Cloudera Developer Training for Apache Hadoop

Duration: 5.00 Days   Course Code: GK3902

Overview:

Learn to create robust data processing applications using Apache Hadoop.
You will learn to build powerful data processing applications in this course. You will learn about MapReduce, the Hadoop Distributed Files System (HDFS), and how to write MapReduce code, and you will cover best practices for Hadoop development, debugging, and implementation of workflows.
This course covers concepts addressed on the Cloudera Certified Developer for Apache Hadoop (CCDH) exam, and you will receive one CCDH exam voucher at the end of class.
Throughout the course, you will write Hadoop code and perform other hands-on exercises to solidify your understanding of the concepts.

Target Audience:

Developers who need to write and maintain Apache Hadoop applications.

Objectives:

- MapReduce and the HDFS

- Write MapReduce code in Java or other programming languages

- Issues to consider when developing MapReduce jobs

- Implement common algorithms in Hadoop

- Best practices for Hadoop development and debugging

- Use other projects such as Apache Hive, Apache Pig, Sqoop, and Oozie

- Advanced Hadoop API topics required for real-world data analysis

Prerequisites:

- Some programming experience (preferably Java)
- Knowledge of Hadoop is not required

Testing and Certification

- This course is part of the following programs or tracks:
- CCDH: Cloudera Certified Developer for Apache Hadoop (CDH3)

Follow-on-Courses:

- Cloudera Training for Apache HBase
- Cloudera Training for Apache Hive and Pig

Content:

1. Motivation for Hadoop

- Problems with Traditional Large-Scale Systems
- Requirements for a New Approach

2. Hadoop: Basic Concepts

- Hadoop Distributed File System (HDFS)
- MapReduce
- Anatomy of a Hadoop Cluster
- Other Hadoop Ecosystem Components

3. Writing a MapReduce Program

- MapReduce Flow
- Examining a Sample MapReduce Program
- Basic MapReduce API Concepts
- Driver Code
- Mapper
- Reducer
- Streaming API
- Using Eclipse for Rapid Development
- New MapReduce API

4. Integrating Hadoop into the Workflow

- Relational Database Management Systems
- Storage Systems
- Importing Data from a Relational Database Management System with Sqoop
- Importing Real-Time Data with Flume
- Accessing HDFS Using FuseDFS and Hoop

5. Delving Deeper into the Hadoop API

- ToolRunner
- Testing with MRUnit
- Reducing Intermediate Data with Combiners
- Configuration and Close Methods for Map/Reduce Setup and Teardown
- Writing Partitioners for Better Load Balancing
- Directly Accessing HDFS
- Using the Distributed Cache

6. Common MapReduce Algorithms

- Sorting and Searching
- Indexing
- Machine Learning with Mahout
- Term Frequency
- Inverse Document Frequency
- Word Co-Occurrence

7. Using Hive and Pig

- Hive Basics
- Pig Basics

8. Practical Development Tips and Techniques

- Debugging MapReduce Code
- Using LocalJobRunner Mode for Easier Debugging
- Retrieving Job Information with Counters
- Logging
- Splittable File Formats
- Determining the Optimal Number of Reducers
- Map-Only MapReduce Jobs

9. Advanced MapReduce Programming

- Custom Writables and WritableComparables
- Saving Binary Data Using SequenceFiles and Avro Files
- Creating InputFormats and OutputFormats

10. Joining Data Sets in MapReduce

- Map-Side Joins
- Secondary Sort
- Reduce-Side Joins

11. Graph Manipulation in Hadoop

- Graph Techniques
- Representing Graphs in Hadoop
- Implementing a Sample Algorithm: Single Source Shortest Path

12. Creating Workflows with Oozie

- Motivation for Oozie
- Workflow Definition Format

Further Information:

For More information, or to book your course, please call us on 353-1-814 8200

info@globalknowledge.ie

www.globalknowledge.ie

Global Knowledge, 3rd Floor Jervis House, Millennium Walkway, Dublin 1