Iverson Associates Sdn Bhd (303330-M)
Suite T113 – T114, 3rd Floor, Centrepoint, Lebuh Bandar Utama
Bandar Utama, 47800 Petaling Jaya, Selangor Darul Ehsan
Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Module Title : Cloudera Data Engineering: Developing Applications with Apache Spark
Duration : 4 days

Overview

This four-day hands-on training course delivers the key concepts and knowledge developers need to use Apache Spark to develop high-performance, parallel applications on the Cloudera Data Platform (CDP).

Hands-on exercises allow students to practice writing Spark applications that integrate with CDP core components, such as Hive and Kafka. Participants will learn how to use Spark SQL to query structured data, how to use Spark Streaming to perform real-time processing on streaming data, and how to work with "big data" stored in a distributed file system.

After taking this course, participants will be prepared to face real-world challenges and build applications to execute faster decisions, better decisions, and interactive analysis, applied to a wide variety of use cases, architectures, and industries.

What Skills You Will Gain

During this course, you will learn to:

- Distribute, store, and process data in a CDP cluster
- Write, configure, and deploy Apache Spark applications
- Use the Spark interpreters and Spark applications to explore, process, and analyze distributed data
- Query data using Spark SQL, DataFrames, and Hive tables
- Use Spark Streaming together with Kafka to process a data stream

Other Training That Might Interest You

Apache Spark Application Performance Tuning

Audience

This course is designed for developers and data engineers.

Prerequisite

All students are expected to have basic Linux experience, and basic proficiency with either Python or Scala programming languages. Basic knowledge of SQL is helpful. Prior knowledge of Spark and Hadoop is not required.

Course Outline

Introduction to Zeppelin

- Why Notebooks?
- Zeppelin Notes
- Demo: Apache Spark In 5 Minutes

HDFS Introduction

- HDFS Overview
- HDFS Components and Interactions
- Additional HDFS Interactions
- Ozone Overview
- Exercise: Working with HDFS

YARN Introduction

- YARN Overview
- YARN Components and Interaction
- Working with YARN
- Exercise: Working with YARN

Distributed Processing History

- The Disk Years: 2000 ->2010
- The Memory Years: 2010 ->2020
- The GPU Years: 2020 ->

Working with RDDs

- Resilient Distributed Datasets (RDDs)
- Exercise: Working with RDDs

Working with DataFrames

- Introduction to DataFrames
- Exercise: Introducing DataFrames
- Exercise: Reading and Writing DataFrames
- Exercise: Working with Columns
- Exercise: Working with Complex Types
- Exercise: Combining and Splitting DataFrames
- Exercise: Summarizing and Grouping DataFrames
- Exercise: Working with UDFs
- Exercise: Working with Windows

Introduction to Apache Hive

- About Hive

Hive and Spark Integration

- Hive and Spark Integration
- Exercise: Spark Integration with Hive

Data Visualization with Zeppelin

- Introduction to Data Visualization with Zeppelin
- Zeppelin Analytics
- Zeppelin Collaboration
- Exercise: AdventureWorks

Distributed Processing Challenges

- Shuffle
- Skew
- Order

Spark Distributed Processing

- Spark Distributed Processing
- Exercise: Explore Query Execution Order

Spark Distributed Persistence

- DataFrame and Dataset Persistence
- Persistence Storage Levels
- Viewing Persisted RDDs
- Exercise: Persisting DataFrames

Writing, Configuring, and Running Spark Applications

- Writing a Spark Application
- Building and Running an Application
- Application Deployment Mode
- The Spark Application Web UI
- Configuring Application Properties
- Exercise: Writing, Configuring, and Running a Spark Application