# AZURE DATA ENGINEER

## Module 1: Advanced Data Architecture & Storage in Azure

### 1.1 Modern Data Architectures

- Data Warehouse vs Data Lake vs Lakehouse
- When to use each in real-world scenarios
- Hybrid architectures & industry case studies

### 1.2 Azure Data Lake Gen2 Advanced Features

- Hierarchical Namespace for folder-based storage
- Lifecycle management policies
- Zone design (Raw, Curated, Trusted layers)

### 1.3 Partitioning, Indexing & File Optimization

- Partition strategy for large datasets
- Small file problem & compaction strategies
- Indexing data in Azure Synapse & Data Lake

### 1.4 Access Control Lists (ACLs) & Security

- Setting ACLs on ADLS Gen2 folders/files
- RBAC roles & granular security
- Managed identities for secure connections
- 1.5 Data Security & Encryption
- Encryption at rest with SSE & CMK
- Encryption in transit with TLS
- Key rotation policies

---

## Module 2: Azure Data Factory (ADF) – Advanced Orchestration

### 2.1 Designing Complex ETL/ELT Pipelines

- Advanced pipeline orchestration patterns
- Reusable activities & templates

### 2.2 Parameterization & Dynamic Content

- Expressions & functions in pipelines
- Metadata-driven parameterization

### 2.3 Metadata-Driven Pipelines

- Config-driven ETL approach
- Dynamic source-to-target mapping

### 2.4 Error Handling & Logging

- Try-Catch implementation in ADF
- Custom logging using Azure Monitor

### 2.5 Monitoring & Alerts

- Built-in monitoring features
- Alerts with Azure Monitor & Log Analytics

### 2.6 CI/CD with ADF

- Git integration with ADF
- Dev, Test, Prod deployment strategy
- Automated release pipelines

---

## Module 3: Real-Time Data Processing with Event Hubs & Stream Analytics

### 3.1 Event Hubs & IoT Hub Advanced Ingestion

- Partitioning in Event Hubs
- IoT telemetry streaming

### 3.2 Stream Analytics SQL - Aggregations & Joins

- Joining streams & reference data
- Advanced aggregations

### 3.3 Windows in Stream Processing

- Tumbling, Sliding, Hopping windows
- Use cases (real-time fraud detection, IoT monitoring)

### 3.4 Real-Time Dashboards

- Stream output to Power BI
- Stream to Cosmos DB for real-time queries

### 3.5 Performance Optimization

- Scaling Stream Analytics jobs
- Query performance best practices

---

## Module 4: Azure Databricks for Data Engineering

### 4.1 Advanced Spark Optimization

- Shuffle partitions & skew handling
- Caching strategies
- Adaptive query execution

### 4.2 Delta Lake Advanced Features

- ACID transactions
- Data versioning & rollback
- Schema enforcement & evolution

### 4.3 Medallion Architecture (Bronze, Silver, Gold Layers)

- Best practices for pipeline layering
- Incremental data loads

### 4.4 Cluster Optimization

- Autoscaling & job clusters
- Optimized runtime versions

### 4.5 MLflow for Model Management

- Tracking experiments
- Registering & deploying ML models

### 4.6 Integration with ADF & Event Hubs

- Orchestrating Databricks notebooks from ADF
- Streaming data ingestion from Event Hubs

---

## Module 5: Synapse Analytics – Advanced Data Modelling & Queries

### 5.1 Dedicated vs Serverless SQL Pools

- When to use each
- Cost & performance comparison

### 5.2 Materialized Views & Caching

- Performance boost with materialized views
- Managing cache refresh policies

### 5.3 Workload Management

- Resource classes & workload groups
- Isolating workloads for performance

### 5.4 Query Optimization

- Statistics & distribution strategies
- Best practices for large-scale queries

### 5.5 PolyBase & External Tables

- Loading external big data
- Querying from external sources

### 5.6 Real-Time Analytics

- Synapse + Power BI integration
- Near real-time pipelines with Synapse

## Module 6: Security, Governance & Compliance

### 6.1 Encryption & Key Management

- Transparent Data Encryption
- Azure Key Vault integration

### 6.2 Advanced Access Controls

- RBAC vs ABAC
- Hierarchical security models

### 6.3 Data Masking & Row-Level Security

- Dynamic data masking
- Row-level & column-level security

### 6.4 Data Cataloging with Azure Purview

- Data lineage
- Glossary & metadata management

### 6.5 Compliance Frameworks

- GDPR, HIPAA, SOC 2 in Azure
- Implementing compliance controls

---

## Module 7: Monitoring, Optimization & Automation

### 7.1 Monitoring Data Workloads

- Azure Monitor dashboards
- Log Analytics queries

### 7.2 Query Performance Tuning

- Synapse queries optimization
- Spark jobs performance tuning

### 7.3 Cost Optimization

- Reserved instances vs pay-as-you-go
- Storage lifecycle management

### 7.4 Infrastructure as Code

- ARM templates
- Bicep deployment scripts

### 7.5 CI/CD for Data Pipelines

- Azure DevOps pipelines
- Automated testing for data flows

### 7.6 Automation of Deployments

- Scheduling deployments
- Blue-green deployment strategies

## Module 8: Capstone Project – Enterprise Data Engineering Solution

### 8.1 Requirement Gathering & Design

- Understand business use case
- Choose right architecture

### 8.2 Data Ingestion (Batch + Real-Time)

- Batch from SQL, APIs
- Real-time from IoT/Event Hubs

### 8.3 Lakehouse Implementation

- Using Databricks Delta Lake
- Building medallion layers

### 8.4 Orchestrating with ADF

- Pipeline for batch + real-time flows
- Monitoring & logging

### 8.5 Security & Governance

- Apply security policies
- Purview for data lineage

### 8.6 Real-Time Analytics

- Synapse + Power BI dashboards
- Low latency reporting

### 8.7 CI/CD & Documentation

- Automating deployments
- Creating architecture documentation

*****Note:** This outline is comprehensive and can be tailored based on course duration, depth of coverage, and the participants expertise levels. As technology continues to evolve, it is crucial to review and update the content regularly to incorporate emerging tools, practices, and industry best standards.