Analyzing Data With Apache Spark (For Developers)

Course Summary

Description
This course will introduce Apache Spark. The students will learn how Spark fits into the Big Data ecosystem, and how to use Spark for data analysis.

This class is taught with either Python language or Scala language.

A language primer can be offered if needed.

Objective
What you will learn:

- Spark ecosystem
- Spark Shell
- Spark Data structures (RDD / Dataframe / Dataset)
- Spark SQL
- Modern data formats and Spark
- Spark API
- Spark & Hadoop & Hive
- Spark ML overview
- GraphX
- Spark Streaming

Topics

- Spark Introduction
- First look at Spark
- Spark Data structures
- Caching
- Dataframes / Datasets
- Spark SQL
- Spark and Hadoop
- Spark API
- Spark ML Overview
- GraphX
- Spark Streaming
- Workshops

Audience
This course is designed for Developers and Architects.

Prerequisites
Developer background

Duration
Three days

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically.
Analyzing Data With Apache Spark (For Developers)

Course Outline

I. Spark Introduction
   A. Big data, Hadoop, Spark
   B. Spark concepts and architecture
   C. Spark components overview
      • Labs: installing and running Spark

II. First look at Spark
   A. Spark shell
   B. Spark web UIs
   C. Analyzing dataset – part 1
      • Labs: Spark shell exploration

III. Spark Data structures
   A. Partitions
   B. Distributed execution
   C. Operations: transformations and actions
      • Labs: Unstructured data analytics using RDDs

IV. Caching
   A. Caching overview
   B. Various caching mechanisms available in Spark
   C. In memory file systems
   D. Caching use cases and best practices
      • Labs: Benchmark of caching performance

V. Dataframes / Datasets
   A. Dataframes Intro
   B. Loading structured data (json, CSV) using Dataframes
   C. Using schema
   D. Specifying schema for Dataframes
      • Labs: Dataframes, Datasets, Schema

VI. Spark SQL
   A. Spark SQL concepts and overview
   B. Defining tables and importing datasets
   C. Querying data using SQL
   D. Handling various storage formats: JSON / Parquet / ORC
      • Labs: querying structured data using SQL; evaluating data formats

VII. Spark and Hadoop
   A. Hadoop Primer: HDFS / YARN
   B. Hadoop + Spark architecture
   C. Running Spark on Hadoop YARN
   D. Processing HDFS files using Spark
   E. Spark & Hive

VIII. Spark API
   A. Overview of Spark APIs in Scala / Python
   B. Life cycle of an Spark application
   C. Spark APIs
   D. Deploying Spark applications on YARN
      • Labs: Developing and deploying an Spark application

IX. Spark ML Overview
   A. Machine Learning primer
   B. Machine Learning in Spark: MLib / ML
   C. Spark ML overview (newer Spark2 version)
   D. Algorithms overview: Clustering, Classifications, Recommendations
      • Labs: Writing ML applications in Spark

X. GraphX
   A. GraphX library overview
   B. GraphX APIs
   C. Create a Graph and navigating it
   D. Shortest distance
   E. Pregel API
      • Labs: Processing graph data using Spark

XI. Spark Streaming
   A. Streaming concepts
   B. Evaluating Streaming platforms
   C. Spark streaming library overview
   D. Streaming operations
   E. Sliding window operations
   F. Structured Streaming
   G. Continuous streaming
   H. Spark & Kafka streaming
      • Labs: Writing spark streaming applications

XII. Workshops
   A. These are group workshops
   B. Attendees will work on solving real world data analysis problems using Spark

Due to the nature of this material, this document refers to numerous hardware and software products by their trade names. References to other companies and their products are for informational purposes only, and all trademarks are the properties of their respective companies. It is not the intent of ProTech Professional Technical Services, Inc. to use any of these names generically.