ROVER CONSULTING & TRAINING SERVICES

Mastering Databricks

Comprehensive Training Programs

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.

16 Hours Training

Duration: 16 Hours

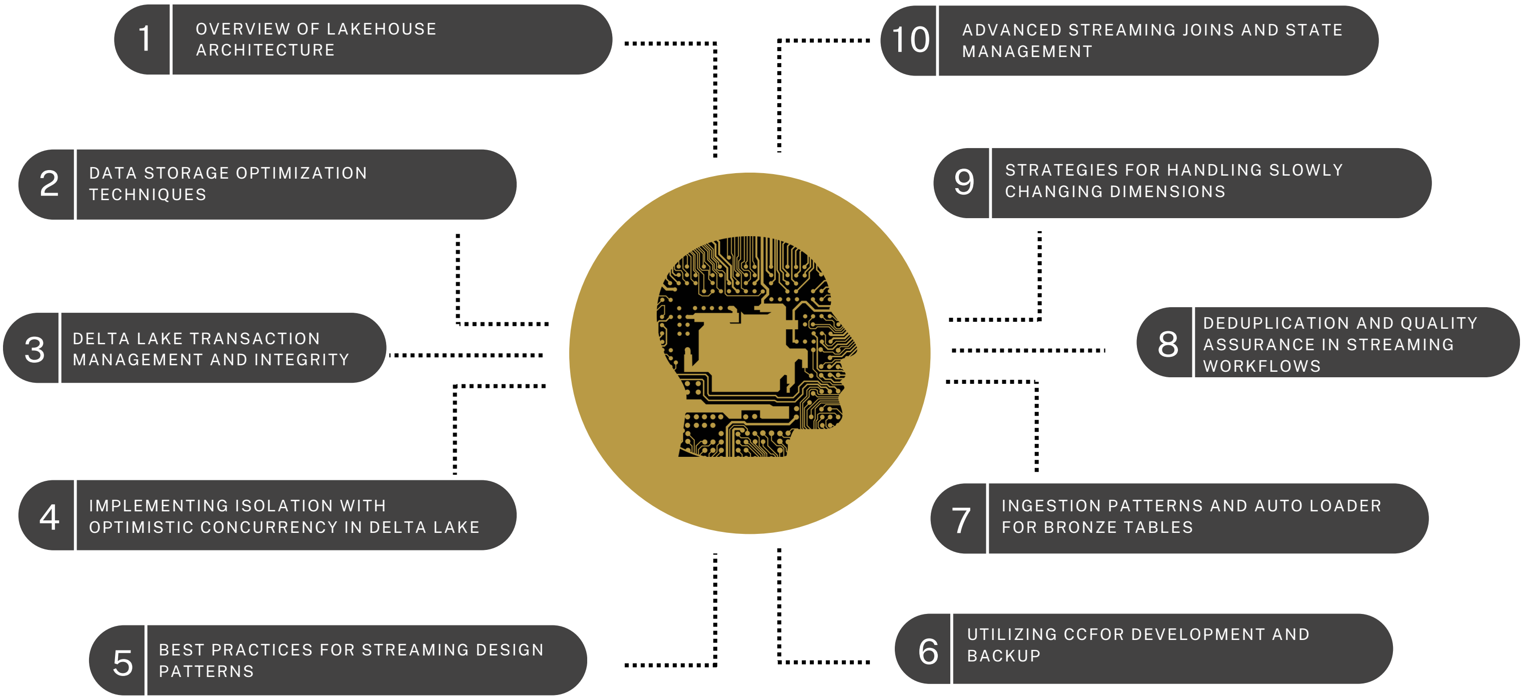
Data Engineering Pro

A 2-Day Expert-Level Training on Databricks

Advanced DataEngineering with Databricks provides an in-depth exploration of advanced data processing, modeling, and Databricks tools. It focuses on optimizing partitioning strategies, implementing incremental data processing with Delta Lake and Structured Streaming, and applying best practices for data transformation and quality. Participants will gain expertise in leveraging Databricks tools for robust security, performance monitoring, and efficient testing and deployment of data pipelines, ensuring effective management and governance within a Databricks environment.

