Non-Relational Data Stores on Azure

Created Date: 2022-08-17

> Metadata 🗃️
>
> - Title: Section 3: Non-Relational Data Stores
> - Author: Eshant Garg
> - Reference: https://www.udemy.com/course/dp200exam/

> Links & Tags 🔗
>
> - Index: Course Note Index
> - Atomic Tag: #datascience
> - Subatomic Tags: #dataengineering

Non-Relational Data Stores on Azure (NoSQL)

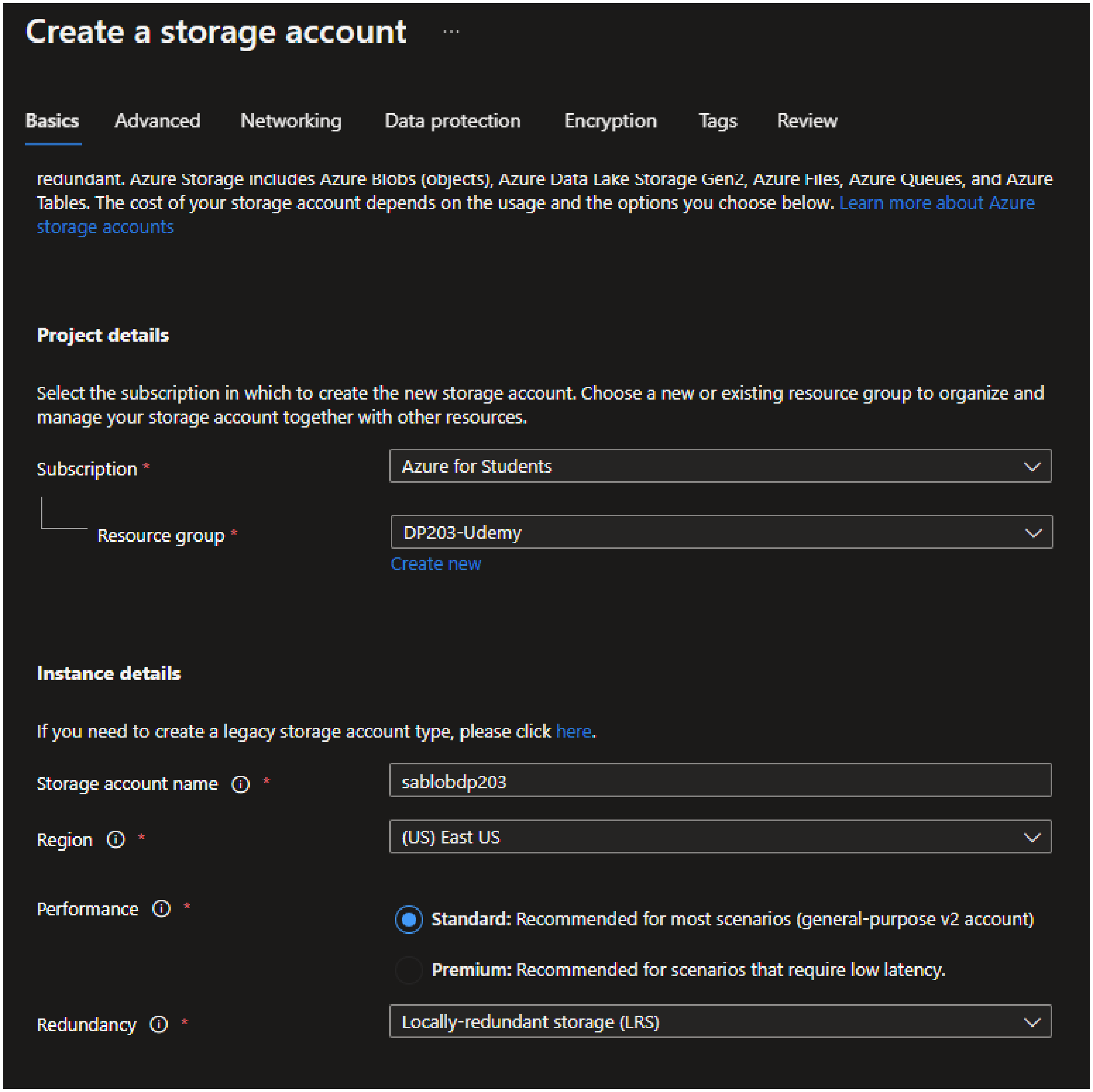
- Primary Storage Services
  - Azure Storage Account (Blob Storage)
  - Azure Data Lake Gen2
  - Azure Cosmos DB (Not Included in Exam)

Azure Storage Account

- Different Types of Data
  - Relational
  - Non-Relational (NoSQL)
  - Datasheets
  - Images
  - Videos
  - Backups

- Data Storage Requirements
  - Storage
  - Access
  - Security
  - Availability
  - Latency
  - Processing
  - Backup
- Types of Azure Data Storage Services (all included when a *standard* storage account is created)
  - Azure Blobs (Containers): Text and binary data
  - Azure Files: Managed file shared (SMB protocol)
  - Azure Queues: Messaging
  - Azure Tables: NoSQL store
  - Azure Disks: Block-level storage volumes for Azure VMs
- Data Storage Features
  - Durable and highly available (redundancy across datacenters or regions)
  - Secure (encryption)
  - Scalable
  - Managed (Azure handles hardware maintenance, updates, and critical issues)
  - Accessible (HTTP, HTTPS, client libraries in many languages, PowerShell or Azure CLI scripting)

Creating a Storage Account



Create a storage account ···

redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. Learn more about Azure storage accounts

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * : Azure for Students

Resource group * : DP203-Udemy

Instance details

If you need to create a legacy storage account type, please click here.

