Course Outline

Spark V2 for Data Analysts

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 Python language and using Jupyter environment

Objective

At the completion of the course, students will know:

•Spark ecosystem
•Spark Shell
•Spark Data structures (RDD)
(Dataframe/Dataset)
•Spark SQL
•Modern data formats and Spark
•Spark & Hadoop & Hive

Topics

- Spark Introduction
- First Look at Spark
- Spark Data structures
- Caching
- Dataframes / Datasets
- Spark SQL
- Spark and Hadoop
- Spark and Hadoop

Audience

This course is designed for Data Analysts and Business Analysts.

Prerequisites

Analyst background (familiarity with SQL, Scripting .etc)

Duration

Three Days

‘

Course Outline

Spark V2 for Data Analysts

Course Outline

I. ***Spark Introduction***
- A. Big Data, Hadoop, Spark
- B. Spark concepts and architecture
- C. Spark components overview
- D. Labs : Installing and running Spark

II. ***First Look at Spark***
- A. Spark shell
- B. Spark web UIs
- C. Analyzing dataset – part 1
- D. Labs : Spark shell exploration

III. ***Spark Data structures***
- A. Partitions
- B. Distributed execution
- C. Operations : transformations and actions
- D. 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
- E. 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
- E. 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
- E. 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. **Workshops**
- A. These are group workshops
- B. Attendees will work on solving real world data analysis problems using Spark.