

**Introduction to Databricks on Azure**

**Course Number:** AZR-152WA  
**Duration:** 2 days

**Overview**

In this Azure Databricks course, participants explore data lake storage integration, database management, Delta Lake fundamentals, and advanced data analysis techniques. The course covers pipeline and job automation and monitoring strategies for optimized performance. Attendees delve into fundamental Big Data principles and practical applications of Apache Spark.  Students also get hands-on Azure Databricks experience for data engineering and analysis.

**Prerequisites**

A basic understanding of SQL and Python is helpful but not necessary.

**Materials**

All students receive comprehensive courseware covering all topics in the course.

**Software Needed on Each Student PC**

Attendees will not need to install any software on their computers for this class. The class will be conducted in a remote environment that Accelebrate will provide; students will only need a local computer with a web browser and a stable Internet connection. Any recent version of Microsoft Edge, Mozilla Firefox, or Google Chrome will work well.

**Objectives**

* Understand the fundamental principles of Big Data and its significance in modern data management.
* Navigate the Azure Databricks platform effectively, including its architecture, portal, and cluster management functionalities.
* Develop practical skills for working with databases and tables within Azure Databricks, utilizing both SQL and PySpark for data manipulation.
* Learn advanced data analysis techniques, including querying, visualization, and exploratory data analysis (EDA), to derive meaningful insights from large datasets.
* Explore pipeline and workflow automation strategies to streamline data processing tasks.
* Implement effective monitoring techniques to optimize performance and ensure reliable data processing workflows.

**Outline**

* Cloud Data Engineering Fundamentals
  + Big Data Overview
  + On-Premises vs. Cloud Data Management Contrasts
  + Data Engineering Essentials
  + Business-driven Data Processing
  + Introduction to Apache Spark
  + Spark's Practical Applications
* Azure Databricks Basics
  + Spark and Azure Databricks
  + Azure Databricks Architecture Overview
  + Navigating the Azure Databricks Portal
  + Cluster Creation Process
  + Cluster Management Essentials
* Azure Databricks Development Environment
  + Overview of Development Environment
  + Notebooks Functionality
  + Practical Notebook Utilization
* File Systems and Data Lake Integration
  + Understanding DBFS
  + Accessing DBFS via Databricks UI
  + Uploading Data to DBFS
  + dbutils for DBFS Interaction
  + Azure Data Lake Storage Integration
  + Utilizing dbutils for Data Lake Mounting
* Database and Table Management in Azure Databricks
  + Understanding Databases and Tables
  + Creating and Managing Databases
  + Working with Tables
  + Using SQL with Tables
  + Using PySpark with Tables
  + Table Features Exploration
  + Understanding Partitioned Tables
* Views in Azure Databricks
  + Understanding Views
  + Using SQL with Views
  + Temporary and Global Views
  + Using PySpark with Views
* Data Analysis in Azure Databricks
  + Querying, Visualizing, and EDA
  + SQL Data Querying
  + PySpark Data Querying
  + Multi-Table Joins
  + Exploratory Data Analysis
  + Table Visualization Techniques
  + Using Charts
  + Data Profiling
* JDBC Integration in Azure Databricks
  + Advantages of JDBC Usage
  + Data Source Addition via JDBC
  + JDBC URL and Connection Parameters
  + Query Execution via JDBC
* Delta Lake in Azure Databricks
  + Introduction to Delta Lake
  + Delta Lake Architecture
  + Features and Advantages of Delta Lake
  + Using Delta Lake for Reliable Data Lakes
* Pipeline and Workflow Automation in Azure Databricks
  + Introduction to Pipelines and Workflow Automation
  + Creating and Managing Pipelines
  + Defining Dependencies and Triggers
  + Incorporating Data Processing
  + Implementing Error Handling
  + Scheduling Execution
* Monitoring and Optimization
  + Spark UI Monitoring
  + Storage Performance Analysis
  + Worker Node and Executor Evaluation
  + Performance Metrics Utilization