TRAINING SHEET

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

> “Cloudera has not only prepared us for success today, but has also trained us to face and prevail over our big data challenges in the future by using Hadoop.”
>
> Persado

## Take your knowledge to the next level

Cloudera University’s four-day data analyst training course will teach you to apply traditional data analytics and business intelligence skills to big data tools like Apache Impala (incubating), Apache Hive, and Apache Pig. Cloudera presents the tools data professionals need to access, manipulate, transform, and analyze complex data sets using SQL and familiar scripting languages.

## Learn a modern toolset

Students will have the chance to learn and work with modern tools, such as:

- Apache Impala (incubating) enables instant interactive analysis of the data stored in Hadoop via a native SQL environment.
- Apache Hive provides a SQL-like query language with HiveQL that makes 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.

## Get hands-on experience

Through instructor-led discussion and interactive, hands-on exercises, participants will navigate the Hadoop ecosystem, learning how to:

- Acquire, store, and analyze data using features in Pig, Hive, and Impala
- Perform fundamental ETL (extract, transform, and load) tasks with Hadoop tools
- Use Pig, Hive, and Impala to improve productivity for typical analysis tasks
- Join diverse datasets to gain valuable business insight
- Perform interactive, complex queries on datasets

## What to expect

This course is designed for data analysts, business intelligence specialists, developers, system architects, and database administrators. *Prior knowledge of Apache Hadoop is not required.*

- Knowledge of SQL is assumed
- Basic familiarity with the Linux command line is expected
- Knowledge of a scripting language (such as Bash scripting, Perl, Python, or Ruby) is helpful but not essential.

## Get certified

Upon completion of the course, attendees are encouraged to continue their study and register for the CCA Data Analyst 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.

## Course details:

### Introduction

### Hadoop Fundamentals

- The Motivation for Hadoop
- Hadoop Overview
- Data Storage: HDFS
- Distributed Data Processing: YARN, MapReduce, and Spark
- Data Processing and Analysis: Pig, Hive, and Impala
- Database Integration: Sqoop
- Other Hadoop Data Tools
- Exercise Scenarios

### 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

### Processing Complex Data with Pig

- Storage Formats
- Complex/Nested Data Types
- Grouping
- Built-In Functions for Complex Data
- Iterating Grouped Data

### Multi-Dataset Operations with Pig

- Techniques for Combining Datasets
- Joining Datasets in Pig
- Set Operations
- Splitting Datasets

### Pig Troubleshooting and Optimization

- Troubleshooting Pig
- Logging
- Using Hadoop's Web UI
- Data Sampling and Debugging
- Performance Overview
- Understanding the Execution Plan
- Tips for Improving the Performance of Pig Jobs

### Introduction to Hive and Impala

- What is Hive?
- What is Impala?
- Why Use Hive and Impala?
- Schema and Data Storage
- Comparing Hive and Impala to Traditional Databases
- Use Cases

### Querying with Hive and Impala

- Databases and Tables
- Basic Hive and Impala Query Language Syntax
- Data Types
- Using Hue to Execute Queries
- Using Beeline (Hive's Shell)
- Using the Impala Shell

### Hive and Impala Data Management

- Data Storage
- Creating Databases and Tables
- Loading Data
- Altering Databases and Tables
- Simplifying Queries with Views
- Storing Query Results

### Data Storage and Performance

- Partitioning Tables
- Loading Data into Partitioned Tables
- When to Use Partitioning
- Choosing a File Format
- Using Avro and Parquet File Formats

### Relational Data Analysis with Hive and Impala

- Joining Datasets
- Common Built-In Functions
- Aggregation and Windowing

### Complex Data with Hive and Impala

- Complex Data with Hive
- Complex Data with Impala

### Analyzing Text with Hive and Impala

- Using Regular Expressions with Hive and Impala
- Processing Text Data with SerDes in Hive
- Sentiment Analysis and *n*-grams

### Hive Optimization

- Understanding Query Performance
- Bucketing
- Indexing Data
- Hive on Spark

### Impala Optimization

- How Impala Executes Queries
- Improving Impala Performance

### Extending Hive and Impala

- Custom SerDes and File Formats in Hive
- Data Transformation with Custom Scripts in Hive
- User-Defined Functions
- Parameterized Queries

### Choosing the Best Tool for the Job

- Comparing Pig, Hive, Impala, and Relational Databases
- Which to Choose?

### Conclusion