Apache Spark (Advanced) on Hadoop

Overview:

The purpose of this 2-day training course is to acquaint you with Spark 2.0 functionality and Performance Tuning techniques. It will cover the fundamentals of the Spark project, covering the basics of RDDs (Resilient Distributed Datasets) and the various operators used (Transformations and Actions).

Duration

2 days

Who is the course for

Data Analysts and Software Developers

Prerequisites

To get the most out of this training, you should have the following knowledge or experience as is builds the foundation for the advanced course:

- ✓ Apache Spark (Basic) on Hadoop – Classroom (ILT#56317)

Course Outline

Module 0. Introduction and Setup:

- ✓ Zeppelin note

Module 1. Datasets and Catalog:

- ✓ What is a Dataset?
- ✓ When to use which object
- ✓ Encoders and semi-structured data
- ✓ Common ways to create DS
- ✓ Cannot create DS these ways
- ✓ Casting DS and convert DS to DF to RDD
- ✓ Review questions: Datasets / Catalog
- ✓ Dataset versus SQL/DataFrames
- ✓ Serialization performance using Encoders
- ✓ Dataset caching
- ✓ Creating DS from an RDD
- ✓ Casting DS and convert DS these ways
- ✓ Hive list Catalog
- ✓ In Review: Datasets / Catalog

Apache Spark (Advanced) on Hadoop

Module 2. Catalyst and Tungsten functionalities:

- ✓ Before we begin: Open Zeppelin note
- ✓ DataFrames, Datasets and Views use Catalyst / Tungsten
- ✓ Catalyst optimizer overview
- ✓ Catalyst: Join on 2 Spark views demo
- ✓ But RDDs can't use Catalyst
- ✓ Loading data in Spark 2.x and Catalyst
- ✓ Loading data in Spark 2.x and Catalyst
  - Load data (old way), then join
  - Execution Plan from 'old way' loading
  - DataFrameReader: Load / Execution plan
- ✓ Dropping hints to Catalyst
- ✓ Catalyst: column pruning demo
- ✓ Catalyst: Column (& Partition) pruning
- ✓ Catalyst: Predicate pushdown concepts
- ✓ Tungsten overview
  - Binary processing
  - Improved Memory usage
  - Improved caching demo
  - Whole-stage code gen
  - Whole-stage code gen demo
  - Whole-stage code gen Vectorization
- ✓ Review questions: Catalyst / Tungsten
- ✓ In Review: Catalyst / Tungsten

Module 3. Performance Tuning:

- ✓ 2 types of Machine Learning
- ✓ How Models are Created
- ✓ Four Common MLlib functions
- ✓ What is Supervised Learning?
- ✓ Spark Superbised Learning Workflow
- ✓ Unsupervised Learning
- ✓ RDD – Machine Learning (MLlib)
- ✓ KMeans scenario
  - Load data
  - Create Model and Predict
  - Compare Actual to Predict
- ✓ Collaborative Filtering (CF) recommender
- ✓ Lab: Will We like Star Wars?
- ✓ Classification Functions (Supervised)
  - Classification uses LabelPoint
- ✓ CASTing X-var and Y-vars for LabelPoint
- ✓ Logistic regression, Support Vector Machines, NaiveBayes and Decision Tree (Supervised)
- ✓ ML Pipeline terminology
- ✓ How ML Pipeline works
- ✓ Cleaning the data
- ✓ Train ML pipeline – The Big Picture
- ✓ Imprving the Model
- ✓ Lab: Predict Titanic Survivors (Random Forest)
- ✓ Review Questions: Machine Learning
- ✓ In Review: Machine Learning
- ✓ But wait, there's more (for MLlib) (Appendix)
- ✓ Linear Regression on scenario (Supervised)

Apache Spark (Advanced) on Hadoop

Format

Lecture / Lab

Additional Information

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