TEKCRUX
We ReDefine IT

Spark

SPARK SCALA DEVELOPMENT COURSE CURRICULUM

Prerequisite: Scala for Spark

Scala Basics

- What is Scala?
- Introduction to Scala REPL
- Installing Scala IDE
- Basic Operations
- Type Inference
- Block expression
- Lazy values
- Defining Functions
- Defining Procedures
- Control Structures in Scala
- Loops – ForEach, For, While, Do-While, For Comprehension
- Collections – Array, ArrayBuffer, Map, Tuples, Lists, ListBuffer, Sets, Sequnce,Vector
- Conditional Operators
- Enumerations

Object Oriented Programming

- Class and Object Basics
- Inheritance in Scala
- Scala Constructors (Auxiliary & Primary)
- Singletons
- Companion Objects
- Nested Classes
- Case Classes
- Packages & Visibility Rules
- Overriding Methods
- Traits
    - Interfaces
    - Layered Traits

Functional Programming

- Functional programming Approach
- Higher Order Functions
  - map
  - filter
  - reduceLeft and reduceRight
  - sortWith
  - folding and scanning etc…
- Anonymous Functions
- Function Currying

Prerequisite: Bigdata and Hadoop Framework Overview

- Introduction to BigData
- Challenges with Bigdata
- Batch Vs. Realtime processing
- Overview- Hadoop Ecosystem
  - HDFS
  - Review of MapReduce
  - Hive
  - HBase
  - Sqoop
  - Flume
  - Kafka

APACHE SPARK

Introduction to Spark

- What is Spark?
- Spark Overview
- Setting up environment
- Build a simple Spark project with Eclipse & Maven
- Using Spark Shell

Spark Basics

- Resilient Distributed Datasets (RDDs)
- Spark Context
- Spark Ecosystem
- In-Memory Computations in Spark

Working with RDDs

- Creating, Loading and Saving RDD
- Transformations on RDD
- Actions on RDD
- Key-Value Pair Transformation on RDDs
- RDD Partitioning
- RDD Persistence

Writing and Deploying on Cluster

- Spark Applications vs. Spark Shell
- Spark Runtime Architecture
  - Executors
  - Driver
  - Cluster Managers
- Creating Spark Context
- Building a Spark Application
- Deploying Spark Applications using Spark-Submit

Spark Job Execution

- RDD Lineage
- Jobs, Stages and Tasks
- Partition and Shuffles
- Data Locality
- Join with or without Partitioner, stages and tasks, etc
- Spark Web UI

Spark SQL

- Overview on Hive
- Spark SQL Architecture
- SparkSession in Spark SQL
- Working with DataFrames
- Integrating Spark SQL with Hive
- Integrating Spark SQL with JDBC Sources (MySQL)
- Integrating Spark SQL with NoSQL DB (Cassandra)
- Handling CSV, JSON and Parquet File Formats
- Loading and Saving Data

Spark Streaming

- Spark Streaming Architecture
- Spark Streaming Transformations
  - Stateless and Stateful Transformations
- Rolling Window and Check pointing
- Integrating Spark with Kafka Streaming Data
- Integrating Spark with Twitter Streaming Data
- Spark Streaming Performance Considerations

Spark MLlib

- What is Machine Learning?
- ML library for Spark
- ML Concepts and Algorithms
  - Classification
  - Regression
  - Clustering
  - Collaborative Filtering
- Typical Steps in ML Pipeline – Executors and Transformers
- ML using Pipelines and DataFrames
- Recommendation Engine – Practical Use Case

Spark GraphX

- Overview of GraphX
- Components of GraphX – VertexRDD, EdgeRDD and Triplets
- Develop simple application with GraphX
- Transformations on GraphX
- Hands on – PageRank, TriangleCount Algorithms
- Common Spark Use-cases

Performance Tuning and Debugging

- Shared Variables: Broadcast Variables
- Shared Variables: Accumulators
- Common Performance Issues
- Performance Tuning Tips
- Spark WebUI
- Monitoring Driver and Executor Logs

Course Deliverables

- Workshop style coaching
- Interactive approach
- Course material
- POC Implementation
- Hands on practice exercises for each topic
- Quiz at the end of each major topic
- Tips and techniques on Cloudera Certification Examination