

**Intermediate Data Engineering with Python**

**Course Number:** PYTH-244WA
**Duration:** 2 days

**Overview**

Data engineering is a fast-growing field that requires advanced knowledge of modern data management and analytics. This Intermediate Data Engineering training course goes beyond the basics to teach attendees how to use Python, the Spark platform, AWS Glue, and more to construct sophisticated data workflows.

**Prerequisites**

All attendees must have taken [Intro to Data Engineering with Python](file:////training/python-data-engineering-intro), or have the equivalent knowledge.

**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**

* Work with the Databricks community cloud lab environment
* Perform data visualization and EDA with pandas and seaborn
* Correlate cause and effect
* Understand the PySpark shell environment
* Understand Spark DataFrames
* Work with the PySpark DataFrame API
* Perform data repair and normalization in PySpark
* Work with Parquet file format in PySpark and pandas
* Use AWS Glue crawlers and classifiers
* Create an S3 Bucket for AWS Glue ETL script output
* Create and Work with Glue scripts using dev endpoints
* Using PySpark API directly
* Understand AWS Glue ETL jobs

**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
* Spark RDDs
	+ The Resilient Distributed Dataset (RDD)
	+ Ways to Create an RDD
	+ Supported Data Types
	+ RDD Operations
	+ RDDs are Immutable
	+ Spark Actions
	+ RDD Transformations
	+ Other RDD Operations
	+ Chaining RDD Operations
	+ RDD Lineage
	+ The Big Picture
	+ What May Go Wrong
	+ Miscellaneous Pair RDD Operations
	+ RDD Caching
* 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
* Overview of the Amazon Web Services (AWS)
	+ Amazon Web Services
	+ The History of AWS
	+ The Initial Iteration of Moving amazon.com to AWS
	+ The AWS (Simplified) Service Stack
	+ Accessing AWS
	+ Direct Connect
	+ Shared Responsibility Model
	+ Trusted Advisor
	+ The AWS Distributed Architecture
	+ AWS Services
	+ Managed vs. Unmanaged Amazon Services
	+ Amazon Resource Name (ARN)
	+ Compute and Networking Services
	+ Elastic Compute Cloud (EC2)
	+ AWS Lambda
	+ Auto Scaling
	+ Elastic Load Balancing (ELB)
	+ Virtual Private Cloud (VPC)
	+ Route53 Domain Name System
	+ Elastic Beanstalk
	+ Security and Identity Services
	+ Identity and Access Management (IAM)
	+ AWS Directory Service
	+ AWS Certificate Manager
	+ AWS Key Management Service (KMS)
	+ Storage and Content Delivery
	+ Elastic Block Storage (EBS)
	+ Simple Storage Service (S3)
	+ Glacier
	+ CloudFront Content Delivery Service
	+ Database Services
	+ Relational Database Service (RDS)
	+ DynamoDB
	+ Amazon ElastiCache
	+ Redshift
	+ Messaging Services
	+ Simple Queue Service (SQS)
	+ Simple Notifications Service (SNS)
	+ Simple Email Service (SES)
	+ AWS Monitoring with CloudWatch
	+ Other Services Example
* Introduction to AWS Glue
	+ What is AWS Glue?
	+ AWS Glue Components
	+ Managing Notebooks
	+ Putting it Together: The AWS Glue Environment Architecture
	+ AWS Glue Main Activities
	+ Additional Glue Services
	+ When To Use AWS Glue?
	+ Integration with other AWS Services
* Introduction to Apache Spark
	+ What is Apache Spark
	+ The Spark Platform
	+ Uniform Data Access with Spark SQL
	+ 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
	+ Interfaces with Data Storage Systems
	+ The Resilient Distributed Dataset (RDD)
	+ Datasets and DataFrames
	+ Data Partitioning
	+ Data Partitioning Diagram
* AWS Glue PySpark Extensions
	+ AWS Glue and Spark
	+ The DynamicFrame Object
	+ The DynamicFrame API
	+ The GlueContext Object
	+ Glue Transforms
	+ A Sample Glue PySpark Script
	+ Using PySpark
	+ AWS Glue PySpark SDK
* Conclusion