Tel +44 20 4571 2395
info@siliconbeachtraining.co.uk

The Apache Spark and Scala Certification course provide in-depth theoretical knowledge as well as practical skills to enhance your competence in Big Data Spark. The course provides an overview of Spark and its ecosystem, Spark Streaming, Spark SQL, RDD and Scala. The course enables the delegates to become a successful Big Data & Spark Developer. During the training, the delegates will also get an opportunity to work on various industry-based use-cases and projects. These project will include big data and spark tools as a part of solution strategy.

The training course will be conducted by industry experts and help the delegates to become a Spark developer. The industry experts have multiple years of experience in this field. The course is also designed by industry experts as per market standards. Spark is one of the most extensively used tools for Big Data & Analytics. It has been used by many large companies across the globe. The demand for Big Data and Spark Developers is increasing rapidly as many organisations are showing their interest in Big Data and are adopting Spark as a part of solution strategy.

Prerequisites

No formal prerequisites are required to attend the training program. However, basic knowledge of Core Java, database, query language and SQL would be beneficial.

Course Objectives

- Get insights into Apache Spark and Scala programming
- Learn Scala and its programming implementation
- Write Spark Applications using Python, Java and Scala
- Understand the difference between Apache Spark and Hadoop
- Implement Spark on a cluster
- Define and explain Spark Streaming
- Learn Scala Java Interoperability and other Scala operations
- Understand RDD and its operation
- Learn the implementation of Spark Algorithms
- Work on Projects using Scala to run on Spark applications
- Learn about the Scala classes concept and execute pattern matching

Scala Course Content

Introduction of Scala

Tel +44 20 4571 2395
info@siliconbeachtraining.co.uk

Pattern Matching

- The importance of Scala
- The concept of REPL (Read Evaluate Print Loop)
- Deep dive into Scala pattern matching
- Type interface
- Higher order function
- Currying
- Traits
- Application space
- Scala for data analysis

Executing the Scala code

- Learning about the Scala Interpreter
- Static object timer in Scala
- Testing String equality in Scala
- Implicit classes in Scala
- The concept of currying in Scala
- Various classes in Scala

Classes concept in Scala

- Learning about the Classes concept
- Understanding the constructor overloading
- The various abstract classes
- The hierarchy types in Scala
- The concept of object equality
- The val and var methods in Scala

Case classes and pattern matching

- Understanding Sealed traits
- Wild
- Constructor
- Tuple
- Variable pattern
- Constant pattern

Concepts of traits with example

- Understanding traits in Scala
- The advantages of traits
- Linearization of traits
- The Java equivalent
- Avoiding of boilerplate code

Scala java Interoperability

- Implementation of traits in Scala and Java
- Handling of multiple traits extending

Scala collections

Tel +44 20 4571 2395
info@siliconbeachtraining.co.uk

- Example of list sequence in Scala

Mutable collections vs. Immutable collections

- The two types of collections in Scala
- Mutable and Immutable collections
- Understanding lists and arrays in Scala
- The list buffer and array buffer
- Queue in Scala
- Double-ended queue Deque
- Stacks
- Sets
- Maps
- Tuples in Scala

Use Case bobsrockets package

- Introduction to Scala packages and imports
- The selective imports
- The Scala test classes
- Introduction to JUnit test class
- JUnit interface via JUnit 3 suite for Scala test
- Packaging of Scala applications in Directory Structure
- Example of Spark Split and Spark Scala

Spark Course Content

Introduction to Spark

- Introduction to Spark
- How Spark overcomes the drawbacks of working MapReduce
- Understanding in-memory MapReduce
- Interactive operations on MapReduce
- Spark stack
- Fine vs. coarse grained update
- Spark stack
- Spark Hadoop YARN
- HDFS Revision
- YARN Revision
- The overview of Spark
- How it is better Hadoop
- Deploying Spark without Hadoop
- Spark history server
- Cloudera distribution

Spark Basics

- Spark installation guide
- Spark configuration
- Memory management
- Executor memory vs. driver memory
- Working with Spark Shell
- The concept of Resilient Distributed Datasets (RDD)
- Learning to do functional programming in Spark

Tel +44 20 4571 2395
info@siliconbeachtraining.co.uk

- Spark RDD
- Creating RDDs
- RDD partitioning
- Operations and transformation in RDD
- Deep dive into Spark RDDs
- The RDD general operations
- A read-only partitioned collection of records
- Using the concept of RDD for faster and efficient data processing
- RDD action for Collect
- Count
- Collectsmap
- Saveastextfiles
- Pair RDD functions

Aggregating Data with Pair RDDs

- Understanding the concept of Key-Value pair in RDDs
- Learning how Spark makes MapReduce operations faster
- Various operations of RDD
- MapReduce interactive operations
- Fine & coarse grained update
- Spark stack

Writing and Deploying Spark Applications

- Comparing the Spark applications with Spark Shell
- Creating a Spark application using Scala or Java
- Deploying a Spark application
- Scala built application
- Creation of mutable list
- Set & set operations
- List
- Tuple
- Concatenating list
- Creating an application using SBT
- Deploying application using Maven
- The web user interface of Spark application
- A real world example of Spark
- Configuring of Spark

Parallel Processing

- Learning about Spark parallel processing
- Deploying on a cluster
- Introduction to Spark partitions
- File-based partitioning of RDDs
- Understanding of HDFS
- Data locality
- Mastering the technique of parallel operations
- Comparing repartition & coalesce
- RDD actions

Spark RDD Persistence

Tel +44 20 4571 2395
info@siliconbeachtraining.co.uk

- Distribution shared memory vs. RDD
- RDD limitations
- Spark shell arguments
- Distributed persistence
- RDD lineage
- Key/Value pair for sorting implicit conversion like CountByKey
- ReduceByKey
- SortByKey
- AggregataeByKey

Spark Streaming & Mlib

- Spark Streaming Architecture
- Writing streaming programcoding
- Processing of spark stream,
- Processing Spark Discretized Stream (DStream)
- The context of Spark Streaming
- Streaming transformation
- Flume Spark streaming
- Request count and Dstream
- Multi batch operation
- Sliding window operations
- Advanced data sources
- Different Algorithms
- The concept of iterative algorithm in Spark
- Analyzing with Spark graph processing
- Introduction to K-Means
- Machine learning
- Various variables in Spark like shared variables
- Broadcast variables
- Learning about accumulators

Spark SQL and Data Frames

- Describe Spark SQL
- The context of SQL in Spark
- Working with XML data
- Parquet files
- JSON support in Spark SQL
- Creating HiveContext
- Writing Data Frame to Hive
- Reading JDBC files
- Understanding the Data Frames in Spark
- Creating Data Frames
- Manual inferring of schema
- Working with CSV files
- Reading JDBC tables
- Data Frame to JDBC
- User defined functions in Spark SQL
- Shared variable and accumulators
- Learning to query and transform data in Data Frames

Improving Spark Performance

Tel +44 20 4571 2395
info@siliconbeachtraining.co.uk

- Troubleshooting the performance problems

Scheduling/ Partitioning

- Learning about the scheduling and partitioning in Spark
- Hash partition
- Range partition
- Scheduling within and around applications
- Static partitioning
- Dynamic sharing
- Fair scheduling
- Map partition with index
- The Zip
- GroupByKey
- Spark master high availability
- Standby Masters with Zookeeper
- Single Node Recovery with Local File System
- High Order Functions

The Apache Spark and Scala Certification course provide in-depth theoretical knowledge as well as practical skills to enhance your competence in Big Data Spark. The course provides an overview of Spark and its ecosystem, Spark Streaming, Spark SQL, RDD and Scala.