
CLOUDERA SEARCH TRAINING

Take your knowledge to the next level

“I immediately began using lessons from my Cloudera class to address and revisit several real-world issues and use cases that had been problematic, and was able to quickly create working code that yielded desired results.”

Intel

Cloudera University's three-day Search training course is for developers and data engineers who want to index data in Hadoop for more powerful real-time queries. Participants will learn to get more value from their data by integrating Cloudera Search with external applications.

Learn a modern toolset

Cloudera Search brings full-text, interactive search and scalable, flexible indexing to Hadoop and an enterprise data hub. Powered by Apache Solr, Search delivers scale and reliability for a new generation of integrated, multi-workload queries.

Get hands-on experience

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

- _ Perform batch indexing of data stored in HDFS and HBase
- _ Perform indexing of streaming data in near-real-time with Flume
- _ Index content in multiple languages and file formats
- _ Process and transform incoming data with Morphlines
- _ Create a user interface for your index using Hue
- _ Integrate Cloudera Search with external applications
- _ Improve the Search experience using features such as faceting, highlighting, spelling correction

What to expect

This course is intended for developers and data engineers with at least basic familiarity with Hadoop and experience programming in a general-purpose language such as Java, C, C++, Perl, or Python. Participants should be comfortable with the Linux command line and should be able to perform basic tasks such as creating and removing directories, viewing and changing file permissions, executing scripts, and examining file output. No prior experience with Apache Solr or Cloudera Search is required, nor is any experience with HBase or SQL.

Course Outline: Cloudera Search Training

Introduction

Overview of Cloudera Search

- _What is Cloudera Search?
- _Helpful Features
- _Use Cases
- _Basic Architecture

Performing Basic Queries

- _Executing a Query in the Admin UI
- _Basic Syntax
- _Techniques for Approximate Matching
- _Controlling Output

Writing More Powerful Queries

- _Relevancy and Filters
- _Query Parsers
- _Functions
- _Geospatial Search
- _Faceting

Preparing to Index Documents

- _Overview of the Indexing Process
- _Generating Configuration Files
- _Schemaless Mode
- _Schema Design
- _Collection Management
- _Using Morphlines to Extract, Transform, and Load Data into Solr

Batch Indexing HDFS Data with MapReduce and Spark

- _Overview of the MapReduce HDFS Batch Indexing Process
- _Using the MapReduce Indexing Tool
- _Testing and Troubleshooting
- _Batch Indexing of Data in HDFS with Spark

Near-Real-Time Indexing with Flume

- _Overview of the Near-Real-Time Indexing Process
- _Introduction to Apache Flume
- _How to Perform Near-Real-Time Indexing with Flume
- _Testing and Troubleshooting

Indexing HBase Data with Lily

- _What is HBase?
- _Batch Indexing for HBase
- _Indexing HBase Tables in Near-Real-Time

Understanding Language and File Type Support

- _Field Types and Analyzer Chains
- _Word Stemming, Character Mapping, and Language Support
- _Schema and Analysis Support in the Admin UI
- _Metadata and Content Extraction with Apache Tika
- _Indexing Binary File Types with SolrCell

Improving Search Quality and Performance

- _Delivering Relevant Results
- _Helping Users Find Information
- _Query Performance and Troubleshooting

Building User Interfaces for Search

- _Search UI Overview
- _Building a User Interface with Hue
- _Integrating Search into Custom Applications

Considerations for Deployment

- _Planning for Deployment
- _Determining Hardware Needs
- _Security Overview
- _Collection Aliasing

Conclusion