Storage account name * : sablobdp203

Region * : (US) East US

Performance * :

● **Standard:** Recommended for most scenarios (general-purpose v2 account)

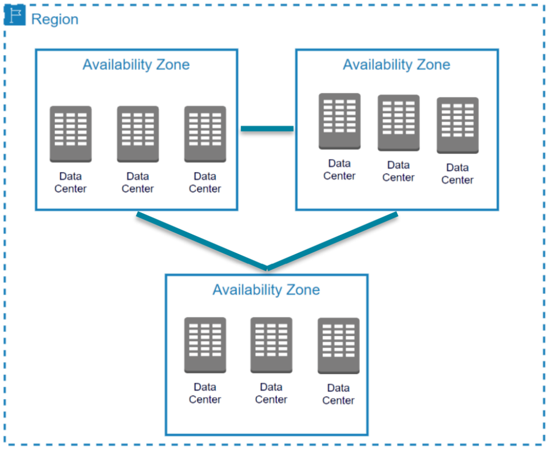
○ **Premium:** Recommended for scenarios that require low latency.

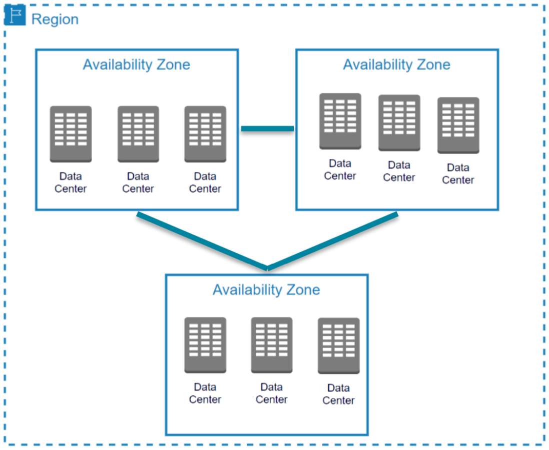
Redundancy * : Locally-redundant storage (LRS)

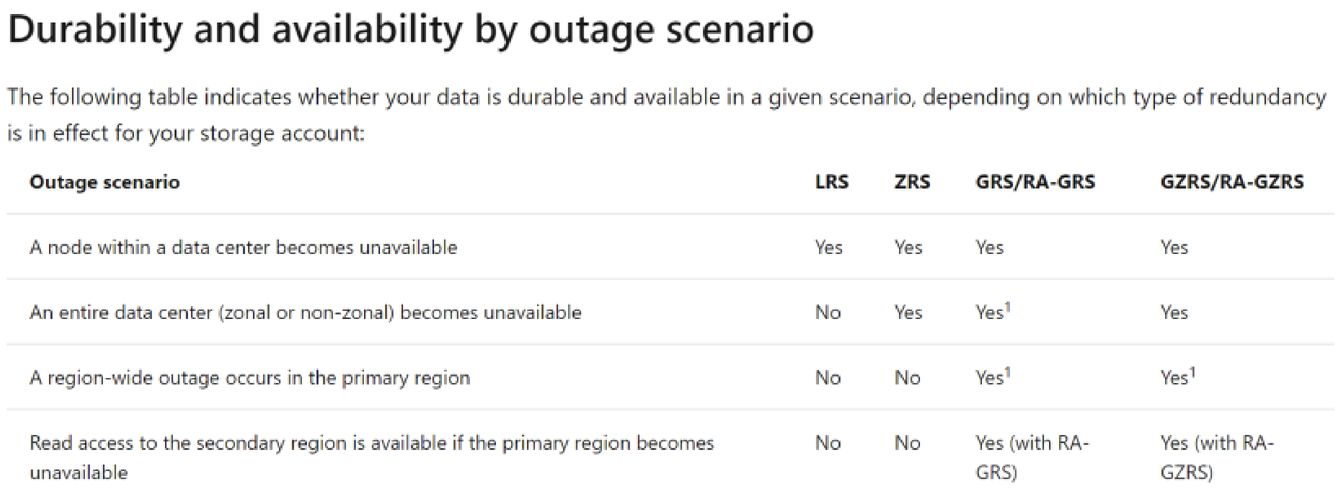
- Standard Performance will create a storage account for all storage services (blobs, files, queues, tables, disks)
- In the "Advanced" tab, you have the option to convert storage account to a Data Lake Gen2 account

Data Redundancy (SA and DL)

- Protect data from hardware failures, network or power outages, and natural disasters

- Redundancy ensures storage account's availability and durability int he event of a failure
- Tradeoff's between lower costs and higher availability
- Redundancy Options;
    - Locally Redundant Storage (LRS)
        - Three synchronous copies in the same data center
    - Zone-Redundant Storage (ZRS)
        - Three synchronous copies in three availability zones (AZs)
    - Geo-Redundant Storage (GRS)
        - LRS + Three asynchronous copies in a single data center of a secondary region (read only)
        - GRS or RA-GRS are recommended by Microsoft
    - Geo-Zone-Redundant Storage (GZRS)
        - ZRS + Three asynchronous copies in a single data center of a secondary region (read only)
    - 





Durability and availability by outage scenario

The following table indicates whether your data is durable and available in a given scenario, depending on which type of redundancy is in effect for your storage account:

Outage scenario	LRS	ZRS	GRS/RA-GRS	GZRS/RA-GZRS
A node within a data center becomes unavailable	Yes	Yes	Yes	Yes
An entire data center (zonal or non-zonal) becomes unavailable	No	Yes	Yes[1]	Yes
A region-wide outage occurs in the primary region	No	No	Yes[1]	Yes[1]
Read access to the secondary region is available if the primary region becomes unavailable	No	No	Yes (with RA-GRS)	Yes (with RA-GZRS)

Azure Blob Storage

- Blob: Binary Large Object
  - Any type of format
  - Text, images, audio, video, excel, backup files, etc
- Use cases;
  - Storing files for shared access
  - Video and audio streaming
  - Storing data for analysis (Data Lake Gen2)
  - Writing to the log file
  - Storing data for disaster recovery, backup, and archiving
- Data stored in a flat structure
  - Directory: Account/Container/Blob
    - Containers are essentially parent folders
    - Blobs are files within the container
    - Blobs cannot contain another container (folder)
- Access levels (via URL https://storage_account_name.blob.core.windows.net/container/file_name)

