cloudera®

TRAINING SHEET

> "Developer training was great. I believe Cloudera is the best vendor evangelizing the Big Data movement and doing a great service in promoting Hadoop in the industry. Thanks for all your help getting me started on this journey."
>
> Cisco

Cloudera Developer Training for Apache Hadoop

Take Your Knowledge to the Next Level with Cloudera's Apache Hadoop Training and Certification

Cloudera University's four-day developer training course delivers the key concepts and expertise participants need to create robust data processing applications using Apache Hadoop. From workflow implementation and working with APIs through writing MapReduce code and executing joins, Cloudera's training course is the best preparation for the real-world challenges faced by Hadoop developers.

Hands-On Hadoop

Through instructor-led discussion and interactive, hands-on exercises, participants will navigate the Hadoop ecosystem, learning topics such as:

- The internals of MapReduce and HDFS and how to write MapReduce code
- Best practices for Hadoop development, debugging, and implementation of workflows and common algorithms
- How to leverage Hive, Pig, Sqoop, Flume, Oozie, and other Hadoop ecosystem projects
- Creating custom components such as WritableComparables and InputFormats to manage complex data types
- Writing and executing joins to link data sets in MapReduce
- Advanced Hadoop API topics required for real-world data analysis

Audience & Prerequisites

This course is best suited to developers and engineers who have programming experience. Knowledge of Java is strongly recommended and is required to complete the hands-on exercises.

Developer Certification

Upon completion of the course, attendees are encouraged to continue their study and register for the Cloudera Certified Developer for Apache Hadoop (CCDH) exam. Certification is a great differentiator; it helps establish you as a leader in the field, providing employers and customers with tangible evidence of your skills and expertise.

cloudera®

Course Outline: Cloudera Developer Training for Apache Hadoop

Introduction

The Motivation for Hadoop

- Problems with Traditional Large-Scale Systems
- Introducing Hadoop
- Hadoopable Problems

Hadoop: Basic Concepts and HDFS

- The Hadoop Project and Hadoop Components
- The Hadoop Distributed File System

Introduction to MapReduce

- MapReduce Overview
- Example: WordCount
- Mappers
- Reducers

Hadoop Clusters and the Hadoop Ecosystem

- Hadoop Cluster Overview
- Hadoop Jobs and Tasks
- Other Hadoop Ecosystem Components

Writing a MapReduce Program in Java

- Basic MapReduce API Concepts
- Writing MapReduce Drivers, Mappers, and Reducers in Java
- Speeding Up Hadoop Development by Using Eclipse
- Differences Between the Old and New MapReduce APIs

Writing a MapReduce Program Using Streaming

- Writing Mappers and Reducers with the Streaming API

Unit Testing MapReduce Programs

- Unit Testing
- The JUnit and MRUnit Testing Frameworks
- Writing Unit Tests with MRUnit
- Running Unit Tests

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

Practical Development Tips and Techniques

- Strategies for Debugging MapReduce Code
- Testing MapReduce Code Locally by Using LocalJobRunner
- Writing and Viewing Log Files
- Retrieving Job Information with Counters
- Reusing Objects
- Creating Map-Only MapReduce Jobs

Partitioners and Reducers

- How Partitioners and Reducers Work Together
- Determining the Optimal Number of Reducers for a Job
- Writing Customer Partitioners

Data Input and Output

- Creating Custom Writable and WritableComparable Implementations
- Saving Binary Data Using SequenceFile and Avro Data Files
- Issues to Consider When Using File Compression
- Implementing Custom InputFormats and OutputFormats

Common MapReduce Algorithms

- Sorting and Searching Large Data Sets
- Indexing Data
- Computing Term Frequency — Inverse Document Frequency
- Calculating Word Co-Occurrence
- Performing Secondary Sort

Joining Data Sets in MapReduce Jobs

- Writing a Map-Side Join
- Writing a Reduce-Side Join

Integrating Hadoop into the Enterprise Workflow

- Integrating Hadoop into an Existing Enterprise
- Loading Data from an RDBMS into HDFS by Using Sqoop
- Managing Real-Time Data Using Flume
- Accessing HDFS from Legacy Systems with FuseDFS and HttpFS

An Introduction to Hive, Imapala, and Pig

- The Motivation for Hive, Impala, and Pig
- Hive Overview
- Impala Overview
- Pig Overview
- Choosing Between Hive, Impala, and Pig

An Introduction to Oozie

- Introduction to Oozie
- Creating Oozie Workflows

Conclusion

Cloudera Certified Developer for Apache Hadoop (CCDH)

Establish yourself as a trusted and valuable resource by completing the certification exam for Apache Hadoop developers. CCDH certifies your technical knowledge, skill, and ability to write, maintain, and optimize Apache Hadoop development projects. The exam can be demanding and will test your fluency with concepts and terminology in the following areas:

Core Hadoop Concepts

Recognize and identify Apache Hadoop daemons and how they function both in data storage and processing. Understand how Apache Hadoop exploits data locality. Determine the challenges to large-scale computational models and how distributed systems attempt to overcome various challenges posed by the scenario.

Storing Files in Hadoop

Analyze the benefits and challenges of the HDFS architecture, including how HDFS implements file sizes, block sizes, and block abstraction. Understand default replication values and storage requirements for replication. Determine how HDFS stores, reads, and writes files.

Job Configuration and Submission

Construct proper job configuration parameters, including using JobConf and appropriate properties. Identify the correct procedures for MapReduce job submission.

Job Execution Environment

Determine the lifecycle of a Mapper and the lifecycle of a Reducer in a MapReduce job. Understand the key fault tolerance principles at work in a MapReduce job. Identify the role of Apache Hadoop Classes, Interfaces, and Methods. Understand how speculative execution exploits differences in machine configurations and capabilities in a parallel environment and how and when it runs.

Job Lifecycle

Analyze the order of operations in a MapReduce job, how data moves from place to place, how partitioners and combiners function, and the sort and shuffle process.

Data Processing

Analyze and determine the relationship of input keys to output keys in terms of both type and number, the sorting of keys, and the sorting of values. Identify the number, type, and value of emitted keys and values from the Mappers as well as the emitted data from each Reducer and the number and contents of the output file.

Key and Value Types

Analyze and determine which of Hadoop's data types for keys and values are appropriate for a job. Understand common key and value types in the MapReduce framework and the interfaces they implement.

Common Algorithms and Design Patterns

Evaluate whether an algorithm is well-suited for expression in MapReduce. Understand implementation, limitations, and strategies for joining datasets in MapReduce. Analyze the role of DistributedCache and Counters.

The Hadoop Ecosystem

Analyze a workflow scenario and determine how and when to leverage ecosystems projects, including Apache Hive, Pig, Sqoop, and Oozie. Understand how Hadoop Streaming might apply to a job workflow.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc., 1001 Page Mill Road, Palo Alto, CA 94304, USA