TRAINING OFFERING

APACHE HADOOP ECOSYSTEM FULL STACK ARCHITECTURE

SUBJECT MATTER EXPERT

This 2 day training course is designed for developers who need to create applications to analyze Big Data stored in Apache Hadoop using Apache Pig and Apache Hive. Topics include: Essential understanding of HDP & its capabilities, Hadoop, YARN, HDFS, MapReduce/Tez, data ingestion, using Pig and Hive to perform data analytics on Big Data.

PREREQUISITES

Students should be familiar with programming principles and have experience in software development. SQL and light scripting knowledge is also helpful. No prior Hadoop knowledge is required.

TARGET AUDIENCE

Developers and data engineers who need to understand and develop Hive applications on HDP.

FORMAT

50% Lecture/Discussion
50% Hands-0n Labs

COURSE OBJECTIVES

- Describe the Case for Hadoop
- Describe the Trends of Volume, Velocity and Variety
- Discuss the Importance of Open Enterprise Hadoop
- Describe the Hadoop Ecosystem Frameworks Across the Following Five Architectural Categories:
    - Data Management
    - Data Access
    - Data Governance & Integration
    - Security
    - Operations
- Describe the Function and Purpose of the Hadoop Distributed File System (HDFS)

- List the Major Architectural Components of HDFS and their Interactions
- Describe Data Ingestion
- Describe Batch/Bulk Ingestion Options
- Describe the Streaming Framework Alternatives
- Describe the Purpose and Function of MapReduce
- Describe the Purpose and Components of YARN
- Describe the Major Architectural Components of YARN and their Interactions
- Define the Purpose and Function of Apache Pig
- Work with the Grunt Shell
- Work with Pig Latin Relation Names and Field Names
- Describe the Pig Data Types and Schema
- Demonstrate Common Operators Such as:
    - ORDER BY
    - CASE
    - DISTINCT
    - PARALLEL
    - FOREACH
- Understand how Hive Tables are Defined and Implemented
- Use Hive to Explore and Analyze Data Sets
- Explain and Use the Various Hive File Formats
- Understand benefits from a Hive Table that Uses ORC File Formats
- Use Hive to Run SQL-like Queries to Perform Data Analysis
- Use Hive to Join Datasets Using a Variety of Techniques
- Write Efficient Hive Queries
- Explain the Uses and Purpose of HCatalog
- Use HCatalog with Pig and Hive

HANDS-ON LABS

- Starting an HDP Cluster
- Using HDFS Commands
- Exploring a MapReduce Program
- Getting Started with Apache Pig
- Exploring Data with Pig
- Splitting a Dataset
- Joining Datasets
- Preparing Data for Apache Hive
- Understanding Apache Hive Tables
- Demonstration: Understanding Partitions and Skew
- Analyzing Big Data with Apache Hive

- Analyzing Big Data with Apache Hive
- Joining Datasets in Apache Hive
- Using HCatalog with Apache Pig