

**Advanced Data Analytics with PySpark**

**Course Number:** PYTH-252WA
**Duration:** 2 days

**Overview**

This Advanced Data Analytics with PySpark course teaches attendees how to develop and run more sophisticated Spark analytics tasks using PySpark (the Python API for Apache Spark). Participants learn to manipulate and analyze data using the Spark Shell to structure data with Spark SQL, Pandas, and Seaborn. By the end of this course, attendees are ready to successfully tackle their large-scale data set projects.

**Prerequisites**

* Successful completion of class [PYTH-100, Introduction to Python Programming](file:////training/python) (or equivalent Python knowledge)
* Knowledge of SQL

**Materials**

All Data Analytics training students receive comprehensive courseware.

**Software Needed on Each Student PC**

* Windows, Mac, or Linux with at least 8 GB RAM
	+ Most class activities will create Spark code and visualizations in a browser-based notebook environment. The class also details how to export these notebooks and how to run code outside of this environment.
* A current version of Anaconda for Python 3.x
* Related lab files that Accelebrate will provide
* Internet access

**Objectives**

* Work with the PySpark Shell Environment
* Understand Spark DataFrames
* Process data with the PySpark DataFrame API
* Work with Pivot Tables in PySpark
* Perform data visualization and Exploratory Data Analysis (EDA) in PySpark

**Outline**

* 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
	+ Project Tungsten
	+ The Resilient Distributed Dataset (RDD)
	+ Datasets and DataFrames
	+ Spark SQL, DataFrames, and Catalyst Optimizer
	+ Spark Machine Learning Library
	+ GraphX
	+ Extending Spark Environment with Custom Modules and Files
* 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
	+ Summary
* Introduction to Spark SQL
	+ What is Spark SQL?
	+ Uniform Data Access with Spark SQL
	+ Hive Integration
	+ Hive Interface
	+ Integration with BI Tools
	+ What is a DataFrame?
	+ Creating a DataFrame in PySpark
	+ Commonly Used DataFrame Methods and Properties in PySpark
	+ Grouping and Aggregation in PySpark
	+ The "DataFrame to RDD" Bridge in PySpark
	+ The SQLContext Object
	+ Examples of Spark SQL/DataFrame (PySpark Example)
	+ Converting an RDD to a DataFrame Example
	+ Example of Reading/Writing a JSON File
	+ Using JDBC Sources
	+ JDBC Connection Example
	+ Performance, Scalability, and Fault-tolerance of Spark SQL
	+ Summary
* Practical Introduction to Pandas
	+ What is pandas?
	+ The Series Object
	+ Accessing Values and Indexes in Series
	+ Setting Up Your Own Index
	+ Using the Series Index as a Lookup Key
	+ Can I Pack a Python Dictionary into a Series?
	+ The DataFrame Object
	+ The DataFrame's Value Proposition
	+ Creating a pandas DataFrame
	+ Getting DataFrame Metrics
	+ Accessing DataFrame Columns
	+ Accessing DataFrame Rows
	+ Accessing DataFrame Cells
	+ Using iloc
	+ Using loc
	+ Examples of Using loc
	+ DataFrames are Mutable via Object Reference!
	+ Deleting Rows and Columns
	+ Adding a New Column to a DataFrame
	+ Appending/Concatenating DataFrame and Series Objects
	+ Example of Appending/Concatenating DataFrames
	+ Re-indexing Series and DataFrames
	+ Getting Descriptive Statistics of DataFrame Columns
	+ Getting Descriptive Statistics of DataFrames
	+ Applying a Function
	+ Sorting DataFrames
	+ Reading From CSV Files
	+ Writing to the System Clipboard
	+ Writing to a CSV File
	+ Fine-Tuning the Column Data Types
	+ Changing the Type of a Column
	+ What May Go Wrong with Type Conversion
	+ Summary
* 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
* Quick Introduction to Python for Data Engineers (Optional)
	+ What is Python?
	+ Additional Documentation
	+ Which version of Python am I running?
	+ Python Dev Tools and REPLs
	+ IPython
	+ Jupyter
	+ Jupyter Operation Modes
	+ Jupyter Common Commands
	+ Anaconda
	+ Python Variables and Basic Syntax
	+ Variable Scopes
	+ PEP8
	+ The Python Programs
	+ Getting Help
	+ Variable Types
	+ Assigning Multiple Values to Multiple Variables
	+ Null (None)
	+ Strings
	+ Finding the Index of a Substring
	+ String Splitting
	+ Triple-Delimited String Literals
	+ Raw String Literals
	+ String Formatting and Interpolation
	+ Boolean
	+ Boolean Operators
	+ Numbers
	+ Looking Up the Runtime Type of a Variable
	+ Divisions
	+ Assignment-with-Operation
	+ Relational Operators
	+ The if-elif-else Triad
	+ An if-elif-else Example
	+ Conditional Expressions (a.k.a. Ternary Operator)
	+ The While-Break-Continue Triad
	+ The for Loop
	+ try-except-finally
	+ Lists
	+ Main List Methods
	+ Dictionaries
	+ Working with Dictionaries
	+ Sets
	+ Common Set Operations
	+ Set Operations Examples
	+ Finding Unique Elements in a List
	+ Enumerate
	+ Tuples
	+ Unpacking Tuples
	+ Functions
	+ Dealing with Arbitrary Number of Parameters
	+ Keyword Function Parameters
	+ The range Object
	+ Random Numbers
	+ Python Modules
	+ Importing Modules
	+ Installing Modules
	+ Listing Methods in a Module
	+ Creating Your Own Modules
	+ Creating a Runnable Application
	+ List Comprehension
	+ Zipping Lists
	+ Working with Files
	+ Reading and Writing Files
	+ Reading Command-Line Parameters
	+ Accessing Environment Variables
	+ What is Functional Programming (FP)?
	+ Higher-Order Functions
	+ Lambda Functions in Python
	+ Lambdas in the Sorted Function
	+ Other Examples of Using Lambdas
	+ Regular Expressions
	+ Using Regular Expressions Examples
	+ Python Data Science-Centric Libraries
* Conclusion