Advanced Hadoop for Developers

Course Summary

Description

Apache Hadoop is one of the most popular frameworks for processing Big Data on clusters of servers. This course delves into data management in HDFS, advanced Pig, Hive, and HBase. These advanced programming techniques will be beneficial to experienced Hadoop developers.

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

Topics

- Data Management in HDFS
- Advanced Pig
- Advanced Hive
- Advanced 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)
- Have attended “Hadoop for Developers” or have similar knowledge

Duration

Three days
Advanced Hadoop for Developers

Course Outline

I. Data Management in HDFS
   A. Various Data Formats (JSON / Avro / Parquet)
   B. Compression Schemes
   C. Data Masking
   D. Labs: Analyzing different data formats; enabling compression

II. Advanced Pig
   A. User-defined Functions
   B. Introduction to Pig Libraries (ElephantBird / Data-Fu)
   C. Loading Complex Structured Data using Pig
   D. Pig Tuning
   E. Labs: advanced pig scripting, parsing complex data types

III. Advanced Hive
   A. User-defined Functions
   B. Compressed Tables
   C. Hive Performance Tuning
   D. Labs: creating compressed tables, evaluating table formats and configuration

IV. Advanced HBase
   A. Advanced Schema Modelling
   B. Compression
   C. Bulk Data Ingest
   D. Wide-table / Tall-table comparison
   E. HBase and Pig
   F. HBase and Hive
   G. HBase Performance Tuning
   H. Labs: tuning HBase; accessing HBase data from Pig & Hive; Using Phoenix for data modeling