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 Developer Training for Spark and Hadoop
Duration : 4 days

What you will learn

Take your knowledge to the next level

This four-day hands-on training course delivers the key concepts and expertise participants need to ingest and process data on a Hadoop cluster using the most up-to-date tools and techniques. Employing Hadoop ecosystem projects such as Spark (including Spark Streaming and Spark SQL), Flume, Kafka, and Sqoop, this training course is the best preparation for the real-world challenges faced by Hadoop developers. With Spark, developers can write sophisticated parallel applications to execute faster decisions, better decisions, and interactive actions, applied to a wide variety of use cases, architectures, and industries.

Get hands-on experience

Through expert-led discussion and interactive, hands-on exercises, participants will learn how to:

- Distribute, store, and process data in a Hadoop cluster
- Write, configure, and deploy Apache Spark applications on a Hadoop cluster
- Use the Spark shell for interactive data analysis
- Process and query structured data using Spark SQL
- Use Spark Streaming to process a live data stream
- Use Flume and Kafka to ingest data for Spark Streaming

What to expect

This course is designed for developers and engineers who have programming experience, but prior knowledge of Hadoop is not required

- Apache Spark examples and hands-on exercises are presented in Scala and Python. The ability to program in one of those languages is required
- Basic familiarity with the Linux command line is assumed
- Basic knowledge of SQL is helpful

Get certified

Upon completion of the course, attendees are encouraged to continue their study and register for the CCA Spark and Hadoop Developer exam. Certification is a great differentiator. It helps establish you as a leader in the field, providing employers and customers with tangible evidence of your skills and expertise.

Course details

Introduction

Introduction to Apache Hadoop and the Hadoop Ecosystem

- Apache Hadoop Overview
- Data Storage and Ingest
- Data Processing
- Data Analysis and Exploration
- Other Ecosystem Tools
- Introduction to the Hands-On Exercises

Apache Hadoop File Storage

- Problems with Traditional

Large-Scale Systems

- HDFS Architecture
- Using HDFS
- Apache Hadoop File Formats

Data Processing on an Apache Hadoop Cluster

- YARN Architecture
- Working With YARN

Importing Relational Data with Apache Sqoop

- Apache Sqoop Overview
- Importing Data
- Importing File Options
- Exporting Data

Apache Spark Basics

- What is Apache Spark?
- Using the Spark Shell
- RDDs (Resilient Distributed Datasets)
- Functional Programming in Spark

Working with RDDs

- Creating RDDs
- Other General RDD Operations

Aggregating Data with Pair RDDs

- Key-Value Pair RDDs
- Map-Reduce

- Other Pair RDD Operations

Writing and Running Apache Spark Applications

- Spark Applications vs. Spark Shell
- Creating the SparkContext
- Building a Spark Application

(Scala and Java)

- Running a Spark Application
- The Spark Application Web UI

Configuring Apache Spark Applications

- Configuring Spark Properties
- Logging

Parallel Processing in Apache Spark

- Review: Apache Spark on a Cluster
- RDD Partitions
- Partitioning of File-Based RDDs
- HDFS and Data Locality
- Executing Parallel Operations
- Stages and Tasks

RDD Persistence

- RDD Lineage
- RDD Persistence Overview
- Distributed Persistence

Common Patterns in Apache Spark Data Processing

- Common Apache Spark Use Cases
- Iterative Algorithms in Apache Spark
- Machine Learning
- Example: k-means

DataFrames and Spark SQL

- Apache Spark SQL and the SQL Context
- Creating DataFrames
- Transforming and Querying DataFrames
- Saving DataFrames
- DataFrames and RDDs
- Comparing Apache Spark SQL, Impala, and Hive-on-Spark

- Apache Spark SQL in Spark 2.x

Message Processing with Apache Kafka

- What is Apache Kafka?
- Apache Kafka Overview
- Scaling Apache Kafka
- Apache Kafka Cluster Architecture
- Apache Kafka Command Line Tools

Capturing Data with Apache Flume

- What is Apache Flume?
- Basic Flume Architecture
- Flume Sources
- Flume Sinks
- Flume Channels
- Flume Configuration

Integrating Apache Flume and Apache Kafka

- Overview
- Use Cases
- Configuration

Apache Spark Streaming:
Introduction to DStreams

- Apache Spark Streaming Overview
- Example: Streaming Request Count
- DStreams
- Developing Streaming Applications

Apache Spark Streaming:
Processing Multiple Batches

- Multi-Batch Operations
- Time Slicing
- State Operations
- Sliding Window Operations

Apache Spark Streaming: Data Sources

- Streaming Data Source Overview
- Apache Flume and Apache Kafka

Data Sources

- Example: Using a Kafka Direct Data Source

Conclusion