From 0 to 1: Spark for Data Science with Python

Modality: Self-Paced Learning

Duration: 8 Hours

SATV Value:

CLC:

NATU:

SUBSCRIPTION: Learn, Master

Get your data to fly using Spark for analytics, machine learning and data science

Let’s parse that.

What's Spark? If you are an analyst or a data scientist, you're used to having multiple systems for working with data. SQL, Python, R, Java, etc. With Spark, you have a single engine where you can explore and play with large amounts of data, run machine learning algorithms and then use the same system to productionize your code.

Analytics: Using Spark and Python you can analyze and explore your data in an interactive environment with fast feedback. The course will show how to leverage the power of RDDs and Dataframes to manipulate data with ease.

Machine Learning and Data Science : Spark's core functionality and built-in libraries make it easy to implement complex algorithms like Recommendations with very few lines of code. We'll cover a variety of datasets and algorithms including PageRank, MapReduce and Graph datasets.

Course Objective:

- Music Recommendations using Alternating Least Squares and the Audioscrobbler dataset
- Dataframes and Spark SQL to work with Twitter data
- Using the PageRank algorithm with Google web graph dataset
- Using Spark Streaming for stream processing
- Working with graph data using the Marvel Social network dataset

Audience:

- Analysts who want to leverage Spark for analyzing interesting datasets
- Data Scientists who want a single engine for analyzing and modelling data as well as productionizing it.
- Engineers who want to use a distributed computing engine for batch or stream processing or both

Prerequisite:

- The course assumes knowledge of Python. You can write Python code directly in the PySpark

shell. If you already have IPython Notebook installed, we'll show you how to configure it for Spark
- For the Java section, we assume basic knowledge of Java. An IDE which supports Maven, like IntelliJ IDEA/Eclipse would be helpful
- All examples work with or without Hadoop. If you would like to use Spark with Hadoop, you'll need to have Hadoop installed (either in pseudo-distributed or cluster mode).

Course Outline:

You, This Course and Us

Introduction to Spark

- What does Donald Rumsfeld have to do with data analysis?
- Why is Spark so cool?
- An introduction to RDDs - Resilient Distributed Datasets
- Built-in libraries for Spark
- Installing Spark
- The PySpark Shell
- Transformations and Actions
- See it in Action : Munging Airlines Data with PySpark - I
- [For Linux/Mac OS Shell Newbies] Path and other Environment Variables

Resilient Distributed Datasets

- RDD Characteristics: Partitions and Immutability
- RDD Characteristics: Lineage, RDDs know where they came from
- What can you do with RDDs?
- Create your first RDD from a file
- Average distance travelled by a flight using map() and reduce() operations
- Get delayed flights using filter(), cache data using persist()
- Average flight delay in one-step using aggregate()
- Frequency histogram of delays using countByValue()
- See it in Action : Analyzing Airlines Data with PySpark - II

Advanced RDDs: Pair Resilient Distributed Datasets

- Special Transformations and Actions
- Average delay per airport, use reduceByKey(), mapValues() and join()
- Average delay per airport in one step using combineByKey()
- Get the top airports by delay using sortBy()
- Lookup airport descriptions using lookup(), collectAsMap(), broadcast()
- See it in Action : Analyzing Airlines Data with PySpark - III

Advanced Spark: Accumulators, Spark Submit, MapReduce , Behind The Scenes

- Get information from individual processing nodes using accumulators
- See it in Action : Using an Accumulator variable

- Long running programs using spark-submit
- See it in Action : Running a Python script with Spark-Submit
- Behind the scenes: What happens when a Spark script runs?
- Running MapReduce operations
- See it in Action : MapReduce with Spark

Java and Spark

- The Java API and Function objects
- Pair RDDs in Java
- Running Java code
- Installing Maven
- See it in Action : Running a Spark Job with Java

PageRank: Ranking Search Results

- What is PageRank?
- The PageRank algorithm
- Implement PageRank in Spark
- Join optimization in PageRank using Custom Partitioning
- See it Action : The PageRank algorithm using Spark

Spark SQL

- Dataframes: RDDs + Tables
- See it in Action : Dataframes and Spark SQL

MLlib in Spark: Build a recommendations engine

- Collaborative filtering algorithms
- Latent Factor Analysis with the Alternating Least Squares method
- Music recommendations using the Audioscrobbler dataset
- Implement code in Spark using MLlib

Spark Streaming

- Introduction to streaming
- Implement stream processing in Spark using Dstreams
- Stateful transformations using sliding windows
- See it in Action : Spark Streaming

Graph Libraries

- The Marvel social network using Graphs