Cognixia®

COURSE CONTENT

www.cognixia.com

COURSE NAME

Data Engineering with Databricks

DURATION

32 Hours

PREREQUISITES

- Beginner familiarity with basic cloud concepts (virtual machines, object storage, identity management)
- Ability to perform basic code development tasks (create compute, run code in notebooks, use basic notebook operations, import repos from git, etc.)
- Intermediate familiarity with basic SQL concepts (CREATE, SELECT, INSERT, UPDATE,
- DELETE, WHILE, GROUP BY, JOIN, etc.)
- Basic knowledge of Python programming, jupyter notebook interface, and PySpark fundamentals

COURSE OUTLINE

Module 1: Introduction to Data Engineering and Data Ingestion Strategies

- What is Data Engineering
- Role of a Data Engineer in modern data architecture
- Conceptual, Logical, and Physical Data Models
- OLTP vs. OLAP Systems
- ETL vs. ELT in Modern Data Architectures
- Data Ingestion Methods
- Batch Processing vs. Streaming Processing
- Extracting Data from APIs, Databases, Object Storage (S3, HDFS, ADLS, GCS)
- Working with Data Formats: CSV, JSON, XML, Avro, Parquet, ORC, Protobuf, Thrift
- Real-time Data Ingestion with Apache Kafka, AWS Kinesis, Google Pub/Sub
- Data Lake Architecture (Raw, Processed, Curated Layers)
- Incremental Load & Change Data Capture (CDC) Strategies
- Log-Based CDC
- Trigger-Based CDC
- Watermarking & Checkpointing in Streaming Pipelines
- Data Validation & Quality Checks
- Handling Corrupt Records in Large Datasets
- Schema Evolution & Enforcement in Avro & Parquet

Module 2: Introduction to Apache Spark & PySpark

- What is Apache Spark? Why is it Used for Big Data Processing?
- Spark Core Components & Execution Flow

- RDDs (Resilient Distributed Datasets) vs. DataFrames vs. Datasets
- Transformations vs. Actions in Spark
- Understanding PySpark
- PySpark vs. Pandas vs. Dask
- Schema Inference & Explicit Schema Definition in PySpark
- Reading & Writing Data in PySpark (CSV, JSON, Avro, Parquet, ORC, Delta)
- Creating Spark DataFrames from Multiple Sources (relevant will be covered)
- RDDs, Lists, Databases, CSV, JSON, Avro, Parquet, ORC, Google BigQuery, Snowflake
- Essential DataFrame Operations

Module 3: Databricks Lakehouse & Medallion Architecture

- Introduction to Databricks Lakehouse
  - Lakehouse vs. Data Lake vs. Data Warehouse
  - Key Features and Benefits

- Medallion Architecture
  - Bronze Layer (Raw Data Ingestion)
  - Silver Layer (Data Cleaning & Transformations)
  - Gold Layer (Aggregated Data for Analytics)

- Performance Optimization in Medallion Architecture
  - Best Practices for Data Organization
  - Incremental Data Processing for Each Layer

- Lab:
  - Ingest raw data into Bronze, clean it in Silver, and aggregate it in Gold.

Module 4: Managing Data with Delta Lake

- Delta Lake Overview
  - What is Delta Lake?
  - Delta vs. Traditional Data Lakes
  - ACID Transactions & Schema Enforcement
  - Delta Lake in Medallion Architecture

- Schema Evolution & Enforcement
  - Schema-on-Read vs. Schema-on-Write
  - Handling Evolving Data Schemas
  - Auto Evolution vs. Explicit Schema Management

- Time Travel & Data Versioning
  - How to Query Previous Versions of Data
  - Implementing Rollback & Recovery

  - Tracking Data Modifications for Audit Purposes

- Optimizing Delta Tables
  - Optimize & Compact Small Files
  - Auto Optimize, Vacuum & Retention Policies
  - Handling Large-Scale Data Growth with OPTIMIZE

- Handling Late Arriving Data
  - Strategies to Handle Late Events
  - Watermarking & Checkpointing in Streaming Pipelines
  - MERGE INTO for Handling Late Data Updates

- Lab:
  - Implement a Delta Lake-based Data Pipeline with Time Travel & Schema Evolution
  - Optimize queries using Z-Ordering and Auto Optimize

- Scalable Data Pipelines & Job Orchestration

Module 5: Delta Live Table Basics

- Introduction to Delta Live Tables
  - What is Delta Live Tables (DLT)?
  - Why Use DLT for Data Pipelines?
  - DLT vs. Standard Data Pipelines

- Key Features of DLT
  - Continuous vs. Triggered Processing
  - Data Quality Enforcements & Expectations
  - Understanding the Event Log in DLT

Module 6: Deploy Workloads with Databricks Workflows

- Introduction to Databricks Workflows
  - What are Workflows & Why Use Them?
  - Jobs vs. Notebooks Execution
  - Best Practices for Productionizing Data Pipelines

- Job Orchestration Strategies
  - Triggering Workflows
  - Managing Job Dependencies
  - Handling Job Failures & Retries

- Advanced Workflow Features
  - Multi-Task Workflows
  - Task Parameterization & Job Chaining

  - Notifications & Monitoring in Workflows

- Lab:
  - Automate an ELT Pipeline using Databricks Workflows
  - Schedule a Pipeline & Monitor Jobs

Module 7: Data Governance with Unity Catalog

- Introduction to Unity Catalog
  - Multi-Cloud Data Governance in Databricks
  - Key Benefits of Unity Catalog
  - Comparison with Other Governance Tools
- Access Control & Permissions
  - Role-Based Access Control (RBAC)
  - Managing Users, Groups, and Roles
  - Row-Level Security (RLS) & Column-Level Security (CLS)

- Data Lineage & Auditing
  - Tracking Data Lineage
  - Using Unity Catalog for Audit Logging
  - Compliance Considerations (GDPR, HIPAA, SOC 2)

- Lab:
  - Configure Unity Catalog & Set Access Controls
  - Implement Row & Column Level Security Policies

Module 8: Performance Tuning & Cost Optimization

- Optimizing Spark Performance
  - Understanding the Catalyst Optimizer
  - Predicate Pushdown & Join Optimization
  - Caching Strategies for Performance Gains

- Databricks Cluster Optimization
  - Choosing the Right Cluster for Workloads
  - Auto Scaling vs. Fixed Cluster Sizing
  - Optimizing Costs with Spot Instances

- Optimizing Delta Lake Queries
  - Z-Ordering vs. Clustering
  - Auto Optimize & File Compaction Strategies
  - Impact of Large File Sizes on Query Performance

- Lab:
  - Run performance benchmarks on optimized vs. non-optimized queries

- Tune cluster configuration for cost and performance efficiency

Module 9: Capstone Project - Real-World Implementation

Note: The Capstone project will take place post-training and falls outside the designated training hours.

- End-to-End Data Pipeline Implementation
  - Define Business Use Case & Data Requirements
  - Design and Implement an ETL Pipeline with Delta Live Tables
  - Use Workflows to Automate & Optimize the Pipeline
  - Implement Security & Governance using Unity Catalog

- Final Review & Certification Preparation
  - Mock Test for Databricks Data Engineer Associate Certification
  - Best Practices for Databricks Deployment in Production

- Capstone Project:
  - Build a Full Data Pipeline from S3 to Delta Lake to Gold Tables in Databricks
  - Optimize, Secure, and Monitor the Entire Data Pipelin