**Giới thiệu chi tiết:**

**1.1. Doanh nghiệp tạo ra dữ liệu trong suốt quá trình hoạt động:**

* Dữ liệu được tạo ra từ nhiều nguồn khác nhau trong doanh nghiệp, bao gồm:
  + Giao dịch bán hàng
  + Hoạt động của khách hàng
  + Marketing
  + Tài chính
  + Nhân sự
  + V.v.
* Dữ liệu có thể ở nhiều dạng khác nhau, bao gồm:
  + Cấu trúc (structured): dữ liệu được tổ chức theo một định dạng cụ thể, ví dụ như bảng trong cơ sở dữ liệu.
  + Bán cấu trúc (semi-structured): dữ liệu có một số cấu trúc, nhưng không hoàn toàn, ví dụ như dữ liệu XML.
  + Không cấu trúc (unstructured): dữ liệu không có cấu trúc, ví dụ như văn bản, hình ảnh, video.

**1.2. Dữ liệu này có thể được lưu trữ ở dạng thô (ingested) hoặc qua xử lý để lưu vào kho lưu trữ dữ liệu (data store) hoặc kho dữ liệu (data warehouse):**

* Dữ liệu thô (ingested data): là dữ liệu được lưu trữ ở dạng ban đầu, chưa qua xử lý.
* Kho lưu trữ dữ liệu (data store): là nơi lưu trữ dữ liệu cho một mục đích cụ thể, ví dụ như hệ thống quản lý quan hệ khách hàng (CRM) hoặc hệ thống quản lý kho hàng (WMS).
* Kho dữ liệu (data warehouse): là nơi lưu trữ dữ liệu cho mục đích phân tích. Dữ liệu trong kho dữ liệu thường được tích hợp từ nhiều nguồn khác nhau và được xử lý để phù hợp cho việc phân tích.

**1.3. Microsoft Azure cho phép doanh nghiệp thực hiện cả hai kịch bản:**

* Azure cung cấp nhiều dịch vụ cho phép doanh nghiệp lưu trữ dữ liệu thô, ví dụ như Azure Blob Storage và Azure Data Lake Storage.
* Azure cũng cung cấp nhiều dịch vụ cho phép doanh nghiệp xử lý dữ liệu và lưu trữ vào kho lưu trữ dữ liệu hoặc kho dữ liệu, ví dụ như Azure Synapse Analytics, Azure Data Factory và Azure Databricks.

**Ví dụ:**

* Một công ty bán lẻ có thể lưu trữ dữ liệu giao dịch bán hàng ở dạng thô trong Azure Blob Storage.
* Công ty này cũng có thể sử dụng Azure Synapse Analytics để xử lý dữ liệu giao dịch bán hàng và lưu trữ vào kho dữ liệu.
* Sau đó, công ty có thể sử dụng kho dữ liệu để phân tích xu hướng bán hàng và đưa ra quyết định kinh doanh tốt hơn.

**2.1. Azure Synapse Analytics:**

* Dịch vụ phân tích tổng hợp, cho phép:
  + Đọc dữ liệu từ nhiều nguồn khác nhau (Azure Blob Storage, Azure Data Lake Storage, Hadoop storage, databases, data warehouses, v.v.).
  + Xử lý dữ liệu bằng Transact-SQL (T-SQL) hoặc Apache Spark.
  + Tạo các phân tích và mô hình khác nhau (machine learning, statistical analysis, linear and non-linear modeling, predictive analytics, graphics).
  + Lưu kết quả vào kho lưu trữ dữ liệu hoặc kho dữ liệu.
* Lợi ích:
  + Mạnh mẽ và linh hoạt.
  + Khả năng mở rộng cao.
  + Dễ sử dụng.

**2.2. Công nghệ xử lý dữ liệu:**

**2.2.1. Transact-SQL (T-SQL):**

* Tương tự SQL dùng trong Azure SQL Database, nhưng có thêm các tiện ích mở rộng để đọc dữ liệu từ các nguồn ngoài.
* Lợi ích:
  + Hiệu quả cao cho các truy vấn dữ liệu có cấu trúc.
  + Dễ học và sử dụng.

**2.2.2. Apache Spark:**

* Nền tảng mã nguồn mở dùng để xử lý dữ liệu song song.
* Lợi ích:
  + Khả năng xử lý dữ liệu nhanh chóng và hiệu quả.
  + Hỗ trợ nhiều ngôn ngữ lập trình (Python, Scala, Java, SQL).
  + Thư viện phong phú cho các tác vụ phân tích dữ liệu.

**2.3. Azure Data Bricks:**

* Nền tảng phân tích tối ưu cho dịch vụ đám mây Microsoft Azure, được thiết kế với những người sáng lập Apache Spark.
* Lợi ích:
  + Cài đặt đơn giản chỉ với một cú nhấp chuột.
  + Quy trình công việc được sắp xếp hợp lý.
  + Không gian làm việc tương tác.
  + Hỗ trợ nhiều ngôn ngữ lập trình (Python, Scala, Java, SQL).

**2.4. Azure HDInsight:**

* Dịch vụ phân tích được quản lý trên đám mây.
* Dựa trên Apache Hadoop, bộ công cụ mã nguồn mở cho phép chạy tác vụ xử lý trên lượng dữ liệu lớn.
* Lợi ích:
  + Dễ sử dụng và quản lý.
  + Khả năng mở rộng cao.
  + Hỗ trợ nhiều frameworks cho các tác vụ phân tích dữ liệu (Hadoop MapReduce, Apache Spark, Apache Hive, Apache Kafka, Apache Storm, R, v.v.).