cloudera®

TRAINING SHEET

> "Cloudera University is establishing a level of structure in an emerging field. Effectively framing the role helps hiring, assigning, and developing the data professionals that are going to be required by industry as big data strategies become pervasive."
>
> Samsung

Designing and Building Big Data Applications

Take Your Knowledge to the Next Level and Solve Real-World Problems with Training for Hadoop and the Enterprise Data Hub

Cloudera University's four-day course for designing and building big data applications prepares you to analyze and solve real-world problems using Apache Hadoop and associated tools in the enterprise data hub (EDH).

You will work through the entire process of designing and building solutions, including ingesting data, determining the appropriate file format for storage, processing the stored data, and presenting the results to the end-user in an easy-to-digest form. Go beyond MapReduce to use additional elements of the EDH and develop converged applications that are highly relevant to the business.

Hands-On Hadoop

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

- Creating a data set with Kite SDK
- Developing custom Flume components for data ingestion
- Managing a multi-stage workflow with Oozie
- Analyzing data with Crunch
- Writing user-defined functions for Hive and Impala
- Transforming data with Morphlines
- Indexing data with Cloudera Search

Audience and Prerequisites

This course is best suited to developers, engineers, and architects who want to use Hadoop and related tools to solve real-world problems. Participants should have already attended Cloudera Developer Training for Apache Hadoop or have equivalent practical experience. Good knowledge of Java and basic familiarity with Linux are required. Experience with SQL is helpful.

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: Designing and Building Big Data Applications

Introduction

Application Architecture
- Scenario Explanation
- Understanding the Development Environment
- Identifying and Collecting Input Data
- Selecting Tools for Data Processing and Analysis
- Presenting Results to the User

Defining and Using Data Sets
- Metadata Management
- What is Apache Avro?
- Avro Schemas
- Avro Schema Evolution
- Selecting a File Format
- Performance Considerations

Using the Kite SDK Data Module
- What is the Kite SDK?
- Fundamental Data Module Concepts
- Creating New Data Sets Using the Kite SDK
- Loading, Accessing, and Deleting a Data Set

Importing Relational Data with Apache Sqoop
- What is Apache Sqoop?
- Basic Imports
- Limiting Results
- Improving Sqoop's Performance
- Sqoop 2

Capturing Data with Apache Flume
- What is Apache Flume?
- Basic Flume Architecture
- Flume Sources
- Flume Sinks
- Flume Configuration
- Logging Application Events to Hadoop

Developing Custom Flume Components
- Flume Data Flow and Common Extension Points
- Custom Flume Sources
- Developing a Flume Pollable Source
- Developing a Flume Event-Driven Source
- Custom Flume Interceptors
- Developing a Header-Modifying Flume Interceptor
- Developing a Filtering Flume Interceptor
- Writing Avro Objects with a Custom Flume Interceptor

Managing Workflows with Apache Oozie
- The Need for Workflow Management
- What is Apache Oozie?
- Defining an Oozie Workflow
- Validation, Packaging, and Deployment
- Running and Tracking Workflows Using the CLI
- Hue UI for Oozie

Processing Data Pipelines with Apache Crunch
- What is Apache Crunch?
- Understanding the Crunch Pipeline
- Comparing Crunch to Java MapReduce
- Working with Crunch Projects
- Reading and Writing Data in Crunch
- Data Collection API
- Functions
- Utility Classes in the Crunch API

Working with Tables in Apache Hive
- What is Apache Hive?
- Accessing Hive
- Basic Query Syntax
- Creating and Populating Hive Tables
- How Hive Reads Data
- Using the RegexSerDe in Hive

Developing User-Defined Functions
- What are User-Defined Functions?
- Implementing a User-Defined Function
- Deploying Custom Libraries in Hive
- Registering a User-Defined Function in Hive

Executing Interactive Queries with Impala
- What is Impala?
- Comparing Hive to Impala
- Running Queries in Impala
- Support for User-Defined Functions
- Data and Metadata Management

Understanding Cloudera Search
- What is Cloudera Search?
- Search Architecture
- Supported Document Formats

Indexing Data with Cloudera Search
- Collection and Schema Management
- Morphlines
- Indexing Data in Batch Mode
- Indexing Data in Near Real Time

Presenting Results to Users
- Solr Query Syntax
- Building a Search UI with Hue
- Accessing Impala through JDBC
- Powering a Custom Web Application with Impala and Search

Conclusion

cloudera.com

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