Overview

Apache Spark is the next-generation successor to MapReduce. Spark is a powerful, open source processing engine for data in the Hadoop cluster, optimized for speed, ease of use, and sophisticated analytics. The Spark framework supports streaming data processing and complex, iterative algorithms, enabling applications to run up to 100x faster than traditional Hadoop MapReduce programs.

This course will provide you an excellent kick start in building your fundamentals in developing big data solutions using Apache Spark platform. The course is well balanced between theory and hands-on labs spread on real world uses cases.

What participants will learn?

The attendees will learn below topics through lectures and hands-on exercises

- Deep Dive into Apache Spark 1.6 Architecture
- Understand Spark APIs, RDDs, Data frames, Spark SQL
- How to do parallel programming and develop Spark applications
- How Spark run on standalone and Cluster including Hadoop?
- Understand Advanced Features and spark internals
- Develop Spark Streaming Applications
- Write advanced algorithms using Spark Machine Learning(ML) Library
- Optimizing and tuning spark applications
- End to End Use Case Implementation

Duration

2 Days

Intended Audience

Architects, developers & data scientists who wish to write, build and maintain Apache Spark jobs.

Prerequisites

All the programming will be done using **Python**, hence the participants should have basic programming knowledge of **Python**. It is advised to refresh these skills to obtain maximum benefit from this workshop.

Detailed Course Outline

- **Big Data & Spark Overview**
    - Overview of Big data and its challenges
    - Spark Architecture Overview
    - Installing and Configuring Spark
- **Quick Overview of Python Language**
- **Spark Architecture – Deep Dive**
    - Using Spark Shell
    - Understanding Resilient Distributed Datasets (RDDs), Types of RDDs
    - Working with RDD Actions & Transformations
    - Complete Flow of a spark program
    - Deploying to Spark Standalone & Hadoop Cluster
    - Using Web UI for monitoring & managing Spark Applications
    - **Hands On**
- **Spark APIs & Usages**
    - Working with Key-value pairs using Spark APIs
    - Overview of RDD lineage, Caching and Persistence
    - Share Variables: Accumulators and Broadcast Variables
    - Integrating with different data sources including HDFS
    - Logging & Unit Testing
    - Track Spark jobs stages for Investigation and Troubleshooting
    - **Hands On**
- **Working with Advanced Spark Features**
    - Working with DataFrames & Spark SQLs
    - Hive & RDD Integrations
    - Working with different data formats: Structured and Unstructured
    - **Hands On**
- **Writing Spark Streaming Applications**
    - Spark Streaming Overview
    - Understanding Streaming Operations
    - Sliding Window Operations
    - Developing Spark Streaming Applications

- **Hands On**

- **Using Spark Machine Learning Algorithms**
  - Understanding ML APIs
  - Applying ML – Classification Algorithm
  - **Hands On**

Lab Requirements

- Desktops or laptops with 64-bit Hardware – CPU should be VT enabled
- 64-bit operating System – Windows 7 & above or Mac
- Minimum 8 GB RAM & 10 GB of Hard disk space
- VMWare Player 6.0 + or VMWare Fusion for Mac Machines
- Download and Install the following software on your desktop or laptop
  -

Instructor Profile

Manaranjan Pradhan has about 15+ years of industry experience working on enterprise java, SOA and Cloud computing platforms. He has worked with TCS, HP, and iGATE and worked on large scale projects for customers like Motorola, Home Depot, CKWB Bank, P&G in the roles of solution and technical architect. He is a freelance who provides consulting and training on Cloud Computing, Big data & Hadoop. He has been teaching Hadoop for 2 years and has trained more than 500 people in Hadoop from large MNCs like EMC, CISCO, HP, YODLEE, YAHOO, SAMSUNG, VeriSign, Success Factors etc.

Manaranjan Pradhan is a Cloudera Certified Developer for Apache Hadoop CDH4.

He writes his blog at http://www.awesomestats.in

Connect with him on Linked in http://in.linkedin.com/pub/manaranjan-pradhan/a/6bb/314