    - Private: No anonymous access
    - Blob: Anonymous read access for blobs only
    - Container: Anonymous read access for containers and blobs

- Types of blob storage

    - Block blobs
        - Large objects that do not use random read and write operations
        - Files that are read from beginning to end (media files, image files for websites, etc)
    - Page blobs
        - Optimized for random read and write operations
        - Provides durable disks for Azure Virtual Machines (VMs)
    - Append blobs
        - Optimized for append operations (logs)
        - When you modify an append blob, blocks are added to the end of the blob only
        - Updating or deleting existing blocks is not supported

Storage Access Tiers

- Data stored in he cloud can be different based on how it's generated, processed, and accessed over its lifetime. Selecting the proper blob storage access tier can save money on data storage

- Data can be set to automatically move to different access tiers after a certain amount of time has passed

- Settings can be changed in the Configuration tab within the storage account

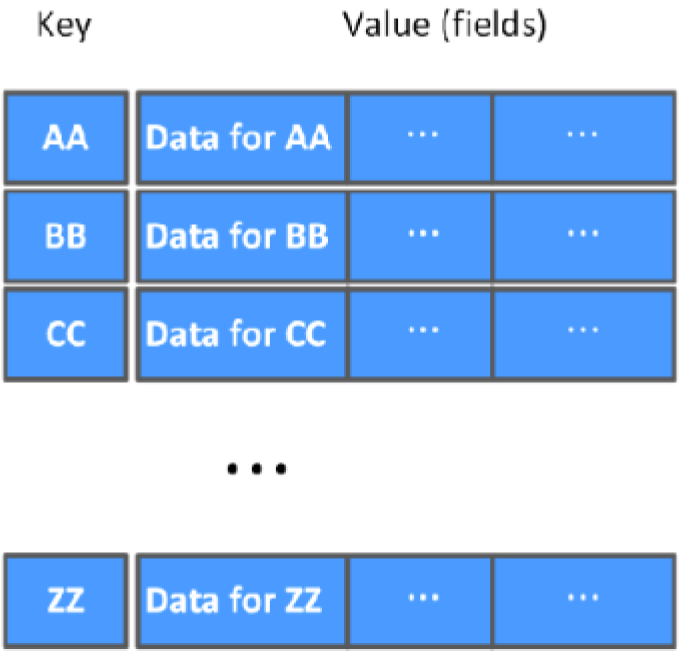
- Types of blob storage access tiers

    - Hot

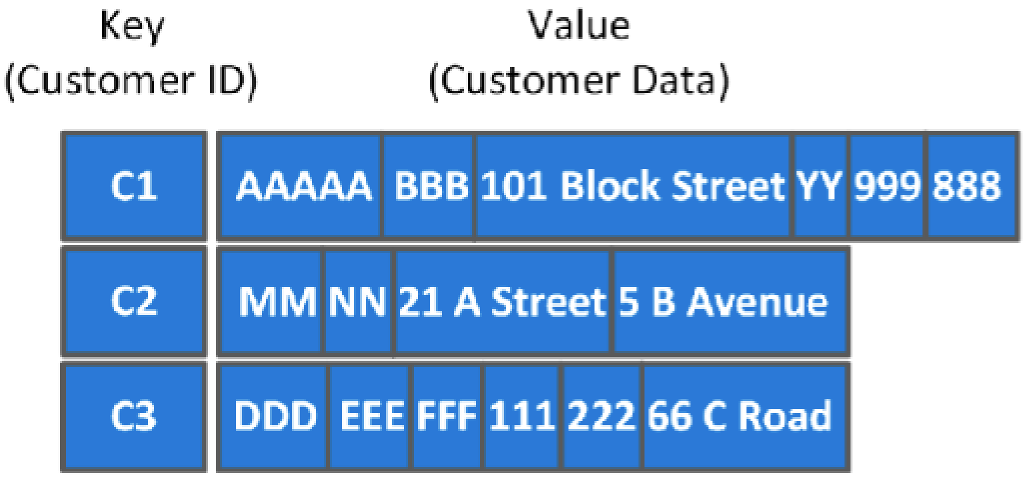
    - Frequently accessed
    - Low latency
    - High cost
  - Cool
    - Infrequently accessed
    - High latency
    - Low cost
    - Data must be stored for at least 30 days
  - Archive
    - Rarely accessed
    - High latency
    - Lowest cost
    - Data must be stored for at least 180 days
    - *Can only be set at the blob level, not the account level*

Azure Table Storage

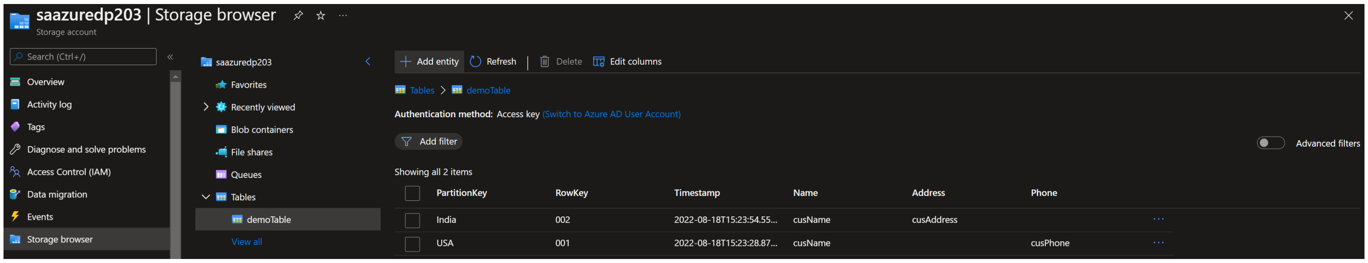
- NoSQL key-value storage
- Items are referred to as rows, fields are known as columns
- All rows in a table must have a key
- No concept of relationships, stored procedures, secondary indexes, or foreign keys
- Data will usually be de-normalized
- Tables split into partitions to ensure fast access
- Supports very large volume of data
- Consider Cosmos DB for new deployment
- Advantages;
  - Easy to scale
  - Holds semi-structured data (fields may not all be the same for each row)
  - No complex relationships
  - Data insertion and retrieval is fast



Key	Value (fields)		
AA	Data for AA	...	...
BB	Data for BB	...	...
CC	Data for CC	...	...
• • •			
ZZ	Data for ZZ	...	...



