Hadoop for Administrators

Course Summary

Description

Apache Hadoop is the most popular framework for processing Big Data on clusters of servers. In this three (optionally, four) days course, attendees will learn about the business benefits and use cases for Hadoop and its ecosystem, how to plan cluster deployment and growth, how to install, maintain, monitor, troubleshoot and optimize Hadoop. They will also practice cluster bulk data load, get familiar with various Hadoop distributions, and practice installing and managing Hadoop ecosystem tools. The course finishes off with discussion of securing cluster with Kerberos.

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

Topics

- Introduction
- Planning and Installation
- HDFS Operations
- Data Ingestion
- MapReduce Operations and Administration
- YARN: New Architecture and New Capabilities
- Advanced Topics
- Optional Tracks

Audience

This course is designed for Hadoop administrators.

Prerequisites

Before taking this course, students should have the following skills:

- Be comfortable with basic Linux system administration
- Basic scripting skills
- Knowledge of Hadoop and Distributed Computing is not required, but will be introduced and explained in the course.

Duration

Three to four days

ProTech protechtraining.com *"Charting the Course ... ... to Your Success!"*

Hadoop for Administrators

Course Outline

I. Introduction
- A. Hadoop history, concepts
- B. Ecosystem
- C. Distributions
- D. High level architecture
- E. Hadoop myths
- F. Hadoop challenges (hardware / software)
- G. Labs: discuss your Big Data projects and problems

II. Planning and Installation
- A. Selecting software, Hadoop distributions
- B. Sizing the cluster, planning for growth
- C. Selecting hardware and network
- D. Rack topology
- E. Installation
- F. Multi-tenancy
- G. Directory structure, logs
- H. Benchmarking
- I. Labs: cluster install, run performance benchmarks

III. HDFS Operations
- A. Concepts (horizontal scaling, replication, data locality, rack awareness)
- B. Nodes and daemons (NameNode, Secondary NameNode, HA Standby NameNode, DataNode)
- C. Health monitoring
- D. Command-line and browser-based administration
- E. Adding storage, replacing defective drives
- F. Labs: getting familiar with HDFS command lines

IV. Data Ingestion
- A. Flume for logs and other data ingestion into HDFS
- B. Sqoop for importing from SQL databases to HDFS, as well as exporting back to SQL
- C. Hadoop data warehousing with Hive
- D. Copying data between clusters (distcp)
- E. Using S3 as complementary to HDFS
- F. Data ingestion best practices and architectures
- G. Labs: setting up and using Flume, the same for Sqoop

V. MapReduce Operations and Administration
- A. Parallel computing before mapreduce: compare HPC vs Hadoop administration
- B. MapReduce cluster loads
- C. Nodes and Daemons (JobTracker, TaskTracker)
- D. MapReduce UI walk through
- E. Mapreduce configuration
- F. Job config
- G. Optimizing MapReduce
- H. Fool-proofing MR: what to tell your programmers
- I. Labs: running MapReduce examples

VI. YARN: New Architecture and New Capabilities
- A. YARN design goals and implementation architecture
- B. New actors: ResourceManager, NodeManager, Application Master
- C. Installing YARN
- D. Job scheduling under YARN
- E. Labs: investigate job scheduling

VII. Advanced Topics
- A. Hardware monitoring
- B. Cluster monitoring
- C. Adding and removing servers, upgrading Hadoop
- D. Backup, recovery and business continuity planning
- E. Oozie job workflows
- F. Hadoop high availability (HA)
- G. Hadoop Federation
- H. Securing your cluster with Kerberos
- I. Labs: set up monitoring

VIII. Optional Tracks
- A. Cloudera Manager for cluster administration, monitoring, and routine tasks; installation, use. In this track, all exercises and labs are performed within the Cloudera distribution environment (CDH5)
- B. Ambari for cluster administration, monitoring, and routine tasks; installation, use. In this track, all exercises and labs are performed within the Ambari cluster manager and Hortonworks Data Platform (HDP 2.0)