Maslak Mahallesi, Maslak Meydan Sk. No:5. Spring Giz Plaza. Maslak/İstanbul
+90 212 282 7700 - info@bilginc.com

HDP Developer: Quick Start

Learn via: **Classroom**

Duration: **4 Day**

Overview

This 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, and developing applications on Apache Spark.

Topics include: Essential understanding of HDP and its capabilities, Hadoop, YARN, HDFS, MapReduce/Tez, data ingestion, using Pig and Hive to perform data analytics on Big Data and an introduction to Spark Core, Spark SQL, Apache Zeppelin, and additional Spark features.

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.

Who Should Attend

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

Outline

Part I: High Level Overview

- Describe the Case for Hadoop
- Identify the Hadoop Ecosystem via architectural categories

Part II: Deeper Look & Demos (2 hrs)

- Detail the HDFS architecture
- Describe data ingestion options and frameworks for batch and real-time streaming
- Explain the fundamentals of parallel processing
- Detail the architecture and features of YARN
- Understand backup and recovery options
- Describe how to secure Hadoop

Live Demonstrations

- Operational overview with Ambari
- Loading data into HDFS

Objectives

- Use Pig to explore and transform data in HDFS
- Transfer data between Hadoop and a relational database
- Understand how Hive tables are defined and implemented
- Use Hive to explore and analyze data sets
- Explain and use the various Hive file formats
- Create and populate 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

- Use HDFS commands to add/remove files and folders

- Use Sqoop to transfer data between HDFS and a RDBMS
- Explore, transform, split and join datasets using Pig
- Use Pig to transform and export a dataset for use with Hive
- Use HCatLoader and HCatStorer
- Use Hive to discover useful information in a dataset
- Describe how Hive queries get executed as MapReduce jobs
- Perform a join of two datasets with Hive
- Use advanced Hive features: windowing, views, ORC files

Objectives

- Describe Spark and Spark specific use cases
- Explore data interactively through the spark shell utility
- Explain the RDD concept
- Understand concepts of functional programming
- Use the Python or Scala Spark APIs
- Create all types of RDDs: Pair, Double, and Generic
- Use RDD type-specific functions
- Explain interaction of components of a Spark Application
- Explain the creation of the DAG schedule
- Build and package Spark applications
- Use application configuration items
- Deploy applications to the cluster using YARN
- Use data caching to increase performance of applications
- Understand join techniques
- Learn general application optimization guidelines/tips
- Create applications using the Spark SQL library
- Create/transform data using dataframes
- Read, use, and save to different Hadoop file formats

Spark Python or Scala Hands-On Labs

- Create a Spark "Hello World" word count application
- Use advanced RDD programming to perform sort, join, pattern matching and regex tasks
- Explore partitioning and the Spark UI
- Increase performance using data caching
- Build/package a Spark application using Maven
- Use a broadcast variable to efficiently join a small dataset to a massive dataset
- Create a data frame and perform analysis
- Load/transform/store data using Spark with Hive tables