Hadoop/Big Data and Spark Syllabus

This course has three pre-requisite – Linux, Java and SQL.

Linux: All commands that are relevant to Hadoop will be taught. No need to learn separately.

Java: Core Java knowledge is required. Trainee can join separate Java batch.

SQL: Basic SQL queries and joins required.

The syllabus includes Spark and Scala, which will require extra sessions.

Hadoop and Big Data Syllabus

Hadoop Course Objectives/highlights

- What is Big Data
- The core technologies of Hadoop
- How Hadoop Distributed File System (HDFS) and MapReduce work
- What other projects exist in the Hadoop ecosystem
- How to develop MapReduce jobs
- Algorithms for common MapReduce tasks
- How to create large workflows using multiple MapReduce jobs
- Best practices for debugging Hadoop jobs

Hadoop/Big Data and Spark Syllabus

- Advanced features of the Hadoop API
- YARN
- Detailed Hadoop Ecosystem – Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, HCatalog, HBase and YARN
- Introduction to Apache Spark
- Hadoop on Amazon Web Services

Introduction:

- Motivation for Hadoop
- Big Data Characteristics, Challenges with traditional system
- Hadoop's History
- Core Hadoop Concepts
- Hadoop Clusters, Installation and Configuration

Linux and Hadoop Basic Comands

- Linux Commands
- HDFS Commands
- Hands-On for All Commands

Hadoop Basic Concepts

Hadoop/Big Data and Spark Syllabus

- What Hadoop is?
- What features the Hadoop Distributed File System (HDFS) provides
  - Architecture
  - Features, Goals and Advantages of HDFS
  - Name Nodes
  - Data Nodes
  - Secondary Name Node
- The concepts behind MapReduce
  - How Map Reduce Works?
  - Data Type
  - Input & Output Formats
- How a Hadoop cluster operates
  - Cluster sizing
  - Capacity planning
  - Replication
  - Blocks
  - Heartbeat Mechanism
  - Data Organization

VM Installation

- Providing Hadoop VM and configuring it
- Learning Eclipse and creating MapReduce JAR

Hadoop/Big Data and Spark Syllabus

Writing a Map Reduce Program

- The Driver Code
- The Mapper
- The Reducer
- The Streaming API
- Develop a MapReduce program for WhatsApp Message Analytics project

The Hadoop Ecosystem

- Introduction
- **Hive**
  - SQL Basics
  - Hive Basics
  - Internal & External Tables
  - Partitioning
  - Buckets
  - DDL,DML
  - Joins, Index and Views
  - 3 Projects – transferring data from one table to multiple tables, convert unstructured data into structured data and perform analytics and alter/rename/drop commands in Hive
- **Pig**

Hadoop/Big Data and Spark Syllabus

- Motivation
- Pig Basics
- PigLatin Language
- Statement Execution Steps
- Data Types
- Loading data files
- Writing queries – SPLIT, FILTER, JOIN, GROUP, SAMPLE, ILLUSTRATE etc.
- Multi Query Execution
- Debugging in Pig
- Pig UDF
- 3 Projects – WordCount using PigLatin, Batting Data Analytics, Production Example

- **HBase**
  - Overview
  - HBase vs HDFS
  - Data Model
  - Key Value
  - Common Commands in HBase
  - HBase Basics
  - Region Server

- **Flume**
  - Flume Basics

Hadoop/Big Data and Spark Syllabus

- Features
- Architecture
- Agent Architecture
- Example where we ingest files in real-time into HDFS
- Flume Use Cases

- **HCatalog**
  - Objective
  - Supported Projects and Formats
- **Sqoop**
  - Motivation
  - Sqoop Features
  - Architecture
  - Hive Import
  - Sqooop Import
  - Sqoop Export
- **ZooKeeper**
  - Fault Tolerant
  - Zookeeper Service
- **Oozie**

Hadoop/Big Data and Spark Syllabus

- Overview
- Features
- Sample Workflow
- Action Nodes
- Decision Nodes
- Workflow Design
- Workflow Scheduler
- Example of MapReduce task

Hadoop 2.X

- Classic MapReduce Architecture
- Challenges with Hadoop 1
- YARN
- Daemons
- Architecture
- Resource Manager
- Node Manager
- Application Master
- Hadoop 1.X Vs Hadoop 2.x

Introduction to Apache Spark

- Spark Details
- DAG
- Scala
- MLLib

Hadoop/Big Data and Spark Syllabus

- GraphX

Hadoop on Amazon Web Services

- Introduction to AWS cloud infrastructure.
- Amazon SaaS, Paas and IaaS.
- Creating EC2 instance for processing.
- Creating S3 buckets
- Deploying data on to the cloud.
- Choosing size of our instance.
- Configuration of EMR instance
- Creating a virtual cluster on Amazon Web Services

Spark And Scala

Candidates can opt for separate Spark and Scala course.

Module 0 - Scala

- **What is Scala?**
  - Why Scala for Spark?
  - Scala in other frameworks
  - Introduction to Scala REPL
  - Basic Scala operations
  - Variable Types in Scala
  - Control Structures in Scala
  - Foreach loop

Hadoop/Big Data and Spark Syllabus

- Functions
- Procedures
- Collections in Scala- Array, ArrayBuffer, Map, Tuples, Lists, and more.

Module 1 - Spark Core

- **Introrduction**
  - Introduction to big data,
  - Challenges with big data
  - Batch Vs. Real Time big data analytics
  - Batch Analytics - Hadoop Ecosystem Overview
  - Real-time Analytics
- **What is Spark?**
  - Spark Ecosystem
  - Modes of Spark
  - Spark installation demo
  - Overview of Spark on a cluster
  - Spark Standalone cluster
  - Spark Web UI
  - Some configurations.
- **Components of Spark Unified stack**
  - Spark Streaming
  - MLlib
  - Core
  - Spark SQL
- **RDD - The core concept of Spark**

Hadoop/Big Data and Spark Syllabus

- RDDs,
- Transformations in RDD,
- Actions in RDD,
- Loading data in RDD,
- Saving data through RDD,
- Key-Value Pair RDD,
- MapReduce and Pair RDD Operations

➢ **Scala and Python shell**

- Word count example

➢ **Shared Variables with examples**

➢ **Submitting jobs in cluster**

➢ **Hands on examples**

Module 2 - Spark SQL

➢ **Overview**

- Hive and Spark SQL architecture
- sqlContext in spark sql

➢ **Dataframes API**

- Understanding concept of data frame
- Loading data in dataframe
- Operations on dataframes.

➢ **Interaction with Hive**

➢ **Reading various data formats**

➢ **Hands on Examples**

Module 3 - Spark Streaming

Hadoop/Big Data and Spark Syllabus

- **Overview of streaming**
  - Spark Streaming Architecture,
  - First Spark Streaming Program,
  - Transformations in Spark Streaming,
  - checkpointing,
  - Parallelism level
- **Introduction to queuing systems. Eg. Kafka**
- **Hands on examples**

Module 4 - Spark MLlib

- **Supervised Learning**
  - Classification - logistic regression, decision trees, random forests, naive Bayes
  - Regression - linear least squares, Lasso, ridge regression, decision trees
- **Unsupervised learning :**
  - Clustering - K-means, Gaussian Mixture
- **Dimensionality reduction**
  - PCA
- **Hands on examples**

Projects

- Web Log Analytics and report generation on real web logs data
- Twitter sentiment Analytics using actual Tweeter data

Note: Hands-on sessions will be conducted for all the topics mentioned above.