DP-203 Microsoft Azure Data Engineer

Day 1 – Data Lake Gen2

25th July 2021

Vinodkumar Bhovi

databoq®

Introduction – what to expect from us

- Structured course designed to pass DP 203
- We deliver cheat sheets at the end of the program
- Slack channel access, PPT’s
- We (databag) have put lot of effort in designing the course
- No donations/fees will ever be asked from databag.ai

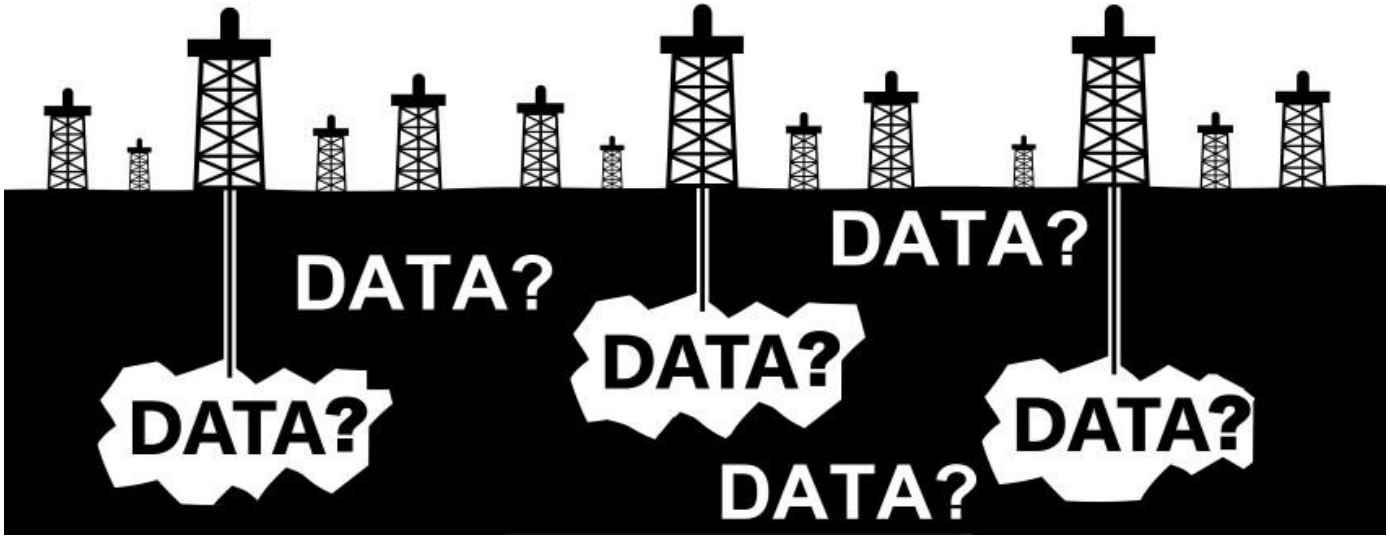
Introduction – what we expect from you

- Free means you've got nothing to lose, keeps you in comfort zone
- We take these trainings serious and expect same from you
- Setup Free Azure Account
- Schedule exam in 21 days – on 14-08-2021
- Consistency
- If you ever used Google drive or facebook or Amazon that is more than enough to learn Azure

databag®

Data is new oil

We need to find it, extract it, refine it, distribute it and monetize it
– David Buckinham, Big data expert



databag®

Data

Data Storage

Azure Storage Accounts

Azure Cosmos DB

Azure Data Lake

Azure SQL

Azure Synapse Analytics

Data Transformation

Azure Data Factory

Azure Stream Analytics

Azure Databricks

Azure HDInsight

databag®

14 days schedule

- Day1: Azure Data Lake - Vinod
- Day2: Azure SQL Databases - Lakshay
- Day3: Azure Cosmos DB - Vinod
- Day4: Azure Cosmos DB - Vinod
- Day5: Azure Data Factory - Vishwamitra
- Day6: Azure Data Factory - Vishwamitra
- Day7: Azure Databricks - Vinod

databag®

14 days schedule *(continued #)*

- Day8: Azure Databricks - Vinod
- Day9: Azure HDInsight - Vinod
- Day10: Azure Stream Analytics - Vinod
- Day11: Azure Synapse Analytics - Vinod
- Day12: Azure Synapse Analytics - Vinod
- Day13: Azure Synapse Analytics - Vinod
- Day14: Practice Exam(61) & Q/A – databag team

Cloud 101

The practice of using a network of remote servers hosted on the internet to store, manage, and process data, rather than a local server or a personal computer.

- **Infrastructure as a service** (IaaS)
  - You rent a virtual server
  - Amazon, Azure, GCP, etc.
- **Platform as a service** (PaaS)
  - You rent an abstract machine
  - Google app engine, Salesforce, etc.
- **Software as a service** (SaaS)
  - You rent a capability
  - Azure SQL, ADF, etc.

Data

Data Storage

Azure Storage Accounts

Azure Cosmos DB

Azure Data Lake

Azure SQL

Azure Synapse Analytics

Data Transformation

Azure Data Factory

Azure Stream Analytics

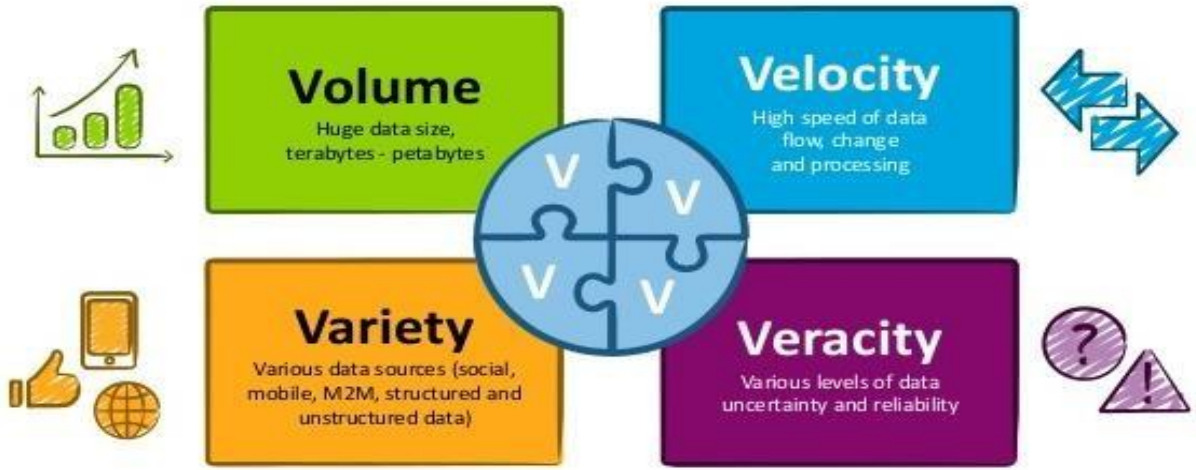
Azure Databricks

Azure HDInsight

databag®

Problem statement

- Need a solution which can handle below 4 V's of data.



