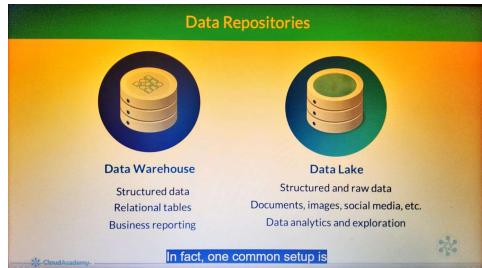


Using Azure Data Lake Storage Gen2

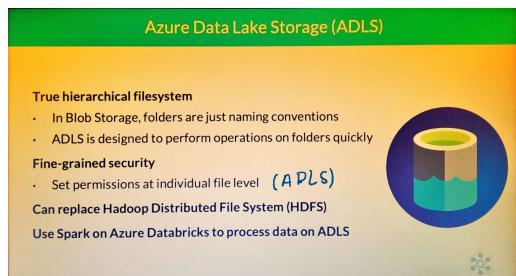
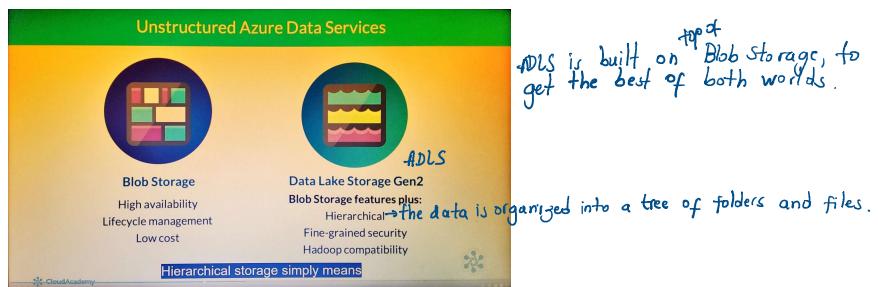
Friday, January 6, 2023 7:29 PM

Overview



In fact, one common setup is

In fact, one common set up is to process data in the data lake and then export it to the data warehouse.



for Blob storage to operate on a simulated folder, it has to perform a separate operation on each file.
→ BS can only restrict access at the container level, rather than at the individual blob level.

Directory → Folder
Container → filesystem
↓
BS ADLS perfectly
→ ADLS can seamlessly integrate with the huge ecosystem of Hadoop SW.

Security



Security Layers

Authentication methods

Azure Active Directory (AAD) verifies a user's identity

- Users must be in AAD to access Azure Data Lake Store

Shared Access Signature

- Only has access to specific data and has an expiry date and time

Shared Key

- Not recommended (older)

Security Layers

Authentication methods

Azure Active Directory (AAD) verifies a user's identity

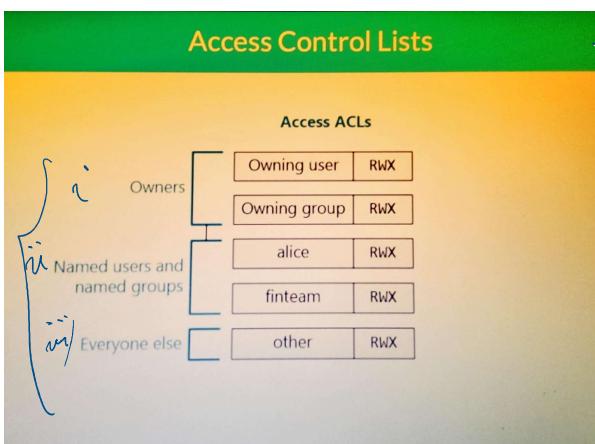
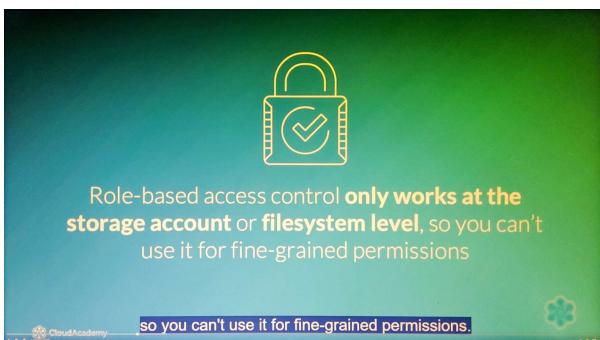
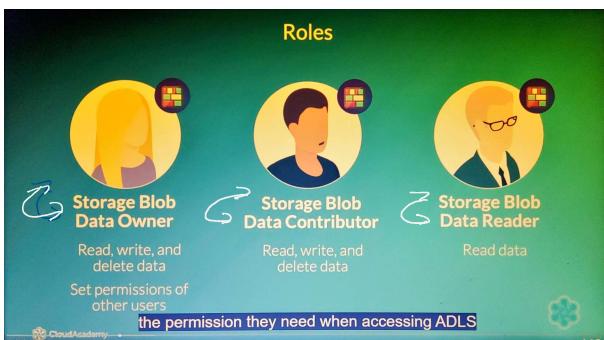
- Users must be in AAD to access Azure Data Lake Store

Shared Access Signature

- Only has access to specific data and has an expiry date and time

Shared Key

- Not recommended



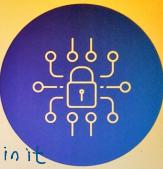
→ to handle permissions for files and folders
 → Each entry in an ACL specifies the read, write, and execute permissions for a specific user or group.

