

**Introduction to Data Engineering with Python**

**Course Number:** PYTH-242WA
**Duration:** 3 days

**Overview**

Python is a popular programming language used by data scientists, analysts, and engineers for its flexibility and robust data science libraries. In this Data Engineering with Python training course, attendees learn how to apply the Python programming language to practical data engineering using popular Python libraries and related platforms such as NumPy, pandas, matplotlib, scikit-learn, Apache Spark, and more.

**Prerequisites**

Participants must have coding experience in at least one modern programming language. Python programming knowledge is beneficial, but not necessary.

**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**

* Perform A/B testing
* Understand data availability and consistency
* Use Jupyter Notebooks
* Understand Python
* Work with NumPy
* Work with files in pandas
* Perform data grouping and aggregation
* Repair and normalize data
* Work with data visualization in Jupyter Notebooks using matplotlib
* Work with Exploratory Data Analysis (EDA)
* Correlate cause and effect
* Use the Parquet data format

**Outline**

* Introduction
* Defining Data Engineering
	+ Data is King
	+ Translating Data into Operational and Business Insights
	+ What is Data Engineering
	+ The Data-Related Roles
	+ The Data Science Skill Sets
	+ The Data Engineer Role
	+ Core Skills and Competencies
	+ An Example of a Data Product
	+ What is Data Wrangling (Munging)?
	+ The Data Exchange Interoperability Options
* Distributed Computing Concepts for Data Engineers
	+ The Traditional Client–Server Processing Pattern
	+ Enter Distributed Computing
	+ Data Physics
	+ Data Locality (Distributed Computing Economics)
	+ The CAP Theorem
	+ Mechanisms to Guarantee a Single CAP Property
	+ Eventual Consistency
* Data Processing Phases
	+ Typical Data Processing Pipeline
	+ Data Discovery Phase
	+ Data Harvesting Phase
	+ Data Priming Phase
	+ Exploratory Data Analysis
	+ Model Planning Phase
	+ Model Building Phase
	+ Communicating the Results
	+ Production Roll-out
	+ Data Logistics and Data Governance
	+ Data Processing Workflow Engines
	+ Apache Airflow
	+ Data Lineage and Provenance
	+ Apache NiFi
* Introduction to Python for Data Engineers
	+ 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
	+ Dates and Times
	+ 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
* Introduction to NumPy
	+ SciPy
	+ NumPy
	+ The First Take on NumPy Arrays
	+ Getting Help
	+ Understanding Axes
	+ Indexing Elements in a NumPy Array
	+ NumPy Arrays
	+ Understanding Types
	+ Re-Shaping
	+ Commonly Used Array Metrics
	+ Commonly Used Aggregate Functions
	+ Sorting Arrays
	+ Vectorization
	+ Broadcasting
	+ Filtering
	+ Array Arithmetic Operations
	+ Array Slicing
	+ 2-D Array Slicing
	+ The Linear Algebra Functions
* 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
* Descriptive Statistics Computing Features in Python
	+ Descriptive Statistics
	+ Non-uniformity of a Probability Distribution
	+ Using NumPy for Calculating Descriptive Statistics Measures
	+ Finding Min and Max in NumPy
	+ Using pandas for Calculating Descriptive Statistics Measures
	+ Correlation
	+ Regression and Correlation
	+ Covariance
	+ Getting Pairwise Correlation and Covariance Measures
	+ Finding Min and Max in pandas DataFrame
* Data Grouping and Aggregation with Pandas
	+ Data Aggregation and Grouping
	+ Sample Data Set
	+ The pandas.core.groupby.SeriesGroupBy Object
	+ Grouping by Two or More Columns
	+ Emulating SQL's WHERE Clause
	+ The Pivot Tables
	+ Cross-Tabulation
* Repairing and Normalizing Data
	+ Repairing and Normalizing Data
	+ Dealing with the Missing Data
	+ Sample Data Set
	+ Getting Info on Null Data
	+ Dropping a Column
	+ Interpolating Missing Data in pandas
	+ Replacing the Missing Values with the Mean Value
	+ Scaling (Normalizing) the Data
	+ Data Preprocessing with scikit-learn
	+ Scaling with the scale() Function
	+ The MinMaxScaler Object
* Data Visualization in Python using Matplotlib
	+ Data Visualization
	+ What is matplotlib?
	+ Getting Started with matplotlib
	+ The matplotlib.pyplot.plot() Function
	+ The matplotlib.pyplot.scatter() Function
	+ Labels and Titles
	+ Styles
	+ The matplotlib.pyplot.bar() Function
	+ The matplotlib.pyplot.hist () Function
	+ The matplotlib.pyplot.pie () Function
	+ The Figure Object
	+ The matplotlib.pyplot.subplot() Function
	+ Selecting a Grid Cell
	+ Saving Figures to a File
* Parallel Data Processing with PySpark
	+ What is Apache Spark
	+ The Spark Platform
	+ Languages Supported by Spark
	+ Running Spark on a Cluster
	+ The Spark Shell
	+ The High-Level Execution Flow in Stand-alone Spark Cluster
	+ The Spark Application Architecture
	+ The Resilient Distributed Dataset (RDD)
	+ The Lineage Concept
	+ Datasets and DataFrames
	+ Data Partitioning
	+ Data Partitioning Diagram
	+ Finding the Most Frequently Used Words in PySpark
* Python as a Cloud Scripting Language
	+ Python's Value
	+ Python on AWS
	+ AWS SDK For Python (boto3)
	+ What is Serverless Computing?
	+ How Functions Work
	+ The AWS Lambda Event Handler
	+ What is AWS Glue?
	+ PySpark on Glue - Sample Script
* Conclusion