Azure Data Lake Creation and Requirements

Requirements

Functionality

1. Scalable storage starting at 35TB
2. Read/write data through a secure tunnel
3. Encryption at rest
4. Automatic backup to another location
5. Accessible at a minimum on the on-premises network
    a. Could restrict access to specific Ips
6. Restricted access to IT team

Definitions:

Data Interactions

- Data Read/Write - The amount of data that's being interacted with
    - Exclusively for reading and writing data without any other facets (reading/writing metadata, etc.)
- Read/Write Operation (transactions) - The action of interacting with the data
    - Used to read/write data while modifying or accessing some other facet of the data (analytics, metadata, etc.)

Essentially, a write operation is involved when the data is being subtly modified in some way. This could be the metadata or some other facet. While a data write is strictly for copying data from one source to the data lake.

Default chunk size of 4MB.

Although these are technically different, they are tied together. For example, when writing 1GB of data from an on-premises machine to the (Azure Data Lake) ADL:
Let's say this 1GB of data consists of 1,000 files and that most of these files are below the 4MB default chunk size. However, let's also say that 100 of these files are at 10MB. The 900 files below 4MB will be uploaded using one transaction operation. The 100 files at 10MB will each be uploaded using 3 transactions. So, this will equate to 1,200 operations for 1,000 files.

Redundancy

- LRS - Locally Redundant Storage
    - Replicated at the same data center
- ZRS - Zone Redundant Storage
    - Replicated across three data centers in the same region
- GRS - Geo-Redundant Storage

    - Replicated to another region, only allowed to read from the other region in the case of a failure
- RA-GRS
    - Replicated to another region and allows read access from that other region.

Access Tiers

- Hot
    - High storage costs, very low read/write access costs
- Cool
    - Lower storage costs, higher read/write access costs
- Archive
    - Very low storage costs, very high read/write access costs
    - Can only be defined at container level
    - Files cannot be defaulted to this tier

Components

Azure Subscription

The subscription that your organization has created under their Directory. This is just for the organization to coordinate individual teams or departments and their pricing for those departments.

Resource Group

A way for a user to organize their creations. For this purpose, the resource group allows us to containerize the Data Lake in the case that we want to expand our cloud services.

Azure Data Lake

Azure Data Lake storage will house blob data blocks in a hierarchical format. This data can be accessed via the Azure web portal or a downloadable program, Azure Data Factory.

Requirement Fulfillment:

1, 3, 4

- Expanding storage
- Up to 500TB storage
- Encryption at rest
- Automatic backup

Data Transfer

Azure Data Factory (ADF)

Not cost effective if you just want to transfer the data as is. However, if you want to run analytics and other functions to better organize the data this is ideal.

Azure Data Factory allows read/write operations similar to the file explorer. ADF supports more functionality than just data transfers. It allows you to transfer data from multiple sources, run machine learning on the data. This can filter, sort, and aggregate data.

Requirement Fulfillment

2

- Read/write data easily

Azure Storage Explorer

Ideal solution for direct transfer of data without any modifications.

Azure Storage Explorer functions very similarly to a file explorer. It allows you to transfer files from one source to another. Azure Storage Explorer is a GUI for AzCopy

AzCopy

AzCopy is a CLI (command-line interface) used to copy data to/from Azure Storage. Azure Storage Explorer uses AzCopy for its data transfer operations.

Security

Azure Private Endpoint

This option is only necessary if we restrict specific IP addresses to only be accessible from the company network.

Azure Private Endpoint will provide an endpoint connection to the Azure Data Lake. This will restrict access to only those who are connected to the ExpressRoute connection or Azure itself.

Requirement Fulfillment:

5, 5a

- Connect directly to Azure

NSG (Network Security Groups)

This option is ideal if we keep the desired computer IP addresses open to the public.

Typical firewall type settings. Can allow a group of IP addresses, single IP addresses, etc.

Requirement Fulfillment:

5

Azure AD

Azure Active Directory acts as a cloud-based identity access management service.

6

- Can assign RBAC (Role-based access control) for specific people

Estimated Costs

There are a few components that will cost money, the rest are included in the subscription and just a necessary part of Azure.

