Syllabus

Lesson 1 - Introduction to Spark - Getting started

- Understand and be able to explain the purpose of Spark
- List and describe the components of the Spark unified stack
- Understand the basics of Resilient Distributed Dataset (RDD)
- Download and install Spark standalone
- Scala and Python overview
- Launch and use Spark's Scala and Python shell ©

Lesson 2 - Resilient Distributed Dataset and DataFrames

- Understand how to create parallelized collections and external datasets
- Work with Resilient Distributed Dataset (RDD) operations
- Utilize shared variables and key-value pairs

Lesson 3 - Spark application programming

- Understand the purpose and usage of the SparkContext
- Initialize Spark with the various programming languages
- Describe and run some Spark examples
- Pass functions to Spark
- Create and run a Spark standalone application
- Submit applications to the cluster

Lesson 4 - introduction to Spark libraries

- Understand and use the various Spark libraries such as SparkSQL, Spark Streaming, MLlib, and GraphX

Lesson 5 - Spark configuration, monitoring and tuning

- Understand components of the Spark cluster
- Configure Spark to modify the Spark properties, environmental variables, or logging properties
- Monitor Spark using the web UIs, metrics, and external instrumentation
- Understand performance tuning considerations