Azure Databricks for R-Based Data Analysis & Engineering

Duration: 3 Days (8 hours/ day)

Labs Requirement: Participants must use their own Azure subscription / Databricks environment.

Note: Azure provides a free subscription valid for 30 days.

Course Overview

This course is designed for R users who want to work with large-scale data using Azure Databricks without moving away from familiar R-based workflows. It introduces Databricks from an R user's perspective, focusing on how distributed computing and Delta Lake can be leveraged for data preparation, analytics, and research use cases, especially in healthcare and scientific domains. Participants will learn how to run R code in Databricks notebooks, prepare and manage large datasets, build analysis-ready data pipelines, and perform advanced analytics efficiently at scale. The course also emphasizes reproducibility, governance, and best practices to ensure that research and analytics workflows remain reliable, auditable, and easy to maintain in collaborative environments.

Course Pre-requisites

Participants should have a working knowledge of R, including experience with data frames or tibbles, data cleaning, and basic data analysis. Familiarity with common R packages such as dplyr, tidyr, and ggplot2 is recommended. Basic understanding of data concepts such as tables, files, and datasets will be helpful. Prior experience with SQL, cloud platforms, or distributed systems is beneficial as minimal and practical exposure to these topics will be covered where necessary for data access and workflow integration.

Course Agenda

Module 1: Azure Databricks Through an R User's Lens

- What Azure Databricks is — explained for R users
- How Databricks complements R & RStudio workflows
- When to use:
    - R
    - Minimal SQL
    - Distributed compute
- High-level Databricks architecture (only what R users need)
- Typical healthcare & research analytics workflows

- What R users must know — and what they can safely ignore

Module 2: Working with R in Databricks Notebooks

- Databricks notebooks for R users
- Writing and running R code in Databricks
- Working with data frames and tibbles at scale
- Reading data from tables and files into R
- Minimal SQL exposure (only for data access)

Module 3: Data Preparation & Engineering Using R

- Data ingestion patterns from an R perspective
- Cleaning and transforming data using R
- Handling missing values, inconsistencies, and outliers
- Creating reusable, analysis-ready datasets
- Structuring datasets for long-term use

Module 4: Advanced Dataset Management with Delta

- Delta tables explained for R users
- Reading and writing Delta data from R
- Versioning and time travel for research use cases
- Updating datasets safely without breaking analyses
- Performance considerations for large datasets in R

Module 5: Advanced Analytics in R on Databricks

- Advanced aggregations and summaries in R
- Feature engineering for statistical analysis
- Preparing datasets for modelling and inference
- Working with large datasets efficiently in R
- Exporting results for downstream reporting

Module 6: Reproducible R Workflows, Governance & Best Practices

- Designing reproducible R workflows in Databricks
- Organising R projects for team use
- Dataset documentation and data dictionaries
- Version control concepts for data & analysis (practical, light)
- Refreshing datasets without breaking analyses
- Common failure patterns and how to avoid them
- Practical checklist for “production-ready” research datasets
- Final Q&A and next steps
