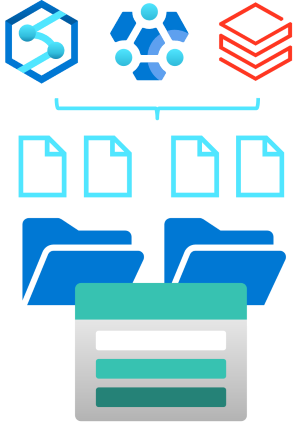
**Understand Azure Data Lake Storage Gen2**

Completed100 XP

* 5 minutes

A data lake is a repository of data that is stored in its natural format, usually as blobs or files. Azure Data Lake Storage is a comprehensive, massively scalable, secure, and cost-effective data lake solution for high performance analytics built into Azure.



Azure Data Lake Storage combines a file system with a storage platform to help you quickly identify insights into your data. Data Lake Storage builds on Azure Blob storage capabilities to optimize it specifically for analytics workloads. This integration enables analytics performance, the tiering and data lifecycle management capabilities of Blob storage, and the high-availability, security, and durability capabilities of Azure Storage.

**Benefits**

Data Lake Storage is designed to deal with this variety and volume of data at exabyte scale while securely handling hundreds of gigabytes of throughput. With this, you can use Data Lake Storage Gen2 as the basis for both real-time and batch solutions.

**Hadoop compatible access**

A benefit of Data Lake Storage is that you can treat the data as if it's stored in a Hadoop Distributed File System (HDFS). With this feature, you can store the data in one place and access it through compute technologies including Azure Databricks, Azure HDInsight, and Azure Synapse Analytics without moving the data between environments. The data engineer also has the ability to use storage mechanisms such as the parquet format, which is highly compressed and performs well across multiple platforms using an internal columnar storage.

**Security**

Data Lake Storage supports access control lists (ACLs) and Portable Operating System Interface (POSIX) permissions that don't inherit the permissions of the parent directory. In fact, you can set permissions at a directory level or file level for the data stored within the data lake, providing a much more secure storage system. This security is configurable through technologies such as Hive and Spark or utilities such as Azure Storage Explorer, which runs on Windows, macOS, and Linux. All data that is stored is encrypted at rest by using either Microsoft or customer-managed keys.

**Performance**

Azure Data Lake Storage organizes the stored data into a hierarchy of directories and subdirectories, much like a file system, for easier navigation. As a result, data processing requires less computational resources, reducing both the time and cost.

**Data redundancy**

Data Lake Storage takes advantage of the Azure Blob replication models that provide data redundancy in a single data center with locally redundant storage (LRS), or to a secondary region by using the Geo-redundant storage (GRS) option. This feature ensures that your data is always available and protected if catastrophe strikes.

**Tip**

Whenever planning for a data lake, a data engineer should give thoughtful consideration to structure, data governance, and security. This should include consideration of factors that can influence lake structure and organization, such as:

* Types of data to be stored
* How the data will be transformed
* Who should access the data
* What are the typical access patterns

This approach will help determine how to plan for access control governance across your lake. Data engineers should be proactive in ensuring that the lake doesn't become the proverbial data swamp which becomes inaccessible and non-useful to users due to the lack of data governance and data quality measures. Establishing a baseline and following best practices for Azure Data Lake will help ensure a proper and robust implementation that will allow the organization to grow and gain insight to achieve more.

**Hiểu về Azure Data Lake Storage Gen2**

Azure Data Lake Storage Gen2 là một giải pháp hồ dữ liệu toàn diện, có khả năng mở rộng lớn, an toàn và tiết kiệm chi phí, được xây dựng trên nền tảng Microsoft Azure để phục vụ phân tích hiệu suất cao.

**Azure Data Lake Storage Gen2 là gì?**

Data Lake (hồ dữ liệu) là một kho lưu trữ dữ liệu ở định dạng gốc của nó, thường là các đối tượng blob hoặc file. Azure Data Lake Storage Gen2 kết hợp hệ thống file với nền tảng lưu trữ để giúp bạn nhanh chóng xác định insights (sự kiện) từ dữ liệu của mình. Data Lake Storage Gen2 xây dựng trên nền tảng lưu trữ Azure Blob nhằm tối ưu hóa cho các khối lượng công việc phân tích. Sự tích hợp này cho phép đạt được hiệu suất phân tích, khả năng phân tầng và quản lý vòng đời dữ liệu của Blob storage, đồng thời đảm bảo tính sẵn sàng cao, bảo mật và độ bền của Azure Storage.

**Lợi ích**

Data Lake Storage Gen2 được thiết kế để xử lý lượng dữ liệu đa dạng và khổng lồ lên đến exabyte (exa bytu) trong khi vẫn bảo mật an toàn cho hàng trăm gigabyte dữ liệu đầu ra. Do đó, bạn có thể sử dụng Data Lake Storage Gen2 làm nền tảng cho cả các giải pháp thời gian thực và theo batch (xử lý hàng loạt).

**Truy cập tương thích với Hadoop**

Một lợi ích của Data Lake Storage Gen2 là bạn có thể xử lý dữ liệu như thể nó được lưu trữ trong Hệ thống tập tin phân tán Hadoop (HDFS). Với tính năng này, bạn có thể lưu trữ dữ liệu ở một vị trí và truy cập nó thông qua các công nghệ tính toán như Azure Databricks, Azure HDInsight và Azure Synapse Analytics mà không cần di chuyển dữ liệu giữa các môi trường. Kỹ sư dữ liệu cũng có khả năng sử dụng các cơ chế lưu trữ như định dạng parquet, có khả năng nén cao và hoạt động tốt trên nhiều nền tảng khác nhau bằng cách sử dụng lưu trữ cột bên trong.