Day - 1

High Level Overview



1 OVERVIEW OF LAKEHOUSE ARCHITECTURE

2 DATA STORAGE OPTIMIZATION TECHNIQUES

3 DELTA LAKE TRANSACTION MANAGEMENT AND INTEGRITY

4 IMPLEMENTING ISOLATION WITH OPTIMISTIC CONCURRENCY IN DELTA LAKE

5 BEST PRACTICES FOR STREAMING DESIGN PATTERNS

6 UTILIZING CCFOR DEVELOPMENT AND BACKUP

7 INGESTION PATTERNS AND AUTO LOADER FOR BRONZE TABLES

8 DEDUPLICATION AND QUALITY ASSURANCE IN STREAMING WORKFLOWS

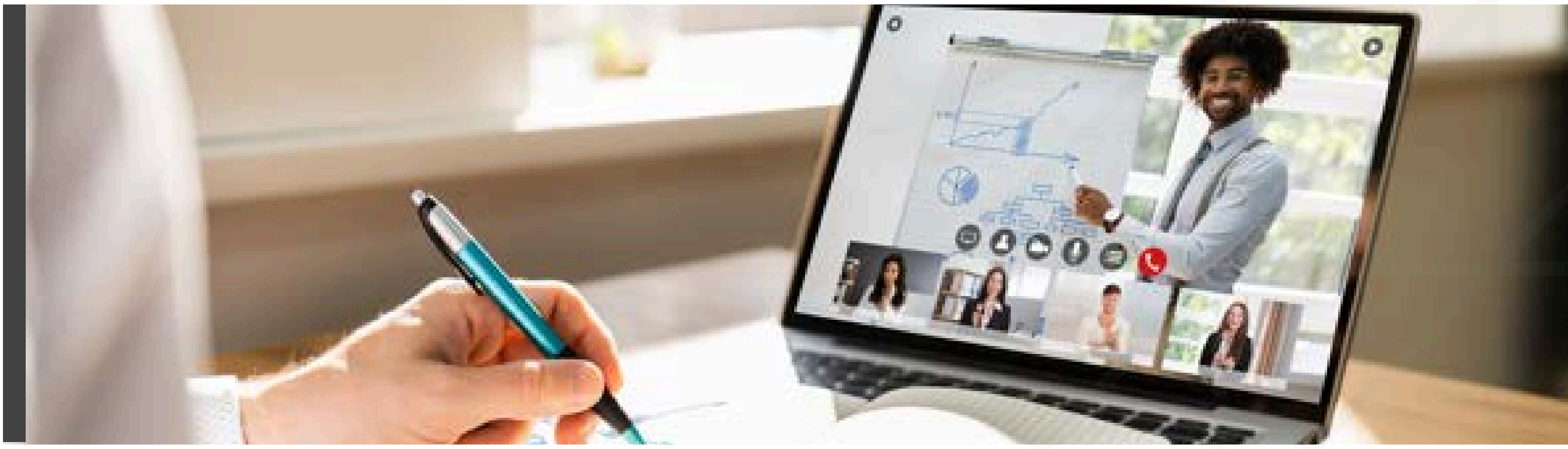
9 STRATEGIES FOR HANDLING SLOWLY CHANGING DIMENSIONS

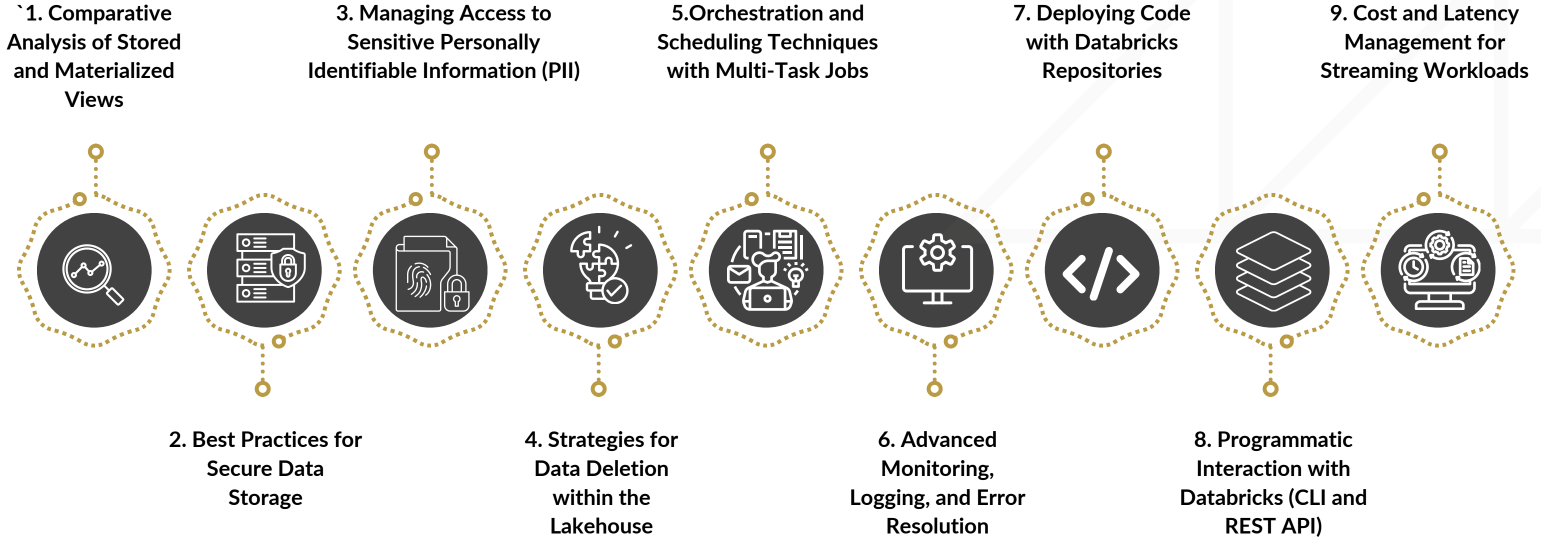
10 ADVANCED STREAMING JOINS AND STATE MANAGEMENT

In the evolving landscape of data management, understanding and implementing advanced techniques is crucial for optimizing performance and ensuring data integrity. This Course delves into Lakehouse architecture, focusing on innovative data storage optimization methods and Delta Lake's transaction management and integrity. It covers strategies for implementing optimistic concurrency and best practices for streaming design patterns. Additionally, the guide explores the benefits of utilizing cloning for development and backup, efficient ingestion patterns with Auto Loader for Bronze tables, and essential techniques for deduplication and quality assurance in streaming workflows. Key strategies for handling slowly changing dimensions and advanced streaming joins and state management are also discussed to provide a comprehensive understanding of modern data practices.