Data Classification

- **Structured data**

Examples: SQL data, Tabular data, csv, spreadsheets

- **Semi - structured data**

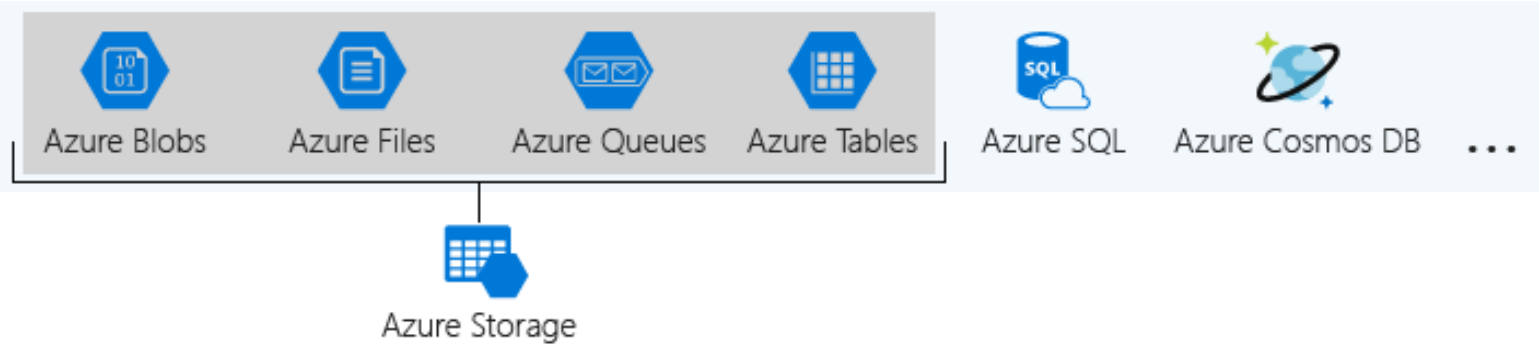
Examples: NoSQL, key/value pairs, JSON, XML, YAML

- **Unstructured data**

Examples: Media files, Office files, Text files, Log files

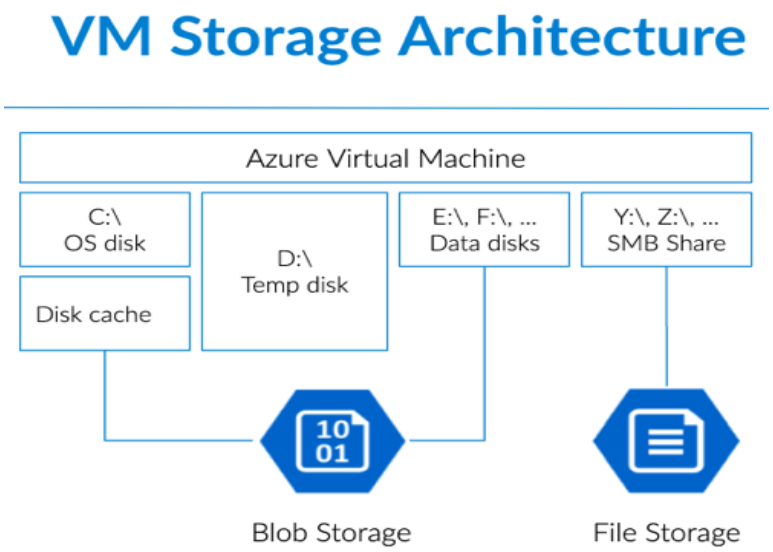
Azure Storage

Azure Storage is a Microsoft-managed cloud service that provides storage that is highly available, secure, durable, scalable and redundant. Within Azure there are two types of storage accounts, four types of storage, four levels of data redundancy and three tiers for storing files



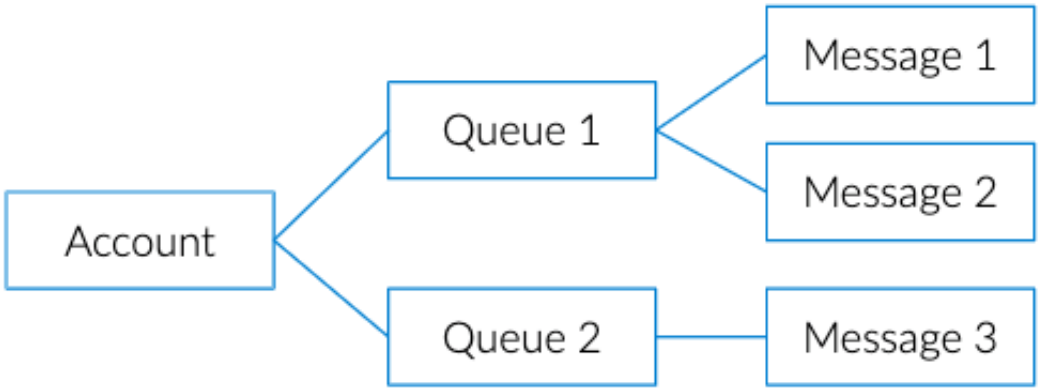
Azure File Storage

Azure Files is a shared network file storage service that provides administrators a way to access native SMB file shares in the cloud



