Apache Spark for Machine Learning & Data Science

Overview:

This hands-on, 3-day Apache Spark training targets experienced Data Scientists wishing to perform data analysis at scale using Apache Spark. This course covers an overview of Apache Spark, hands-on projects utilizing extract-transform-load operations (ETL), employing exploratory data analysis (EDA), building machine learning models, evaluating models and performing cross validation.

The course is written using Scala 2.11, Python 2.x and Spark 2.0. All hands-on labs are run on Databricks Community Edition, a free cloud based Spark environment. This allows the participants to maximize their time using open source Apache Spark cluster installations. Labs can easily be ported to run on open source Apache Spark after class.

Learning Objectives:

General Apache Spark:

- ✓ Improve performance through judicious use of caching and applying best practices.
- ✓ Troubleshoot slow running DataFrame queries using explain-plan and the Spark UI.
- ✓ Visualize how jobs are broken into stages and tasks and executed within Spark.
- ✓ Troubleshoot errors and program crashes using executor logs, driver stack traces, and local-mode runtimes.
- ✓ Troubleshoot Spark jobs using the administration UIs and logs inside Databricks.
- ✓ Find answers to common Spark and Databricks questions using the documentation and other resources.

Extracting, Processing and Analyzing Data:

- ✓ Extract, transform and load (ETL) data from multiple federated data sources (JSON, relational database, etc.) with DataFrames.
- ✓ Extract structured data from unstructured data sources by parsing using Datasets (where possible) or RDDs (if not possible with Datasets), with transformations and actions (map, flatMap, filter, reduce, reduceByKey).
- ✓ Extend the capabilities of DataFrames using user defined functions (UDFs and UDAFs) in Python and Scala.
- ✓ Resolve missing fields in DataFrame rows using filtering and imputation.
- ✓ Apply best practices for data analytics using Spark.
- ✓ Perform exploratory data analysis (EDA) using DataFrames and Datasets to:
  - Compute descriptive statistics
  - Identify data quality issues
  - Better understand a dataset

Visualizing Data:

- ✓ Integrate visualizations into a Spark application using Databricks and popular visualization libraries (d3, ggplot, matplotlib).
- ✓ Develop dashboards to provide “at-a-glance” summaries and reports.

Apache Spark for Machine Learning & Data Science

Machine Learning:

- ✓ Learn to apply various regression and classification models, both supervised and unsupervised.
- ✓ Train analytical models with Spark MLlib's DataFrame-based estimators including: linear regression, decision trees, logistic regression, and k-means.
- ✓ Use Spark MLlib transformers to perform pre-processing on a dataset prior to training, including: standardization, normalization one-hot encoding, and binarization.
- ✓ Create Spark MLlib pipelines to create a processing pipeline including transformations, estimations, evaluation of analytical models.
- ✓ Evaluate model accuracy by dividing data into training and test datasets and computing metrics using Spark MLlib evaluators.
- ✓ Tune training hyper-parameters by integrating cross-validation into Spark MLlib Pipelines.
- ✓ Compute using RDD-based Spark MLlib functionality not present in the MLlib DataFrame API, by converting DataFrames to RDDs and applying RDD transformations and actions. (Optional Module).
- ✓ Troubleshoot and tune machine learning algorithms in Spark.
- ✓ Understand and build a general machine learning pipeline for Spark.

Target Audience

Data Scientists and Software engineers with some machine learning background.

Prerequisites

- ✓ Basic Scala or Python - Required
- ✓ Some machine learning background - Required
- ✓ SQL is helpful, but not required

Lab Requirements

- ✓ Chrome or Firefox browser. Internet Explorer, Edge and Safari are not supported.
- ✓ Internet (web access).

Course Outline

Day 1: Spark Overview:

- ✓ Spark intro and ecosystem
- ✓ Lab: Getting connected and learning the environment
- ✓ RDDs, DAGs, Executors, and Spark Architecture
- ✓ Lab: Extract-Transform-Load Operations (Map transformation)
- ✓ DataFrames and Spark SQL
- ✓ Lab: Exploring data w/ Spark SQL + simple visualizations
- ✓ Lab: DataFrames
- ✓ Spark machine learning (DataFrame pipelines & the legacy RDD API)
- ✓ Lab: Linear regression with Spark MLlib pipelines

Day 2:

- ✓ Review of previous day
- ✓ Your first machine learning example (regression / logistic regression & lab)
  - Featurizing a DataFrame using transformers
  - Training a linear regression model using estimators
  - Evaluating the model using evaluators
  - Putting it all together using MLlib pipelines
- ✓ Selecting and extracting features & lab
  - Wiki EDA (Exploratory Data Analysis)
  - Wiki data prep
  - Tokenization
  - TF/IDF
  - Vectorization

✓ Latent semantic analysis & lab

- Decision trees & lab
- Visualize test / train data
- Metadata for decision trees
- Build a decision tree
- Evaluate
- Cross-validation
- Random forest

Day 3:

✓ Review of previous day
✓ K-means clustering & lab

- Load data
- Creating a DataFrame from RDD/Dataset data
- Transform the data
- Save intermediate work as a parquet file
- Document / topic clustering lab
- Visualization

✓ Graphs/GraphFrames - Co-occurrence networks & lab

- Edges & vertices
- Graph operators
    - Subgraph
    - Reverse
    - oin
    - Counting degrees (in-& out-)
- Graph algorithms
    - Page rank
    - Connected components
- Visualization

✓ Optional/future: collaborative filtering (recommender systems & lab
✓ Evaluation metrics
✓ Model selection / cross validation
✓ Model optimization
✓ Model parallel (vs. data parallel) lab (optional)

- Model parallel processing using RDDs and map to call scikit-learn
- Grid search
- Cross validation

Reading homework is required the first and second evenings of the course.

Contact: info@unicom.co.uk | + 44 (0) 1895 256 484 | www.unicom.co.uk