Hadoop for Business Analysts

Course Summary

Description

Apache Hadoop is the most popular framework for processing Big Data. Hadoop provides rich and deep analytics capability, and it is making in-roads in to tradional BI analytics world. This course will introduce an analyst to the core components of Hadoop eco system and its analytics.

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

Topics

- Introduction to Hadoop
- HDFS Overview
- Map Reduce Overview
- Pig
- Hive
- BI Tools for Hadoop
- Conclusion

Audience

This course is designed for Business Analysts.

Prerequisites

Before attending this course, students should have a programming background with databases / SQL, and basic knowledge of Linux (be able to navigate Linux command line, editing files with vi / nano).

Duration

Three days

Hadoop for Business Analysts

Course Outline

I. Introduction to Hadoop
- A. Hadoop history, concepts
- B. Eco system
- C. Distributions
- D. High level architecture
- E. Hadoop myths
- F. Hadoop challenges
- G. Hardware / software
- H. Labs : first look at Hadoop

II. HDFS Overview
- A. Concepts (horizontal scaling, replication, data locality, rack awareness)
- B. Architecture (Namenode, Secondary namenode, Data node)
- C. Data integrity
- D. Future of HDFS : Namenode HA, Federation
- E. Labs : Interacting with HDFS

III. Map Reduce Overview
- A. Mapreduce concepts
- B. Daemons : jobtracker / tasktracker
- C. Phases : driver, mapper, shuffle/sort, reducer
- D. Thinking in map reduce
- E. Future of mapreduce (yarn)
- F. Labs : Running a Map Reduce program

IV. Pig
- A. Pig vs java map reduce
- B. Pig latin language
- C. User defined functions
- D. Understanding pig job flow
- E. Basic data analysis with Pig
- F. Complex data analysis with Pig
- G. Multi datasets with Pig
- H. Advanced concepts
- I. Lab : writing pig scripts to analyze / transform data

V. Hive
- A. Hive concepts
- B. Architecture
- C. Data types
- D. Hive data management
- E. Hive vs sql
- F. Labs (multiple) : creating Hive tables and running queries, joins , using partitions, using text analytics functions

VI. BI Tools for Hadoop
- A. BI tools and Hadoop
- B. Overview of current BI tools landscape

VII. Conclusion
- A. Choosing the best tool for the job