CAP - Introduction to Hadoop and Big Data

Course Description

The Big Data landscape is one which is continuously evolving as new technologies emerge and existing technologies mature. This is a comprehensive course covering the Hadoop architecture and the Hadoop ecosystem of tools. These technologies are at the foundation of the Big Data movement, and they facilitate scalable management and processing of vast quantities of data.

Students who complete this course will understand the architecture of Hadoop clusters at both the hardware and system software levels. Students will learn to apply Hadoop and related Big Data technologies such as MapReduce, Hive, Impala, and Pig in developing analytics and solving the types of problems faced by enterprises today.

Prerequisites

A course in Operating Systems; programming experience in Java, Python, or C/C++ for assignments; and a general understanding of networking and distributed systems. Familiarity with Linux and databases will be helpful.

Texts

Recommended

- Hadoop: The Definitive Guide (third edition), by Tom White

Optional

- Hadoop Operations, by Eric Sammer
- Programming Pig, by Alan Gates
- Programming Hive, by Capriolo, Wampler, and Rutherglen

Tools

Cloudera's fully configured Hadoop VM (virtual machine), including datasets for homework labs.

Detailed Course Outline

Lecture 1

1. **Course Introduction**
   - About This Course
   - About Apache Hadoop

2. **The Motivation for Hadoop**
   - Problems with traditional large-scale systems
   - Requirements for a New Approach
   - Hadoop!
   - Hadoop-able problems

3. **Hadoop Basic Concepts**
   - What is Hadoop?
   - The Hadoop Distributed File System (HDFS)
   - How MapReduce Works

4. **Hadoop Solutions**
   - Some Common Hadoop Applications
   - Other Interesting Hadoop Use Cases

Homework:
Setup and access Hadoop environment
Lab: Using HDFS
Readings

Lecture 2

1. **Review**

2. **The Hadoop Ecosystem**
   - Introduction
   - Data Storage: HBase
   - Data Integration: Flume and Sqoop
   - Data Processing: Spark
   - Data Analysis: Hive, Pig and Impala
   - Workflow Engine: Oozie
   - Machine Learning: Mahout

3. **Managing Your Hadoop Solution**
   - Hadoop in the Data Center
   - Cluster Hardware

4. **Introduction to MapReduce**
   - Mapreduce Overview
   - Example: WordCount
   - Mappers
   - Reducers

5. **Hadoop Clusters**
   - Hadoop Cluster Overview
   - Hadoop Jobs and Tasks

Homework:
Lab: Running a MapReduce Job
Readings

Lecture 3

1. **Review**

2. **Writing a MapReduce Program in Java**
   - Basic MapReduce API Concepts
   - Writing a MapReduce Program in Java
   - Speeding up Hadoop Development by Using Eclipse
   - Differences Between the Old and New MapReduce APIs

3. **Writing a MapReduce Program Using Streaming**
   - Writing Mappers and Reducers with the Streaming API

4. **Unit Testing MapReduce Programs**
   - Unit testing
   - The JUnit and MRUnit testing frameworks
   - Writing Unit Tests with MRUnit
   - Running Unit Tests

Homework:
Lab: Writing a MapReduce Java Program
Lab: More Practice with MapReduce Java Programs
Lab: Writing a MapReduce Streaming Program
Lab: Unit Testing with the MRUnit Framework
Readings

Lecture 4

1. **Review**

2. **Delving Deeper into the Hadoop API**
   - Using the ToolRunner Class
   - Setting Up and Tearing Down Mappers and Reducers
   - Decreasing the Amount of Intermediate Data with Combiners
   - Accessing HDFS programmatically
   - Using the Distributed Cache
   - Using the Hadoop API's Library of Mappers, Reducers and Partitioners

   Homework:
   Lab: Using ToolRunner and Passing Parameters
   Lab: Using a Combiner
   Readings

Lecture 5

1. **Review**

2. **Practical Development Tips and Techniques**
   - Strategies for Debugging MapReduce Code
   - Testing MapReduce Code Locally Using LocalJobRunner
   - Writing and Viewing Log Files
   - Retrieving Job Information with Counters
   - Reusing Objects
   - Creating Map - only MapReduce Jobs

   Homework:
   Lab: Testing with LocalJobRunner
   Lab: Logging
   Lab: Using Counters and a Map - Only Job
   Readings

Lecture 6

1. **Review**

2. **Partitioners and Reducers**
   - How Partitioners and Reducers Work Together
   - Determining the Optimal Number of Reducers for a Job

- Writing Custom Partitioners

3. **Data Input and Output**
   - Custom Writable and WritableComparable Implementations
   - Saving binary data using SequenceFiles and Avro data files
   - Issues to Consider When Using File Compression

Homework:
Lab: Writing a Partitioner
Lab: Implementing a Custom WritableComparable
Lab: Using SequenceFiles and File Compression
Readings