Key (Customer ID)	Value (Customer Data)					
C1	AAAAA	BBB	101 Block Street	YY	999	888
C2	MM	NN	21 A Street	5 B Avenue		
C3	DDD	EEE	FFF	111	222	66 C Road



saazuredp203 | Storage browser

Storage account

Authentication method: Access key (Switch to Azure AD User Account)

Showing all 2 items

PartitionKey	RowKey	Timestamp	Name	Address	Phone
India	002	2022-08-18T15:23:54.55...	cusName	cusAddress	
USA	001	2022-08-18T15:23:28.87...	cusName		cusPhone

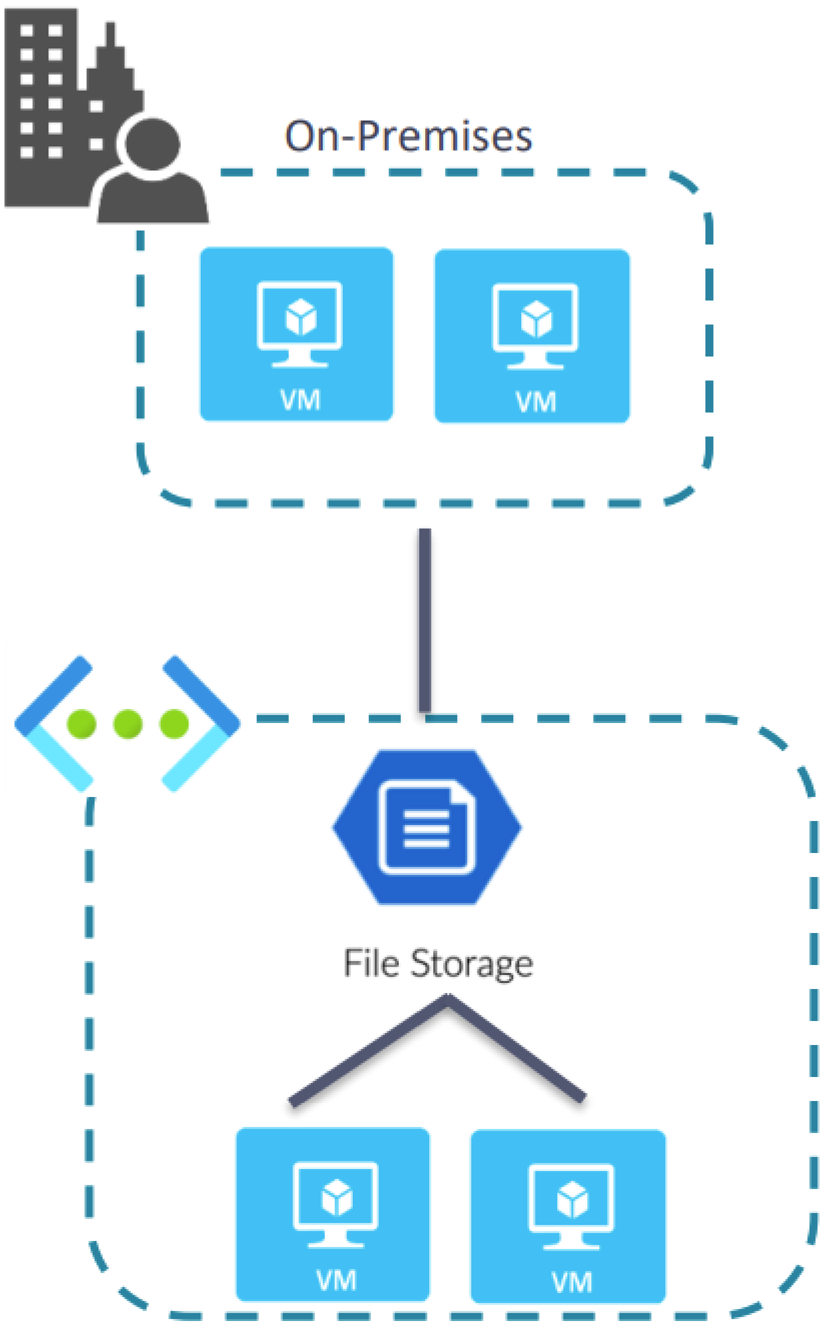
Azure Queue Storage

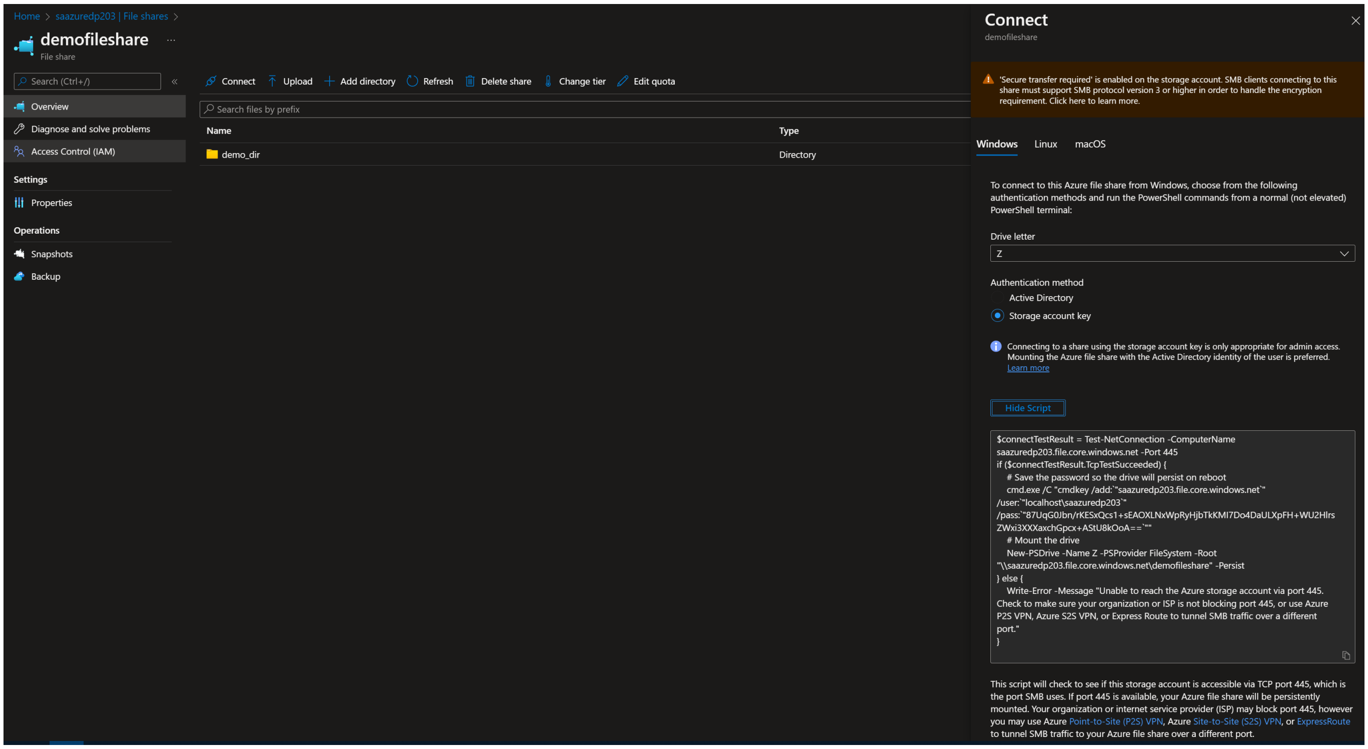
- Message queuing service to store large number of messages
- Access messages via authenticated calls using HTTP or HTTPS

- Commonly used to create a backlog of work to process asynchronously

Azure File Share Storage

- Enables you to create files shares in the cloud, and access these file shares from anywhere with an internet connection
- Mounted concurrently by cloud or on-premises deployments
- Accessible from Windows, Linux, and macOS clients
- Accessible Server Message Block (SMB) protocol or Network File System (NFS) protocol
  - *For exam purposes: If SMB is in the question, Azure File Share is likely the answer*
- Azure Files ensures the data is encrypted at rest, and the SMB protocol ensures the data is encrypted in transit
- Use Cases
  - Replace or supplement on-premises file servers
  - Share application settings
  - Dev/Test/Debug
- Key Benefits
  - Shared access: Replace on-premises file shares with Azure file shares without application compatibility issues
  - Fully managed: Azure will manage hardware or an OS
  - Resiliency: you don’t have to deal with local power and network issues.


On-Premises
VM
VM
File Storage
VM
VM



Name	Type
demo_dir	Directory

Connect

demofileshare

'Secure transfer required' is enabled on the storage account. SMB clients connecting to this share must support SMB protocol version 3 or higher in order to handle the encryption requirement. Click here to learn more.

To connect to this Azure file share from Windows, choose from the following authentication methods and run the PowerShell commands from a normal (not elevated) PowerShell terminal:

Drive letter: Z

Authentication method

○ Active Directory

● Storage account key

Connecting to a share using the storage account key is only appropriate for admin access. Mounting the Azure file share with the Active Directory identity of the user is preferred. Learn more

```
$connectTestResult = Test-NetConnection -ComputerName
saazuredp203.file.core.windows.net -Port 445
if ($connectTestResult.TcpTestSucceeded) {
