DATA ENGINEERING | ADVANCED | 2 DAYS

Databricks Data Engineering and Pipelines

Build Databricks pipelines with notebooks, PySpark, and Delta Lake - from ADLS Gen2 ingest through medallion layers to Auto Loader, MERGE patterns, and production Jobs.

Why Databricks Pipelines?

Reliable lakehouse pipelines

Design Bronze-to-Gold flows on Delta that stay consistent from ingest to reporting.

Production orchestration

Schedule multi-task Databricks Jobs with dependencies, retries, and alerts.

Incremental by default

Use Auto Loader and MERGE INTO patterns instead of brittle full reloads.

Secure operations

Manage secrets, identities, monitoring, and cost controls for pipeline workloads.

What you'll be able to do

- ✓ Build data transformation pipelines using Databricks notebooks and PySpark
- ✓ Implement incremental data loading patterns using Delta Lake MERGE INTO and Auto Loader
- ✓ Design and schedule multi-task workflows using Databricks Jobs
- ✓ Apply the medallion architecture (Bronze, Silver, Gold) to structure data engineering pipelines
- ✓ Ingest data from Azure Data Lake Storage Gen2 into Delta tables across pipeline layers
- ✓ Handle schema evolution, bad records, and data quality in pipelines
- ✓ Manage secrets and credentials securely in Databricks pipeline code
- ✓ Monitor, troubleshoot, and maintain production pipelines

Who this is for

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

Curriculum

01 Databricks Notebooks and PySpark for Pipeline Work

- Notebook cell types, magic commands (%sql, %python, %md), and widgets for parameterisation
- PySpark DataFrames for analysts: transformations versus actions
- Common DataFrame operations used in pipeline notebooks
- Organising notebook code for reusable Bronze-to-Gold steps
- **Hands-on:** Build and parameterise a pipeline notebook with SQL and Python cells

02 Ingesting Data from ADLS Gen2 into Delta

- Reading from ADLS Gen2 with the abfss:// protocol
- Loading CSV, JSON, and Parquet into DataFrames
- Writing to Delta tables with append, overwrite, and merge modes
- Choosing write patterns for raw landing versus curated layers
- Validating row counts and table schemas after ingest

03 Medallion Architecture: Bronze, Silver, and Gold

- Medallion design principles for lakehouse data engineering
- Bronze: land raw ADLS Gen2 data into Delta with minimal transformation
- Silver: cleanse, deduplicate, standardise, and cast types
- Gold: business aggregations and wide tables ready for reporting and Power BI
- **Hands-on:** Implement a Bronze-to-Silver-to-Gold pipeline on a sample ADLS Gen2 dataset

04 Schema Evolution and Bad-Record Handling

- Schema enforcement versus schema evolution: when to allow change
- Configuring schema behaviour for evolving source files
- Corrupt-record modes: PERMISSIVE, DROPMALFORMED, and FAILFAST
- Quarantining bad records without failing the entire pipeline
- Practical rules for production schema change management

05 Auto Loader for Incremental Ingestion

- Streaming ingestion from ADLS Gen2 into Delta with Auto Loader
- Auto Loader versus batch COPY INTO: when to use each
- Schema inference and evolution with cloudFiles.schemaLocation
- Using the rescue column for unexpected fields
- Operational considerations for continuous and triggered streaming jobs

06 Incremental Loads and MERGE INTO for CDC

- Full load versus incremental versus Change Data Capture patterns
- MERGE INTO for CDC-style upserts in a single statement
- Handling inserts, updates, and deletes against Delta targets
- Idempotent pipeline design to avoid duplicate Gold metrics
- Validating MERGE outcomes with before/after row checks

07 Orchestrating Pipelines with Databricks Jobs

- Creating Jobs, adding tasks, and applying cluster policies
- Scheduling pipelines for recurring production runs
- Multi-task workflows: dependencies, parallel tasks, and conditional branching
- Failure handling with retries, on-failure tasks, and email alerts
- Passing parameters and task values across notebook jobs

08 Secrets, Identity, and Secure Storage Access

- Creating secret scopes and storing credentials safely
- Referencing secrets from notebooks and Jobs without hardcoding
- Managed identities and service principals for ADLS Gen2 access
- Least-privilege patterns for pipeline identities
- Checklist for removing secrets from source-controlled notebooks

09 Monitoring, Cost Control, and Production Ops

- Using job run history, cluster logs, and event logs for troubleshooting
- Job clusters versus all-purpose clusters for pipeline cost efficiency
- Sizing clusters and enabling auto-termination for idle workloads
- Detecting common failure modes: schema drift, bad files, and permission errors
- Operational checklist for handing pipelines to production support

10 Capstone: Scheduled Incremental Pipeline to Gold

- Ingest new files from ADLS Gen2 with Auto Loader into Bronze
- Apply Silver cleansing and Gold aggregations with MERGE where needed
- Orchestrate the flow as a multi-task Databricks Job on a schedule
- Secure credentials with secrets or managed identity patterns
- **Hands-on:** Deliver a scheduled incremental pipeline from ADLS Gen2 to a Gold Delta table