Hadoop Uncovered: Foundational Training for Admins and Developers

Course Summary

Description

Get ready for big data! You will learn the fundamentals of setting up a Hadoop cluster as well as the "soup" of related technologies like Hive, Pig and Oozie. Come prepared to learn how to access the Hadoop file system, write MapReduce jobs using java, Pig and Hive. as well as how to use Pig, Hive and Oozie. Every participant will work with their own installation of a Hadoop 2, single node cluster. Each topic is reinforced with hands-on workshops.

Objectives
At the end of this course, students will be able to:

- Understand the Hadoop File System (HDFS)
- Understand what MapReduce is and why you should care
- Write a MapReduce job with Java, Pig and Hive
- Understand how the different Hadoop technologies interoperate to provide a cohesive big data solution
- Understand basic management of a Hadoop cluster
- Learn how to perform basic unit testing of your MapReduce jobs
- Understand the different modes that Hadoop can be run in to support massive amounts of data as
- well as your MapReduce jobs during development

And Best of All: You will walk away with a fully configured virtual machine (that can run under VirtualBox or VMWare) with Hadoop and all the related technologies installed, configured and ready to run. Also included in the VM is the necessary development environment using Eclipse so when you return home after your training you can be immediately productive in extending your Hadoop knowledge by using a live environment without the hassle of having to set one up from scratch!

Topics

- Hadoop Overview
- HDFS
- HBASE
- Map Reduce on YARN
- Hadoop Streaming
- MapReduce Workflows
- Oozie
- Pig
- Hive

Prerequisites

Before taking this course, students should have basic Java Knowledge. (Experience with Eclipse a plus).
**This course should be considered an intermediate level course (Intermediate Java knowledge not required).

Duration

Four days

Hadoop Uncovered: Foundational Training for Admins and Developers

Course Outline

I. Hadoop Overview

What is Big Data? How did we get to this point? How does Hadoop compare to a relational database system? This section answers all those questions and more.

A. Big Data Introduction
B. History
C. Comparison to Relational Databases
D. Hadoop Ecosystem

II. HDFS

The hadoop file system is a core component of Hadoop. It is the magic sauce that makes big data processing possible. Covered in this section is how the file system relates to the underlying OS File system, the different processes that manage the file system and how to access HDFS to put data in and pull data out.

A. Architecture/Concepts
B. Access
C. Namenodes
D. Filesystem Shell
E. Accessing HDFS with Java
F. Reading/Writing/Browsing file system

III. HBASE

HBASE is a non-relational database that is similar to Google's Big Table. Intended for massive amounts of data, this section covers the structure and data model and how to access/use hbase. You will use the java API as well as look at how MapReduce jobs can access from an HBASE data source.

A. Overview
B. Architecture
C. Data Model
D. Installation and Shell
E. Access via Java API
F. Administration access via Java
G. Scan API
H. Filters
I. Storage Model
J. Table Design

IV. Map Reduce on YARN

The guts of big data processing is MapReduce. This section will provide all the fundamental learning to get you well on your way to writing MapReduce jobs in java. You will learn about and use YARN to submit your jobs and learn how HDFS works with YARN to achieve processing on a massive scale. You will also dive into more detail on MapReduce theory and learn about the foundational concepts behind processing big data.

A. Introduction
B. Processing Model
C. Command line tools
D. MapReduce framework
E. Submitting MapReduce Jobs
F. Writing MapReduce jobs in Java
G. MapReduce Theory
H. Distributive Cache
I. Speculative Executin
J. YARN Components
K. Counters
L. Details of MapReduce Job Execution

V. Hadoop Streaming

Many developers as well as data analysts do not know java. No problem, Hadoop streaming to the rescue. This section will cover how to write MapReduce jobs in any language that has access to the stdout and stdin streams. Examples and exercises are done in python.

A. Implementing a streaming job
B. Counters in streaming jobs
C. Contrast with Java Jobs

VI. MapReduce Workflows

MapReduce jobs are often chained together to accomplish a larger processing goal; each job just being a step in the overall flow. This section covers how to put together workflows using java. We also discuss breaking up larger jobs into smaller "reusable" chunks that can then be reused in many workflows.

A. Problem decomposition into MapReduce Jobs
B. Coding workflows
C. Using the JobControl Class

Hadoop Uncovered: Foundational Training for Admins and Developers

Course Outline (cont'd)

VII. Oozie

Writing a workflow job using java is great and all, but there has to be another way. Oozie to the rescue! Oozie is a workflow system that sits on top of the Hadoop MapReduce ecosystem to allow analysts to compose workflows from existing jobs using XML.

A. Oozie Installation
B. Writing Oozie workflows
C. Deploying and running Oozie jobs

VIII. Pig

Just when you thought you knew enough scripting languages along comes another one. Pig is a powerful technology in the Hadoop ecosystem that allows you to write MapReduce jobs using PigLatin. You will learn a little PigLatin as well as use the Grunt interactive shell.

A. Installation
B. Pig Latin
C. Writing Pig Scripts
D. User Defined functions
E. Data set joins

IX. Hive

Not everyone wants to give up their SQL skills. Hive provides a way to overlay some structure on all the semistructured data stored in Hadoop. We will cover table creation and the SQL 92 type syntax that can be used to execute MapReduce jobs on the Hadoop cluster.

A. Installation
B. Table creation and deletion
C. Partitioning
D. Loading data into Hive
E. Joins
F. Bucketing