    # Save the password so the drive will persist on reboot
    cmd.exe /C "cmdkey /add:`"saazuredp203.file.core.windows.net`"
/user:`"localhost\saazuredp203`"
/pass:`"87UqG0Jbn/rKESxQcs1+sEAOXLNxWpRyHjbTkKMI7Do4DaULXpFH+WU2Hlrs
ZWxi3XXXaxchGpcx+AStU8kOoA==`""
    # Mount the drive
    New-PSDrive -Name Z -PSProvider FileSystem -Root
"\\saazuredp203.file.core.windows.net\demofileshare" -Persist
} else {
    Write-Error -Message "Unable to reach the Azure storage account via port 445.
Check to make sure your organization or ISP is not blocking port 445, or use Azure
P2S VPN, Azure S2S VPN, or Express Route to tunnel SMB traffic over a different
port."
}
```

This script will check to see if this storage account is accessible via TCP port 445, which is the port SMB uses. If port 445 is available, your Azure file share will be persistently mounted. Your organization or internet service provider (ISP) may block port 445, however you may use Azure Point-to-Site (P2S) VPN, Azure Site-to-Site (S2S) VPN, or ExpressRoute to tunnel SMB traffic to your Azure file share over a different port.

Azure Disk Storage

- VM uses disks as a place to store an operating system, applications, and data in Azure
- One virtual machine can have one OS disk and multiple data disk but one data disk can only be link with one VM
- Both the OS disk and the data disk are virtual hard disks (VHDs) stored in an Azure storage account
- The VHDs used in Azure are .vhd files stored as page blobs in a standard or premium storage account in Azure
- Unmanaged disks: We can create a storage account and specify it when we create the disk
    - Not recommended, previous unmanaged disks should migrate to managed disk
- Managed disk
    - Azure creates and manages storage accounts in the background
    - We don't have to worry about scalability issues
    - Azure creates and manages the disk for us based on the size and performance tier we specify

- Managed Disk types:
  - Standard HDD: Backup, non-critical, infrequent access
  - Standard SSD: Lightly used production applications or dev/test environments
  - Premium SSD disks: Super fast and high performance, very low latency, recommended for production and performance sensitive workloads
  - Ultra disks (SSD): For most demanding IO-intensive workloads such as SAP HANA, top tier databases (for example, SQL, Oracle), and other transaction-heavy workloads

Azure Data Lake Gen 2 (Storage Account)

> “If you think of a DataMart as a store of bottled water – clean and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.”

