INSTRUCTOR-LED TRAINING COURSE

Hadoop Decoded Applications

HADOOP DECODED APPLICATIONS

Lecture/Lab

ILT 57755	3 Days

COURSE DESCRIPTION

Hadoop is more popular than ever and is generating data-driven business value across every industry. This course gives attendees the essential skills to build Big Data applications using Hadoop technologies such as HDFS, YARN, Apache Kafka, Apache Hive and Apache Spark, in an analytical ecosystem with Teradata components such as Teradata Database, Teradata Viewpoint, and Teradata QueryGrid.

In this course, students will have access to their own cluster to gain hands-on experience. Students will use the Hadoop's distributed file system and process distributed datasets with Hive. In addition, students will develop applications in Spark using Scala and Python via RDDs and DataFrames.

Students will write applications using Hive and Spark and learn about common issues encountered when processing vast datasets in distributed systems.

A discussion of additional tools, Hadoop distributions, and the opportunity to ask questions of experts in Hadoop technology make this popular course an essential grounding for companies looking to implement Hadoop effectively within their enterprise.

AUDIENCE

Hive Developers, Spark Developers, Hadoop Developers, Data Scientists, Business Analysts/Data Analysts, and Data Engineers

COURSE OBJECTIVES

After successfully completing this course, you will be able to:

- Describe the issues of "Big Data" and how they are remedied using Hadoop.
- Describe the Hadoop architecture and its core components (HDFS, YARN).
- Load data into Hadoop from various sources (Flume, Sqoop, Kafka).
- Use Hive to analyze unstructured and structured data at a large scale.
- Explain the importance of the Hive Metastore.
- Write applications with Spark using RDDs and Spark SQL using DataFrames.
- Use Spark SQL to analyze datasets from Hive using Hive Metastore.
- Use Spark Streaming and Structured Streaming for near-real-time analysis.
- Integrate Hadoop with Teradata (Teradata Unified Data Architecture, Teradata Viewpoint, Teradata QueryGrid).

PREREQUISITES

To get the most out of this training, you should have the following knowledge or experience:

- Students are expected to have some prior programming experience and can use basic Linux commands.
- Experience in SQL, Scala and Python will be a distinct advantage.
- Prior Hadoop experience is a bonus.

OURSE OUTLINE*

DAY 1	DAY 2	DAY 3
~ Introduction and Setup ~ Hadoop Architecture ~ Ingesting Data into Hadoop HDFS ~ Apache Hive Architecture	~ Hive Query Language ~ Spark Architecture and Concepts ~ Spark Core	~ Spark SQl ~ Integrating Teradata in Hadoop

* Timing and topics covered by day may vary.

TERADATA.

COURSE TABLE OF CONTENTS

COURSE CONTENT

Module 00 – Introduction and Setup:

- Introduction to course
- Setup
- Connect to the Cloud Lab Environment

Module 01 – Hadoop Architecture:

- What is Hadoop
- Hadoop Nodes
- Three Core Processing Components
  - Yarn (Yet Another Resource Manager)
  - HDFS (Hadoop Distributed File System)
  - MapReduce (MR Processing Engine)
- Lab Exercise
- Review Question
- Summary

Module 02 – Ingesting Data into Hadoop HDFS:

- How to load data into Hadoop using several popular ingest utilities
  - Hadoop Command Line (hdfs dfs-put)
  - Ambari (Files View)
  - Sqoop
  - WebHDFS
  - Flume
  - distcp
  - 3rd Party Utilities
    - Apache Pig
    - Apache Hive
    - Apache Spark
    - Apache Presto
    - Teradata Connector for Hadoop (TDCH)
- Lab Exercises
- Review Questions
- Summary

Module 03 – Apache Hive Architecture and Concepts:

- Introduction
  - What is Hive
  - Hive vs. JavaMR
  - Hive vs. RDBMS
  - Feature
- Architecture
  - Components
  - How it Works
  - The Hive Metastore
- Lab Exercise
- Review Questions
- Summary

Module 04 – Hive Query Language:

- Accessing Hive
- About HQL
  - Hive Objects and Data Types
- Data Definition Language
  - Databases
  - Tables
  - Partition, Skew, and Bucket Tables
  - Encoding and File Formats
- Data Manipulation Language
  - Loading/Inserting Data
  - Data Retrieval: Queries
  - Built in Functions and UDFs
- Lab Exercise
- Review Questions
- Summary

Module 05 – Spark Architecture and Concepts:

- Architecture
  - Berkley Data Analysis Stack
  - Spark vs. MapReduce
  - Spark Building Blocks
  - Component Locations
  - Execution Speed
- Why Spark?
- Deployment Options
  - Spark on Hadoop
- Lab Exercises
- Review Questions
- References and Summary

Module 06 – Spark Core:

- About Spark
  - Spark Shell and Zeppelin
- Scala/Python for Spark
  - Immutable and Mutable
  - Anonymous Functions
- Spark RDDs
  - RDD Creation
  - RDD Operations
  - RDD Persistence
- Lab Exercise
- Review Questions
- Summary

Module 07 – Spark SQL:

- About Spark SQL
- Spark DataFrames
  - DataFrames
  - DataFrame Creation
  - DataFrame API
- Spark SQL
  - Querying Hive Tables
  - Querying DataFrames
  - Spark-sql Shell
- Speed and Ease of Use
- Lab Exercise
- Review Questions
- Summary

Module 08 – Integrating Teradata in Hadoop:

- Teradata Unified Data Architecture
- Teradata Viewpoint
- Teradata QueryGrid
- How to configure TD QueryGrid for Apache Hive and Query
- How to configure TD QueryGrid for Apache Spark and Query
- Review Questions
- Quiz
- Summary