DATA ENGINEERING | INTERMEDIATE | 2 DAYS

Azure Data Lake Architecture and Implementation

Design and operate Azure Data Lake Storage Gen2 - zone architecture, RBAC and ACLs, Databricks integration, and lifecycle controls for analytics-ready lake foundations.

Why Azure Data Lake?

Architecture that scales

Design ADLS Gen2 lakes with zones, partitions, and naming that stay maintainable.

Analytics-ready storage

Connect Databricks with abfss:// and Unity Catalog external locations.

Secure by design

Combine RBAC, ACLs, service principals, and managed identities for enterprise access.

Cost and lifecycle control

Apply tiering, retention, versioning, and monitoring for production lakes.

What you'll be able to do

- Understand Data Lake architecture principles and how ADLS Gen2 implements them
- Design a well-structured Data Lake using zones, folder hierarchies, and naming conventions
- Configure access control using both RBAC and ACLs on ADLS Gen2
- Integrate Azure Data Lake Storage Gen2 with Azure Databricks for analytics workloads
- Implement data lifecycle management: tiering, retention policies, and cost optimisation
- Apply data governance and security best practices on a Data Lake
- Register ADLS Gen2 paths as governed storage using Unity Catalog external locations
- Understand the role of the Data Lake in a migration from an on-premise data warehouse

Who this is for

- Data Engineers
- Data Analysts
- Analytics Engineers
- BI Professionals
- Cloud Engineers
- Data Platform Teams

Curriculum

01 Data Lake Principles and Architecture Choices

- What a Data Lake is: schema-on-read, raw data preservation, and scalability
- Data Lake versus Data Warehouse versus Lakehouse: when to use each
- How ADLS Gen2 fits a migration from an on-premise warehouse
- Success criteria for a well-governed analytics lake

02 ADLS Gen2 Hierarchical Namespace and Storage Structure

- How ADLS Gen2 extends Blob Storage with a hierarchical namespace
- Storage account structure: containers, directories, and files
- abfss:// path patterns used by analytics engines
- Provisioning checklist for a training or project ADLS Gen2 account

03 Zone Architecture: Raw, Curated, and Consumption

- Designing Raw (Bronze), Curated (Silver), and Consumption (Gold) zones
- What belongs in each zone and how data should move between them
- Separating landing, cleansing, and serving concerns for multi-team lakes
- Mapping zones to lakehouse Bronze, Silver, and Gold language

04 Folder Design, Partitioning, and Naming Conventions

- Partitioning by date, entity, and source system: best practices and anti-patterns
- Folder hierarchy patterns that keep discovery and jobs predictable
- Naming conventions and metadata standards for a governed Data Lake
- Documenting layout so pipelines and analysts share one mental model

05 File Formats: CSV, JSON, Parquet, and Delta

- Trade-offs for storage size, schema, and query performance
- When CSV or JSON is acceptable in landing versus curated zones
- Why Parquet and Delta dominate analytics and lakehouse workloads
- Choosing formats per zone without locking future consumers out
- **Hands-on:** Provision an ADLS Gen2 account and design a zone-based folder structure

06 Identity, RBAC, and ACLs on ADLS Gen2

- Entra ID (Azure AD) integration for identity and access
- RBAC built-in roles: Storage Blob Data Reader, Contributor, Owner, and when to use each
- ACLs at directory and file level, and how they differ from RBAC
- Combining RBAC and ACLs for enterprise Data Lake access patterns

07 Service Principals, Managed Identities, and Databricks Access

- Service principals and managed identities for pipelines and Databricks clusters
- Mounting or accessing ADLS Gen2 with service principal auth and credential passthrough
- Reading and writing from Databricks using the abfss:// protocol
- Unity Catalog external locations for governed ADLS Gen2 paths

08 Lifecycle Management, Retention, and Monitoring

- Access tier policies: Hot, Cool, Archive, and automated tiering rules
- Retention policies, soft delete, and versioning for data protection
- Monitoring with Azure Monitor, Storage Insights, and diagnostic logging
- Cost and operations checklist for ongoing lake ownership

09 Capstone: Secure Zone-Based Lake Connected to Databricks

- Finalise zone layout, partitions, and naming for a sample analytics domain
- Apply RBAC and ACL patterns for engineers versus analysts
- Connect Databricks, read partitioned data, and validate access boundaries
- **Hands-on:** Configure RBAC and ACLs, connect Databricks to ADLS Gen2, and read partitioned data