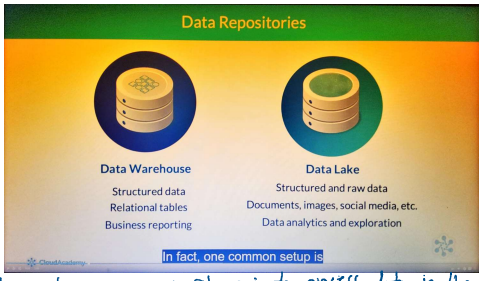
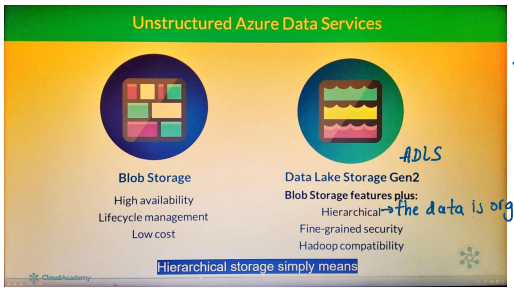
Using Azure Data Lake Storage Gen2

Friday, January 6, 2023 7:29 PM

Overview

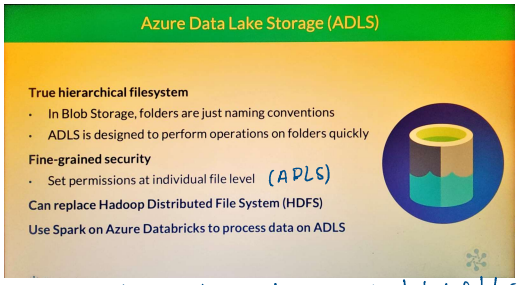


In fact, one common set up is to process data in the data lake and then export it to the data warehouse.



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

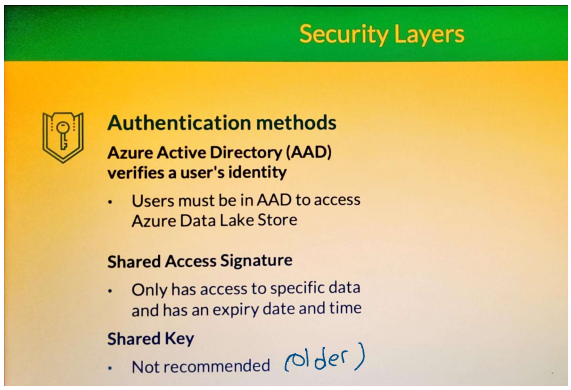
→ the data is organized into a tree of folders and files.

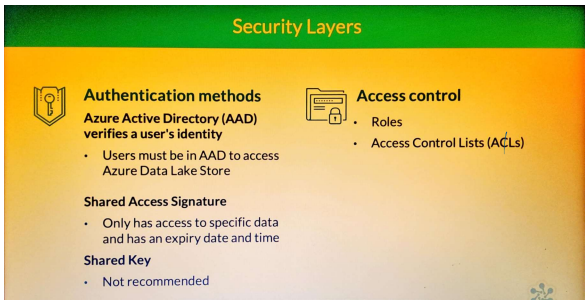


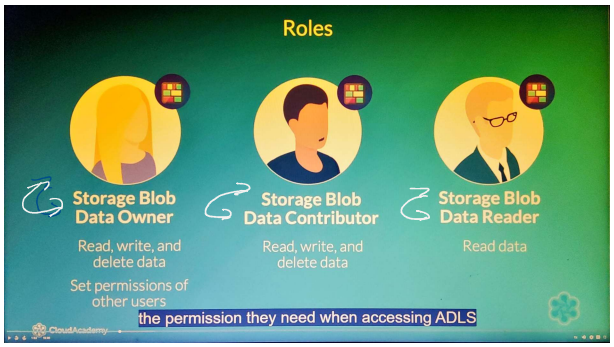
for Blob storage to operate on a simulated folder it has to perform a separate operation on each file.
→ BS can only restrict access at the container level, rather than at the individual blob level.

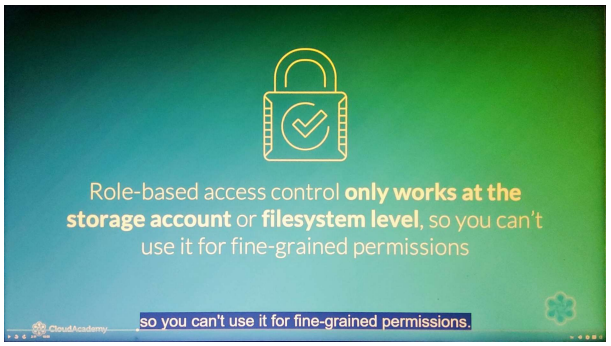
Directory ↔ Folder
Container ↔ filesystem
BS ADLS perfecta!
→ ADLS can seamlessly integrate with the huge ecosystem of Hadoop SW.

