

Introduction to Distributed Computing with Spark

Learn via: **Classroom**

Duration: **3 Day**

Overview

Spark is at the forefront of distributed computing. This module is taught using Python and includes the basics of critical functional programming constructs such as map, flatmap, list comprehensions, and data structures.

We cover core concepts of Spark-like resilient distributed datasets, memory caching, actions, transformations, tuning, and optimization. Students get to build functioning applications from end to end and learn critical tooling to enhance productivity. They apply that knowledge directly to developing, building, and deploying Spark jobs that they will run on large, real-world datasets in the cloud (AWS and Google Cloud Platform).

Prerequisites

Basic to intermediate Python, basic to intermediate programming, and/or successful completion of the Introduction to Machine Learning course

Who Should Attend

Data analysts or data scientists with some programming experience looking to utilize advances in cloud computing to more effectively work with big data.

What You Will Learn

- Basics of the Spark API
- Big data development considerations and techniques