Apache Spark (Basic) on Hadoop

Overview:

This three-day course covers the essentials for developers who need to create applications to analyze big data stored in Apache Hadoop using Spark. Each student has their own Spark cluster and have access to dozens of hands-on labs.

Duration

3 days

Who is the course for

Data Scientists and Software Developers

Prerequisites

To get the most out of this training, you should have the following knowledge or experience as they will not be discussed during class:

- ✓ Hadoop Distributed File System (HDFS), YARN (Yet Another Resource Manager) and MapReduce processing engine
- ✓ Scala or Python coding
- ✓ Linux command line experience

Course Outline

Module 0. Introduction and Setup:

- ✓ How to start Spark and Zeppelin services in Ambari
- ✓ How to login to Spark using Python and Scala

Module 1. Spark Architecture:

- ✓ What is Apache Spark?
- ✓ Spark processing (Jobs, Stages, Tasks)
- ✓ Spark components (Driver, Context, Yarn, HDFS, Workers, Executors)

Module 2. Getting Started with RDDs:

- ✓ Running queues in Python, Scala and Zeppelin
- ✓ Queries using most popular Transformations and Actions
- ✓ Creating RDDs

Module 3. Pair RDDs:

- ✓ Difference between RDDs and Pair RDD
- ✓ 1 Pair Actions, 1 Pair Transformations and 2 Pair Transformations

Apache Spark (Basic) on Hadoop

Module 4. Spark SQL:

- ✓ Working with DataFrames and Tables and DataSets
- ✓ Catalyst optimizer overview

Module 5. Spark Streaming):

- ✓ Working with DStreams
- ✓ Stateless and Stateful Streaming labs using HDFS and Sockets

Module 6. Visualizations using Zeppelin:

- ✓ Creating various Charts using DataFrames and Tables
- ✓ How to create Pivot charts and Dynamic forms

Module 7. Spark UI

- ✓ Overview of Job, Stage and Tasks
- ✓ Monitoring Spark jobs in Spark UI

Module 8. Performance Tuning:

- ✓ Caching, Checkpoint, Accumulators and Broadcast Variables
- ✓ Hashed Partitions, Tungsten, Executor memory and Serialization

Module 9. Spark Applications

- ✓ Creating an application via spark-submit
- ✓ Parameter configurations (number executors, driver memory, executor cores, etc.)

Module 10. Spark 2.0 Machine Learning (ML)

- ✓ How ML Pipelines work
- ✓ Making Predictions using Decision Tree

Format

Lecture / Lab

The course content can be customised to cover any specialised material you may require for your specific training needs.

This course can be offered as private on-site training hosted at your offices. For more information, please contact us at **info@unicom.co.uk**

Contact Us

UNICOM

www.unicom.co.uk
@UNICOMSeminars
info@unicom.co.uk
+44 1895 256 484 (UK)
www.youtube.com/unicomseminars
www.linkedin.com/UNICOMSeminars