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.

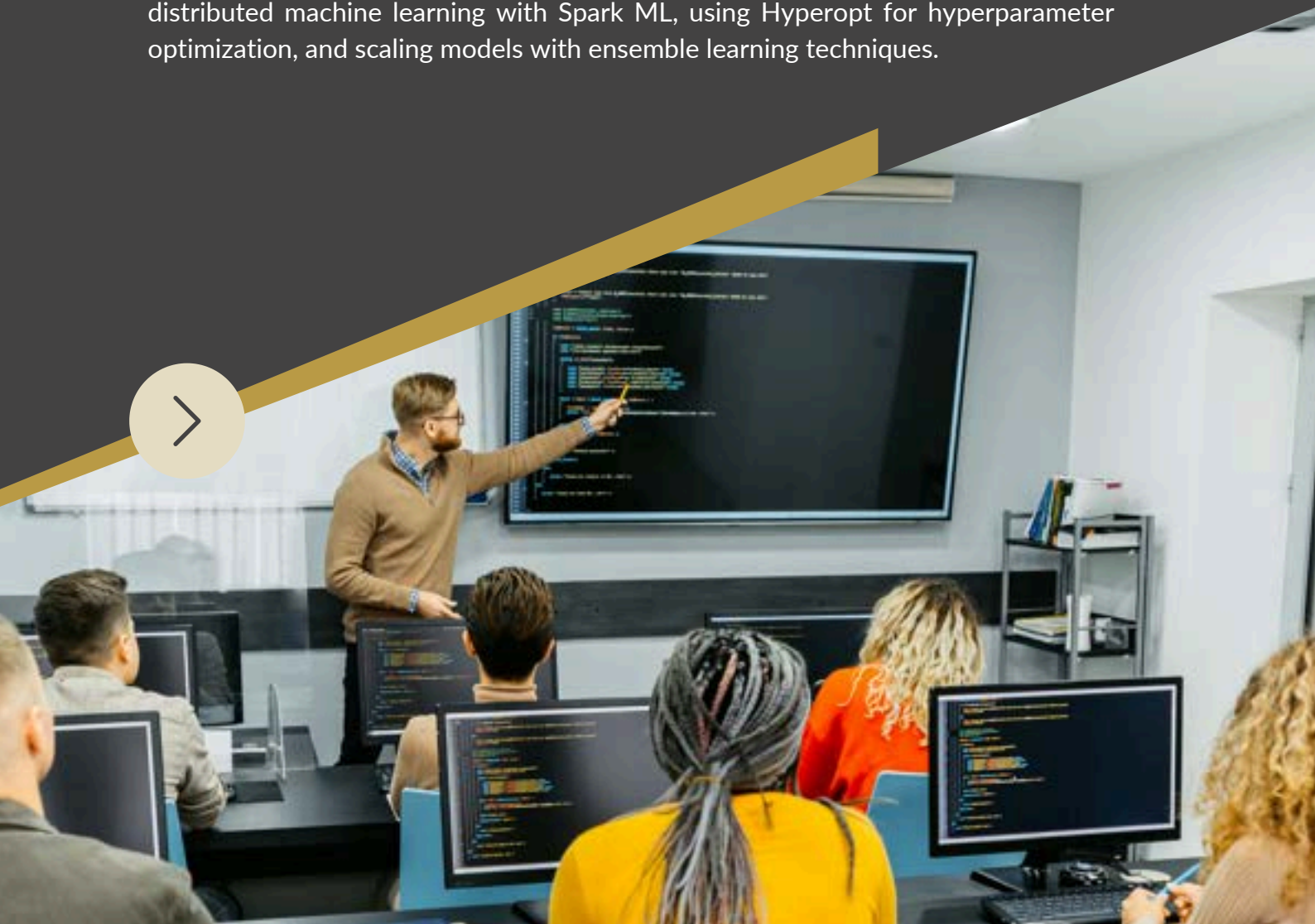
ROVER

Duration: 16 Hours

Harnessing Machine Learning

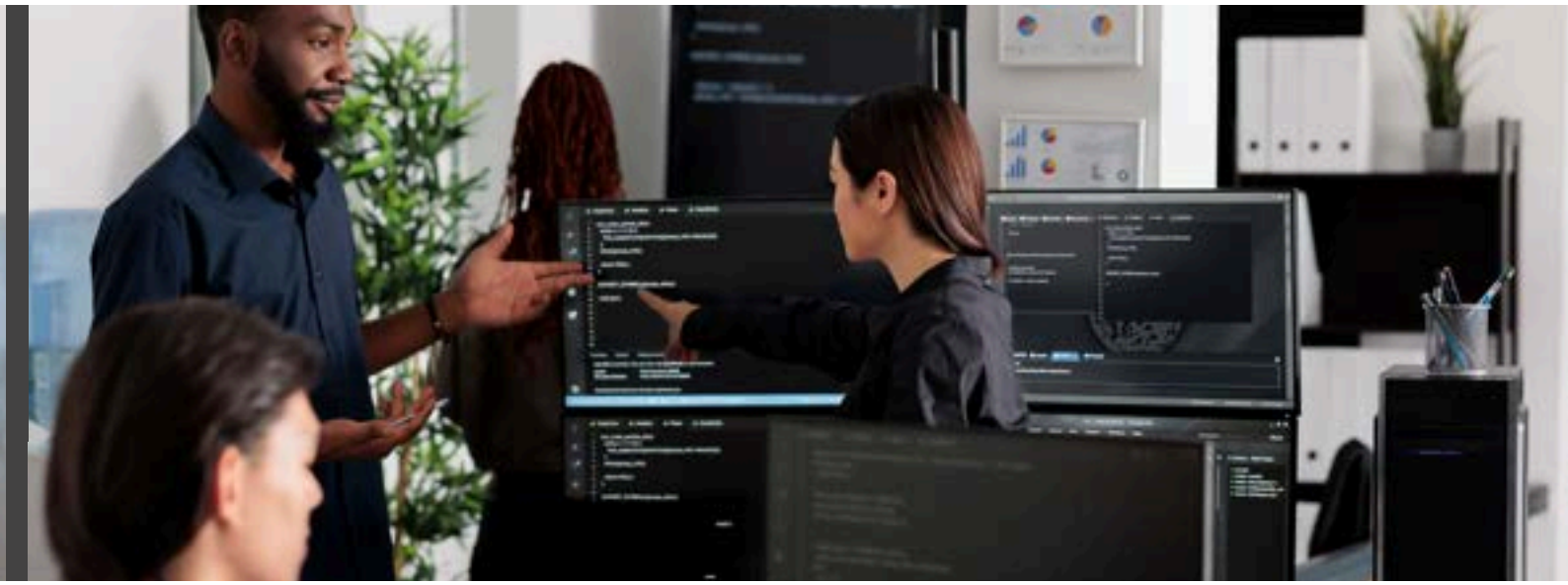
A 2-Day Immersive Training on Databricks

Machine Learning with Databricks provides a thorough examination of advanced machine learning techniques using Databricks, focusing on optimizing machine learning workflows and leveraging Databricks' capabilities. It covers configuring and managing machine learning clusters, integrating Git repositories, and orchestrating multi-task workflows. Participants will explore AutoML for automating pipeline creation, utilize the Feature Store for feature management, and apply MLflow for experiment tracking. The program also addresses machine learning workflows, including exploratory data analysis, feature engineering, hyperparameter tuning, and model evaluation. Additionally, it delves into distributed machine learning with Spark ML, using Hyperopt for hyperparameter optimization, and scaling models with ensemble learning techniques.



