

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 focuses on advanced programming techniques that will be beneficial to experienced Hadoop developers.

Topics

- Data Management in HDFS
- Advanced Pig
- Advanced Hive
- Advanced HBase

Audience

This course is designed for developers.

Prerequisites

- Comfortable with Java programming language (most programming exercises are in java)
- Comfortable in Linux environment (be able to navigate Linux command line, edit files using vi / nano)
- Attended “Hadoop for Developers” or has working knowledge of Hadoop.

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
 - I. Browse our courses
 - J. We offer instructor-led courses (onsite and online)
 - K. Looking for team training?
 - L. Up-skill your team with a customized, private training
 - M. Public Classes
 - N. Suitable for small teams and individuals