Lecture 7

1. **Review**

2. **Common MapReduce Algorithms**
   - Sorting and Searching Large Data Sets
   - Indexing Data
   - Computing Term Frequency - Inverse Document Frequency (TF - IDF)
   - Calculating Word Co-Occurrence
   - Performing a Secondary Sort

3. **Joining Data Sets in MapReduce Jobs**
   - Writing a Map-Side Join
   - Writing a Reduce-Side Join

Homework:
Lab: Creating an Inverted Index
Lab: Calculating Word Co-Occurrence
Readings

Lecture 8

1. **Review**

2. **Hadoop Tools for Data Acquisition**
   - Loading Data from an RDBMS into HDFS by Using Sqoop
   - Managing Real-Time Data Using Flume

3. **Creating Workflows with Oozie**
   - Introduction to Oozie

- Creating Oozie Workflows

4. **Introduction to Pig**
   - What is Pig?
   - Pig's Features
   - Pig Use Cases
   - Interacting with Pig

5. **Midterm Exam**

Homework:
Lab: Importing Data with Sqoop
Lab: Running an Oozie Workflow
Lab: Exploring a Secondary Sort Example
Study for midterm exam
Readings

Lecture 9

1. **A Brief Review**
   - Hadoop Review
   - Pig Review

2. **Basic Data Analysis with Pig**
   - Pig Latin Syntax
   - Loading Data
   - Simple Data Types
   - Field Definitions
   - Data Output
   - Viewing the Schema
   - Filtering and Sorting Data
   - Commonly-used Functions

3. **Processing Complex Data with Pig**
   - Storage Formats
   - Complex/Nested Data Types
   - Grouping
   - Built-in Functions for complex Data
   - Iterating Grouped Data

Homework:
Lab: Data Ingest with Hadoop Tools
Lab: Using Pig for ETL Processing
Lab: Analyzing Ad Campaign Data with Pig

Readings

Lecture 10

1. **Review**

2. **Multi-Dataset Operations with Pig**
   - Techniques for Combining Data Sets
   - Joining Data Sets in Pig
   - Set Operations
   - Splitting Data Sets

3. **Extending Pig**
   - Adding Flexibility with Parameters
   - Macros and Imports
   - UDFs
   - Contributed Functions
   - Using Other Languages to Process Data with Pig

4. **Pig Troubleshooting and Optimization**
   - Troubleshooting Pig
   - Logging
   - Using Hadoop's Web UI
   - Data Sampling and Debugging
   - Performance Overview
   - Understanding the Execution Plan
   - Tips for Improving the Performance of your Pig Jobs

Homework:
Lab: Analyzing Disparate Data Sets with Pig
Lab: Extending Pig with Streaming and UDFs
Readings

Lecture 11

1. **Review**

2. **Introduction to Hive**
   - What Is Hive?
   - Hive Schema and Data Storage
   - Comparing Hive to Traditional Databases
   - Hive Use Cases
   - Interacting with Hive

3. **Relational Data Analysis with Hive**
   - Hive Databases and Tables
   - Basic HiveQL Syntax
   - Data Types
   - Joining Datasets
   - Common Built-in Functions

4. **Hive Data Management**
   - Hive Data Formats
   - Creating Databases and Hive-Managed Tables
   - Loading Data into Hive
   - Altering Databases and Tables
   - Self-Managed Tables
   - Simplifying Queries with Views
   - Storing Query Results
   - Controlling Access to Data

Homework:
Lab: Running Hive Queries from the Shell, Scripts, Hue
Lab: Data Management with Hive
Readings

Lecture 12

1. **Review**

2. **Text Processing with Hive**
   - Overview of Text Processing
   - Important String Functions
   - Using Regular Expressions in Hive
   - Sentiment Analysis and n-grams

3. **Hive Optimization**
   - Understanding Query Performance
   - Controlling Job Execution
   - Partitioning
   - Bucketing
   - Indexing Data

4. **Extending Hive**
   - SerDes
   - Data Transformation with Custom Scripts
   - User-Defined Functions
   - Parameterized Queries

Homework:
Lab: Gaining Insight with Sentiment Analysis
Lab: Data Transformation with Hive
Readings

Lecture 13

1. **Review**

2. **Introduction to Impala**
   - What is Impala?
   - How Impala differs from Hive and Pig
   - How Impala differs from Relational Databases
   - Limitations and Future Directions
   - Using the Impala shell

3. **Analyzing Data with Impala**
   - Basic Syntax
   - Data Types
   - Filtering, Sorting and Limiting Results
   - Joining and Grouping Data
   - User-Defined Functions
   - Improving Impala Performance

4. **Conclusion: Choosing the Best Tool for the Job**
   - Comparing MapReduce, Pig, Hive, Impala, and Relational Databases
   - Which to Choose?

5. **Final Exam Review**

Homework:
Lab: Interactive Analysis with Impala
Study for final exam