

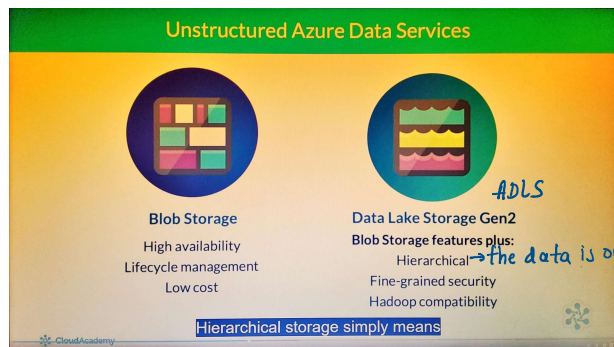
Using Azure Data Lake Storage Gen2

Friday, January 6, 2023 7:29 PM

Overview



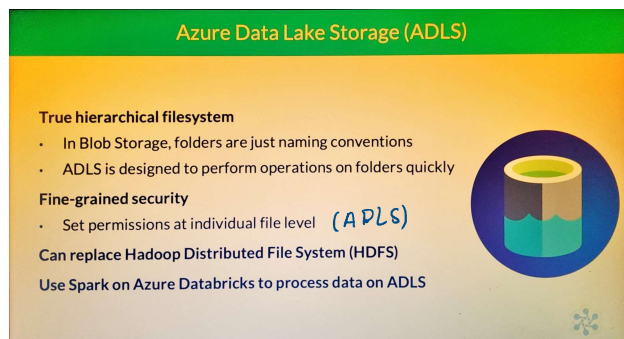
In fact, one common set up is to process data in the data lake and then export it to the data warehouse.



ADLS is built on top of Blob Storage, to get the best of both worlds.

ADLS

the data is organized into a tree of folders and files.



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 (6 layers)

- Authentication
- Access control
- Network isolation
- Data protection
- Advanced threat protection
- Auditing



Cloud Academy

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

Access control

- Roles
- Access Control Lists (ACLs)


Roles



Storage Blob Data Owner


Read, write, and delete data

Set permissions of other users



Storage Blob Data Contributor

Read, write, and delete data

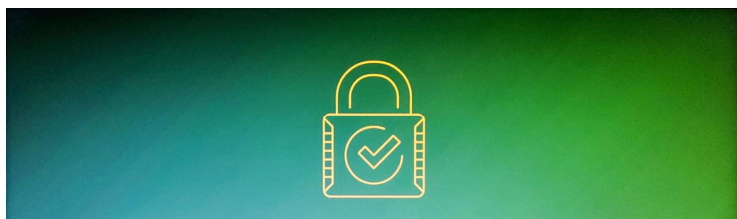



Storage Blob Data Reader

Read data

the permission they need when accessing ADLS

Cloud Academy





Role-based access control **only works at the storage account or filesystem level**, so you can't use it for fine-grained permissions

so you can't use it for fine-grained permissions.

Access Control Lists

Access ACLs

Category	Entity	Permissions
Owners	Owning user	RWX
	Owning group	RWX
Named users and named groups	alice	RWX
	finteam	RWX
Everyone else	other	RWX

Handwritten notes: i, ii, iii

→ 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

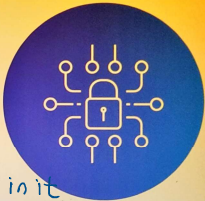
Assign permissions to groups instead of users

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

Set default ACLs on folders, when possible

Handwritten note: 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.



	Regular Access ACLs	Default ACLs
Owners	Owning user RWX	Owning user RWX
	Owning group RWX	Owning group RWX
Named users and named groups	alice RWX	alice RWX
	finteam RWX	finteam RWX
Everyone else	other RWX	other RWX
	mask RWX	mask RWX

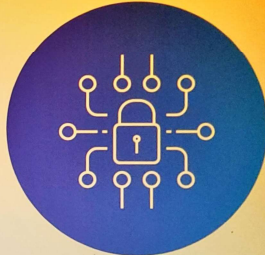
Default ACLs have exactly

- iii) Network Isolation: you can actually set up a firewall just for your data lake.
- 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