Azure Queue Storage

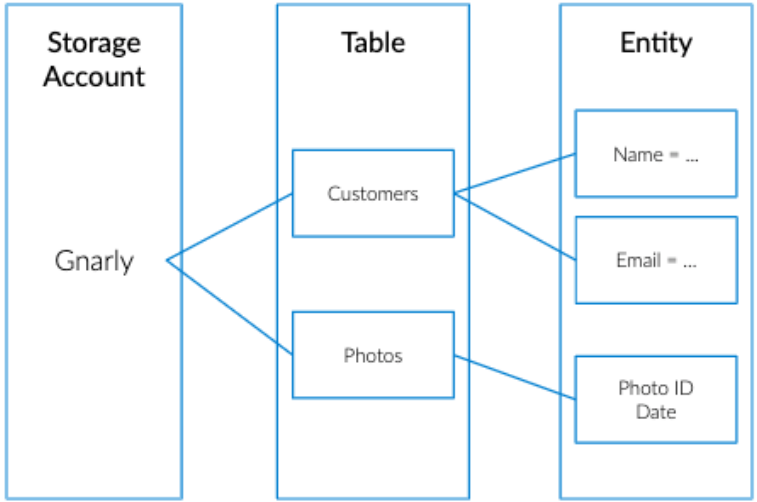
Azure Queue Storage is a service that allows users to store high volumes of messages, process them asynchronously and consume them when needed



databag®

Azure Table Storage

Azure Table Storage is a scalable, NoSQL, key-value data storage system that can be used to store large amounts of data in the cloud. This storage offering has a schema less design, and each table has rows that are composed of key-value pairs

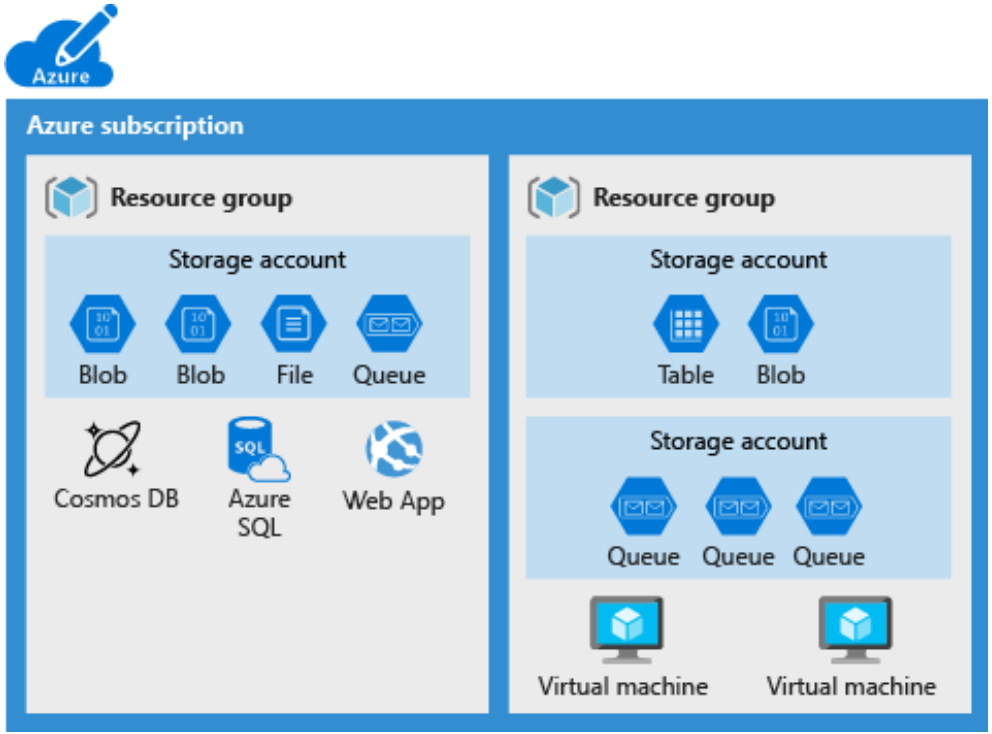


Azure Blob Storage

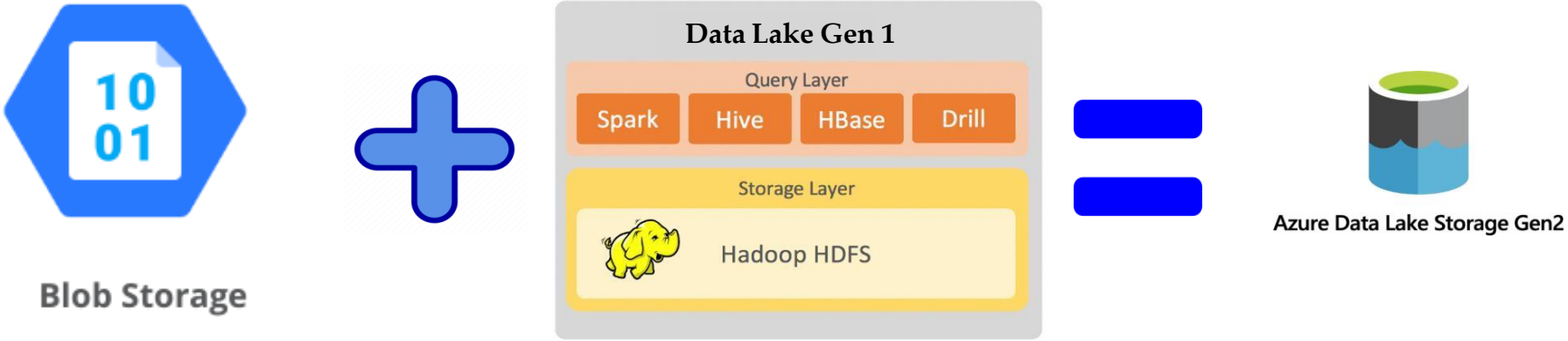
Azure Blob Storage is Microsoft Azure's service for storing binary large objects or blobs which are typically composed of unstructured data such as text, images, and videos, along with their metadata. Blobs are stored in directory-like structures called "containers."