Day - 1

High Level Overview



01

Databricks Machine Learning Integration

02

Databricks Runtime for Machine Learning

03

AutoML Capabilities

04

Feature Store Utilization

05

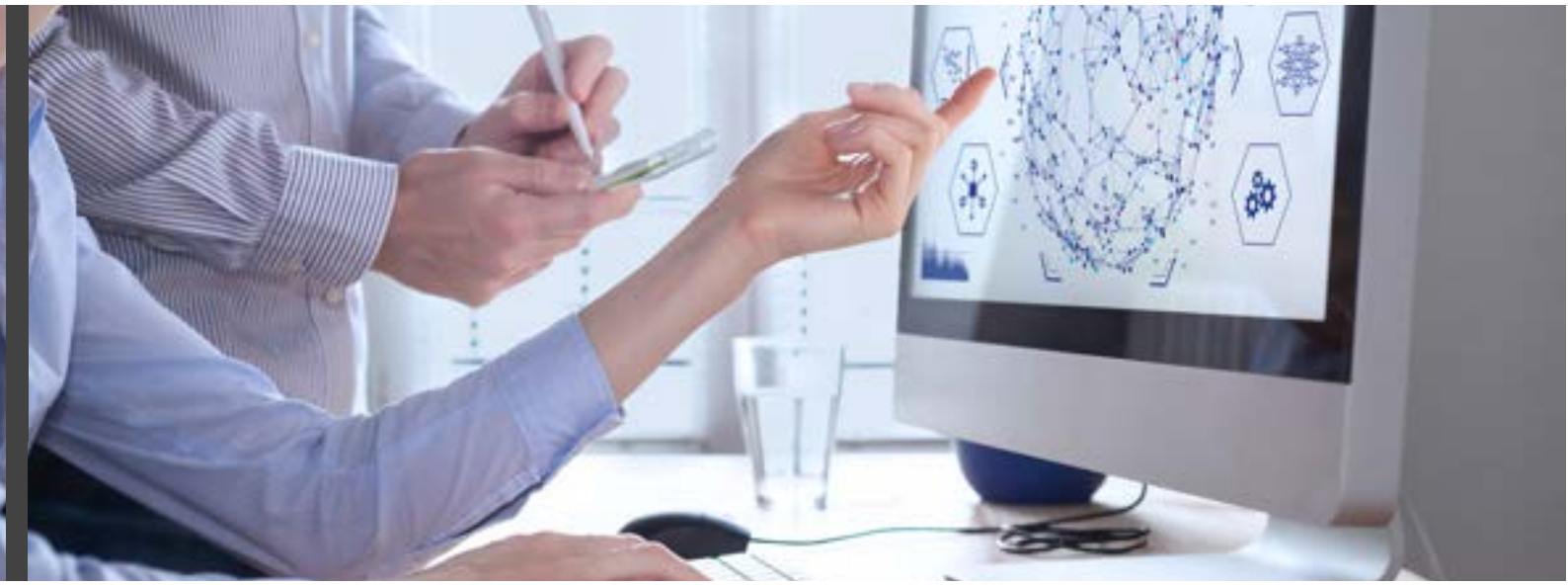
Managed MLflow Operations

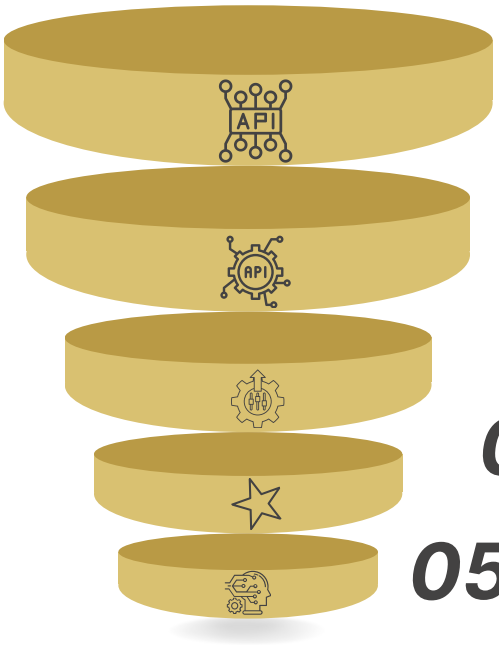
This overview explores the seamless integration of machine learning within the Databricks platform, focusing on key components that enhance data-driven insights. It begins with an examination of Databricks Machine Learning Integration and the specialized Databricks Runtime for Machine Learning, which optimizes model training and deployment.

The discussion also covers the platform's AutoML capabilities, enabling automated model selection and hyperparameter tuning. Additionally, it highlights the utilization of Feature Store for consistent and reusable features, along with the managed operations of MLflow, which streamline the entire machine learning lifecycle from experimentation to production.

Day - 2

High Level Overview





01 Pandas UDFs and Function APIs

02 Pandas API on Spark

03 Hyperopt for Hyperparameter Tuning

04 Spark ML Modeling APIs

05 Distributed Machine Learning Concepts

This overview delves into key concepts and tools essential for distributed machine learning and data processing. It begins with an exploration of distributed machine learning concepts, emphasizing the scalability and efficiency provided by such architectures. The discussion extends to Spark ML modeling APIs, which facilitate the development of machine learning models on large datasets. Additionally, the use of Hyperopt for hyperparameter tuning is highlighted, offering a powerful method for optimizing model performance. The guide also covers the Pandas API on Spark, enabling seamless integration of Pandas operations in distributed environments, and the use of Pandas UDFs and function APIs, which enhance the flexibility and performance of data transformations in Spark.

