Data Engineering with Databricks - Extended

Duration : 4 Days

Introduction to Databricks

- Overview of Databricks and the Lakehouse Platform
- Data Engineer Role in Databricks
- Key Components of Databricks Workspace

Delta Lake Fundamentals

- Introduction to Delta Lake
- Delta Lake Architecture and File Structure
- Key Features of Delta Lake
    - ACID Transactions
    - Time Travel and Versioning
    - Schema Evolution
    - Data Skipping and Caching
- Creating and Managing Delta Tables
- Updating, Deleting, and Merging Data
- Delta Lake Optimization Techniques

Relational Entities on Databricks

- Introduction to Relational Entities
- Tables, Views, and Temporary Views
- Managed and External Tables
- Catalogs and Databases
- Data Cataloging and Metadata Management

ETL with Spark SQL

- Introduction to Spark SQL
- Querying Data with Spark SQL
- ETL Process with Databricks Notebooks
- Extract data from a single file and from a directory of files
- Create a table from a JDBC connection and from an external CSV file
- Working with validations – duplicate values, unique values, missing values etc.
- Working with UDFs

Incremental Data Processing with Structured Streaming and Auto Loader

- Introduction to Structured Streaming
- Data Ingestion with Auto Loader
- Building Stream Processing Pipelines
- Watermarking and Late Data Handling
- Checkpointing and Fault Tolerance
- Streaming to Delta Lake
- ACID transactions in Delta Lake
- Working with Managed Tables

Medallion Architecture in the Data Lakehouse

- Overview of Medallion Architecture
- Bronze, Silver, and Gold Layers
- Best Practices for Medallion Architecture
- Implementing Medallion Architecture in Databricks

Delta Live Tables

- Introduction to Delta Live Tables
- Declarative Data Pipelines
- Creating and Managing Delta Live Tables Pipelines
- Monitoring and Troubleshooting Pipelines
- Incremental Data Processing with Delta Live Tables

Task Orchestration with Databricks Jobs

- Introduction to Databricks Jobs
- Creating and Scheduling Jobs
- Job Dependencies and Workflows
- Identify the benefits of using multiple tasks in Jobs
- Set up a retry policy in case of failure.
- Create an alert in the case of a failed task.
- Identify that an alert can be sent via email

Databricks SQL

- Introduction to Databricks SQL
- Building and Executing SQL Queries
- Creating and Managing Dashboards
- Visualization Techniques

Managing Permissions in the Lakehouse

- Overview of Permissions and Security
- Access Control with Unity Catalog
- Identify Unity Catalog securable
- Implementing Access Control
- Securing Delta Lake with Table Constraints
- Auditing and Compliance in Databricks

Productionizing Dashboards and Queries on Databricks SQL

- Designing for Production Grade Dashboards
- Performance Optimization Techniques
- Scheduled Refresh and Alerts
- Dashboard Sharing and Permissions