- Large object storage in cloud
- Optimized for storing massive amounts of unstructured data
    - Text or Binary Data
- General purpose object storage
- Cost efficient
- Provide multiple Tiers

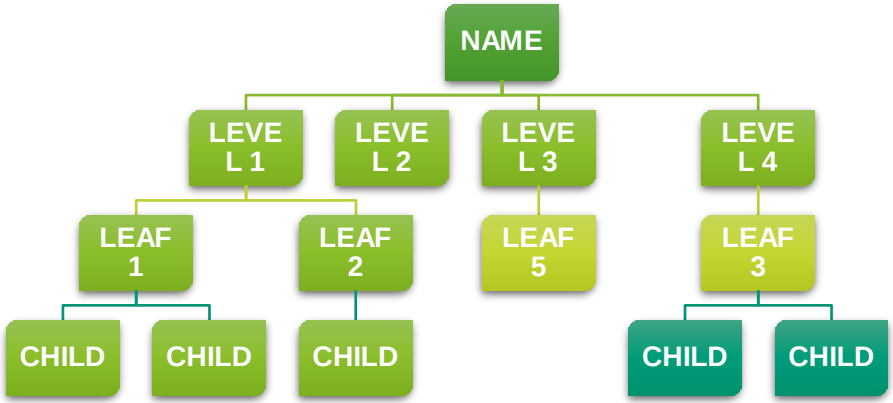
Azure Service Hierarchy

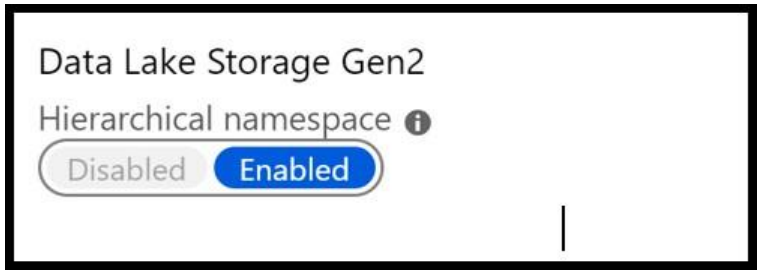


Azure Data lake Gen 2



Hierarchical namespace (demo)

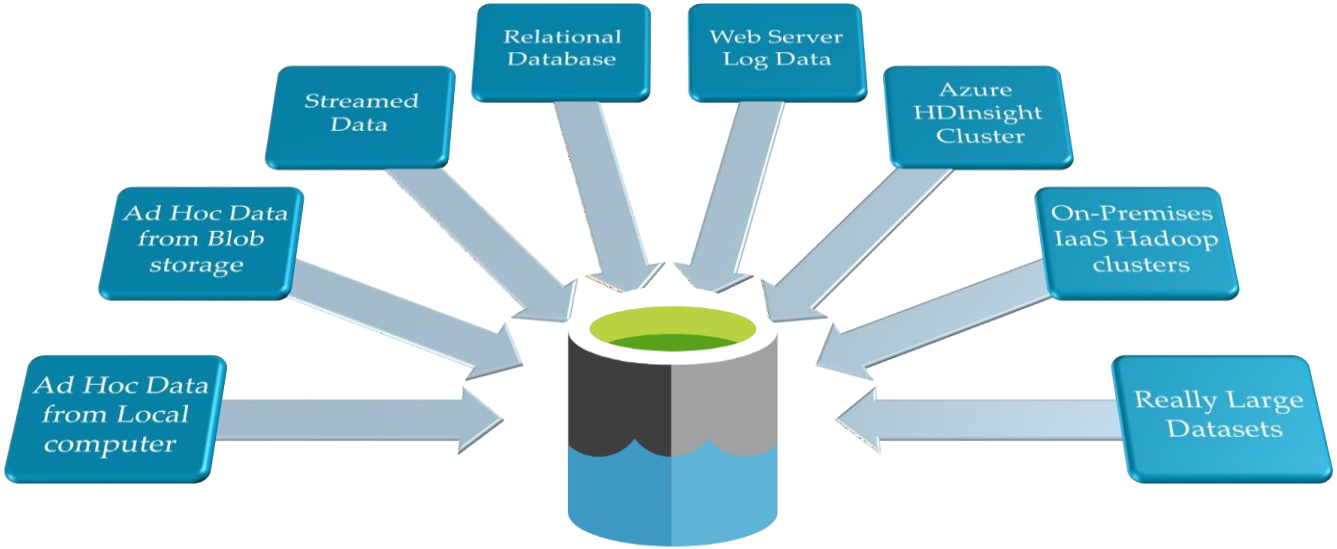




- Hierarchical namespace organizes objects/files into a hierarchy of directories for efficient data access.
- Blob storage is not hierarchical namespace
  - Use slashes in Blob storage file names to stimulate a tree like directory structure
- Blob can't integrate with Hadoop
  - Because Blob doesn't have hierarchical namespace

