

**Introduction to Spark 3 with Python**

**Course Number:** SPRK-112  
**Duration:** 3 days

**Overview**

This Spark 3 with Python training course teaches attendees the basic and higher-level constructs of Spark using Python programming. Attendees receive a solid technical introduction to the Spark architecture and learn how Spark works. Participants learn how to leverage Spark SQL, DataFrames, and DataSets, which are now the preferred programming API. In addition, students explore possible performance issues and strategies for optimization. The course also covers more advanced topics, including the use of Spark Streaming to process streaming data and Kafka server integration.

**Prerequisites**

All attendees must have solid experience programming in Python 3 or later.

**Materials**

All Spark training attendees receive comprehensive courseware.

**Software Needed on Each Student PC**

* Windows, Mac, or Linux PCs with the current Chrome or Firefox browser.
  + 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.
* Internet access

**Objectives**

All students will:

* Understand the need for Spark in data processing
* Understand the Spark architecture and how it distributes computations to cluster nodes
* Be familiar with basic installation/setup/layout of Spark
* Use the Spark shell for interactive and ad-hoc operations
* Understand RDDs (Resilient Distributed Datasets), and data partitioning, pipelining, and computations
* Understand/use RDD ops such as map(), filter(), and others.
* Understand and use Spark SQL and the DataFrame/DataSet API.
* Understand DataSet/DataFrame capabilities, including the Catalyst query optimizer and Tungsten memory/CPU optimizations.
* Be familiar with performance issues, and use the DataSet/DataFrame and Spark SQL for efficient computations
* Understand Spark’s data caching and use it for efficient data transfer
* Write/run standalone Spark programs with the Spark API
* Use Spark Structured Streaming to process streaming (real-time) data
* Ingest streaming data from Kafka, and process via Spark Structured Streaming
* Understand performance implications and optimizations when using Spark

**Outline**

* Introduction to Spark
  + Overview, Motivations, Spark Systems
  + Spark Ecosystem
  + Spark vs. Hadoop
  + Acquiring and Installing Spark
  + The Spark Shell, SparkContext
* RDDs and Spark Architecture
  + RDD Concepts, Lifecycle, Lazy Evaluation
  + RDD Partitioning and Transformations
  + Working with RDDs - Creating and Transforming (map, filter, etc.)
* Spark SQL, DataFrames, and DataSets
  + Overview
  + SparkSession, Loading/Saving Data, Data Formats (JSON, CSV, Parquet, text, etc.)
  + Introducing DataFrames and DataSets (Creation and Schema Inference)
  + Supported Data Formats (JSON, Text, CSV, Parquet)
  + Working with the DataFrame (untyped) Query DSL (Column, Filtering, Grouping, Aggregation)
  + SQL-based Queries
  + Working with the DataSet (typed) API
  + Mapping and Splitting (flatMap(), explode(), and split())
  + DataSets vs. DataFrames vs. RDDs
* Shuffling Transformations and Performance
  + Grouping, Reducing, Joining
  + Shuffling, Narrow vs. Wide Dependencies, and Performance Implications
  + Exploring the Catalyst Query Optimizer (explain(), Query Plans, Issues with lambdas)
  + The Tungsten Optimizer (Binary Format, Cache Awareness, Whole-Stage Code Gen)
* Performance Tuning
  + Caching - Concepts, Storage Type, Guidelines
  + Minimizing Shuffling for Increased Performance
  + Using Broadcast Variables and Accumulators
  + General Performance Guidelines
* Creating Standalone Applications
  + Core API, SparkSession.Builder
  + Configuring and Creating a SparkSession
  + Building and Running Applications - sbt/build.sbt and spark-submit
  + Application Lifecycle (Driver, Executors, and Tasks)
  + Cluster Managers (Standalone, YARN, Mesos)
  + Logging and Debugging
* Spark Streaming
  + Introduction and Streaming Basics
  + Structured Streaming
    - Continuous Applications
    - Table Paradigm, Result Table
    - Steps for Structured Streaming
    - Sources and Sinks
  + Consuming Kafka Data
    - Kafka Overview
    - Structured Streaming - "Kafka" format
    - Processing the Stream
* Conclusion