Hadoop for Developers

Course Summary

Description

Apache Hadoop is the most popular framework for processing Big Data on clusters of servers. This course will introduce a developer to various components (HDFS, MapReduce, Pig, Hive and HBase) Hadoop ecosystem.

Format: Lectures and hands on labs. (50% lecture + 50% labs). Pace of the class is determined by the students.

Topics

- Introduction to Hadoop
- HDFS
- Map Reduce
- Pig
- Hive
- HBase

Audience

This course was designed for developers.

Prerequisites

Before taking this course, students should have the following skills:

- Be comfortable with Java programming language (most programming exercises are in java)
- Be comfortable in Linux environment (be able to navigate Linux command line, edit files using vi / nano)

Duration

Four days
Hadoop for Developers

Course Outline

I. Introduction to Hadoop
   A. Hadoop history, concepts
   B. Eco system
   C. Distributions
   D. High level architecture
   E. Hadoop myths
   F. Hadoop challenges
   G. Hardware / software
   H. Lab : first look at Hadoop

II. HDFS
   A. Design and architecture
   B. Concepts (horizontal scaling, replication, data locality, rack awareness)
   C. Daemons : Namenode, Secondary namenode, Data node
   D. Communications / heart-beats
   E. Data integrity
   F. Read / write path
   G. Namenode High Availability (HA), Federation
   H. labs : Interacting with HDFS

III. Map Reduce
   A. Concepts and architecture
   B. Daemons (mrv1) : jobtracker / tasktracker
   C. Phases : driver, mapper, shuffle/sort, reducer
   D. Map reduce version 1 and version 2 (YARN)
   E. Internals of map reduce
   F. Introduction to java map reduce program
   G. labs : Running a sample MapReduce program

IV. Pig
   A. Pig vs java map reduce
   B. Pig job flow
   C. Pig latin language
   D. Etl with pig
   E. labs : writing Pig scripts to analyze data

V. Hive
   A. Concepts
   B. Architecture
   C. Data types
   D. Hive vs sql
   E. Loading data and querying
   F. Partitions
   G. Joins
   H. Text processing
   I. labs : various labs on processing data with Hive

VI. HBase
   A. Intro
   B. Concepts and architecture
   C. Hbase vs rdbms vs cassandra
   D. Hbase java api
   E. Time series data on hbase
   F. Schema design
   G. labs : Interacting with HBase using shell; programming in HBase Java API ; Schema design exercise