Access Control Lists		
	File	Folder
Read (R)	Can read the contents of a file	Requires Read and Execute to list the contents of the folder
Write (W)	Can write or append to a file	Requires Write and Execute to create child items in a folder
Execute (X)	Does not mean anything in the context of Data Lake Storage	

ACL Best Practices

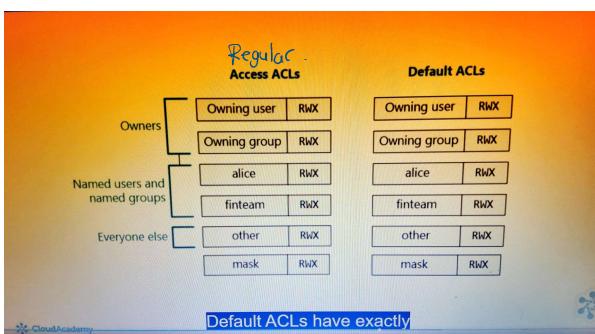
Assign permissions to groups instead of users

- Easier to set up and maintain
- Limit of 9 custom entries per ACL



Set default ACLs on folders, when possible
The every file or folder that gets created in it will have those ACLs too.

It's also a good idea to set default ACLs.



- iii) Network Isolation : you can actually set up a firewall just for your datalake.
 - iv) Encryption: data protected (in transit - or not in transit) with Azure Storage Encryption or your own.
 - v) Defender: potential malicious
 - vi) Auditory: Activity log .
- Ingesting

Ingesting Data

Ways to upload data from your desktop to ADLS

- AzCopy
- Azure Storage Explorer
- PowerShell
- Azure CLI



I'm going to do something a little bit different. I'm going to use AzCopy from

AccessingADLS from Azure Databricks

- ADLS is intended to work with Hadoop- Compatible software.
- A best option in Azure is Databricks, which is a managed Spark service.
- to use ADL and Databrick together; you need to perform some security related steps.

→ to use ADL and Databricks together; you need to perform some security related steps.

Accessing ADLS from Azure Databricks

Ways to authenticate Databricks to ADLS
Premium Databricks Workspace (more expensive).

- Credential passthrough - uses Azure AD credentials (easiest) and secure, ONLY for premium workspace
- Service principal - identity you assign to a service
- Embed storage account access key in code on Databricks - not recommended (there is a security risk to have an account key in plain text in your code)



Demo

- Create an Azure Databricks workspace
- Spin up a Spark cluster and create a notebook
- Access ADLS filesystem from Databricks notebook



Accessing ADLS from Azure Databricks

Authenticate using a service principal

- Create a service principal by registering an app (Azure Databricks instance) in Azure Active Directory
- Assign Storage Blob Data Contributor role to the service principal
- Create a secret in an Azure Key Vault that the service principal can use to authenticate
- In Databricks workspace, create an Azure Key Vault-backed secret scope
- Run mount code, including the service principal ID, the names of the secret scope and secret, and the Azure AD tenant ID



Analyzing data with Azure Databricks

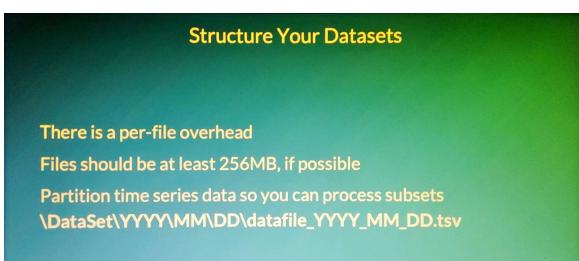
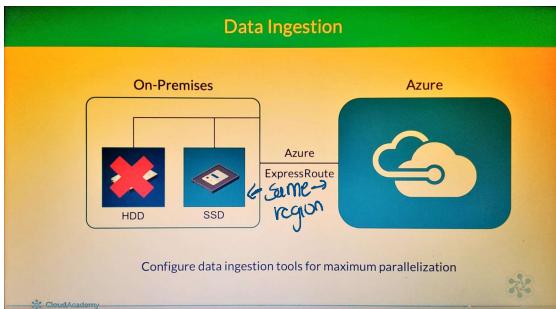
Monitoring and Optimization

ADLS has:

- Insights (not for all services)
- Alerts
- Metrics.

