PT. Floatway Systems
Ruko Citra Mas, Jl. Alternatif Cibubur
Cileungsi Km 1,8 Rt.004 Rw 005
Harjamukti, Cimanggis - Depok
Phone : (021) 84309690

FLOATWAY SYSTEMS
Let's collaborate!

COURSE TITLE
Big Data Introduction with Apache Hadoop Frameworks

COURSE NUMBER
PTFS 0016

INTRODUCTION
Data analysis as part of business intelligence solutions is a growing demanding needs. It has become more and more mission critical than ever before.

The challenges are data is getting bigger in size and complexity, thus make it hard to provide a valid and timely management report. Recency is the key to making the prompt and right decision.

The problem of preparing the data can be solved by a promising and proven big data platform, Hadoop. Based on Google big data platform mechanism, Hadoop successfully grow from a small project into a mainstream big data platform nowadays.

This course is developed to provide you with a basic understanding of using Hadoop, when it can use best, and how to work with Hadoop with many of its frameworks.

WHO SHOULD ATTEND?
- ✓ Database Administrator (DBA).
- ✓ Java Application Developer.

OBJECTIVES
At the completion of this course, attendee should be able to :
- ✓ Understand the concepts of Big Data and Predictive Analytics
- ✓ Understand Hadoop storage, map reduce platform and its utilities
- ✓ Configuring Hadoop Cluster.
- ✓ Setting up, developing and configuring Hadoop, Map Reduce, HBase, Yarn, Tez, Flume, Spark, Zookeeper.

COURSE DURATION
10 days / 60 hours

COURSE PREREQUISITES
- ✓ Basic understanding one or several popular DBMS and of Structured Query Language (SQL).
- ✓ Basic understanding of Java programming language.
- ✓ Basic understanding of Operating System Filesystem (I/O, Type of Filesystem, etc).

COURSE REQUIREMENTS
- ✓ PC or Laptop with minimum of 2.4GHz CPU, 8 GB of RAM, DVD Drive and 200 GB of available hard disk space.
- ✓ Softwares :
  - Microsoft Windows 7 Professional Edition
  - CentOS Linux Operating System
  - VirtualBox
  - Java Runtime Environment (JRE)

COURSE OUTLINE

Day 1

1. Introduction to Hadoop
   - ✓ History of Apache Hadoop and distributed computing
   - ✓ Hadoop File System (HDFS)
   - ✓ Hadoop Map Reduce Framework
   - ✓ Applications that built on Hadoop: Hive, Mahout, etc.
2. Installation and Configuration
   - ✓ Java Runtime Environment / Java Development Kit
   - ✓ Pentaho Data Integration
   - ✓ XAMPP package (Apache HTTP Server and MySQL).
   - ✓ SQLYog – a GUI based mysql client.
   - ✓ Data and Script samples
   - ✓ VirtualBox
   - ✓ Cloudera / Hortonworks Virtual Distribution
3. Hadoop File System (HDFS)
   - ✓ Setting up HDFS
   - ✓ HDFS command line interface (CLI) operations
   - ✓ Monitoring Task using HDFS Monitoring UI
   - ✓ Interfacing with Thrift and WEBDAV

Day 2

4. Map Reduce
   - ✓ Introduction to Map Reduce and How it Works
   - ✓ Running a Hadoop Map Reduce sample
   - ✓ Mapper and Reducer Class
   - ✓ Developing a Simple Hadoop Map Reduce program: Word Counts.
   - ✓ Developing data aggregators using Map Reduce.
   - ✓ Covernting several common algorithms to Map Reduce.
   - ✓ Monitoring Map Reduce job

Day 3

5. Pig
   - ✓ Introduction to Pig
   - ✓ Pig Installation and Set Up
   - ✓ Getting Started with Pig Scripting
   - ✓ Aggregating Data
   - ✓ User Defined Functions

Day 4

6. HBase
   - ✓ Introduction to HBase
   - ✓ Importing data
   - ✓ Queries

Day 5

7. Hadoop YARN
   - ✓ Introduction to Hadoop YARN
   - ✓ Capacity and Fair Scheduler

- ✓ YARN Application Development
- ✓ YARN Commands

8. Apache Tez

- ✓ Introduction to Tez
- ✓ Comparing Pig, Hive and ETL service
- ✓ Tez and YARN
- ✓ Tez Data Flow API
- ✓ Runtime Reconfiguration

Day 5, 6 and 7

9. Data Ingestion Services

Flume

- ✓ Flume Agent Setup and Configuration
- ✓ Starting Agent
- ✓ Data flow model
- ✓ Source, Channels and Sinks
- ✓ Design and executing data flow
- ✓ Multiplexing data flow

Sqoop

- ✓ Setup and Configuration
- ✓ Running Sqoop Server
- ✓ Running Sqooop CLI
- ✓ Sqoop Commands

Kafka

- ✓ Kafka as Hadoop Messaging Service
- ✓ Server and Client Setup and Configuration
- ✓ Broker, Producer and Consumer Setup and Configuration
- ✓ Basic Operations
- ✓ Monitoring

Day 8

10. Apache Spark (In Memory Processing)

- ✓ Setting up Spark Cluster
- ✓ Shell Client
- ✓ Resilient Distributed Database
- ✓ RDD Transformations and Actions
- ✓ Persistency Storage
- ✓ Brief Introduction to Spark Streaming and Spark SQL

Day 9

11. Machine Learning with Apache Mahout

- ✓ Data Mining and Machine Learning
- ✓ Training and Testing Dataset
- ✓ Introduction to Apache Mahout
- ✓ Classification, Clustering and Association Rules
- ✓ Stand alone Apache Mahout and integration with Hadoop

Day 10

12. Zookeper

- ✓ What is Zookeper?
- ✓ Installation and running Zookeper
- ✓ Zookeper Services and their coordination

13. Hadoop Administration
    - ✓ Software Packages Preparation
    - ✓ Basic Cluster Setup
    - ✓ Configuring Environment and Hadoop Daemons
    - ✓ Configuring Slaves
    - ✓ Load Test the Hadoop Environment

NOTES
This Course Description is subject to change due to product design changes and individual attendee needs and experience.