Hadoop For Developers

Course Outline:

Section 1: Introduction to Hadoop

- hadoop history, concepts
- eco system
- distributions
- high level architecture
- hadoop myths
- hadoop challenges
- hardware / software
- Lab : first look at Hadoop

Section 2: HDFS

- Design and architecture
- concepts (horizontal scaling, replication, data locality, rack awareness)
- Daemons : Namenode, Secondary namenode, Data node
- communications / heart-beats
- data integrity
- read / write path
- Namenode High Availability (HA), Federation
- labs : Interacting with HDFS

Section 3 : Map Reduce

- concepts and architecture
- daemons (MRV1) : jobtracker / tasktracker
- phases : driver, mapper, shuffle/sort, reducer
- Map Reduce Version 1 and Version 2 (YARN)
- Internals of Map Reduce
- Introduction to Java Map Reduce program
- labs : Running a sample MapReduce program

Section 4 : Pig

- pig vs java map reduce
- pig job flow
- pig latin language
- ETL with Pig
- Transformations & Joins
- User defined functions (UDF)
- labs : writing Pig scripts to analyze data

Section 5: Hive

- architecture and design
- data types
- SQL support in Hive
- Creating Hive tables and querying
- partitions
- joins

- text processing
- labs : various labs on processing data with Hive

Section 6: HBase

- concepts and architecture
- hbase vs RDBMS vs cassandra
- HBase Java API
- Time series data on HBase
- schema design
- labs : Interacting with HBase using shell; programming in HBase Java API ; Schema design exercise