**Study Guide: Azure Storage for Data Engineering**

**1. Azure Storage Account Types**

* **General Purpose v1 (GPv1)**: Legacy, basic features, limited scalability.
* **General Purpose v2 (GPv2)**: Modern workloads, supports all features, best for analytics & AI.
* **Blob Storage Account**: Optimized for unstructured data (blobs), simple storage for files.

**2. Blob Storage Concepts**

* **Container**: Top-level folder for blobs.
* **Blobs**: Types:
  + **Block Blob** – Store text & binary files, large data upload in blocks.
  + **Append Blob** – Optimized for append-only operations (logs).
  + **Page Blob** – Random read/write, used for VMs.

**3. Azure Data Lake Storage Gen2 (ADLS Gen2)**

* Combines **Blob Storage** + **Hierarchical Namespace**.
* Allows file/folder-like structure in cloud storage.
* Enables efficient analytics on large datasets.
* Recommended folder structure: /bronze, /silver, /gold.

**4. Other Storage Services**

* **Table Storage**: NoSQL, key-value pairs, fast lookup.
* **Queue Storage**: Messaging between apps/services, asynchronous.
* **Azure Files**: Managed file shares accessible via SMB/NFS.

**5. Authentication & Authorization**

* **Shared Key**: Full access with storage account key.
* **SAS Tokens**: Time-bound, granular access to specific resources.
* **Azure AD Integration**: Role-based access control (RBAC).
* **RBAC Roles**:
  + Storage Blob Data Reader → Read-only
  + Storage Blob Data Contributor → Read/Write
  + Owner → Full control

**6. Lifecycle Management & Cost Optimization**

* Move blobs between tiers: **Hot → Cool → Archive** based on usage.
* Reduce cost by automatically archiving unused data.

**7. Storage Redundancy Options**

* **LRS**: Locally Redundant Storage – 3 copies in single region.
* **ZRS**: Zone-Redundant Storage – 3 copies across availability zones.
* **GRS**: Geo-Redundant Storage – LRS + replication to secondary region.
* **RA-GRS**: Read-Access GRS – GRS + read access from secondary region.

**8. Monitoring Storage**

* **Metrics**: Usage, capacity, transactions.
* **Diagnostics**: Enable logging, track operations.
* **Integration**: Send logs to Log Analytics / Power BI for reporting.

**9. Integration with Analytics Services**

* **Azure Data Factory (ADF)**: Pipeline ingestion from storage.
* **Databricks**: Read/write ADLS Gen2 for processing & analytics.

**10. Daily Quiz Key Points**

* Use **GPv2** for modern workloads.
* **Append Blob** for append-only logs.
* Hierarchical namespace enables **folder-like management**.
* **GRS / RA-GRS** for geo-redundancy & disaster recovery.
* Azure AD + RBAC provides **least-privilege secure access**.
* Azure Files supports **SMB & NFS protocols**.

**Quick Lab**

Create Resource Group:

az group create --name RGStorage --location eastus

1. Create Storage Account (GPv2 with Hierarchical Namespace):

az storage account create --name mystorageacc --resource-group RGStorage --location eastus --sku Standard\_LRS --kind StorageV2 --hierarchical-namespace true

1. Create Container:

az storage container create --account-name mystorageacc --name bronze

1. Upload a file to container:

az storage blob upload --account-name mystorageacc --container-name bronze --name sample.csv --file sample.csv

1. Configure Lifecycle Rule (Hot → Cool → Archive):

az storage account management-policy create --account-name mystorageacc --policy @policy.json