**Azure Synapse Analytics**

**Azure Synapse Analytics** cung cấp một bộ công cụ để phân tích và xử lý dữ liệu của tổ chức. Nó kết hợp các công nghệ SQL, khả năng truy vấn Transact-SQL và các công cụ Spark mã nguồn mở để cho phép bạn xử lý nhanh chóng lượng dữ liệu rất lớn. Bây giờ, chúng ta sẽ xem xét kỹ hơn các tính năng của Synapse Analytics và thời điểm bạn nên cân nhắc sử dụng nó.

**Tổng quan về Azure Synapse Analytics**

Azure Synapse Analytics là dịch vụ phân tích tích hợp cho phép tổ chức nhanh chóng thu thập thông tin chi tiết từ tất cả dữ liệu của họ ở mọi quy mô siêu lớn, từ cả kho dữ liệu và hệ thống phân tích dữ liệu lớn. Synapse của bạn được cấu thành từ các yếu tố sau:

* **Synapse SQL Pool:** Một tập hợp các máy chủ chạy Transact-SQL, đây là phương ngữ SQL được sử dụng bởi Azure SQL Database và Microsoft SQL Server. Bạn viết logic xử lý dữ liệu của mình bằng Transact-SQL.
* **Synapse Spark Pool:** Một cụm máy chủ chạy Apache Spark để xử lý dữ liệu. Bạn viết logic xử lý dữ liệu của mình bằng một trong bốn ngôn ngữ được hỗ trợ: Python, Scala, SQL và C Sharp thông qua .NET for Apache Spark. Spark Pool hỗ trợ Azure Machine Learning thông qua tích hợp với các gói Spark ML và Azure ML.
* **Synapse Pipelines:** Một nhóm logic các hoạt động cùng nhau thực hiện một tác vụ. Các hoạt động trong pipeline xác định các hành động cần thực hiện trên dữ liệu của bạn. Ví dụ: bạn có thể sử dụng hoạt động sao chép để chuyển đổi dữ liệu từ tập dữ liệu nguồn sang tập dữ liệu đích. Bạn có thể bao gồm các hoạt động chuyển đổi dữ liệu khi nó được truyền hoặc bạn có thể kết hợp dữ liệu từ nhiều nguồn khác nhau.
* **Synapse Link:** Một thành phần cho phép bạn kết nối với Cosmos DB. Bạn có thể sử dụng nó để thực hiện phân tích gần thời gian thực trên dữ liệu hoạt động được lưu trữ trong cơ sở dữ liệu Cosmos DB.
* **Synapse Studio:** Giao diện người dùng web cho phép kỹ sư dữ liệu truy cập tất cả các công cụ Synapse Analytics. Bạn có thể sử dụng Synapse Studio để tạo pool SQL và Spark, xác định và chạy pipeline và cấu hình liên kết đến các nguồn dữ liệu bên ngoài.

**Lưu ý:** Bất kỳ dữ liệu nào được lưu trữ trong Azure Synapse Analytics đều có thể được sử dụng để xây dựng và huấn luyện mô hình với Azure Machine Learning. Các phần sau đây mô tả chi tiết từng yếu tố này.

**Synapse SQL Pool**

Khi bạn sử dụng Synapse, khối lượng công việc phân tích của bạn chạy bằng một pool SQL. Trong pool Sequel, các nút điều khiển và nút tính toán trong cụm chạy một phiên bản của Azure SQL Database hỗ trợ các truy vấn phân tán. Bạn xác định logic của mình bằng các câu lệnh Transact-SQL. Bạn gửi các câu lệnh Transact-SQL của mình đến nút điều khiển, nút này sẽ chia công việc thành các truy vấn hoạt động trên một tập hợp con của dữ liệu và sau đó gửi các truy vấn nhỏ hơn này đến các nút tính toán. Dữ liệu được chia thành các phần được gọi là phân phối. Phân phối là đơn vị lưu trữ và xử lý cơ bản cho các truy vấn song song chạy trên dữ liệu phân tán. Mỗi truy vấn nhỏ hơn chạy trên một trong các phân phối dữ liệu. Nút điều khiển và nút tính toán sử dụng Dịch vụ di chuyển dữ liệu (DMS) để di chuyển dữ liệu giữa các nút khi cần thiết, để chạy các truy vấn song song và trả về kết quả chính xác.

Synapse Analytics sử dụng một công nghệ gọi là Polybase để làm cho dữ liệu bên ngoài trông giống như các bảng SQL. Bạn có thể chạy các truy vấn trên các bảng này trực tiếp hoặc bạn có thể chuyển dữ liệu thành một loạt các bảng SQL do Synapse Analytics quản lý để truy vấn sau này. Synapse sử dụng Azure Storage để quản lý dữ liệu của bạn trong khi nó đang được xử lý.

**Cấu hình Synapse SQL Pool**

Theo mặc định, một pool SQL theo yêu cầu được tạo trong mỗi không gian làm việc của Azure Synapse Analytics. Sau đó, bạn có thể cung cấp thêm các pool theo yêu cầu hoặc được cung cấp sẵn. Điều quan trọng cần lưu ý là các pool theo yêu cầu chỉ cho phép bạn truy vấn dữ liệu được lưu trữ trong các tệp bên ngoài. Nếu bạn muốn nhập và tải dữ liệu vào Synapse Analytics, bạn phải tạo pool SQL của riêng mình.

Azure Synapse Analytics được thiết kế để chạy các truy vấn trên các tập dữ liệu khổng lồ. Bạn có thể tự động thu phóng pool SQL lên đến