



**ROVER**

**ROVER CONSULTING & TRAINING SERVICES**

# Mastering Databricks

## Comprehensive Training Programs

**16 Hours Training**

Unlock the full potential of Databricks with our expert-led training programs. From advanced data engineering techniques to cutting-edge machine learning applications, our comprehensive courses are designed to elevate your skills and drive data excellence.



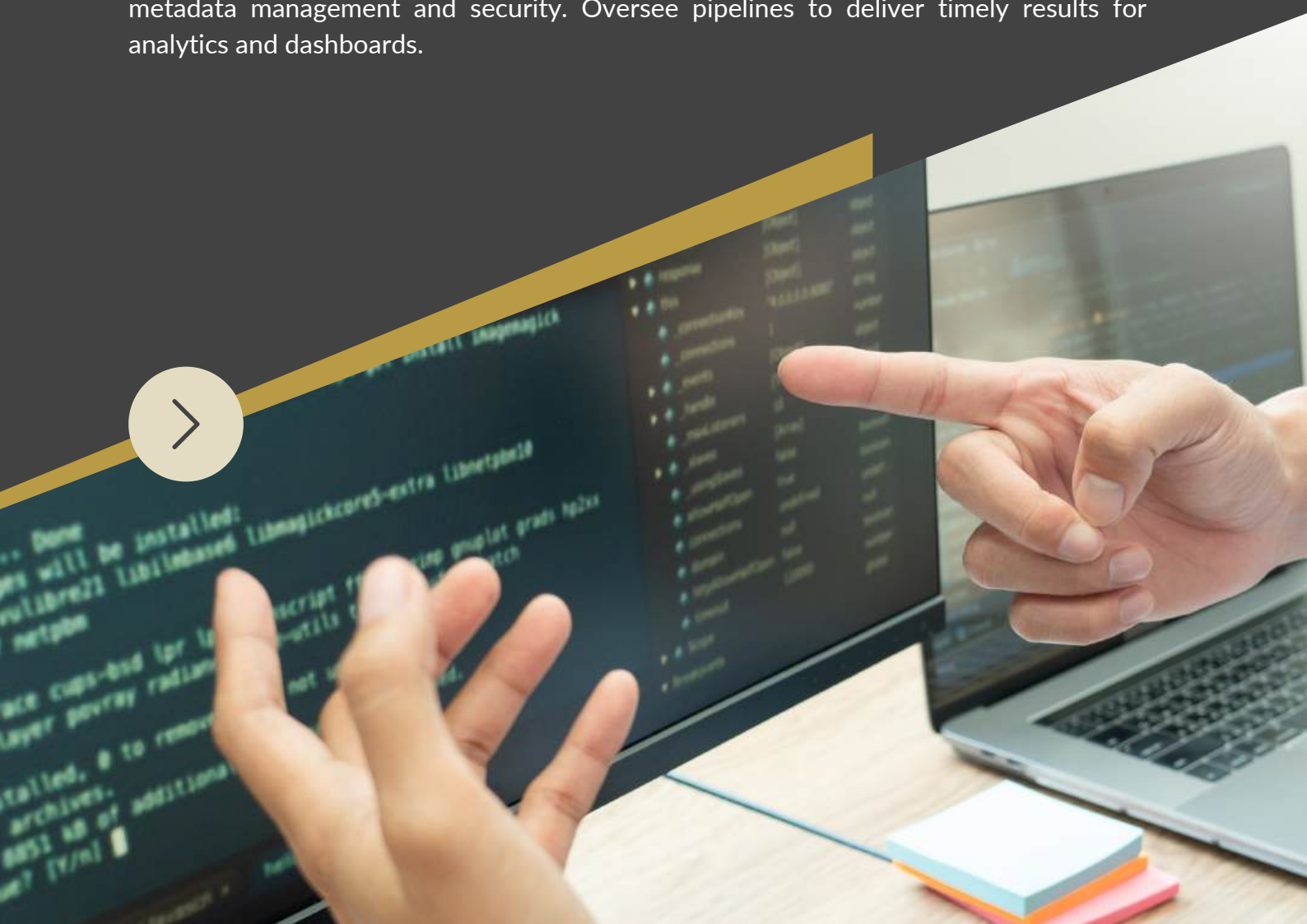
Duration: 16 Hours

# Accelerating Data Engineering Excellence

## A 2-Day Expert-Level Training on Databricks

Optimize data pipeline development with the Databricks Lakehouse Platform. Use SQL and Python for efficient data extraction, transformation, and loading. Leverage Delta Live Tables for streamlined ingestion and incremental updates. Ensure data integrity and performance with Delta Lake's ACID transactions and versioning.

In addition, learn to Implement robust data governance using Unity Catalog for metadata management and security. Oversee pipelines to deliver timely results for analytics and dashboards.



