Azure Data Lake Store Gen2 vs Azure Storage Account

Azure Data Lake Store Gen2 is a massively scalable, secure data lake functionality built on Azure Blob Storage which is designed for big data analytics and offers a hierarchical file system.

Azure Blob Storage

Azure Blob Storage is an object storage solution for the cloud. It is optimized for storing massive amounts of unstructured data, such as text or binary data. You can store large amounts of unstructured data (no restrictions on the kinds of data it can hold), in a single hierarchy, also known as a flat namespace.

Blob Storage can manage thousands of simultaneous uploads, enormous amounts of video data, constantly growing log files, and can be reached from anywhere with an internet connection via HTTP/HTTPS.

Blobs aren’t limited to common file formats. A blob could contain gigabytes of binary data streamed from a scientific instrument, an encrypted message for another application, or data in a custom format for an application. Azure takes care of the physical storage needs on your behalf.

Azure Data Lake Store (ADLS) Gen2

Azure Data Lake Storage is a comprehensive, scalable, and cost-effective data lake solution for high-performance big data analytics built into Azure.

Azure Data Lake Storage Gen1 is an enterprise-wide hyper-scale repository for big data analytic workloads. It enables you to capture data of any size, type, and ingestion speed in one single place for operational and exploratory analytics.

Azure Data Lake Storage Gen2 is a set of capabilities dedicated to big data analytics, built on Azure Blob Storage. It converges the capabilities of Azure Data Lake Storage Gen1 with Azure Blob storage.

ADLS Gen2 = Azure Blob Storage + ADLS Gen1

ADLS Gen2 provides file system semantics, file-level security, and scale, which are inherited from ADLS Gen1. All these capabilities are built on Blob storage resulting in low cost, tiered access, high security, high availability, and durability.

It is designed to manage and process multiple petabytes of information with hundreds of gigabits of throughput. A hierarchical namespace mechanism allows ADLS Gen2 to provide file system performance at object storage scale (optimizes I/O of high-volume data) and prices.

Key Differences

Structure
Blob: Flat namespace object store.
ADLS: Hierarchical namespaces (*much like a File System*).

Purpose
Blob: General purpose object store for a wide variety of storage scenarios, including big data analytics.
ADLS: Optimized storage for big data analytics workloads.

Performance *(Analytics Workload)*
Blob: Good storage retrieval performance.
ADLS: Better storage retrieval performance.

Cost
Blob: High cost for Analysis.
ADLS: Low cost for Analysis.

Hierarchical namespaces *organize blob data into directories and stores metadata about each directory and the files within it. They keep the data organized, which yields better storage and retrieval performance for an analytical use case and lowers the cost of analysis. This structure allows operations, such as directory renames and deletes, to be performed in a single atomic operation.*
Flat namespaces*, by contrast, require several operations proportionate to the number of objects in the structure.*

Use Cases

Blob storage is ideal for:

- Serving images or documents directly to a browser.
- Storing files for distributed access, such as installation.
- Streaming video and audio.
- Storing data for backup and restore, disaster recovery, and archiving.
- Writing to log files.
- Any type of text or binary data, such as application backend, backup data, and general purpose data.

ADLS is ideal for:

- Creating a modern data warehouse.
- Advanced analytics against big data.
- Creating a real-time analytical solution.
- Hadoop compatible access (HDFS, ABFS) is required. Access it through compute technologies including Azure Databricks, Azure HDInsight, and Azure Synapse Analytics without moving the data between environments.
- ACL and POSIX permissions along with some extra granularity support is required.
- Batch, interactive, streaming analytics and machine learning data such as log files, IoT data, click streams, large datasets.