Day - 1

Course Curriculum

MODULE 1: ADVANCED MACHINE LEARNING WITH DATABRICKS

1. Databricks Machine Learning Integration

1.1. Assess scenarios for deploying standard versus single-node clusters
1.2. Integrate Databricks Repos with external Git repositories for version control
1.3. Manage branching, commits, and synchronization between Databricks Repos and external Git platforms
1.4. Orchestrate complex machine learning workflows leveraging Databricks Jobs

2. Databricks Runtime for Machine Learning

2.1. Configure and deploy clusters utilizing Databricks Runtime for Machine Learning
2.2. Implement and manage Python libraries across Databricks notebooks

3. AutoML Capabilities

3.1. Comprehend the machine learning pipeline automated by AutoML
3.2. Retrieve and evaluate source code and performance metrics from AutoML-generated models
3.3. Utilize the AutoML data exploration notebook to analyze dataset attributes

4. Feature Store Utilization

4.1. Articulate the advantages of Feature Store for managing machine learning features
4.2. Create and populate Feature Store tables, and integrate features into model training and scoring

5. Managed MLflow Operations

5.1. Employ the MLflow Client API for experiment tracking and management
5.2. Log metrics, artifacts, and models; implement nested runs for detailed tracking
5.3. Register and transition model stages using MLflow Client API and Model Registry interface

MODULE 2: MACHINE LEARNING WORKFLOWS

1. Exploratory Data Analysis (EDA)

1.1. Execute summary statistics and outlier detection on Spark DataFrames using `.summary()` and dbutils

2. Feature Engineering Techniques

2.1. Implement indicator variables for imputed or replaced missing values
2.2. Analyze and apply methods for handling missing data, including mode, mean, and median imputation
2.3. Conduct one-hot encoding of categorical features and understand its impact on model performance

3. Model Training Strategies

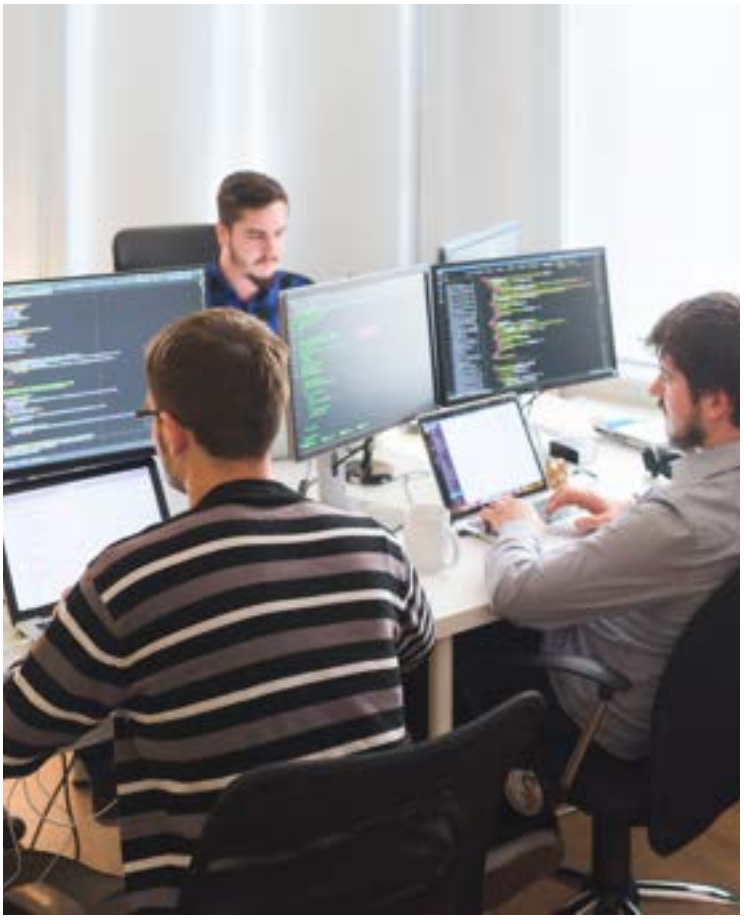
3.1. Apply random search and Bayesian optimization for hyperparameter tuning
3.2. Navigate challenges associated with parallelizing iterative models and leverage Hyperopt with SparkTrials for optimization

4. Model Evaluation and Selection

4.1. Execute cross-validation and grid-search for model evaluation
4.2. Utilize metrics such as Recall, F1 Score, and RMSE, with considerations for log-transformed labels

Day - 2

Course Curriculum



MODULE 4: SCALING AND DISTRIBUTING MACHINE LEARNING MODELS

1. Model Distribution Techniques

1.1. Understand the methodologies for scaling linear regression and decisiaon tree models within Spark

2. Ensemble Learning Distribution

2.1. Explore ensemble learning methodologies including bagging, boosting, and stacking, and their application in distributed environments