# High Level Overview

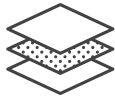


01



FOUNDATIONS  
OF DELTA LAKE

02



RELATIONAL DATA  
MANAGEMENT ON  
DATABRICKS

03



BUILDING ETL  
PIPELINES WITH  
SPARK SQL

04



CORE PYTHON  
SKILLS FOR SPARK  
SQL

05



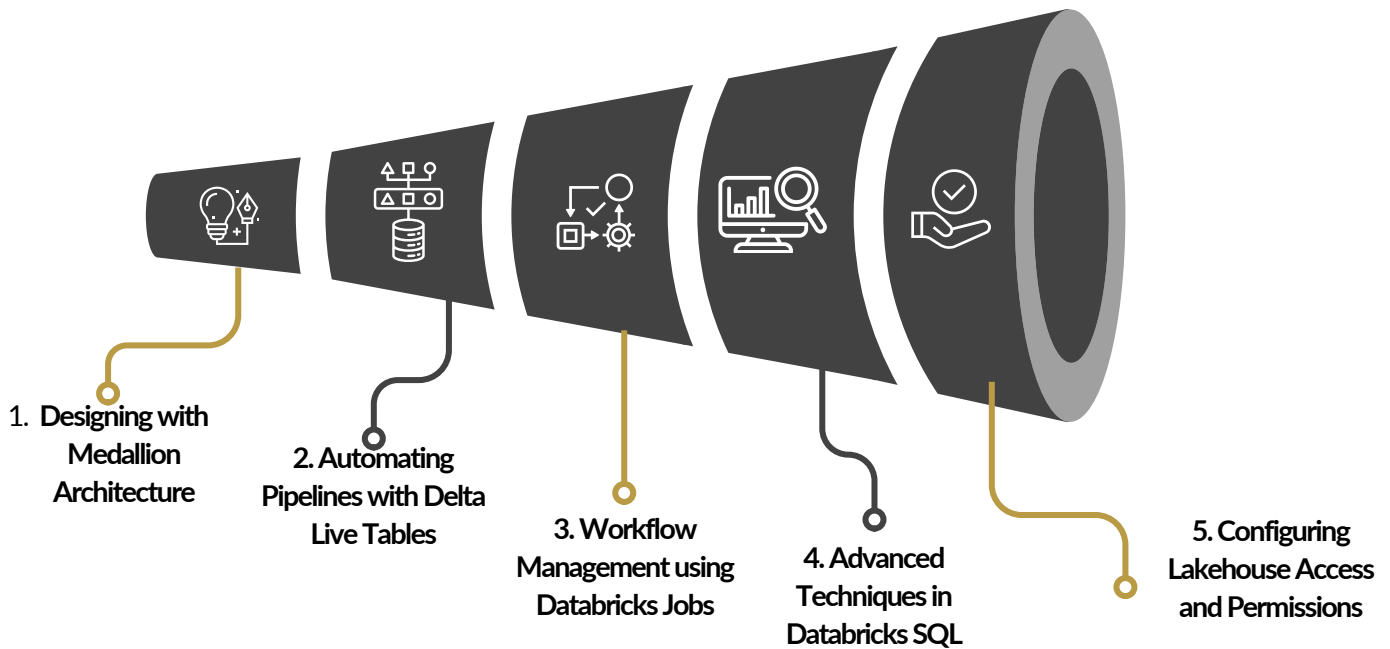
EFFICIENT INCREMENTAL  
PROCESSING WITH  
STRUCTURED STREAMING  
AND AUTO LOADER

Our First day of training covers essential topics to master data engineering with Databricks. Start with the Foundations of Delta Lake to understand the core principles of reliable data management. Dive into Relational Data Management on Databricks to optimize your data storage and retrieval.

Learn to build efficient ETL Pipelines with Spark SQL, and strengthen your Core Python Skills for Spark SQL to enhance data processing.

Finally, explore Efficient Incremental Processing with Structured Streaming and Auto Loader to handle real-time data updates seamlessly.

# High Level Overview



Our second day of course delves into advanced strategies for optimizing data workflows on Databricks. Start by designing with Medallion architecture to structure data layers effectively. Learn Automating Pipelines with Delta Live Tables or seamless, scalable data processing.

Master Workflow Management using Databricks Jobs to streamline and automate tasks. Explore Advanced Techniques in Databricks SQL to enhance your data querying capabilities. Finally, you will understand Configuring Lakehouse Access and Permissions to ensure secure and efficient data access in the Lakehouse environment.

# Course Curriculum



## MODULE 1: DATABRICKS LAKEHOUSE PLATFORM

### 1. Understanding the Databricks Lakehouse

- 1.1 Analyze the synergy between data lakehouses and data warehouses.
- 1.2 Explore the improvements in data integrity and quality within a lakehouse environment compared to traditional data lakes.

### 2. Exploring Databricks Platform Architecture

- 2.1 Gain a comprehensive overview of key architectural components.
- 2.2 Differentiate between general-purpose clusters and job-specific clusters.