When you're transferring large amounts of data from local to Azure, there are 3 potential bottlenecks:

- Speed of your storage (your local disks should be on SSDs rather than spinning disks) and on storage arrays rather than individual disks
- Have a high-speed internal network (network interface cards)
- the network connection between your local infrastructure and the Azure cloud should be fast. (major bottleneck → Azure Express Route)



Question CORRECT

In Azure Data Lake Storage Gen2, if you enable _____ it will watch for attempts to access or exploit your storage accounts. If any suspicious activities are detected, then it will send you alerts through Microsoft Defender for Cloud.

Access Control
 network isolation
 authentication
 Advanced Threat Protection
 I don't know yet

Explanation
The fifth layer of security is Advanced Threat Protection. If you enable this, it will watch for attempts to access or exploit your storage accounts. If any suspicious activities are detected, then it will send you alerts through Microsoft Defender for Cloud.

Learn more: /course/using-azure-data-lake-storage-gen2/security/

Something wrong with this question? Report an issue

Question CORRECT

Which of the following is not a recommendation for optimizing your Azure Data Lake Storage Gen2?

If your data source is also in Azure, then put it in the same region as the data lake, if possible.
 Avoid storing many files smaller than 256 MB in size.
 Store your local data on spinning disks, rather than SSDs.
 Store your local data on storage arrays rather than individual disks.
 I don't know yet

Explanation
Ideally, your local data should be on SSDs rather than spinning disks, and on storage arrays rather than individual disks. If your data source is also in Azure, then put it in the same region as the data lake, if possible. When your data is being processed, there's a per-file overhead, so if you have lots of small files, it can impact the performance of the job. If possible, your files should be at least 256 MB in size.

Learn more: /course/using-azure-data-lake-storage-gen2/monitoring-and-optimization/

Question 3

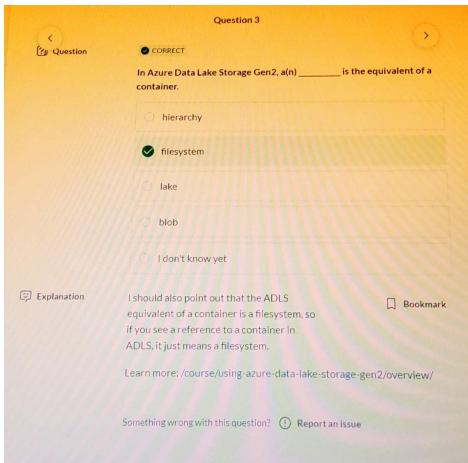
CORRECT

In Azure Data Lake Storage Gen2, a(n) _____ is the equivalent of a container.

hierarchy
 filesystem
 lake
 blob
 I don't know yet

Explanation I should also point out that the ADLS equivalent of a container is a filesystem, so if you see a reference to a container in ADLS, it just means a filesystem.
[Learn more: /course/using-azure-data-lake-storage-gen2/overview/](#)

Something wrong with this question?



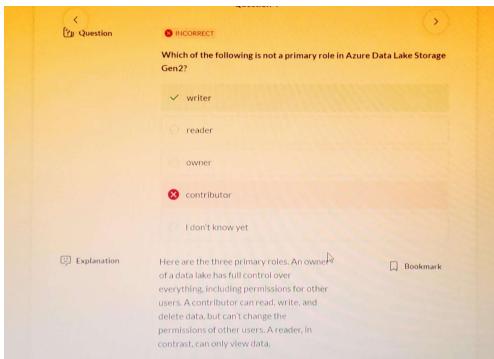
Question 4

INCORRECT

Which of the following is not a primary role in Azure Data Lake Storage Gen2?

writer
 reader
 owner
 contributor
 I don't know yet

Explanation Here are the three primary roles. An owner of a data lake has full control over everything, including permissions for other users. A contributor can read, write, and delete data, but can't change the permissions of other users. A reader, in contrast, can only view data.
[Learn more: /course/using-azure-data-lake-storage-gen2/roles/](#)



Question 5

CORRECT

Which of the following statements about data warehouses and data lakes is false?

Data lakes store any kind of data, whether it's structured or not.
 One common setup is to process data in the data warehouse and then export it to the data lake.
 Data warehouses store data in structured, relational tables.
 Data warehouses have been around for decades, but the term "data lake" was coined more recently.
 I don't know yet

Explanation Data warehouses have been around for decades, but the term "data lake" was only coined in about 2011. While data warehouses store data in structured, relational tables, data lakes store any kind of data, whether it's structured or not. For example, you could store everything from documents to images to social media streams. Data warehouses are generally used for business reporting, while data lakes are more often used for data analytics and exploration. In fact, one common setup is to process data in the data lake and then export it to the data warehouse.
[Learn more: /course/using-azure-data-lake-storage-gen2/data-lake-vs-data-warehouse/](#)