**Bảo mật**

Data Lake Storage Gen2 hỗ trợ danh sách kiểm soát truy cập (ACL) và quyền Portable Operating System Interface (POSIX) không kế thừa quyền của thư mục gốc. Trên thực tế, bạn có thể đặt quyền ở cấp thư mục hoặc cấp file cho dữ liệu được lưu trữ trong hồ dữ liệu, cung cấp hệ thống lưu trữ an toàn hơn nhiều. Bảo mật này được cấu hình thông qua các công nghệ như Hive và Spark hoặc các tiện ích như Azure Storage Explorer, chạy trên Windows, macOS và Linux. Tất cả dữ liệu được lưu trữ đều được mã hóa ở trạng thái nghỉ bằng cách sử dụng khóa được quản lý bởi Microsoft hoặc khách hàng.

**Hiệu suất**

Data Lake Storage Gen2 tổ chức dữ liệu được lưu trữ thành một hệ thống phân cấp gồm các thư mục và thư mục con, giống như một hệ thống file, để dễ dàng điều hướng hơn. Do đó, việc xử lý dữ liệu đòi hỏi ít tài nguyên tính toán hơn, giúp giảm thiểu thời gian và chi phí.

**Dự phòng dữ liệu**

Data Lake Storage Gen2 tận dụng các mô hình sao chép Azure Blob cung cấp sự dự phòng dữ liệu trong một trung tâm dữ liệu duy nhất với lưu trữ dự phòng cục bộ (LRS) hoặc sao chép sang vùng phụ bằng cách sử dụng tùy chọn lưu trữ dự phòng địa lý (GRS). Tính năng này đảm bảo dữ liệu của bạn luôn sẵn sàng và được bảo vệ trong trường hợp thảm họa xảy ra.

**Mẹo**

Bất cứ khi nào lên kế hoạch cho một hồ dữ liệu, kỹ sư dữ liệu nên cân nhắc kỹ lưỡng về cấu trúc, quản trị dữ liệu và bảo mật. Điều này bao gồm xem xét các yếu tố có thể ảnh hưởng đến cấu trúc và tổ chức của hồ, chẳng hạn như:

* Các loại dữ liệu được lưu trữ
* Dữ liệu sẽ được chuyển đổi như thế nào
* Ai nên truy cập dữ liệu
* Các mẫu truy cập điển hình là gì

Cách tiếp cận này sẽ giúp xác định cách lên kế hoạch cho quản trị kiểm soát quyền truy cập trên toàn bộ hồ dữ liệu của bạn. Kỹ sư dữ liệu nên chủ động đảm bảo rằng hồ dữ liệu không trở thành một "vũng lầy dữ liệu" theo nghĩa đen, nghĩa là không thể truy cập và không hữu ích cho người dùng do thiếu quản trị dữ liệu và các biện pháp chất lượng dữ liệu. Thiết lập đường cơ sở và tuân theo các quy trình thực hành tốt nhất cho Azure Data Lake sẽ giúp đảm bảo triển khai phù hợp và mạnh mẽ, cho phép tổ chức phát triển và đạt được insights để đạt được nhiều thành tựu hơn.

**Danh sách kiểm soát truy cập (ACL) và quyền Portable Operating System Interface (POSIX) không kế thừa quyền của thư mục gốc**

**Danh sách kiểm soát truy cập (ACL)** là một cơ chế bảo mật cho phép bạn kiểm soát ai có thể truy cập và thực hiện các hành động trên một đối tượng (file, thư mục, thiết bị, v.v.). ACL cung cấp khả năng kiểm soát chi tiết hơn so với các quyền truy cập cơ bản như "đọc", "ghi" và "thực thi".

**Quyền Portable Operating System Interface (POSIX)** là một tập hợp các tiêu chuẩn định nghĩa cách các hệ thống hoạt động tương tác với người dùng, ứng dụng và phần cứng. Quyền POSIX bao gồm các quyền truy cập cơ bản như "đọc", "ghi", "thực thi" và "chủ sở hữu".

**Tính năng không kế thừa quyền của thư mục gốc** cho phép bạn đặt quyền riêng cho từng file và thư mục trong hệ thống, bất kể quyền truy cập của thư mục gốc. Điều này giúp tăng cường bảo mật và linh hoạt trong việc quản lý quyền truy cập.

**Ví dụ:**

Giả sử bạn có một thư mục /data với quyền truy cập "đọc" và "ghi" cho tất cả người dùng. Bên trong thư mục /data, bạn có một file secret.txt mà bạn chỉ muốn cho phép người dùng admin truy cập.

**Với ACL**, bạn có thể đặt quyền truy cập "đọc" và "ghi" cho file secret.txt chỉ dành cho người dùng admin, ngay cả khi thư mục /data cho phép truy cập cho tất cả người dùng.

**Với quyền POSIX**, bạn có thể thay đổi chủ sở hữu của file secret.txt thành admin và cấp quyền truy cập "đọc" và "ghi" cho chủ sở hữu. Điều này sẽ ngăn cản những người dùng khác truy cập file, ngay cả khi họ có quyền truy cập vào thư mục /data.