PySpark and Machine Learning (3 Days)

Goal - Make Practical Machine Learning Scalable & Easy

Objectives

- Learn about Apache Spark and the Spark 2.0 architecture
- Build and interact with Spark DataFrames using Spark SQL
- Learn how to solve graph and deep learning problems using GraphFrames and TensorFrames respectively
- Read, transform, and understand data and use it to train machine learning models
- Build machine learning models with MLlib and ML
- Learn how to submit your applications programmatically using spark-submit
- ML Algorithms: common learning algorithms such as classification, regression, clustering, and collaborative filtering
- Featurization: feature extraction, transformation, dimensionality reduction, and selection
- Pipelines: tools for constructing, evaluating, and tuning ML Pipelines
- Persistence: saving and load algorithms, models, and Pipelines
- Utilities: linear algebra, statistics, data handling, etc.

Who is the target audience?

- Anyone interested in Machine Learning
- Any intermediate level people who know the basics of machine learning, including the classical algorithms like linear regression or logistic regression, but who want to learn more about it and explore all the different fields of Machine Learning.
- Any people who are not that comfortable with coding but who are interested in Machine Learning and want to apply it easily on datasets.
- Any data analysts who want to level up in Machine Learning.
- Any people who are not satisfied with their job and who want to become a Data Scientist.
- Any people who want to create added value to their business by using powerful Machine Learning tools

Prerequisites:

Knowledge prerequisites

- Basic Python data structures
- Basic knowledge of Pandas dataframes and SQL
- Knowledge of common data storage formats like JSON, delimiter separated files, HDFS, etc
- Entry-level machine learning

Software Prerequisites

- Apache Spark (Downloadable from http://spark.apache.org/downloads.html)
- A Python distribution containing IPython, Pandas and Scikit-learn
- PySpark

Course Agenda

Days	Modules	Course Outline
1	**Module 1**	**Introduction to Spark** What is Apache Spark? Spark Jobs and APIs Spark 2.0 architecture Installation and Configuration
	Module 2	**Resilient Distributed Datasets** Internal workings of an RDD Creating RDDs Global versus local scope Transformations Actions Hands on Session on RDD and Spark Assignments 1 Best Practices 1
Day 2	**Module 3**	**DataFrames** Python to RDD communications Catalyst Optimizer refresh Speeding up PySpark with DataFrames Creating DataFrames Simple DataFrame queries Interoperating with RDDs Querying with the DataFrame API Hands On Session on Pandas DataFrame and PySpark Assignments 2
	Module 4	**Prepare Data for Modeling** Checking for duplicates, missing observations, and outliers Getting familiar with your data Visualization Hands on Session Data Modeling Assignments 3

Days	Modules	Course Outline
Day 3	**Module 5**	**Introducing MLlib** Overview of the package Loading and transforming the data Getting to know your data Creating the final dataset Predicting infant survival Hands on Session using PySpark MLib Assignments 4
		Introducing the ML Package Overview of the package Predicting the chances of infant survival with ML Parameter hyper-tuning Other features of PySpark ML in action Implementation of ML Algorithm • Random Forest • Regression • K-means Assignments 5
		GraphFrames Introducing GraphFrames Installing GraphFrames Preparing your flights dataset Building the graph Executing simple queries Understanding vertex degrees Determining the top transfer airports Understanding motifs Determining airport ranking using PageRank Determining the most popular non-stop flights Using Breadth-First Search Visualizing flights using D3 Assignment 6 **Conclusion and Summary**