databag®

Data Ingestion

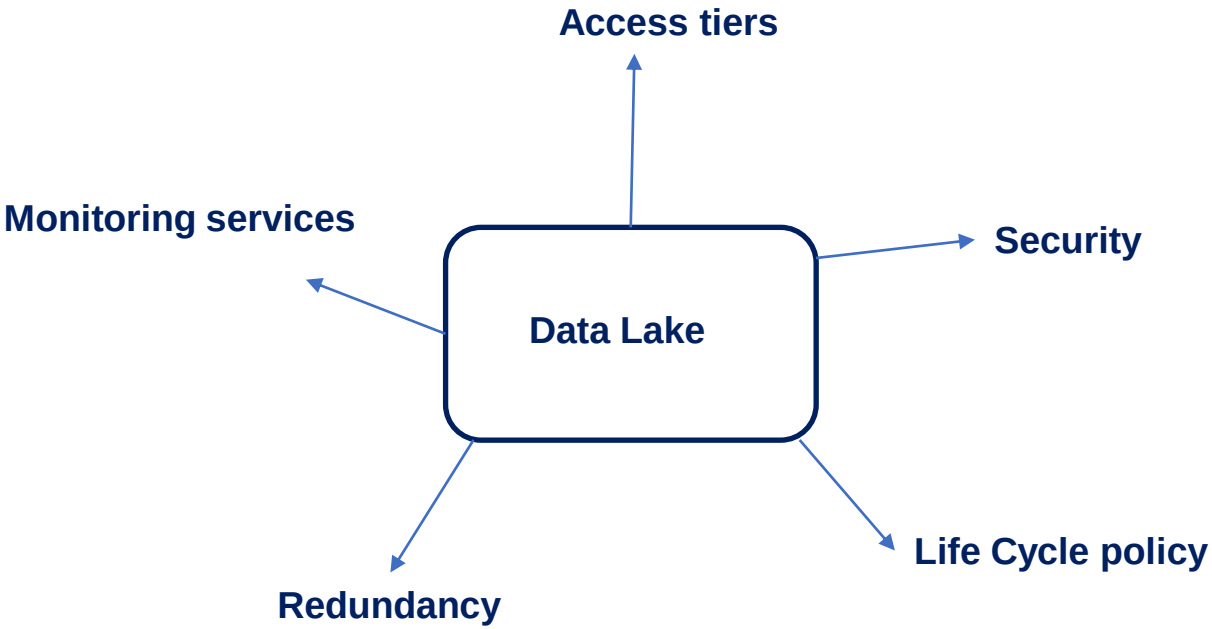


Data Lake

Azure Data Lake

databag®

Azure Data Lake



Access Tiers

- **Hot** - Optimized for storing data that is accessed frequently.
- **Cool** - Optimized for storing data that is infrequently accessed and stored for at least 30 days (early deletion fee).
- **Archive** - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours (early deletion fee).

databag®

Access Tiers

- **Archive** - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours.

  To read data in archive storage, you must first change the tier of the blob to hot or cool. This process is known as **rehydration** and can take hours to complete

  - **Standard priority**: The rehydration request will be processed in the order it was received and may take up to 15 hours.
  - **High priority**: The rehydration request will be prioritized over Standard requests and may finish in under 1 hour for objects under ten GB in size.

Access Tiers

Data storage prices pay-as-you-go

All prices are per GB per month.

	Premium	Hot	Cool	Archive
First 50 terabyte (TB) / month	$0.15 per GB	$0.018 per GB	$0.01 per GB	$0.00099 per GB
Next 450 TB / month	$0.15 per GB	$0.0173 per GB	$0.01 per GB	$0.00099 per GB
Over 500 TB / month	$0.15 per GB	$0.0166 per GB	$0.01 per GB	$0.00099 per GB

databag®

Access Tiers

Operations and data transfer

	Premium	Hot	Cool	Archive
Write operations (per 10,000)[1]	$0.0228	$0.065	$0.13	$0.13
Read operations (per 10,000)[2]	$0.0019	$0.005	$0.013	$6.50
Iterative Read Operations (per 10,000)[3]	N/A	$0.005	$0.013	$6.50
Iterative Write Operations (100's)[4]	N/A	$0.065	$0.13	$0.13
Data Retrieval (per GB)	N/A	N/A	$0.01	$0.02
Data Write (per GB)	Free	Free	Free	Free
Index (GB/month)	N/A	$0.026	N/A	N/A
All other Operations (per 10,000), except Delete, which is free	$0.0019	$0.005	$0.013	$6.50

Security

- **Authentication**
  - Storage Account keys
  - Shared access signature (SAS)
  - Azure Active Directory (Azure AD)
- **Access Control**
  - Role based access control (RBAC)
  - Access control list (ACL)
- **Network access**
  - Firewall and virtual network
- **Data Encryption**

Storage Account Access Keys

Authentication

Shared Access Signature (SAS)

Authentication

Shared Access Signature (SAS)

Shared Access Signature

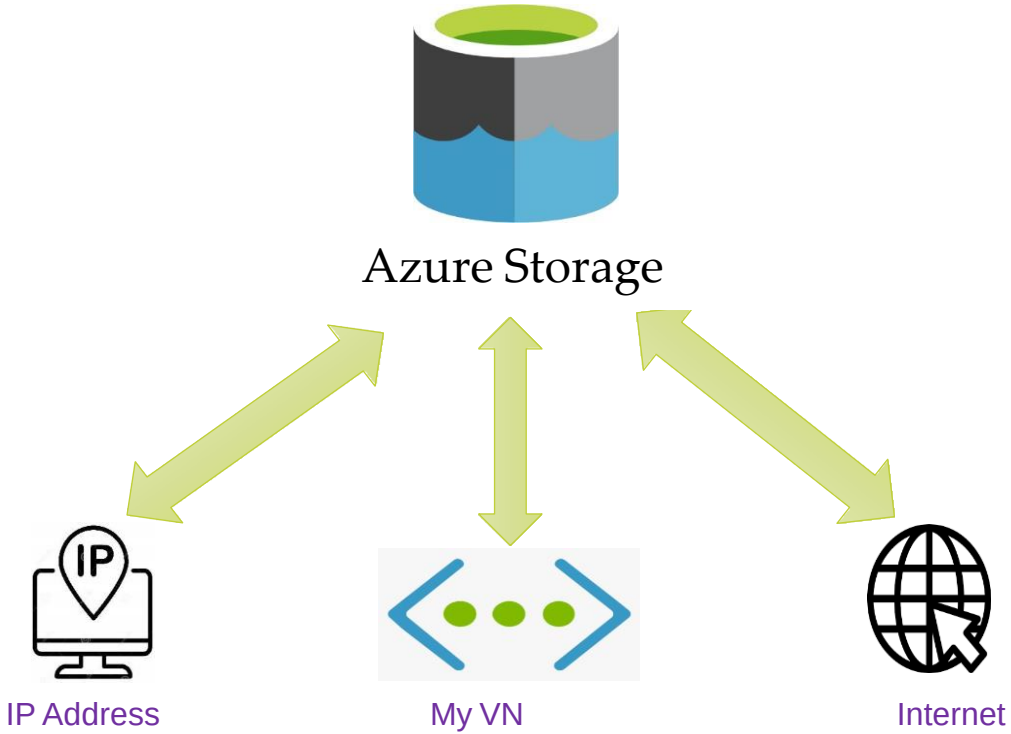
- Security token string
- “SAS Token”
- Contains permission like start and end time
- Azure doesn’t track SAS after creation

To invalidate, regenerate storage account key used to sign SAS

Azure Active Directory (AD)

- Grand access to Azure Active directory (AD) Identities
- AD is an enterprise identity provider, Identity as a Service (IDaaS)
- Globally available from any device
- Identities – user, group or application principle
- Assign role at Subscription, RG, Storage account, container level.
- No longer need to store credentials with application config files
- Like IIS Application pool identity approach
- Role based Access control (RBAC)

Firewalls and Virtual Networks



Security

- **Authentication**
  - Storage Account keys
  - Shared access signature (SAS)
  - Azure Active Directory (Azure AD)
- **Access Control**
  - Role based access control (RBAC)
  - Access control list (ACL)
- **Network access**
  - Firewall and virtual network
- **Data Encryption**

Lifecycle Management

Azure Blob Storage lifecycle management offers a rich, rule-based policy which you can use to transition your data to the best access tier and to expire data at the end of its lifecycle.

