

**Programming on Azure Databricks with PySpark, SQL, and Scala Training**

**Course Number:** AZR-156WA  
**Duration:** 3 days

**Overview**

In this Databricks training, attendees master the Azure Databricks cloud platform. This course allows students to work with multiple programming languages and systems, including PySpark, SQL, and Scala, to determine which language/system best suits specific tasks.

**Prerequisites**

Practical knowledge of data processing and experience using at least one programming language.

**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**

* Create and manage Azure Databricks workspaces and clusters using UI and automation
* Understand the Databricks developer experience, including language choices, notebook environments, and table creation
* Explore the fundamentals of Apache Spark, its architecture, and common use cases
* Utilize Spark SQL and DataFrames for data manipulation and analysis
* Leverage the pandas library for data exploration and visualization in Python
* Create visualizations using Matplotlib and Seaborn to gain insights from data

**Outline**

* Introduction
* Azure Databricks
  + Azure Databricks
  + Creating an Azure Databricks Workspace UI
  + The Azure Databricks Service Blade
  + The Databricks Dashboard
  + Databricks Cluster Creation UI
  + Databricks File System (DBFS)
  + Databricks Integration with Data Lake
  + Automation Jobs
  + Databricks Developer Experience
  + Development Environments
  + Which Databricks-Supported Language Should I Use?
  + Notebook Runtime Flavor Configuration
  + The Notebook UI
  + Creating Tables
  + Create a New Table UI
  + Creating a Table from a DBFS File
  + Creating Your Table Visually with Databricks UI (The Preview Screen)
  + Querying a Databricks Table using SQL
  + A Data Profile Visualization Example
  + Performing Exploratory Data Analysis (EDA) with Data Charts
  + Spark and Databricks
  + Real-time Transformations
  + Databricks Machine Learning (ML)
  + The Cost of Doing Business on Databricks
* Introduction to Apache Spark
  + What is Apache Spark
  + The Spark Platform
  + Spark vs Hadoop's MapReduce (MR)
  + Common Spark Use Cases
  + Languages Supported by Spark
  + Running Spark on a Cluster
  + The Spark Application Architecture
  + The Driver Process
  + The Executor and Worker Processes
  + Spark Shell
  + Jupyter Notebook Shell Environment
  + Spark Applications
  + The spark-submit Tool
  + The spark-submit Tool Configuration
  + Interfaces with Data Storage Systems
  + The Resilient Distributed Dataset (RDD)
  + Datasets and DataFrames
  + Spark SQL, DataFrames, and Catalyst Optimizer
  + Project Tungsten
  + Spark Machine Learning Library
  + Spark (Structured) Streaming
  + GraphX
  + Extending Spark Environment with Custom Modules and Files
  + Spark 3
  + Spark 3 Updates at a Glance
* The Spark Shell
  + The Spark Shell
  + The Spark v.2 + Command-Line Shells
  + The Spark Shell UI
  + Spark Shell Options
  + Getting Help
  + Jupyter Notebook Shell Environment
  + Example of a Jupyter Notebook Web UI (Databricks Cloud)
  + The Spark Context (sc) and Spark Session (spark)
  + Creating a Spark Session Object in Spark Applications
  + The Shell Spark Context Object (sc)
  + The Shell Spark Session Object (spark)
  + Loading Files
  + Saving Files
* Introduction to Spark SQL
  + What is Spark SQL?
  + Uniform Data Access with Spark SQL
  + Using JDBC Sources
  + Hive Integration
  + What is a DataFrame?
  + Creating a DataFrame in PySpark
  + Creating a DataFrame in PySpark (Cont'd)
  + Commonly Used DataFrame Methods and Properties in PySpark
  + Commonly Used DataFrame Methods and Properties in PySpark (Cont'd)
  + Grouping and Aggregation in PySpark
  + The "DataFrame to RDD" Bridge in PySpark
  + The SQLContext Object
  + Converting an RDD to a DataFrame Example
  + Performance, Scalability, and Fault-tolerance of Spark SQL
* Introduction to pandas
  + What is pandas?
  + Conversion Between PySpark and pandas DataFrames
  + Pandas API on Spark
  + The pandas DataFrame Object
  + The DataFrame's Value Proposition
  + Creating a pandas DataFrame
  + Getting DataFrame Metrics
  + Accessing DataFrame Columns
  + Accessing DataFrame Rows
  + Accessing DataFrame Cells
  + Deleting Rows and Columns
  + Adding a New Column to a DataFrame
  + Getting Descriptive Statistics of DataFrame Columns
  + Getting Descriptive Statistics of DataFrames
  + Sorting DataFrames
  + Reading From CSV Files
  + Writing to a CSV File
* Data Visualization with seaborn in Python
  + Data Visualization
  + Data Visualization in Python
  + Matplotlib
  + Getting Started with matplotlib
  + Figures
  + Saving Figures to a File
  + Seaborn
  + Getting Started with seaborn
  + Histograms and KDE
  + Plotting Bivariate Distributions
  + Scatter plots in seaborn
  + Pair plots in seaborn
  + Heatmaps
* Conclusion