Synechron

Data Engineering with Databricks Training

Duration - 5 Day

Content:

Day 1(Session 1 & 2): Introduction to Databricks and Foundational Data Management

1. **Get Started with Databricks Data Science and Data Engineering Workspace**
   - Overview of Databricks architecture
   - Navigating the Databricks UI and managing clusters
   - Notebook basics and collaboration features
   - Integrating Databricks with source control (Git, GitHub)
2. **Transform Data with Spark**
   - Introduction to PySpark and the DataFrame API
   - Performing data transformations and aggregations
   - Working with complex data types (arrays, maps, structs)
   - Joining, filtering, and partitioning data in Spark
   - Working with Spark SQL for data exploration
3. **Manage Data with Delta Lake**
   - Introduction to Delta Lake and its benefits
   - Creating and managing Delta tables
   - Upserts, deletes, and merge operations in Delta Lake
   - Time travel and data versioning
   - Schema evolution and enforcement in Delta Lake
   - **Additional Topic**: Delta Lake best practices for ETL workflows

Day 2(Session 3 & 4): Advanced Data Pipeline Development and Access Control

1. **Build Data Pipelines with Delta Live Tables**
   - Introduction to Delta Live Tables (DLT)
   - Defining and managing data pipeline tasks
   - Incremental data processing with DLT
   - Enforcing data quality with expectations in DLT
   - Using SQL and Python in Delta Live Tables
2. **Deploy Workloads with Databricks Workflows**
   - Understanding Databricks Workflows for pipeline orchestration
   - Scheduling and triggering jobs
   - Configuring and managing multi-task workflows
   - **Additional Topic**: Monitoring and troubleshooting Databricks Workflows
3. **Manage Data Access with Unity Catalog**
   - Introduction to Unity Catalog and its architecture
   - Creating and managing catalogs, schemas, and tables
   - Managing metadata and data lineage in Unity Catalog
   - **Additional Topic**: Unity Catalog integration with external data lakes (Azure, AWS)

Day 3(Session 5 & 6): Advanced Lakehouse Architecture and Real-Time Processing

1. **The Lakehouse Architecture**
   - Understanding the Databricks Lakehouse concept
   - Components of the Lakehouse: Delta Lake, Databricks SQL, Unity Catalog
   - Benefits of combining data lakes and warehouses
2. **Optimizing Data Storage**
   - Partitioning, bucketing, and clustering for performance
   - Data caching and Z-Order optimization

  - **Additional Topic**: Using the OPTIMIZE command and VACUUM for data management

3. **Understanding Delta Lake Transactions**
   - ACID transactions in Delta Lake
   - Schema enforcement and data quality rules
   - Transactional operations for data reliability
4. **Clone for Development and Data Backup**
   - Creating shallow and deep clones in Delta Lake
   - Use cases for cloning (testing, backups, versioning)
5. **Auto Loader and Bronze Ingestion Patterns**
   - Using Auto Loader for incremental data ingestion
   - Bronze, Silver, and Gold layering for data organization
6. **Streaming Deduplication and Quality Enforcement**
   - Deduplication techniques for streaming data
   - Enforcing data quality in real-time processing
7. **Slowly Changing Dimensions (SCD)**
   - Handling SCD Type 1 and Type 2 in Delta Lake
   - Implementing SCD for historical data tracking
8. **Streaming Joins and Statefulness**
   - Stateful operations in streaming (e.g., join, aggregation)
   - Managing stateful transformations in streaming data

Day 4(Session 7 & 8): Secure Data Management, Deployment, and Cost Optimization

1. **Stored and Materialized Views**
   - Creating and managing views for data analytics
   - Using materialized views for optimized queries
   - **Additional Topic**: Managing dependencies between views and tables
2. **Storing Data Securely**
   - Data encryption at rest and in transit
   - Best practices for securing data in Databricks
   - Integrating with key management services (Azure Key Vault, AWS KMS)
3. **Granting Privileged Access to PII**
4. **Deleting Data in the Lakehouse**
5. **Orchestration and Scheduling with Multi-Task Jobs**
   - Building complex workflows with dependencies
   - Managing job clusters and job settings for performance
   - Monitoring job performance and troubleshooting issues
6. **Monitoring, Logging, and Handling Errors**
   - Setting up logging and alerts for Databricks jobs
   - Using Databricks Metrics UI for monitoring
   - Handling errors and retries in workflows

Day 5: (Session 9 & 10):

1. **Different types of clusters (general purpose, DWH, serverless etc..) , and choice of using them for different use cases.**
2. **Creating, deploying, sharing of notebooks**
3. **Promoting Code with Databricks Repos**
   - Integrating Databricks Repos with Git for version control
   - Using branching strategies for development and production
   - CI/CD integration for seamless code promotion
4. **Programmatic Platform Interactions (Databricks CLI and REST API)**
   - Using the Databricks CLI for administrative tasks

   - Automating workflows and deployments with the REST API
   - **Additional Topic**: Accessing Unity Catalog programmatically

5. **Managing Costs and Latency with Streaming Workloads**
   - Cost optimization strategies for streaming
   - Managing cluster usage and minimizing idle time

- Reducing latency in streaming pipelines with Auto Loader and Delta Lake