Day - 2

High Level Overview





`1. Comparative Analysis of Stored and Materialized Views

2. Best Practices for Secure Data Storage

3. Managing Access to Sensitive Personally Identifiable Information (PII)

4. Strategies for Data Deletion within the Lakehouse

5.Orchestration and Scheduling Techniques with Multi-Task Jobs

6. Advanced Monitoring, Logging, and Error Resolution

7. Deploying Code with Databricks Repositories

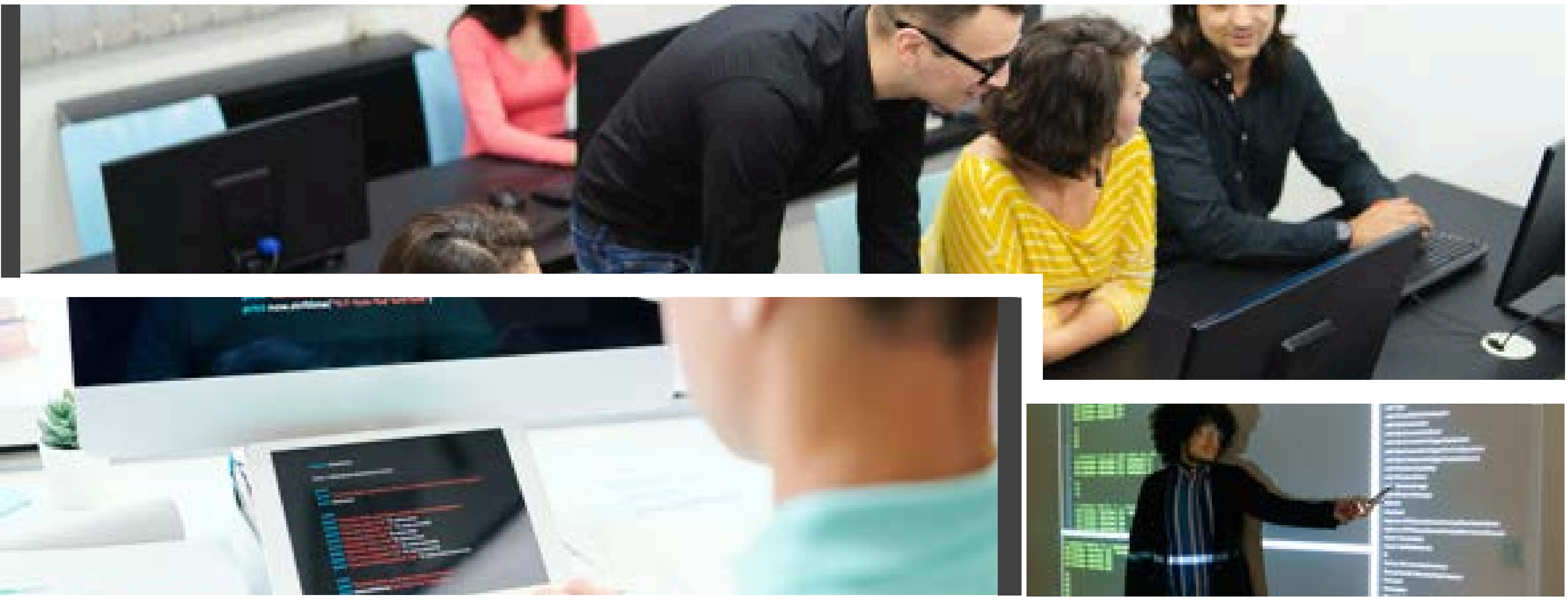
8. Programmatic Interaction with Databricks (CLI and REST API)

9. Cost and Latency Management for Streaming Workloads

This overview offers a deep dive into critical aspects of data management and security within modern architectures. It begins with a comparative analysis of stored and materialized views, highlighting their respective benefits and use cases. The discussion extends to best practices for secure data storage and effective management of access to sensitive Personally Identifiable Information (PII). Strategies for data deletion within the Lakehouse are examined, alongside orchestration and scheduling techniques using multi-task jobs. The guide also covers advanced monitoring, logging, and error resolution practices, deploying code with Databricks Repositories, and programmatic interaction through the CLI and REST API. Finally, it addresses cost and latency management for streaming workloads, ensuring efficient and secure data operations.