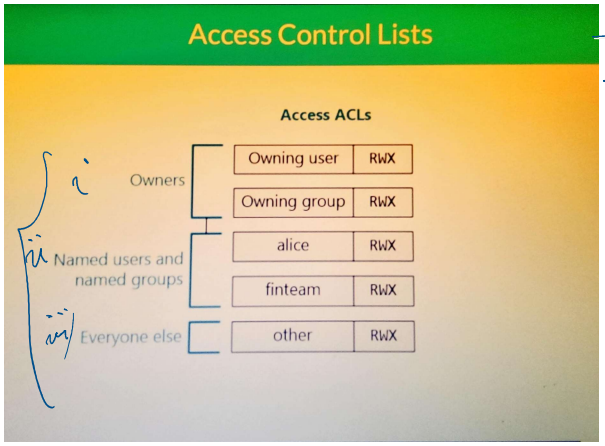
Security



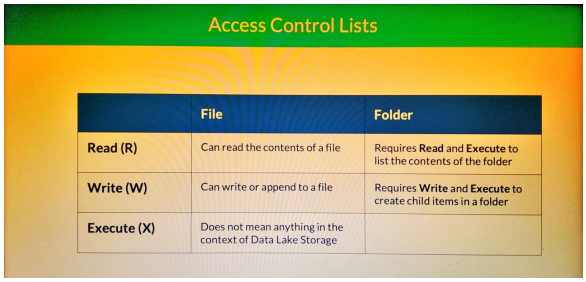


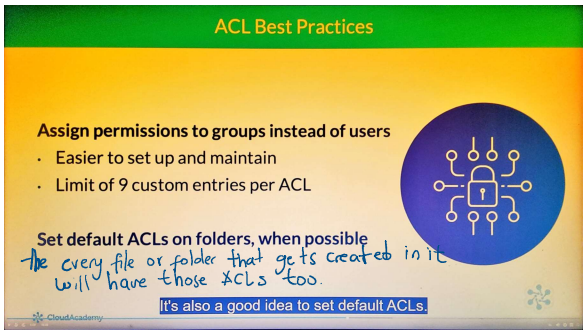


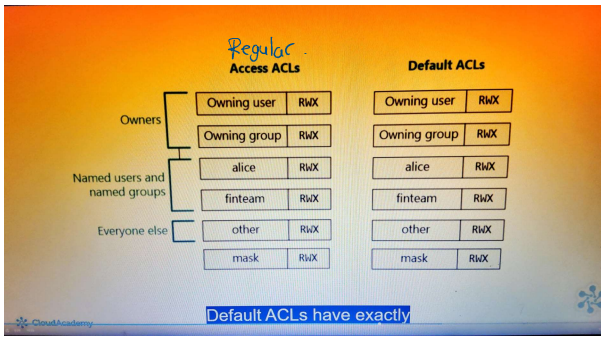




→ to handle permissions for files and folders
→ Each entry in an ACL specifies the read, write, and execute permissions for a specific user or group.

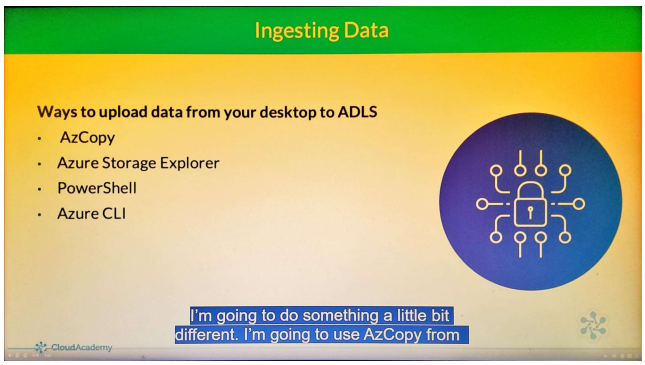






iii) Network Isolation: you can actually set up a firewall just for your datalake.
iv) Encryption: data protected (in transit - or not in transit) with Azure Storage Encryption or your own.
v) Defender: potential malicious
vi) Auditory: Activity log.

