Apache Hadoop with Hive

Course Summary

Description
Hive is the de-facto standard for data warehousing Hadoop. This course starts with standard Hive setup and operations, continues into Advanced Hive use, discusses performance and execution engines, and ends with a practical workshop.

Topics
- Hive Basics
- Hive Advanced
- Impala (for Cloudera track)
- TEZ (for Hortonworks track)
- Bonus Workshop

Audience
This course is intended for data scientists, software engineers and Administrators. It gives them a practical level of experience, achieved through a combination of about 50% lecture, 50% lab work.

Prerequisites
- Familiarity with SQL
- Browser – the course is taught in Cloudera Hue or Hortonworks Views
- Command line helpful but not required

Duration
Two Days
Apache Hadoop with Hive

Course Outline

I. Hive Basics
   A. Defining Hive Tables
   B. SQL Queries over Structured Data
   C. Filtering / Search
   D. Aggregations / Ordering
   E. Partitions
   F. Joins
   G. Text Analytics (Semi-Structured Data)

II. Hive Advanced
   A. Transformation, Aggregation
   B. Working with Dates, Timestamps, and Arrays
   C. Converting Strings to Date, Time, and Numbers
   D. Create new Attributes, Mathematical Calculations, Windowing Functions
   E. Use Character and String Functions
   F. Binning and Smoothing
   G. Processing JSON Data
   H. Execution Engines (Tez, MR, Spark)

III. Impala (for Cloudera track)
   A. Architecture
   B. Impala joins and other SQL specifics

IV. TEZ (for Hortonworks track)
   A. Architecture
   B. Performance and use

V. Bonus Workshop
   A. Students will work in teams to do this end-to-end workshop
   B. Setup a data warehouse with Hive
   C. Query and analyze data with Hive and Spark