- Data Lake is a large container (repository) to store raw data
  - Structured, semi-structured, unstructured, stream, and batch data store (any type of data)
  - There is no limit to the amount of data that can be stored in a data lake
- Data warehouses store transformed data that is ready to be consumed
- Data Lakes store raw data that is ready for exploration
- Data Lake Gen 2 is a combination of Azure Blob Storage and Data Lake Gen 1
  - Data Lake Gen 1 was primarily Hadoop Distributed File System (HDFS) which was revolutionary but had limitations

Blob Storage vs Data Lake Gen 2 Storage

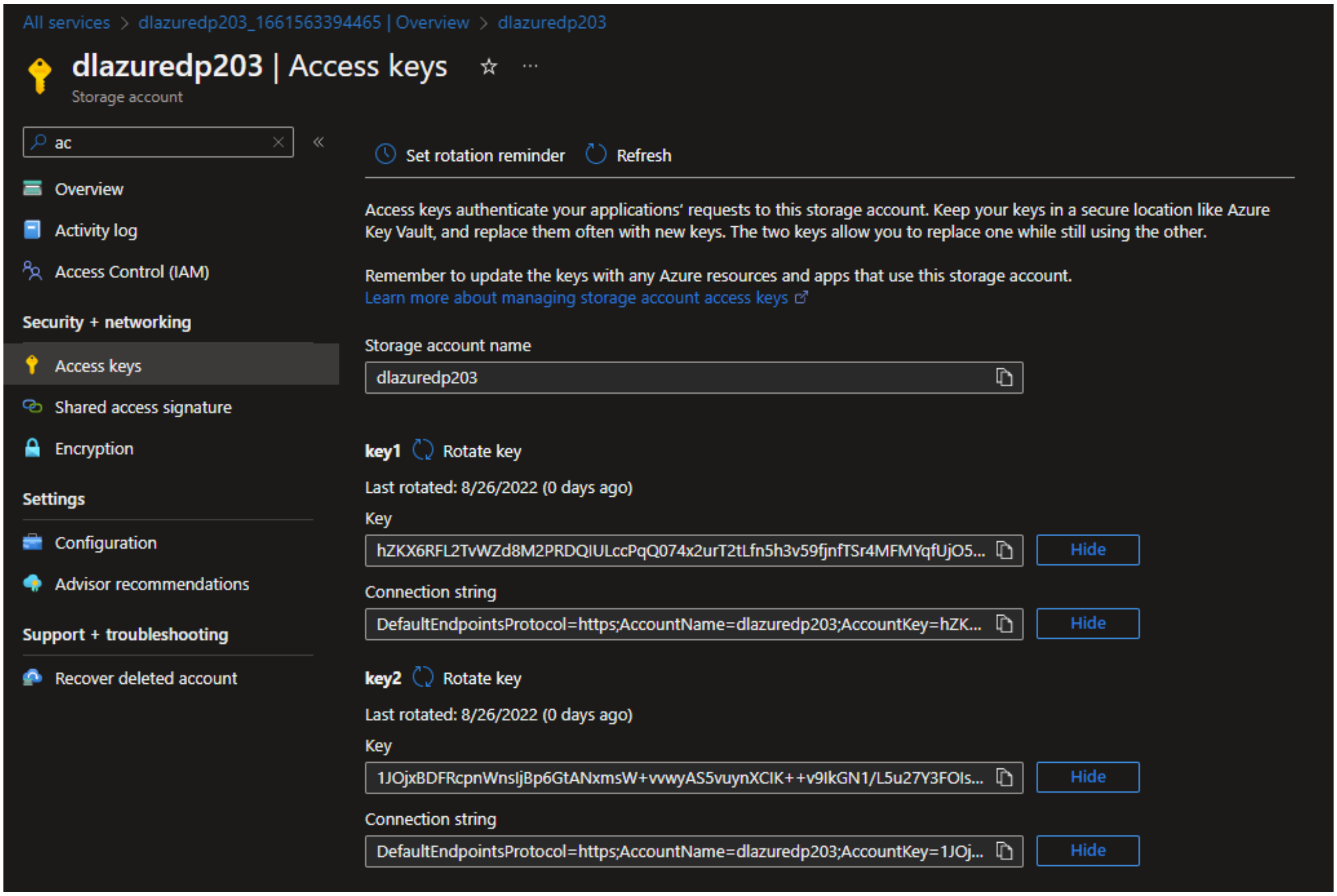
- Azure Blob Storage
  - General purpose data storage (not optimized for big data)
  - Container based object storage
  - Available in every Azure region
  - Local and global redundancy
  - Processing performance limit
- Azure Data Lake Gen 2
  - Optimized for big data analytics
  - Hierarchical namespace on Blob Storage
  - Available in every Azure region
  - Local and global redundancy
  - Supports a subset of Blob storage features
  - Supports multiple Azure integrations (Synapse, Databricks, etc.)
  - Compatible with Hadoop

> Data Lake Gen 2 is built on top of Blob Storage and retains most, but not all, of its features.