Day - 1

Course Curriculum



MODULE 1: ADVANCED DATA PROCESSING TECHNIQUES

1. Partitioning Strategies and Performance Optimization

1.1. Differentiate between partitioning strategies: coalesce, repartition, repartition by range, and rebalance

1.2. Evaluate and select optimal partitioning columns for various data scenarios

1.3. Analyze the effects of file size management and over-partitioning on Spark query performance

2. Data Frame Management and Manipulation

2.1. Configure PySpark DataFrames to control file size during disk writes

2.2. Implement strategies for updating records in Spark tables (Type 1 updates)

3. Streaming and Delta Lake Integration

3.1. Apply design patterns for Structured Streaming and Delta Lake integration

3.2. Optimize state management with stream-static joins and Delta Lake

3.3. Implement stream-static joins and deduplication techniques within Spark Structured Streaming

3.4. Activate Change Data Feed (CDF) on Delta Lake tables and adapt processing workflows for CDC

3.5. Utilize CDF for efficient data propagation and deletion

3.6. Demonstrate effective data partitioning strategies for archiving and data deletion

MODULE 2: DATA MODELING AND TRANSFORMATION

1. Transformation and Quality Assurance

1.1. Outline data transformation objectives during the transition from bronze to silver layers

1.2. Examine how Change Data Feed (CDF) resolves update and delete propagation issues within Lakehouse architecture

1.3. Utilize Delta Lake cloning to understand the interaction of shallow and deep clones with source and target tables

2. Table Design and Implementation

2.1. Design multiplex bronze tables to address challenges in scaling streaming workloads

2.2. Implement best practices for streaming data from multiplex bronze tables

2.3. Apply incremental processing, data quality enforcement, and deduplication from bronze to silver layers

2.4. Assess data quality enforcement methods based on Delta Lake capabilities

2.5. Address the absence of foreign key constraints in Delta Lake tables

2.6. Implement constraints to maintain data integrity in Delta Lake tables

2.7. Develop lookup tables and evaluate trade-offs for normalized data models

2.8. Design architectures for Slowly Changing Dimension (SCD) tables using Delta Lake for both streaming and batch workloads

2.9. Implement SCD Types 0, 1, and 2 tables

MODULE 3: DATABRICKS TOOLS AND OPTIMIZATION

1. Delta Lake Fundamentals

1.1. Explain Delta Lake's transaction log and cloud object storage for ensuring atomicity and durability

1.2. Describe Delta Lake's Optimistic Concurrency Control for transaction isolation and conflict resolution

1.3. Detail the functionality of Delta Lake's cloning features

2. Optimization and Indexing

2.1. Apply Delta Lake indexing techniques, including partitioning, Z-order indexing, bloom filters, and file size management

2.2. Optimize Delta tables for performance in Databricks SQL service

Day - 2

Course Curriculum

MODULE 4: DATA SECURITY AND GOVERNANCE

1. Dynamic Data Access Control

1.1. Implement dynamic views for data masking and access control
1.2. Utilize dynamic views to manage row and column-level access within the data environment

MODULE 5: PERFORMANCE MONITORING AND LOGGING

1. System Performance Analysis

1.1. Analyze Spark UI elements for performance evaluation, application debugging, and optimization
1.2. Monitor event timelines and metrics for job stages within the cluster
1.3. Derive insights from Spark UI, Ganglia UI, and Cluster UI to address performance issues and debug applications

2. Job Management and Deployment

2.1. Design systems to manage cost and latency SLAs for production streaming jobs
2.2. Deploy and oversee streaming and batch job execution

MODULE 6: TESTING AND DEPLOYMENT STRATEGIES

1. Notebook and Code Management

1.1. Adapt notebook dependency patterns to integrate Python file depende
1.2. Convert Python code maintained as Wheels for direct imports using re paths
1.3. Troubleshoot and resolve failed jobs