Lifecycle management policy helps you:

- Transition blobs to a cooler storage tier such as hot to cool, hot to archive, or cool to archive in order to optimize for performance and cost
- Delete blobs at the end of their lifecycles
- Define up to 100 rules
- Run rules automatically once a day
- Apply rules to containers or specific subset of blobs, up to 10 prefixes per rule

databag®

Data redundancy for storage

LRS – Locally-redundant storage: Three copies of your data which is maintained within the same primary data center.

ZRS- Zone-redundant storage: Three copies of your data replicated synchronously to 3 Azure availability zones in a primary region. Zones are in different physical locations or different data centers.

GRS- Geo-redundant storage: This allows your data to be stored in different geographic areas of the country or world. Again, you get three copies of the data within a primary region, but it goes one step further and places three additional asynchronous copies in another region. For example, you can now have a copy in Virginia and in California to protect your data from fires or hurricanes depending on the coast.

RA-GRS- Read Access Geo-redundant storage: This is GRS but adds a read-only element that allows you to have read access for things like reporting.

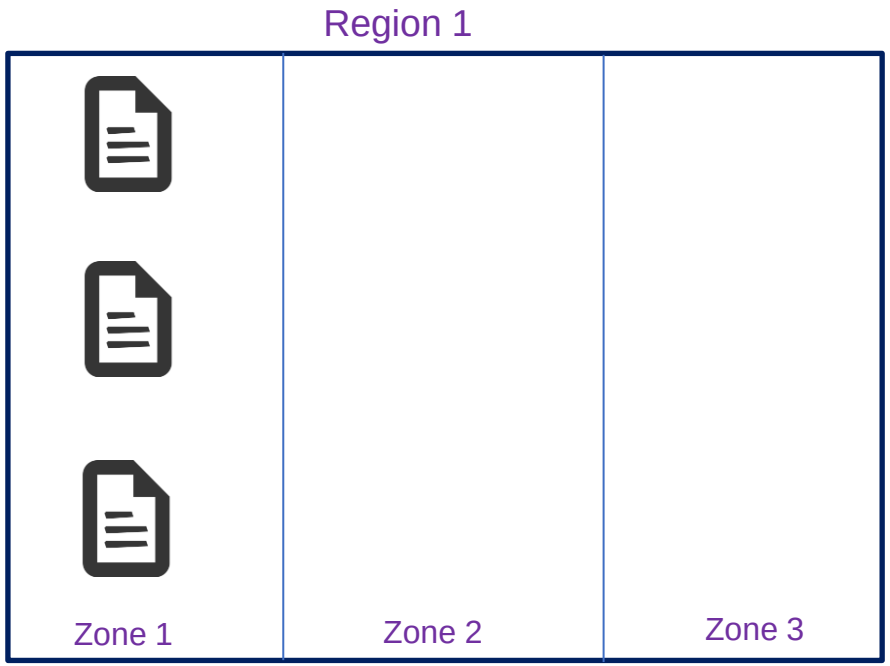
Geo-zone-redundant storage (GZRS): Copy your data synchronously over three primary region Azure availability zones using ZRS. It then asynchronously copies your data to a single physical location within the secondary region.

Read-access geo-zone-redundant storage (RA-GZRS): it adds a layer of readability to your secondaries.

databag®

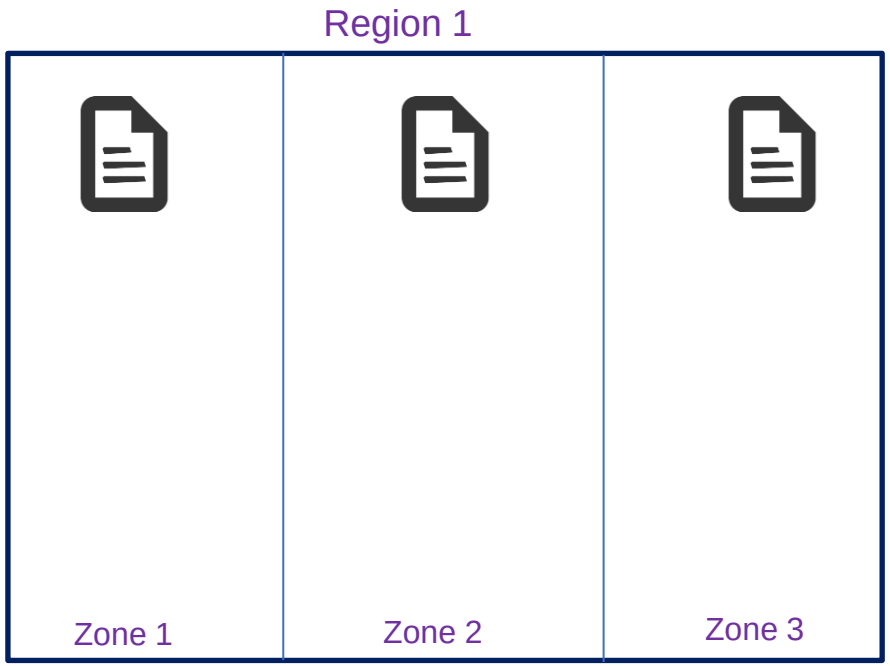
Data redundancy for storage

Locally Redundant Storage (LRS)



