

Advanced Hadoop for Developers

Course Outline:

Section 1: Data Management in HDFS

- Various Data Formats (JSON / Avro / Parquet)
- Compression Schemes
- Data Masking
- Labs : Analyzing different data formats; enabling compression

Section 2: Advanced Pig

- User-defined Functions
- Introduction to Pig Libraries (ElephantBird / Data-Fu)
- Loading Complex Structured Data using Pig
- Pig Tuning
- Labs : advanced pig scripting, parsing complex data types

Section 3 : Advanced Hive

- User-defined Functions
- Compressed Tables
- Hive Performance Tuning
- Labs : creating compressed tables, evaluating table formats and configuration

Section 4 : Advanced HBase

- Advanced Schema Modelling

- Compression
- Bulk Data Ingest
- Wide-table / Tall-table comparison
- HBase and Pig
- HBase and Hive
- HBase Performance Tuning
- Labs : tuning HBase; accessing HBase data from Pig & Hive; Using Phoenix for data modeling