Azure Data Lake

Specs

- Archive Tier
- GRS redundancy
- 40TB storage

Transactions

Initial upload - 35TB of ~15 million files

- ~17 million write operations of 4MB chunks = $464
- Data Write is free
- **Total = $464**

Grand total = $464

Per month costs

- Write - 2TB operations per month
    - Data Write - $0
    - Write operations - $14
    - **Total = $14 per month**
- Read - 1TB operations per month
    - Data Retrieval - $20
    - Read Operations - $169
    - **Total = $189 per month**
- Storage - 40TB
    - **Total = $122 per month**

Grand total = $325 per month

Data Transfer Applications

Azure Data Factory

For the initial upload this would cost an additional $170 just to use the Data Factory read/write operations without any data modification. Pricing varies on the type of modification we would use.

Azure Storage Explorer

Free, cost occurs at Data Lake transaction/operation level.

Total Costs

Without Azure Data Factory

Total initial costs = $464

Total monthly costs = $325/month

Creation and Connection of Components

This will only involve the following components, as the others were deemed unnecessary: Azure Subscription, Azure Resource Group, Azure Data Lake, Azure Storage Explorer, and NSG.

Azure Subscriptions

In this case, the subscription was created by Virginia Tech. However, if you're not under an organization owned Directory you can follow these steps:

1. Log into the Azure Portal under your Directory
2. Navigate to the Subscriptions page
3. Fill in the generic subscription information
    a. Create an appropriate name for the services you plan to use
    b. Associate a billing account or create a new one
    c. Review and create

Each subscription that you create should be organized by the types of services you will create under each subscription. For example, production services should be disjointed from development and QA subscriptions.

Azure Resource Group

Similar to Azure Subscriptions, the resource group should organize your services further and provide fine-grained access.

1. Log into the Azure Portal Under your Directory
2. Navigate to the Resource Groups page
3. Create a new resource group
    a. Select the appropriate subscription
    b. Give the resource group an appropriate name
        i. Ex: company-archives for a cluster of services associated with archiving company materials
    c. Select the region that is closest to the majority of users
    d. Review and create

Azure Data Lake

An Azure Data Lake is built upon Azure Blob Storage. The key difference is that it allows a hierarchical structure.

Creation

1. Log into the Azure Portal Under your Directory
2. Navigate to the Storage Account page
3. Create a new storage account
    a. Basics
        i. Select the appropriate subscription and resource group (see above for examples)
        ii. Name your storage account with the intention of full clarity in the case that you need to expand your cloud services under this resource group
        iii. Select the region that is closest to the most users who will access this service
        iv. Select the performance tier
            1. Standard is the default unless you're working with big analytical data. If you are, select Premium

v. Select a redundancy that is appropriate for your deliverable/security needs
    1. See the "Components" section for examples

b. Advanced
    i. Secure transfer for the REST API should always be enabled unless you're positive that all data being transferred through the API can be accessible to the public
    ii. Enable public access on containers if you think any containers will need public access
        1. This option doesn't enable public access, it only allows the option for public access
    iii. Enable storage account key access if you believe you'll need to use SAS
        1. This option is good for accessing data through URLs and if you want to grant fine-grained permissions
    iv. If you will be copying data to another storage account, the permitted scope can be modified for enhanced security
    v. **Data Lake Storage Gen2 Hierarchical Namespace option must be enabled for this to be considered a Data Lake storage**
        1. Without this option enabled you are just creating a blob storage service
    vi. See the "Components" section for a description of Access Tiers

c. Networking
    i. Public access - allows access for anyone with permission on a public network
    ii. Public access from selected VN and IPs - Allows you to restrict the access to a specific virtual network on Azure or selected public IP addresses
        1. You can assign specific Ips, virtual networks, etc. once the Data Lake has been created. See the NSG in the "Components" section
    iii. Disable public access and use private access - Restricts access to an Azure Private Endpoint. See the "Components" section for more information

d. Data Protection
    i. Here are various features that provide a bit of padding in the case that something goes wrong, or a mistake is discovered
        1. Soft delete is essentially a recycle bin that won't permanently delete data for X amount of time
        2. Versioning is good if you plan to have a timeline for a document that can be restored to any point in the timeline

e. Encryption
    i. MMK is where encryption keys are managed by Microsoft without any needed user interactions. This is the ideal setting for minimal overhead.
        1. If you have organizational standards that you must follow, need to integrate your encryption keys with another service provider, etc. Use CMK
    ii. Infrastructure encryption provides another step of encryption on the infrastructure level. It's ideal for scenarios where one of the keys is compromised

4. Review and create

Post Creation

Please find a few points for post-creation of the Data Lake.

Tiers

- Archive tier cannot be set by default, this must be set on a container/file based level.
    - You can create a Lifecycle Management rule to automatically archive any data after X days. There are also other conditions that you can set here.