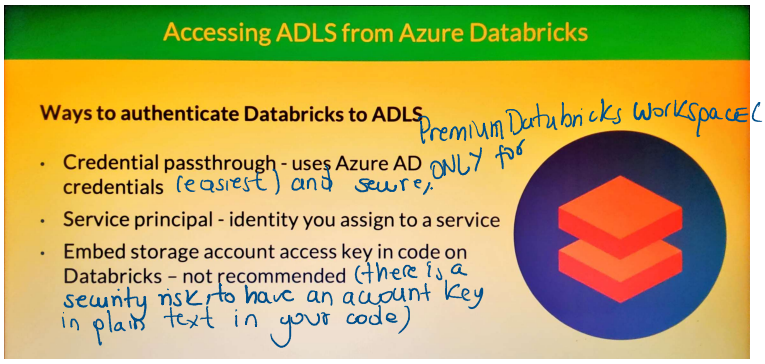
Ingesting

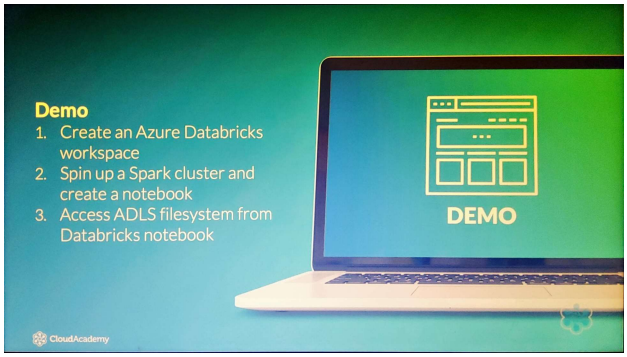


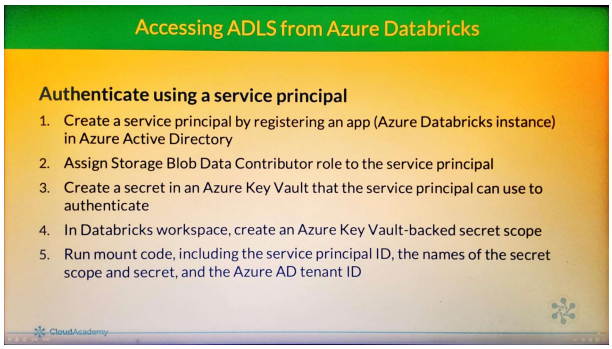
AccessingADLS from Azure Databricks

→ 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 Databrick together, you need to perform some security related steps.







Analyzing data with AzureDatabricks

Monitoring and Optimization

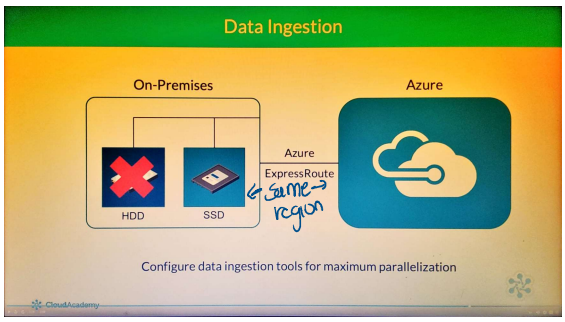
ADLS has:
- → Insights (not for all services)
- → Alerts
- → Metrics.

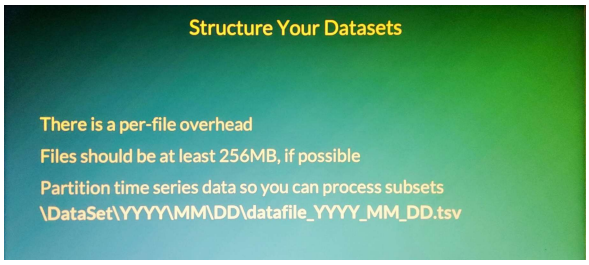
When you're transferring large amounts of data from local to Azure, there are 3 potential bottlenecks:

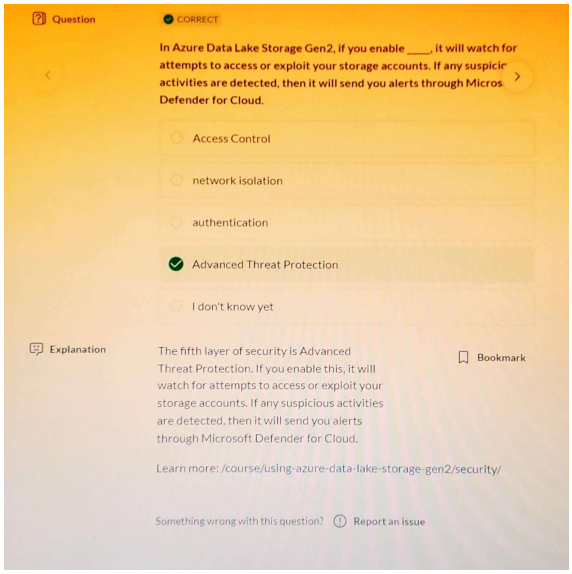
i) Speed of your storage (your local disks should be on SSDs rather than spinning disks) and on storage arrays rather than individual disks)