MODULE 3: SPARK ML FRAMEWORK

1. Distributed Machine Learning Concepts

1.1. Address challenges in scaling machine learning models and identify Spark ML's role in distributed learning

1.2. Differentiate between Spark ML and scikit-learn in the context of distributed versus single-node solutions

2. Spark ML Modeling APIs

2.1. Perform data splitting, model training, and evaluation using Spark ML

2.2. Develop and troubleshoot Spark ML Pipelines, understanding key considerations and potential issues

3. Hyperopt for Hyperparameter Tuning

3.1. Utilize Hyperopt for parallelized and Bayesian hyperparameter optimization in Spark ML models

3.2. Analyze the relationship between the number of trials and model performance accuracy

4. Pandas API on Spark

4.1. Compare Spark DataFrames with Pandas on Spark DataFrames, and address performance considerations

4.2. Convert between PySpark and Pandas on Spark DataFrames and leverage Pandas API for scalable data processing

5. Pandas UDFs and Function APIs

5.1. Implement Apache Arrow for efficient Pandas-to-Spark conversions

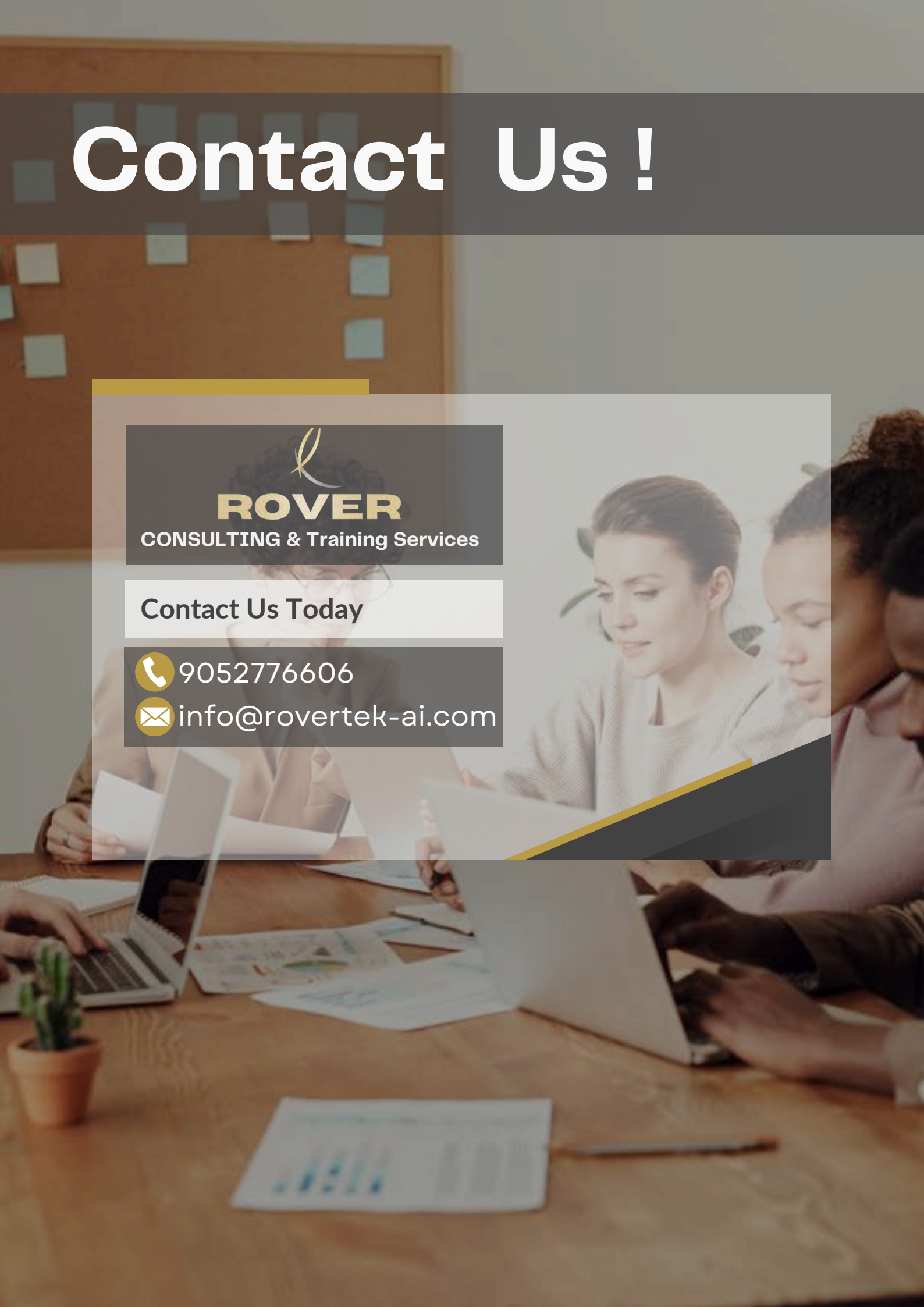
5.2. Utilize Pandas UDFs for parallel model applications and function APIs for group-specific model training

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