**Giới thiệu**

Nhiều tổ chức đã dành hai thập kỷ qua để xây dựng các kho dữ liệu và giải pháp trí tuệ kinh doanh (BI) dựa trên hệ thống cơ sở dữ liệu quan hệ. Nhiều giải pháp BI đã bỏ lỡ các cơ hội để lưu trữ dữ liệu phi cấu trúc do chi phí và tính phức tạp trong các loại dữ liệu này trong cơ sở dữ liệu.

**Hồ dữ liệu** đã trở thành một giải pháp phổ biến cho vấn đề này. Hồ dữ liệu cung cấp lưu trữ dựa trên file, thường nằm trong hệ thống file phân tán hỗ trợ khả năng mở rộng cao cho khối lượng dữ liệu lớn. Các tổ chức có thể lưu trữ các file có cấu trúc, bán cấu trúc và không cấu trúc trong hồ dữ liệu và sau đó sử dụng chúng từ đó trong các công nghệ xử lý dữ liệu lớn, chẳng hạn như Apache Spark.

**Azure Data Lake Storage Gen2** cung cấp một giải pháp dựa trên đám mây cho lưu trữ hồ dữ liệu trong Microsoft Azure và hỗ trợ nhiều giải pháp phân tích quy mô lớn được xây dựng trên Azure.

**Tóm lược nội dung**

Bài viết này giới thiệu về Azure Data Lake Storage Gen2, một giải pháp lưu trữ dữ liệu hồ dựa trên đám mây trong Microsoft Azure. Hồ dữ liệu là một giải pháp lưu trữ cho phép các tổ chức lưu trữ khối lượng lớn dữ liệu có cấu trúc, bán cấu trúc và không cấu trúc. Azure Data Lake Storage Gen2 được sử dụng để hỗ trợ các giải pháp phân tích dữ liệu lớn được xây dựng trên Azure.

**Hệ thống tập tin phân tán (Distributed File System)**

Hệ thống tập tin phân tán (Distributed File System - DFS) là một hệ thống cho phép truy cập dữ liệu từ nhiều máy tính (host) chia sẻ dữ liệu với nhau thông qua một mạng máy tính. Dữ liệu được lưu trữ trên các thiết bị lưu trữ khác nhau và được quản lý bởi một hệ thống thống nhất, cung cấp cho người dùng một cái nhìn tổng quan về tất cả dữ liệu như thể nó được lưu trữ trên một máy tính duy nhất.