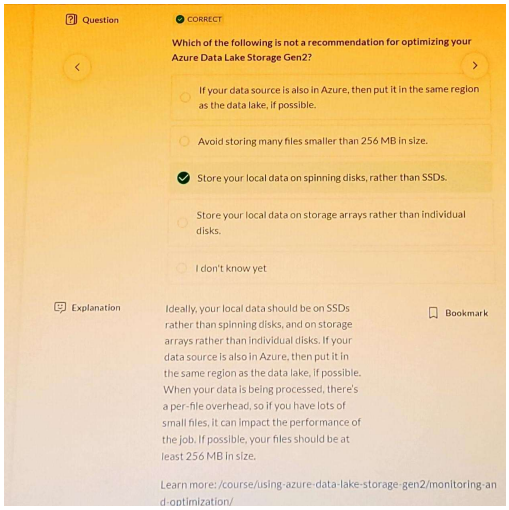
ii) Have a high-speed internal network (network interface cards)

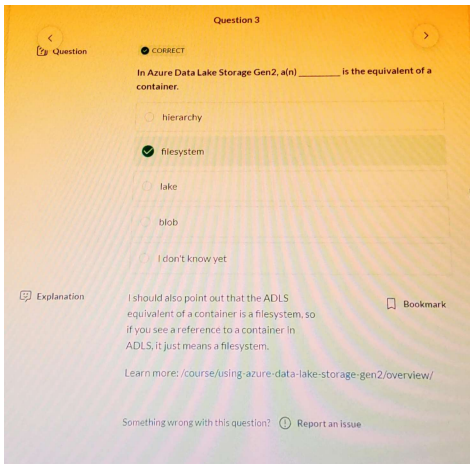
iii) the network connection between your local infrastructure and the Azure cloud should be fast. (major bottleneck → Azure Express Route)

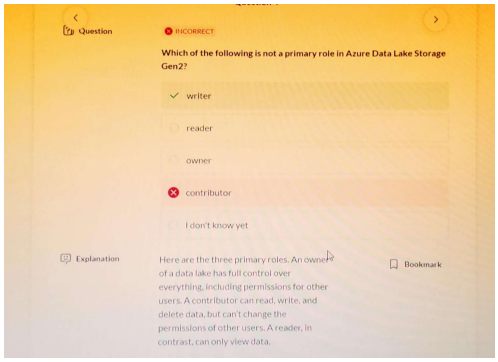

Data Ingestion
On-Premises
Azure
HDD
SSD
Azure
ExpressRoute
same region
Configure data ingestion tools for maximum parallelization
CloudAcademy

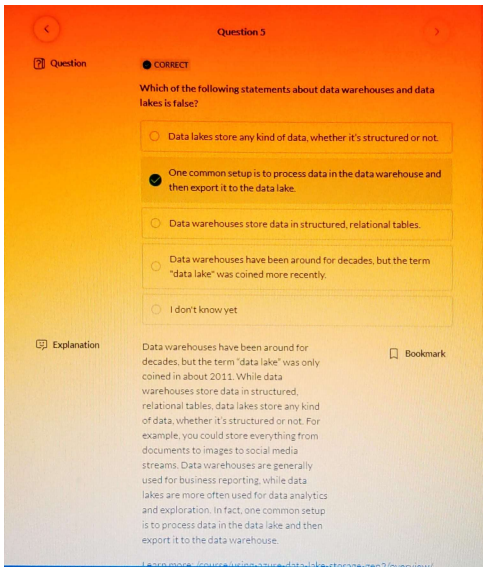

Structure Your Datasets
There is a per-file overhead
Files should be at least 256MB, if possible
Partition time series data so you can process subsets
\DataSet\YYYY\MM\DD\datafile_YYYY_MM_DD.tsv


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 suspicic activities are detected, then it will send you alerts through Micros Defender for Cloud.
Access Control
network isolation
authentication
Advanced Threat Protection
I don't know yet
Explanation
Bookmark
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
Bookmark
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
Question
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.
Bookmark
Learn more: /course/using-azure-data-lake-storage-gen2/overview/
Something wrong with this question?
Report an issue


Question
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.
Bookmark


Question 5
Question
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.
Bookmark