
* Chuyển đổi dữ liệu này thành thông tin có ý nghĩa.
* Cung cấp công cụ ETL (Extract, Transform, Load - Trích xuất, Chuyển đổi, Nạp) và ELT (Extract, Load, Transform - Trích xuất, Nạp, Chuyển đổi) có thể mở rộng và lập trình được.
* Được sử dụng để triển khai các dự án tích hợp dữ liệu.
* Được mô tả như một dịch vụ tích hợp dữ liệu.

**Mục đích của Azure Data Factory:**

* Lấy dữ liệu từ một hoặc nhiều nguồn dữ liệu.
* Chuyển đổi dữ liệu thành định dạng phù hợp để xử lý.

**Ví dụ:**

* Một công ty game thu thập hàng petabyte nhật ký trò chơi được tạo bởi các trò chơi trong code.
* Công ty muốn phân tích các nhật ký này để có được thông tin chi tiết về sở thích, nhân khẩu học và hành vi sử dụng của khách hàng.
* Công ty cũng muốn xác định các cơ hội bán thêm (upsell) và bán chéo (cross-sell), phát triển các tính năng mới hấp dẫn, thúc đẩy tăng trưởng kinh doanh và cung cấp trải nghiệm tốt hơn cho khách hàng.
* Để phân tích các nhật ký này, công ty cần sử dụng dữ liệu tham chiếu, chẳng hạn như thông tin khách hàng, thông tin trò chơi và thông tin chiến dịch marketing, nằm trong kho lưu trữ dữ liệu cục bộ.
* Công ty muốn tận dụng dữ liệu này từ kho lưu trữ dữ liệu cục bộ, kết hợp với dữ liệu nhật ký bổ sung có trong kho lưu trữ dữ liệu đám mây.
* Để trích xuất thông tin chi tiết, công ty muốn xử lý dữ liệu được ghép nối bằng cách sử dụng cụm Spark trên đám mây (sử dụng Azure HDInsight) và xuất bản dữ liệu đã chuyển đổi vào kho dữ liệu đám mây, chẳng hạn như Azure Synapse Analytics.
* Công ty có thể sử dụng thông tin trong kho dữ liệu để tạo và xuất bản báo cáo.

**Azure Data Factory có thể:**

* Tạo và lên lịch các luồng công việc theo hướng dữ liệu được gọi là pipelines, có thể thu thập dữ liệu từ các kho lưu trữ dữ liệu khác nhau được công ty game sử dụng.
* Xây dựng các quy trình ETL phức tạp để chuyển đổi dữ liệu trực quan với các luồng dữ liệu (dataflows) hoặc bằng cách sử dụng các dịch vụ điện toán như Azure HDInsight, Azure Databricks và Azure SQL Database.
* Xuất bản dữ liệu đã chuyển đổi sang Azure Synapse Analytics để các ứng dụng Business Intelligence (BI) sử dụng.

**Pipeline là gì?**

* Một nhóm logic các hoạt động thực hiện một đơn vị công việc.
* Các hoạt động trong một pipeline cùng nhau thực hiện một tác vụ.

**Ví dụ:**

* Một pipeline có thể chứa một loạt các hoạt động thu thập dữ liệu thô từ Azure Blob Storage, sau đó chạy truy vấn Hive trên cụm HDInsight để phân vùng dữ liệu và lưu trữ kết quả trong cơ sở dữ liệu Cosmos DB.

**Azure Data Lake**

* Là một bộ sưu tập các dịch vụ phân tích và lưu trữ mà bạn có thể kết hợp để triển khai giải pháp dữ liệu lớn.
* Gồm ba yếu tố chính: Data Lake Store, Data Lake Analytics và HDInsight.

**Azure Data Lake Store:**

* Cung cấp một hệ thống tập tin có thể lưu trữ lượng dữ liệu gần như không giới hạn.
* Sử dụng tổ chức phân cấp, giống như hệ thống tập tin Windows và Linux, nhưng bạn có thể lưu trữ một lượng lớn dữ liệu thô, đối tượng blob và dữ liệu có cấu trúc.
* Được tối ưu hóa cho khối lượng công việc phân tích.
* Tương thích với Hệ thống tập tin phân tán Hadoop (Hadoop Distributed File System - HDFS).
* Bạn có thể chạy các tác vụ Hadoop bằng cách sử dụng Azure HDInsight, có thể đọc và ghi dữ liệu trong Data Lake Store một cách hiệu quả.
* Cung cấp bảo mật chi tiết cho dữ liệu bằng cách sử dụng danh sách kiểm soát truy cập (ACL).

**Azure Data Lake Analytics:**

* Là dịch vụ công việc phân tích theo yêu cầu mà bạn có thể sử dụng để xử lý dữ liệu lớn.
* Cung cấp một khung và bộ công cụ để bạn sử dụng để phân tích dữ liệu được lưu trữ trong Microsoft Azure Data Lake Store và các kho lưu trữ khác.
* Bạn viết các tác vụ chứa các truy vấn để chuyển đổi dữ liệu và trích xuất thông tin chi tiết.