Security + Networking

- You can restrict access to a private network or specific IP address.
    - If you have a private endpoint, this is where you would connect it
- You can also generate access keys and SAS. This is used to easily grant HTTPS access to resources.

Access Control (IAM)

- Access can be granted to other Azure services or AD users

Data Migration

- The Data Migration section displays Azure options for migrating from common external/internal storages

Storage Browser

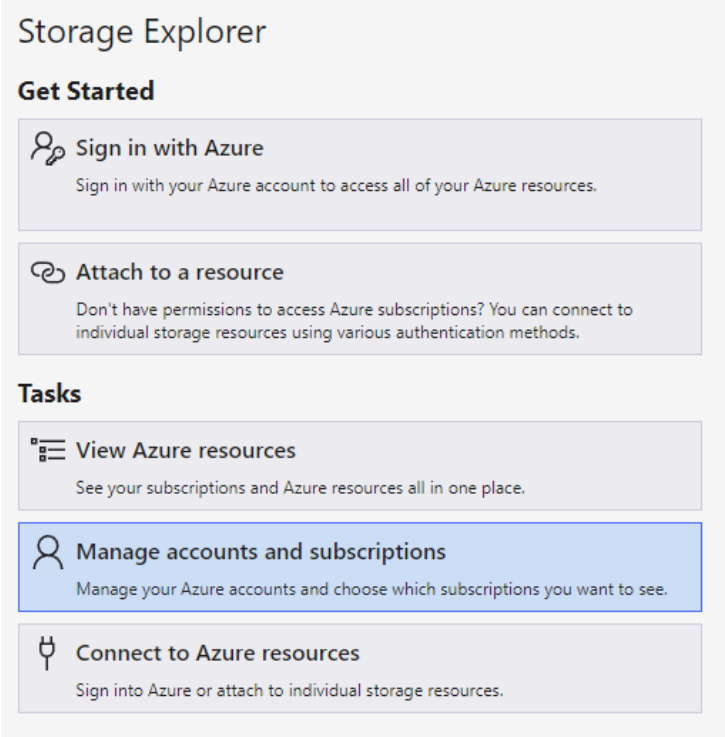
- This lets you view containers and blocks of data from the Azure Portal. It has diminished functionality over non-web based programs, such as Azure Storage Explorer

Azure Storage Explorer

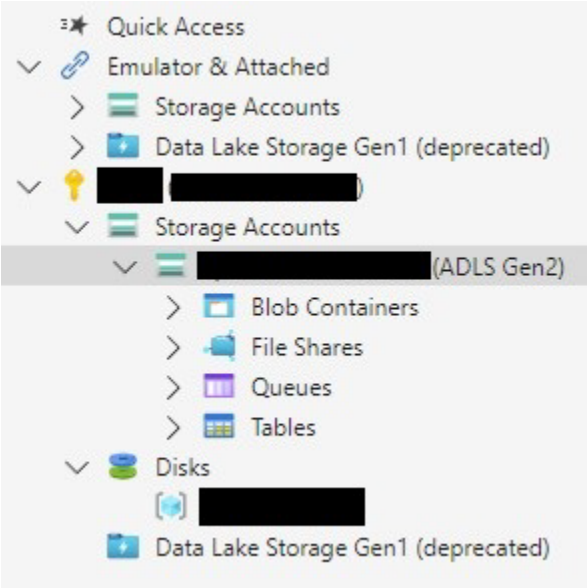
Azure Storage Explorer acts as a GUI for AzCopy. There are many functions of this program. For this use case the instructions to connect to a Data Lake are as follows:

1. Download and install Azure Storage Explorer

2. Open the storage explorer and in the main start window click Sign in with Azure



   a. This will open another window. Select the appropriate option and sign in by clicking next and following the instructions
3. Click "View Azure resources"
4. On the left panel you can now navigate through your data lake, similar to Windows File Explorer



Resources

- Pricing Calculator: https://azure.microsoft.com/en-us/pricing/calculator/
- Azure Data Factory: https://azure.microsoft.com/en-us/products/data-factory/

- Azure Storage Explorer: https://azure.microsoft.com/en-us/products/storage/storage-explorer/
- Azure Private Endpoint: https://learn.microsoft.com/en-us/azure/private-link/private-endpoint-overview
- Azure Data Lake: https://azure.microsoft.com/en-us/solutions/data-lake/
- Azure Blob Storage: https://azure.microsoft.com/en-us/products/storage/blobs/
- AzCopy: https://learn.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-v10