cloudera®

TRAINING SHEET

> “ The professionalism and expansive technical knowledge demonstrated by our instructor were incredible. The quality of the Cloudera training was on par with a university.”
>
> General Dynamics

Data Science at Scale using Spark and Hadoop

Take your knowledge to the next level

Data scientists build information platforms to provide deep insight and answer previously unimaginable questions. Spark and Hadoop are transforming how data scientists work by allowing interactive and iterative data analysis at scale.

Learn how Spark and Hadoop enable data scientists to help companies reduce costs, increase profits, improve products, retain customers, and identify new opportunities.

Cloudera University’s three-day course helps participants understand what data scientists do, the problems they solve, and the tools and techniques they use. Through in-class simulations, participants apply data science methods to real-world challenges in different industries and, ultimately, prepare for data scientist roles in the field

Get hands-on experience

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

- How to identify potential business use cases where data science can provide impactful results
- How to obtain, clean and combine disparate data sources to create a coherent picture for analysis
- What statistical methods to leverage for data exploration that will provide critical insight into your data
- Where and when to leverage Hadoop streaming and Apache Spark for data science pipelines
- What machine learning technique to use for a particular data science project
- How to implement and manage recommenders using Spark’s MLlib, and how to set up and evaluate data experiments
- What are the pitfalls of deploying new analytics projects to production, at scale

What to expect

This course is suitable for developers, data analysts, and statisticians with basic knowledge of Apache Hadoop: HDFS, MapReduce, Hadoop Streaming, and Apache Hive as well as experience working in Linux environments. Students should have proficiency in a scripting language; Python is strongly preferred, but familiarity with Perl or Ruby is sufficient.

Get certified

Upon completion of the course, attendees are encouraged to continue their study and register for the Cloudera Certified Professional: Data Scientist 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 Introduction to Data Science

Introduction
- About This Course
- About Cloudera
- Course Logistics
- Introductions

Data Science Overview
- What Is Data Science?
- The Growing Need for Data Science
- The Role of a Data Scientist

Use Cases
- Finance
- Retail
- Advertising
- Defense and Intelligence
- Telecommunications and Utilities
- Healthcare and Pharmaceuticals

Project Lifecycle
- Steps in the Project Lifecycle
- Lab Scenario Explanation

Data Acquisition
- Where to Source Data
- Acquisition Techniques

Evaluating Input Data
- Data Formats
- Data Quantity
- Data Quality

Data Transformation
- File Format Conversion
- Joining Data Sets
- Anonymization

Data Analysis and Statistical Methods
- Relationship Between Statistics and Probability
- Descriptive Statistics
- Inferential Statistics
- Vectors and Matrices

Fundamentals of Machine Learning
- Overview
- The Three C's of Machine Learning
- Importance of Data and Algorithms
- Spotlight: Naive Bayes Classifiers

Recommender Overview
- What is a Recommender System?
- Types of Collaborative Filtering
- Limitations of Recommender Systems
- Fundamental Concepts

Introduction to Apache Spark and MLlib
- What is Apache Spark?
- Comparison to MapReduce
- Fundamentals of Apache Spark
- Spark's MLlib Package

Implementing Recommenders with MLlib
- Overview of ALS Method for Latent Factor Recommenders
- Hyperparameters for ALS Recommenders
- Building a Recommender in MLlib
- Tuning Hyperparameters
- Weighting

Experimentation and Evaluation
- Designing Effective Experiments
- Conducting an Effective Experiment
- User Interfaces for Recommenders

Production Deployment and Beyond
- Deploying to Production
- Tips and Techniques for Working at Scale
- Summarizing and Visualizing Results
- Considerations for Improvement
- Next Steps for Recommenders

Conclusion

cloudera.com

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

201510
cloudera-training-sheet-introduction-to-data-science-105