### 3. Cluster Management and Configuration

- 3.1 Learn about version management and updates with Databricks Runtime.
- 3.2 Discover techniques for filtering and accessing specific clusters.
- 3.3 Understand the implications of cluster termination and identify optimal restart scenarios.

### 4. Notebook Functionality and Collaboration

- 4.1 Leverage multiple programming languages within notebooks.
- 4.2 Execute notebooks programmatically from within other notebooks.
- 4.3 Explore strategies for sharing and collaborating on notebooks.

### 5. CI/CD Integration with Databricks Repos

- 5.1 Implement continuous integration and deployment workflows using Databricks Repos.
- 5.2 Understand Git operations and their integration with Databricks Repos.
- 5.3 Compare notebooks version control with Databricks Repos.

## MODULE 2: ELT WITH APACHE SPARK

### 1. Data Extraction and Loading Techniques

- 1.1 Extract data from single files and directory structures.
- 1.2 Create views, temporary views, and common table expressions (CTEs) for effective data management.

### 2. Managing External Data Sources

- 2.1 Interact with non-Delta external tables.
- 2.2 Explore methods for creating tables from JDBC connections and external CSV files.

### 3. Data Transformation and Validation

- 3.1 Apply aggregation functions and handle NULL values.
- 3.2 Implement strategies for data deduplication and integrity validation.
- 3.3 Ensure unique values and perform field validations.

### 4. Data Type Conversion and Parsing

- 4.1 Cast columns to timestamps and extract temporal data.
- 4.2 Utilize string operations and dot notation for data extraction.
- 4.3 Understand the benefits of array functions and JSON parsing.

### 5. Advanced SQL Techniques

- 5.1 Analyze join queries and choose between explode and flatten functions.
- 5.2 Pivot data formats and define SQL User-Defined Functions (UDFs).
- 5.3 Utilize CASE/WHEN constructs for advanced SQL logic

## MODULE 3: INCREMENTAL DATA PROCESSING

### 1. Delta Lake ACID Transactions

- 1.1 Understand ACID transaction principles and their advantages.
- 1.2 Evaluate ACID compliance and transaction benefits.

### 2. Data and Metadata Management

- 2.1 Distinguish between data and metadata management.
- 2.2 Compare managed and external tables.

### 3. Table Management and Version Control

- 3.1 Create, manage, and inspect tables.
- 3.2 Analyze Delta Lake directory structures and historical data.
- 3.3 Roll back tables to previous versions and query specific versions.

### 4. Data Optimization and Compaction

- 4.1 Utilize Zordering for data optimization and file compaction.
- 4.2 Implement generated columns and add metadata annotations.

### 5. Data Operations and Commands

- 5.1 Compare CTAS and CREATE OR REPLACE TABLE with INSERT OVERWRITE.
- 5.2 Identify scenarios for using MERGE and COPY INTO commands.
- 5.3 Address COPY INTO command issues and troubleshoot effectively.

### 6. Delta Live Tables (DLT)

- 6.1 Create and manage Delta Live Tables pipelines.
- 6.2 Understand triggered versus continuous pipelines.
- 6.3 Leverage Auto Loader for efficient data ingestion.
- 6.4 Handle constraint violations and change data capture.
- 6.5 Analyze event logs and troubleshoot DLT syntax issues.

# Course Curriculum

## MODULE 4: PRODUCTION PIPELINES

### 1. Task Management and Configuration

- 1.1 Explore the advantages of utilizing multiple tasks within jobs.
- 1.2 Configure predecessor tasks and identify optimal use cases.
- 1.3 Review and analyze task execution history.

### 2. Scheduling and Monitoring Tasks

- 2.1 Employ CRON expressions for task scheduling.
- 2.2 Debug and resolve task failures.
- 2.3 Implement retry policies and notification alerts.
- 2.4 Configure email notifications for task alerts.



## MODULE 5: DATA GOVERNANCE

### 1. Principles of Data Governance

- 1.1 Explore core components and best practices in data governance.
- 1.2 Differentiate between metastores and catalogs.

### 2. Managing Unity Catalog

- 2.1 Gain an overview of Unity Catalog and its security features.
- 2.2 Define and utilize service principals.
- 2.3 Understand security modes compatible with Unity Catalog.

### 3. Best Practices and Access Control

- 3.1 Set up UC-enabled clusters and Databricks SQL warehouses.
- 3.2 Navigate and query three-layer namespaces.
- 3.3 Implement data object access controls and adhere to best practices.
- 3.4 Follow best practices for metastore and workspace colocation.
- 3.5 Use service principals and ensure business unit segregation.

For India Region

# Pricing

**Course fees**

**₹ 3,00,000**

**Travel and Accommodation cost**  
(\* Applicable for In-person training)

**₹1,00,000**




# Contact Us !



CONSULTING & Training Services

Contact Us Today

 9052776606

 [info@rovertek-ai.com](mailto:info@rovertek-ai.com)





**ROVER**

**ROVER CONSULTING INDIA PRIVATE LIMITED**

**Thank  
You**