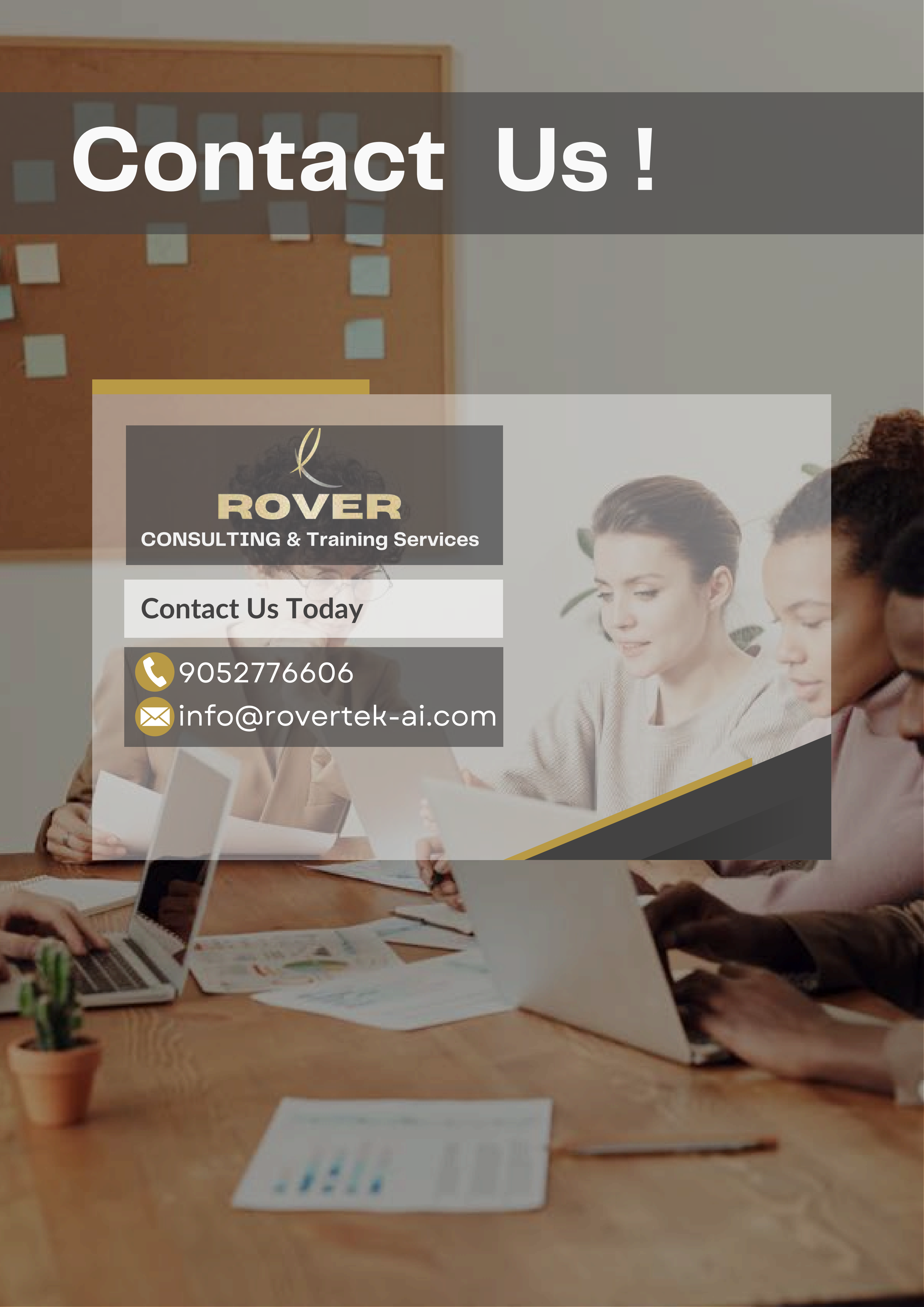
2. Job Creation and CLI Configuration

2.1. Design Jobs based on common use cases and establish multi-task job dependencies
2.2. Configure Databricks CLI for workspace and cluster interaction
2.3. Execute CLI commands for job deployment and monitoring
2.4. Utilize REST API for job cloning, run triggering, and output export

For India Region

Pricing

Course fees	₹ 3,00,000
Travel and Accommodation cost (*Applicable for In-person training)	₹1,00,000



Contact Us !

ROVER

CONSULTING & Training Services

Contact Us Today

☎ 9052776606

✉ info@rovertek-ai.com

ROVER

ROVER CONSULTING INDIA PRIVATE LIMITED

Thank You