springpeople

COURSE NAME

Spark Intermediate

COURSE OVERVIEW

Course on Apache Spark & Scala is a 3 days(24 hours) course which will cover different concepts of Big Data, Challenges in Big Data Processing, Approach to Big Data Problems using Apache Spark, specifics of Spark like it's Components, Installation Steps, RDDs, Transformations, Actions, Lazy Execution, Integration with HDFS.

DURATION

2 Days

TRAINEE PRE-REQUISITES

- Knowledge of Hadoop Eco-system
- Knowledge of Scala

LEARNING OBJECTIVES

After the completion of this course, you will be able to:

☐. Understand Big Data and the challenges associated

☐Find an approach to Big Data problems with Apache Spark

☐Implement Apache Spark Concepts

☐Apply Scala for Spark

☐Understand data frame concept and How to run SQL queries using Spark-SQL

☐Follow latest emerging trends like MLib, GraphX based on Spark

LAB REQUIREMENTS DETAILS

- 8 GB RAM windows machine
- Internet connection for setting up SBT/Maven project
- Virtualization feature on the machine should be enabled

COURSE CONTENT

Day 1

Introduction of Spark

Evolution of distributed systems

Why we need new generation of distributed system?

Limitation with Map Reduce in Hadoop,

Understanding need of Batch Vs. Real Time Analytics

Batch Analytics - Hadoop Ecosystem Overview, Real Time Analytics Options

Introduction to stream and in memory analysis

What is Spark? A Brief History: Spark

Honds-On

1. Installing Spark and sbt

2. Integrating Spark in Eciplse

3. Running Spark in Eclipse and Spark Standalone cluster

Using Scala for creating Spark Application

Invoking Spark Shell

Creating the SparkContext

Loading a File in Shell

Performing Some Basic Operations on Files in Spark Shell

Building a Spark Project with sbt

Running Spark Project with sbt, Caching Overview

Distributed Persistence

Spark Streaming Overview

Example: Streaming Word Count

Testing Tips in Scala

Performance Tuning Tips in Spark

Shared Variables: Broadcast Variables

Shared Variables: Accumulators

springpeople

Day 2

Running SQL queries using Spark SQL
Starting Point: SQLContext
Creating DataFrames
DataFrame Operations
Running SQL Queries Programmatically

Interoperating with RDDs
Inferring the Schema Using Reflection
PInferring the Schema Using Reflection
Data Sources
Generic Load/Save Functions
Save Modes
Saving to Persistent Tables

Parquet Files

Loading Data Programmatically
Partition Discovery
Schema Merging
JSON Datasets
Hive Tables

JDBC To Other Databases
Troubleshooting
Performance Tuning
Caching Data In Memory
Compatibility with Apache Hive
Unsupported Hive Functionality

Honds-On

1. Running SQL Quries with MySql

2. Running Hive queries

3. Reading JSON file and storing it as a Parquet format

Tuning Spark
Data Serialization
Memory Tuning
Determining Memory Consumption
Tuning Data Structures
Serialized RDD Storage
Garbage Collection Tuning
Other Considerations
Level of Parallelism

Memory Usage of Reduce Tasks
Broadcasting Large Variables
Data Locality
Summary

Job Scheduling and Monitoring
Overview
Scheduling Across Applications
Dynamic Resource Allocation
Configuration and Setup
Resource Allocation Policy
Request Policy
Remove Policy
Graceful Decommission of Executors
Scheduling Within an Application
Fair Scheduler Pools
Default Behavior of Pools
Configuring Pool Properties