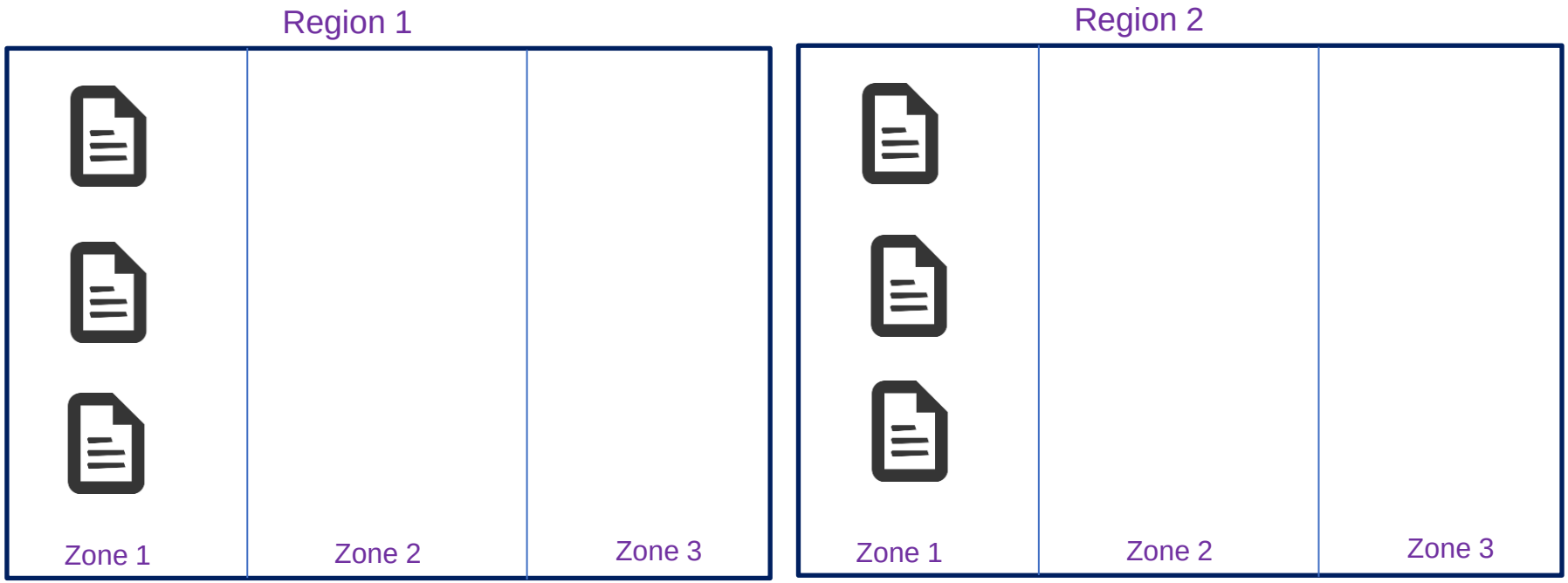
Data redundancy for storage

Zone Redundant Storage (ZRS) --- HA



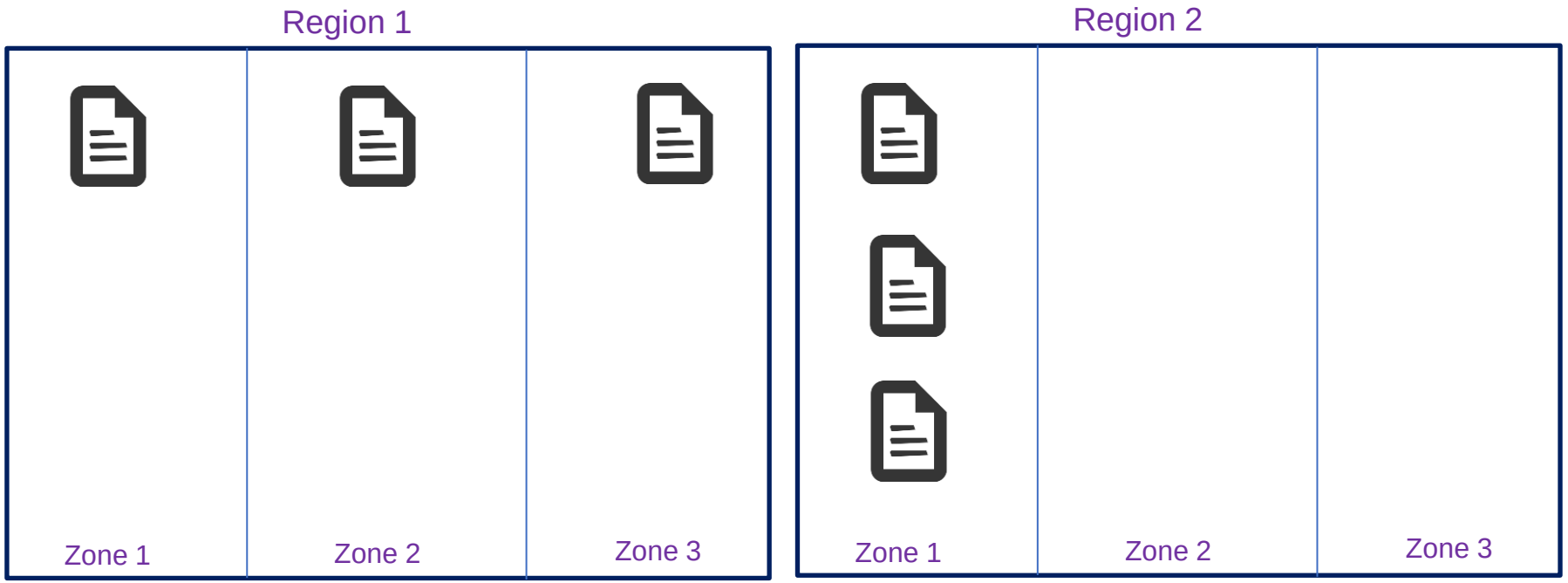
Data redundancy for storage

Geo Redundant Storage (GRS) --- DR



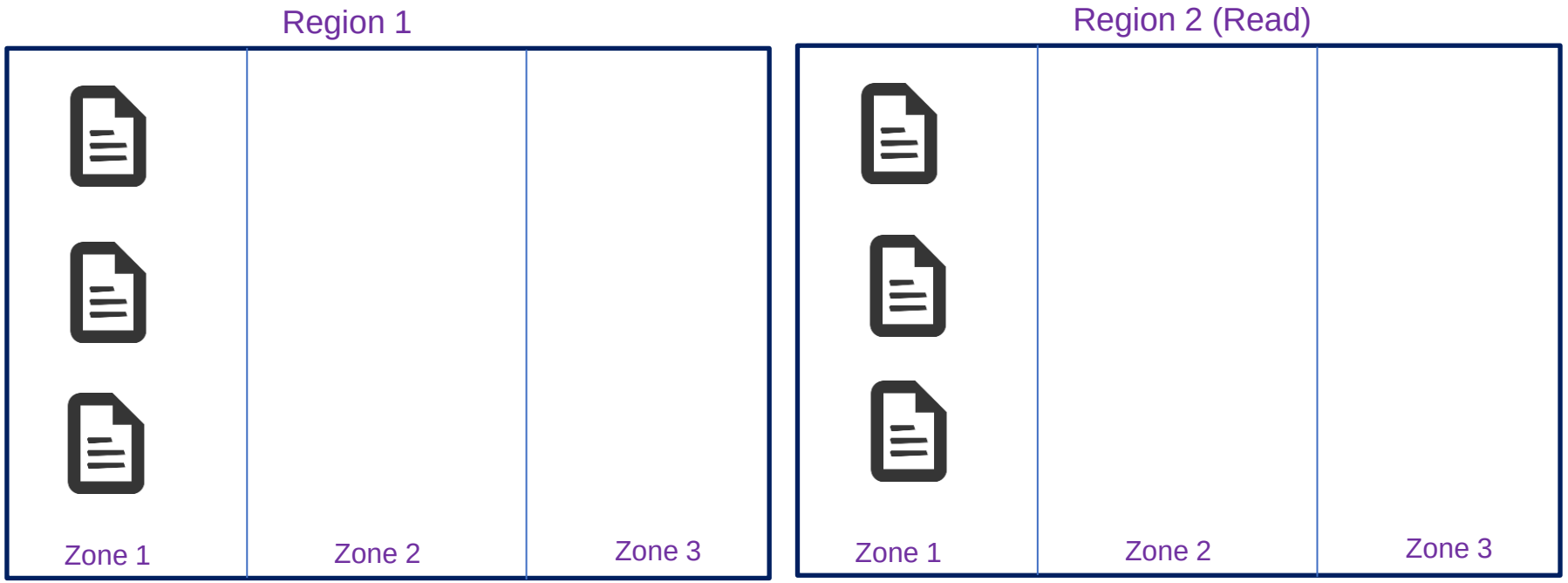
Data redundancy for storage

Geo Zone Redundant Storage (GZRS) – HA/DR



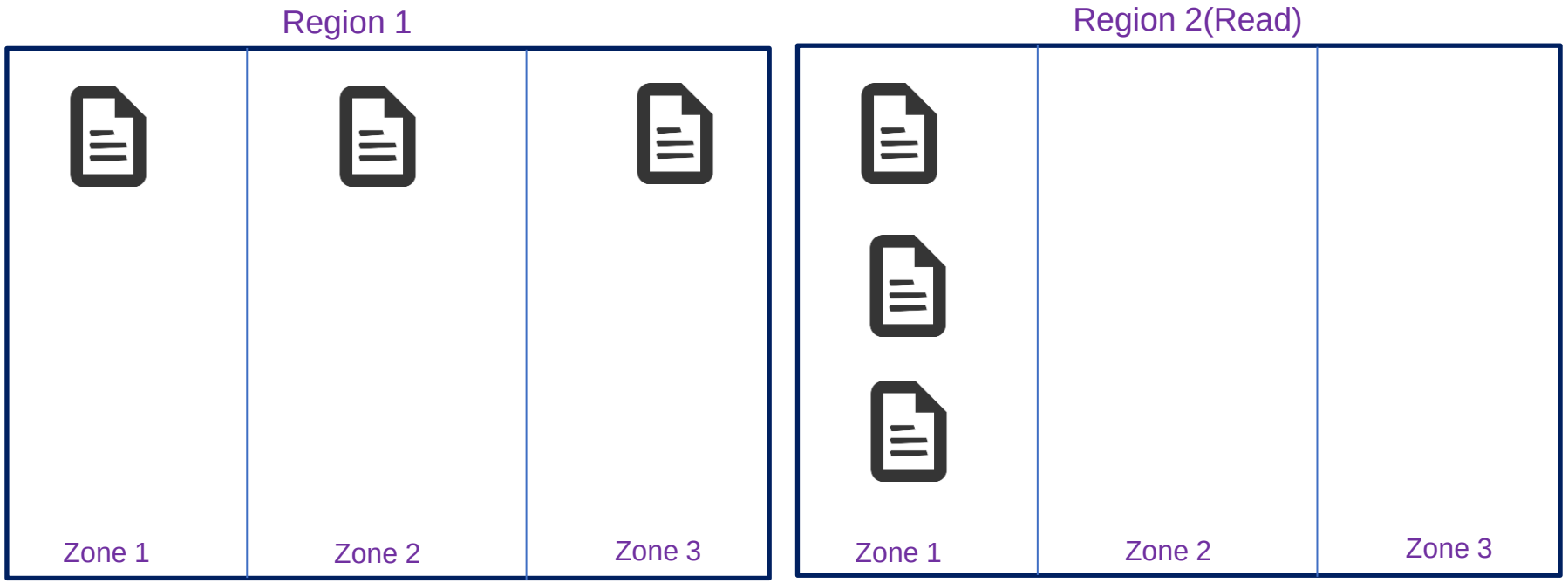
Data redundancy for storage

Read Access Geo Redundant Storage (RA-GRS)



Data redundancy for storage

Read Access Geo Zone Redundant Storage (RA-GZRS) – HA/DR



Data redundancy for storage

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)

databag®

Monitoring services

- Alerts
- Metrics
- Diagnostics
- Logs Analytics

databag®

databag®