

**Data Engineering, ETL, and DataOps**

**Course Number:** PYTH-280WA
**Duration:** 3 days

**Overview**

This Data Engineering training course teaches attendees the basics of  Extract, Transform, and Load (ETL), data processing technologies, data manipulation with Pandas, data visualization using Python, and the principles of DataOps. The course also covers Apache Spark, Spark SQL, and essential Python skills.

**Prerequisites**

Attendees should be familiar with programming concepts (Python is a plus.) An understanding of big data concepts is beneficial but is not mandatory.

**Materials**

All Data Engineering training students receive comprehensive courseware.

**Software Needed on Each Student PC**

* Anaconda Python 3.6 or later
* Spyder IDE and Jupyter notebook (Comes with Anaconda)

**Objectives**

* Explain the benefits of using Apache Spark for big data processing
* Design a Spark application to process a distributed dataset
* Implement Spark SQL to perform complex data analysis and transformations
* Optimize Spark applications for performance and scalability
* Use Spark Streaming to process real-time data streams
* Build machine learning models using Spark MLlib
* Develop graph processing applications using Spark GraphX
* Integrate Spark with other big data frameworks, such as Hadoop and Hive
* Deploy Spark applications to production environments
* Monitor and troubleshoot Spark applications in production
* Apply DataOps principles to improve the efficiency and quality of data engineering projects
* Use Pandas and Seaborn to create and interpret data visualizations in Python

**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
* 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
* 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
* 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
* Intro to DataOps
	+ Problems in the Data & Analytics Industry
	+ Root Cause: Organizational Complexities
	+ Solution: What Is DataOps?
* DataOps Production Pipeline
	+ The Three DataOps Pipelines
	+ Meta-Orchestrate Tools, Teams & Processes
	+ Automate Tests for Error Detection
	+ Types of Tests
	+ Measure Production Processes, Reflect & Improve
* DataOps Development Pipeline
	+ Development Lifecycle Complexities
	+ Data & Analytics Development
	+ How to Achieve Fast Deployments
	+ DataOps Deployments: Beyond DevOps
* DataOps Environment Pipeline
	+ DataOps Environment Challenges
	+ Environment Management: Components & Use Cases
	+ Principles of DataOps Environments
* DataOps Implementation
	+ Lean DataOps Implementation
	+ Four Phases of Lean DataOps
	+ Getting started with DataOps
* 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 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
	+ Comments:
	+ 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)?
	+ Terminology: Higher-Order Functions
	+ Lambda Functions in Python
	+ Example: Lambdas in the Sorted Function
	+ Other Examples of Using Lambdas
	+ Regular Expressions
	+ Using Regular Expressions Examples
	+ Python Data Science-Centric Libraries
* Conclusion