Azure Fundamentals – Day 3 Summary

1. Azure Storage Account

- The core place where all company data lives in Azure.

- Think of it as a cloud hard drive.

- Supports: Blob Storage, Data Lake Gen2, File Shares.

2. Data Lake Gen2 (Most Important for Data Engineers)

- Hierarchical namespace: real folders, fast renames, fast deletes.

- Hadoop-compatible: optimized for big data processing.

- Supports Parquet, CSV, JSON, Delta.

- Preferred by all modern data engineering pipelines.

3. Storage Structure

storage_account/

container/

folder/

file.csv

4. File Formats

- CSV: simple but slow for big data.

- Parquet: columnar, compressed, fast, analytics-friendly (most recommended).

- JSON: used for logs and API data.

- Delta: used in Databricks/Fabric for ACID tables.

5. Authentication & Access Methods

A. Access Keys

- Full access, not secure, dev only.

B. SAS Tokens

- Temporary, restricted, secure links.

- Used for limited-time sharing.

C. RBAC (Role-Based Access Control)

- Assign permissions to users/services.

- Storage Blob Data Reader/Contributor/Owner.

D. Managed Identity (Best Practice)

- No passwords or keys.

- Azure manages authentication.

- Used for ADF, Databricks, Synapse, Fabric.

6. Azure Data Factory (ADF) Authentication

- Uses Linked Services to connect.

- Authenticating via Managed Identity is recommended.

- Pipelines can read/write Data Lake securely.

7. Databricks Access to Data Lake

- Uses 'abfss://' protocol for Data Lake Gen2.

- Reads data via Spark:

df = spark.read.parquet("abfss://container@account.dfs.core.windows.net/path")

- Auth via Managed Identity preferred.

8. Medallion Architecture (Raw → Processed → Curated)

A. Raw Layer (Bronze)

- Exact source data, no changes.

- Append-only.

B. Processed Layer (Silver)

- Cleaned, validated, deduplicated data.

C. Curated Layer (Gold)

- Business-ready, aggregated datasets.

- Used by Power BI, Fabric, analytics.

Summary:

You learned the fundamentals every Azure Data Engineer must know:

- Storage accounts and containers

- Data Lake Gen2 advantages

- Authentication methods

- ADF and Databricks access

- Modern data architecture (raw/processed/curated)

This is the foundation of Azure data engineering.