Data Lake Security Options

- Authentication
  - Storage Account Keys (Access Keys)
    - No longer recommended to use in production
  - Shared Access Signature (SAS Token)
    - Best practice to work on principle of least privilege (provided users with the minimum permissions required to complete assignment)
    - Contains permissions such as start and end time, service restrictions, resource types, and permissions (read, write, etc.)
    - SAS Token's are not tracked by Azure after creation
    - To invalidate a token, simply regenerate the storage account
    - Token are associated with Access Keys, so if an Access Key is regenerated, the SAS Token will no longer work

    - Azure Active Directory (Azure AD)
        - Identity management solution (we can create identities inside the service such as users, groups, service principals, etc.)
        - Users and groups can be assigned roles within a storage account under the Access Control (IAM) option
- Access Control
    - Role Based Access Control (RBAC)
    - Access Control List (ACL)
- Network Access
    - Firewall and Virtual Network
        - Every object in a storage account has a URL endpoint which can be accessed from anywhere in the world with the proper permissions by
            - IP Addresses
            - Virtual Networks
            - Internet
        - Access to storage account can be limited to specific or ranges of IP addresses, and/or specific virtual networks
- Data Protection
    - Data Encryption in Transit
    - Data Encryption at Rest
- Advanced Threat Protection



dlazuredp203 | Access keys

Storage account

Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location like Azure Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the other.

Remember to update the keys with any Azure resources and apps that use this storage account.
Learn more about managing storage account access keys

Storage account name
dlazuredp203

key1

Last rotated: 8/26/2022 (0 days ago)

Key
hZKX6RFL2TvWZd8M2PRDQIULccPqQ074x2urT2tLfn5h3v59fjnfTSr4MFMYqfUjO5...

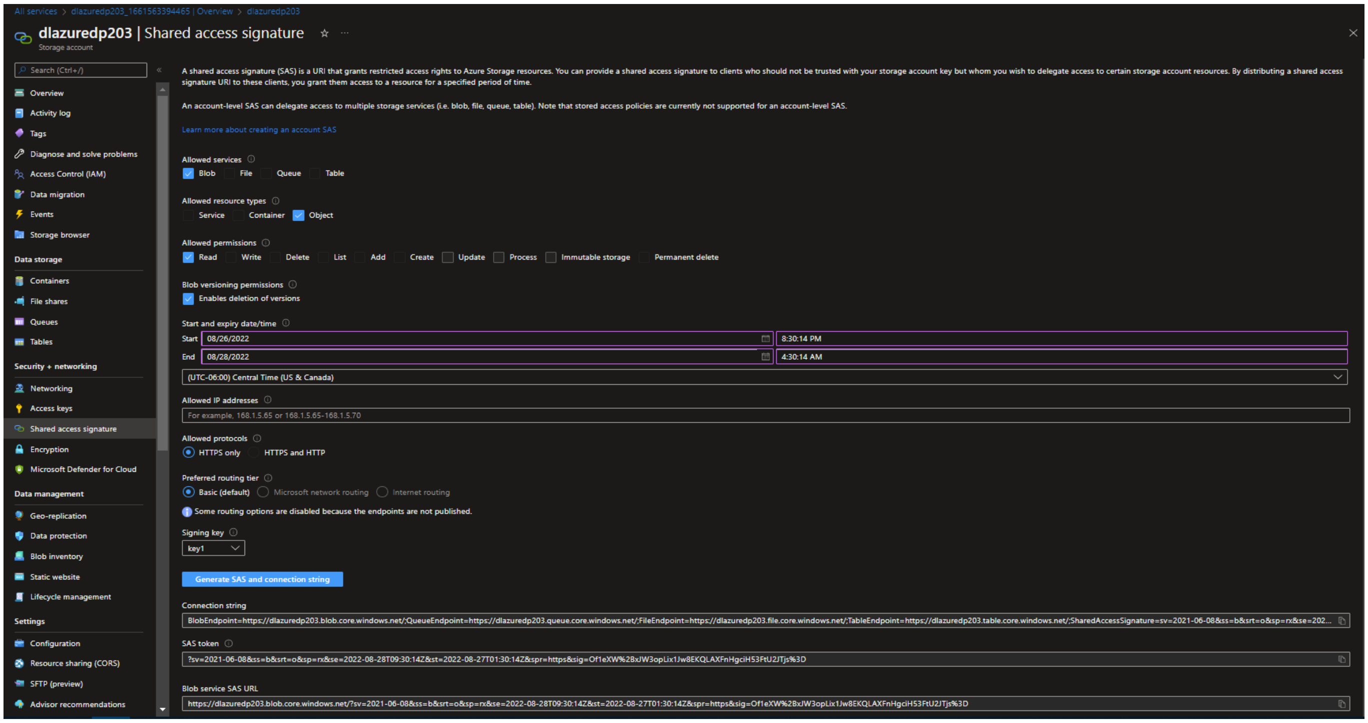
Connection string
DefaultEndpointsProtocol=https;AccountName=dlazuredp203;AccountKey=hZK...

key2

Last rotated: 8/26/2022 (0 days ago)

Key
1JOjxBDFRcpnWnsljBp6GtANxmsW+vwyAS5vuynXCIK++v9lkGN1/L5u27Y3FOls...

Connection string
DefaultEndpointsProtocol=https;AccountName=dlazuredp203;AccountKey=1JOj...



dlazuredp203 | Shared access signature

Storage account

A shared access signature (SAS) is a URI that grants restricted access rights to Azure Storage resources. You can provide a shared access signature to clients who should not be trusted with your storage account key but whom you wish to delegate access to certain storage account resources. By distributing a shared access signature URI to these clients, you grant them access to a resource for a specified period of time.

An account-level SAS can delegate access to multiple storage services (i.e. blob, file, queue, table). Note that stored access policies are currently not supported for an account-level SAS.

