Cloudera Data Analyst Training: Using Pig, Hive, and Impala with Hadoop

Take your knowledge to the next level with Cloudera’s Apache Hadoop Training

Cloudera University’s three-day data analyst training course focusing on Apache Pig and Hive and Cloudera Impala will teach you to apply traditional data analytics and business intelligence skills to Big Data. Cloudera presents the tools data professionals need to access, manipulate, and analyze complex data sets using SQL and familiar scripting languages.

Advance Your Ecosystem Expertise

Apache Hive makes multi-structured data accessible to analysts, database administrators, and others without Java programming expertise. Apache Pig applies the fundamentals of familiar scripting languages to the Hadoop cluster. Cloudera Impala enables real-time interactive analysis of the data stored in Hadoop via a native SQL environment.

Hands-On Hadoop

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

- The fundamentals of Apache Hadoop and data ETL (extract, transform, load), ingestion, and processing with Hadoop tools
- Joining multiple data sets and analyzing disparate data with Pig
- Organizing data into tables, performing transformations, and simplifying complex queries with Hive
- Performing real-time interactive analyses on massive data sets stored in HDFS or HBase using SQL with Impala
- How to pick the best analysis tool for a given task in Hadoop

Audience & Prerequisites

This course is best suited to data analysts, business analysts, developers, and administrators who have experience with SQL and basic UNIX or Linux commands. Prior knowledge of Java and Apache Hadoop is not required.
## Course Outline: Cloudera Data Analyst Training

### Introduction
- About this Course
- About Cloudera
- Course Logistics
- Introductions

### Hadoop Fundamentals
- The Motivation for Hadoop
- Hadoop Overview
- HDFS
- MapReduce
- The Hadoop Ecosystem
- Lab Scenario Explanation
- Hands-On Exercise: Data Ingest with Hadoop Tools

### Introduction to Pig
- What Is Pig?
- Pig’s Features
- Pig Use Cases
- Interacting with Pig

### Basic Data Analysis with Pig
- Pig Latin Syntax
- Loading Data
- Simple Data Types
- Field Definitions
- Data Output
- Viewing the Schema
- Filtering and Sorting Data
- Commonly-Used Functions
- Hands-On Exercise: Using Pig for ETL Processing

### Processing Complex Data with Pig
- Storage Formats
- Complex/Nested Data Types
- Grouping
- Built-in Functions for Complex Data
- Iterating Grouped Data
- Hands-On Exercise: Analyzing Ad Campaign Data with Pig

### Multi-Dataset Operations with Pig
- Techniques for Combining Data Sets
- Joining Data Sets in Pig
- Set Operations
- Splitting Data Sets
- Hands-On Exercise: Analyzing Disparate Data Sets with Pig

### Extending Pig
- Adding Flexibility with Parameters
- Macros and Imports
- UDFs
- Contributed Functions
- Using Other Languages to Process Data with Pig
- Hands-On Exercise: Extending Pig with Streaming and UDFs

### Pig Troubleshooting and Optimization
- Troubleshooting Pig
- Logging
- Using Hadoop’s Web UI
- Optional Demo: Troubleshooting a Failed Job with the Web UI
- Data Sampling and Debugging
- Performance Overview
- Understanding the Execution Plan
- Tips for Improving the Performance of Your Pig Jobs

### Introduction to Hive
- What Is Hive?
- Hive Schema and Data Storage
- Comparing Hive to Traditional Databases
- Hive vs. Pig
- Hive Use Cases
- Interacting with Hive

### Relational Data Analysis with Hive
- Hive Databases and Tables
- Basic HiveQL Syntax
- Data Types
- Joining Data Sets
- Common Built-in Functions
- Hands-On Exercise: Running Hive Queries on the Shell, Scripts, and Hue

### Hive Data Management
- Hive Data Formats
- Creating Databases and Hive-Managed Tables
- Loading Data into Hive
- Altering Databases and Tables
- Self-Managed Tables
- Simplifying Queries with Views
- Storing Query Results
- Controlling Access to Data
- Hands-On Exercise: Data Management with Hive

### Text Processing with Hive
- Overview of Text Processing
- Important String Functions
- Using Regular Expressions in Hive
- Sentiment Analysis and N-Grams
- Hands-On Exercise (Optional): Gaining Insight with Sentiment Analysis

### Hive Optimization
- Understanding Query Performance
- Controlling Job Execution Plan
- Partitioning
- Bucketing
- Indexing Data

### Extending Hive
- SerDes
- Data Transformation with Custom Scripts
- User-Defined Functions
- Parameterized Queries
- Hands-On Exercise: Extending Hive with Streaming and UDFs

### Introduction to Impala
- What is Impala?
- How Impala Differs from Hive and Pig
- How Impala Differs from Relational Databases
- Limitations and Future Directions
- Using the Impala Shell

### Analyzing Data with Impala
- Basic Syntax
- Data Types
- Filtering, Sorting, and Limiting Results
- Joining and Grouping Data
- Improving Impala Performance
- Hands-On Exercise: Interactive Analysis with Impala

### Choosing the Best Tool for the Job
- Comparing MapReduce, Pig, Hive, Impala, and Relational Databases
- Which to Choose?

### Conclusion