Course Outline

Hadoop for Systems Administrators

Course Summary

Description

This course covers the essentials of deploying and managing an Apache Hadoop cluster. The course is lab intensive with each participant creating their own Hadoop cluster using either the CDH (Cloudera's Distribution, including Apache Hadoop) or Hortonworks Data Platform stacks. Core Hadoop services are explored in depth with emphasis on troubleshooting and recovering from common cluster failures. The fundamentals of related services such as Ambari, Zookeeper, Pig, Hive, HBase, Sqoop, Flume, and Oozie are also covered. The course is approximately 60% lecture and 40% labs.

Topics

- Hadoop Overview
- HDFS
- YARN
- MapReduce
- Installing Hadoop with Ambari Lab Tasks
- Data Ingestion
- Data Lineage and Governance
- Data Processing Frameworks
- NoSQL Implementations
- Cluster Management

Audience

This course is designed for those wanting to learn the essentials of deploying and managing an Apache Hadoop cluster.

Prerequisites

Before taking this course, students should be comfortable with the Linux commands and have some systems administration experience, but do not need previous Hadoop experience.

Duration

Three days

Course Outline

Hadoop for Systems Administrators

Course Outline

I. Hadoop Overview

A. Data Analysis
B. Big Data
C. Origins of Hadoop
D. Hadoop Marketplace
E. Hadoop Core
F. Hadoop Ecosystem:
G. Cluster Architecture
H. Hardware/Software Requirements
I. Running Commands on Multiple Systems

Labs:

- Running Commands on Multiple Hosts
- Preparing to Install Hadoop

II. HDFS

A. Design Goals
B. Design
C. Blocks
D. Block Replication
E. Namenode Daemon
F. Secondary Namenode Daemon
G. Datanode Daemon
H. Accessing HDFS
I. Permissions and Users
J. Adding and Removing Datanodes
K. Balancing

Labs

- Single Node HDFS
- Multi-node HDFS
- Files and HDFS
- Managing and Maintaining HDFS

III. YARN

A. YARN Design Goals
B. YARN Architecture
C. Resource Manager
D. Node Manager
E. Containers
F. YARN: Other Important Features
G. Slider

Lab

- YARN

IV. MapReduce

A. MapReduce
B. Terminology and Data Flow

Lab

- Mapreduce

V. Installing Hadoop with Ambari Lab Tasks

A. CDH Uninstall
B. Installing Hadoop with Ambari
C. Tez

VI. Data Ingestion

A. Sqoop
B. Flume
C. Kafka

Lab

D. Sqoop

VII. Data Lineage and Governance

A. Falcon
B. Atlas
C. Oozie

VIII. Data Processing Frameworks

A. The Bane of MapReduce
B. Tez overview
C. Pig
D. Hive
E. Spark
F. Storm
G. Solr
H. Solr (cont)

Lab

- Pig

IX. NoSQL Implementations

A. HBase
B. Phoenix

X. Cluster Management

A. Ambari Metrics System (AMS)
B. Zookeeper