Learn more about creating an account SAS

Allowed services
☑ Blob □ File □ Queue □ Table

Allowed resource types
□ Service □ Container ☑ Object

Allowed permissions
☑ Read □ Write □ Delete □ List □ Add □ Create □ Update □ Process □ Immutable storage □ Permanent delete

Blob versioning permissions
☑ Enables deletion of versions

Start and expiry date/time
Start 08/26/2022 8:30:14 PM
End 08/28/2022 4:30:14 AM
(UTC-06:00) Central Time (US & Canada)

Allowed IP addresses
For example, 168.1.5.65 or 168.1.5.65-168.1.5.70

Allowed protocols
● HTTPS only ○ HTTPS and HTTP

Preferred routing tier
● Basic (default) ○ Microsoft network routing ○ Internet routing

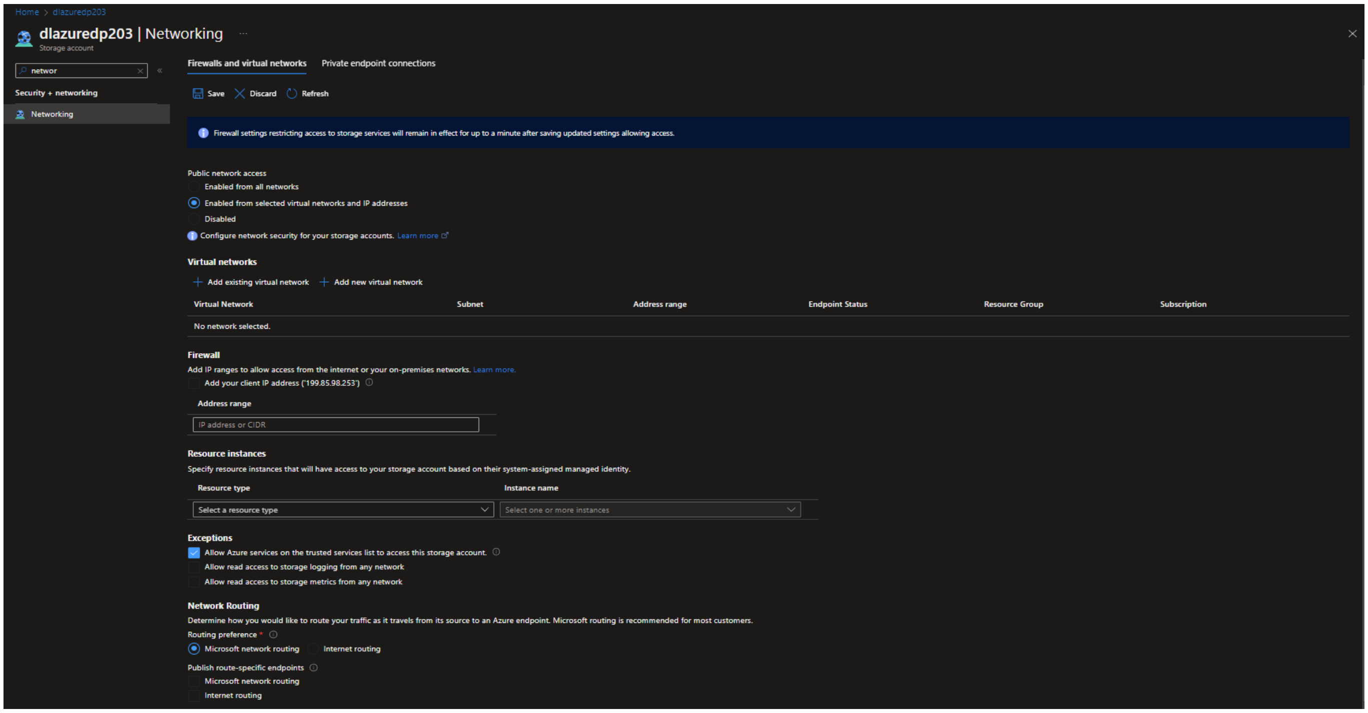
Some routing options are disabled because the endpoints are not published.

Signing key
key1

Connection string
BlobEndpoint=https://dlazuredp203.blob.core.windows.net/;QueueEndpoint=https://dlazuredp203.queue.core.windows.net/;FileEndpoint=https://dlazuredp203.file.core.windows.net/;TableEndpoint=https://dlazuredp203.table.core.windows.net/;SharedAccessSignature=sv=2021-06-08&ss=b&srt=o&sp=rx&se=202...

SAS token
?sv=2021-06-08&ss=b&srt=o&sp=rx&se=2022-08-28T09:30:14Z&st=2022-08-27T01:30:14Z&spr=https&sig=Of1eXW%2BxJW3opLix1Jw8EKQLAXFnHgciH53FtU2JTjs%3D

Blob service SAS URL
https://dlazuredp203.blob.core.windows.net/?sv=2021-06-08&ss=b&srt=o&sp=rx&se=2022-08-28T09:30:14Z&st=2022-08-27T01:30:14Z&spr=https&sig=Of1eXW%2BxJW3opLix1Jw8EKQLAXFnHgciH53FtU2JTjs%3D



Firewall settings restricting access to storage services will remain in effect for up to a minute after saving updated settings allowing access.

Public network access

○ Enabled from all networks
● Enabled from selected virtual networks and IP addresses
○ Disabled

Configure network security for your storage accounts. Learn more

Virtual networks

+ Add existing virtual network + Add new virtual network

Virtual Network	Subnet	Address range	Endpoint Status	Resource Group	Subscription
No network selected.					

Firewall

Add IP ranges to allow access from the internet or your on-premises networks. Learn more.

□ Add your client IP address ('199.85.98.253')

Address range

IP address or CIDR

Resource instances

Specify resource instances that will have access to your storage account based on their system-assigned managed identity.

Resource type	Instance name
Select a resource type	Select one or more instances

Exceptions

☑ Allow Azure services on the trusted services list to access this storage account.
□ Allow read access to storage logging from any network
□ Allow read access to storage metrics from any network

Network Routing

Determine how you would like to route your traffic as it travels from its source to an Azure endpoint. Microsoft routing is recommended for most customers.

Routing preference *

● Microsoft network routing ○ Internet routing

Publish route-specific endpoints

□ Microsoft network routing
□ Internet routing

High Availability and Disaster Recovery

- High Availability
  - Making a service available within a region
  - If one instance goes down, another will pick it up
  - No expected data loss
- Disaster Recovery
  - Recovery from site/region level event
  - Typically some data loss
  - Settings accessible in Configuration, Data Protection, and Geo-Replication menus within Storage Account

Azure CosmosDB

No longer part of DP-203 exam