



Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Course Outline :: Apache Spark Development::

Module Title : Apache Spark Development

Duration : 4 days

Overview

Spark is a fast and general cluster computing system for Big Data. It provides high-level APIs in Scala, Java, Python, and R, and an optimized engine that supports general computation graphs for data analysis. It also supports a rich set of higher-level tools including Spark SQL for SQL and DataFrames, Spark ML for machine learning, GraphX for graph data processing, and Spark Streaming for live data stream processing. With Spark running on Apache Hadoop YARN, developers can create applications to derive actionable insights within a single, shared dataset in Hadoop.

This training course will teach you how to solve Big Data problems using Apache Spark framework. The training will cover a wide range of Big Data use cases such as ETL, DWH, data virtualization, streaming, graph data structure, machine learning. It will also demonstrate how Spark integrates with other well established Hadoop ecosystem products. You will learn the course curriculum through theory lectures, live demonstrations and lab exercises. This course will be taught in Python programming language.

Mode of Delivery: Classroom based instructor led program

Prerequisites

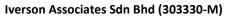
Following are the pre-requisites for the course.

- Programming knowledge in Python is required
- Basic Knowledge of big data use-cases.
- Basic knowledge of databases, OLAP/OTLP use cases, SQL
- Knowledge of Java stack JVM is helpful

Course Outline

Day 1 - Spark Core, Spark Internals, Performance Tuning

Apache Spark Core

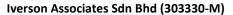




Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Course Outline :: Apache Spark Development::

Module	Topics
Introduction to Spark	 What is Apache Spark- the story of the evolution from Hadoop Advantages of Spark over Hadoop Map Reduce Lambda architecture for enterprise data and analytics services Deployment modes – YARN, Standalone, Mesos Developing on Spark using REPL, Zeppelin, IDE Data sources for Spark application
Lab Exercise	 Install and get started with VM Launching spark REPL and Zeppelin
Resilient Distributed Dataset (RDD)	 RDD – operations – read from the file, transforming and saving persistent Leveraging in memory processing Pair RDD – operations Working with semi structured data formats using regex, json, xml libraries
Lab Exercise	 Explore data from data.sfgov.org using RDD Explore Apache Server logs using regex Join and aggregate data from grouplens.org





Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Course Outline :: Apache Spark Development::

Spark Internals and Performance Tuning

Module	Topics
Spark Internals	 Anatomy of Spark jobs on YARN, Standalone and Mesos RDD partitions
	 Spark literature: Narrow, wide operations, shuffle, DAG, Shuffle, Stages, and Tasks
	Job metrics
	Fault Tolerance
Performance Tuning	Factors that affect performance of spark application
	 Configuring memory and CPU for Spark drivers and executors in standalone and YARN mode
	Controlling logging of Spark daemons and spark applications
	Capturing job metrics using Spark History Server
	Benefits of shared variables – accumulator, broadcast var
	Types of spark caching and their use cases
	Role of checking pointing of Spark RDD
Lab Exercises	Examine spark metrics and logs
	 Evaluate impact of different caching types on memory and processing time
	Use shared variables
Setup Spark Cluster	Set up Spark on YARN
	Set up Spark Standalone





Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Course Outline :: Apache Spark Development::

Day 2: Spark Dataframe, SQL and Building Application

Building Spark Application [Optional]

Module	Topics
Building Application	 Building Spark application using Eclipse IDE Building Spark Application using SBT Building Spark application using Jupyter Notebook (Optional)
Lab Exercise	 Create a project using Eclipse and submit to cluster Create a project using EBT and submit to cluster

Spark Dataframe and SQL





Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Course Outline :: Apache Spark Development::

Module	Topics
Dataframe Basics	Introduction to Dataframe
	Difference between RDD and Dataframe
	 Dataframe internals that makes it fast – Catalyst Optimizer and Tungsten
	Loading and processing data into dataframe
	Saving dataframe to file systems
Lab Exercise	Process data.sfgov.org data using dataframe
Dataframe Advanced	Hive Context vs Spark SQL Context
	Working with Hive Tables
	Working with JDBC data source
	 Data formats – text format such csv, json, xml, binary formats such as parquet, orc
	UDF in Spark Dataframe
	Spark SQL as JDBC service and its benefits and limitations
	 Analytical queries in Spark – windows functions, pivot, rollup and cubes
	Working with Cassandra
	° An introduction to Cassandra
	 Working with HBase using Spark



Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Course Outline :: Apache Spark Development::

Module	Topics
Hands On	 Persist spark temporary tables using Hive Processing Mysql data using Dataframe Working with UDF
	 Working with file formats Text formats – CSV, json, xml Binary formats – ORC, Parquet, Avro Integrating Spark with BI tools

Day 3: Spark Streaming, Kafka Receiver

Spark Streaming

Module	Topics
Spark Streaming	Architecture of streaming application
	Streaming Context – initialization, configuration, characteristics
	Dstream – operations
	Receiver characteristics
	Window operation – batch internal, window length, sliding interval
	Fault Tolerance using checkpointing and replication
	Partition behavior of Dstream
	Kafka Streaming
Lab Exercise	Trending analysis on live twitter stream using Spark stream
	Saving live streams into HDFS and RDBMS
	Saving live streams into HBase using Spark SQL (optional)



Tel: 03-7726 2678 Fax: 03-7727 9737 Website: www.iverson.com.my

Course Outline :: Apache Spark Development::

Kafka Receiver

Module	Topics
Kafka for Streaming	 Overview Kafka Architecture Kafka terminologies – brokers, messages, consumer groups, consumer, producer Configure Kafka cluster using multi nodes
	 Fault tolerance, consistency, schema validation Kafka connectors for JDBC, HDFS
Lab Exercise	Setting up Kafka cluster – multi node brokers
Streaming Advanced Receiver	 Introduction to KAFKA receiver for Spark Lambda architecture
Lab Exercise	Create a spark streaming application using KAFKA

Day 4: Machine Learning

Module	Topics
Introduction to Machine Learning	 Descriptive and Inferential Statistics Overview machine learning use cases Identify machine learning that fits your need Pipeline of machine learning operation Introduction to Spark ML, Spark MLlib Machine Learning Algorithms Introduction to Python SiKit Learn library
Lab Exercises	 Predict power demand using Spark ML Predict trend of stock price