Training: Cloudera
# DANA-262 Analyzing with Cloudera Data Warehouse

CLOUDERA Training Partner

## TRAINING GOALS:

This four-day Analyzing with Data Warehouse course will teach you to apply traditional data analytics and business intelligence skills to big data. This course presents the tools data professionals need to access, manipulate, transform, and analyze complex data sets using SQL and familiar scripting languages.

What you'll learn

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

- Use Apache Hive and Apache Impala to access data through queries
- Identify distinctions between Hive and Impala, such as differences in syntax, data formats, and supported features
- Write and execute queries that use functions, aggregate functions, and subqueries
- Use joins and unions to combine datasets
- Create, modify, and delete tables, views, and databases
- Load data into tables and store query results
- Select file formats and develop partitioning schemes for better performance
- Use analytic and windowing functions to gain insight into their data
- Store and query complex or nested data structures
- Process and analyze semi-structured and unstructured data
- Optimize and extend the capabilities of Hive and Impala
- Determine whether Hive, Impala, an RDBMS, or a mix of these is the best choice for a given task
- Utilize the benefits of CDP Public Cloud Data Warehouse

What to expect

This course is designed for data analysts, business intelligence specialists, developers, system architects, and database administrators. Some knowledge of SQL is assumed, as is basic Linux command-line familiarity.



Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

## CONSPECT:

- Foundations for Big Data Analytics
  - Big Data Analytics Overview
  - Data Storage: HDFS
  - Distributed Data Processing: YARN, MapReduce, and Spark
  - Data Processing and Analysis: Hive and Impala
  - Database Integration: Sqoop
  - Other Data Tools
  - Exercise Scenario Explanation
- Introduction to Apache 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 Apache 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
- Common Operators and Built-In Functions
  - Operators
  - Scalar Functions
  - Aggregate Functions
- 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



Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

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

- Working with Multiple Datasets
  - UNION and Joins
  - Handling NULL Values in Joins
  - Advanced Joins
- Analytic Functions and Windowing
  - Using Analytic Functions
  - Other Analytic Functions
  - Sliding Windows
- Complex Data
  - Complex Data with Hive
  - Complex Data with Impala
- Analyzing Text
  - Using Regular Expressions with Hive and Impala
  - Processing Text Data with SerDes in Hive
  - Sentiment Analysis and n-grams in Hive
- Apache Hive Optimization
  - Understanding Query Performance
  - Cost-Based Optimization and Statistics
  - Bucketing
  - ORC File Optimizations
- Apache Impala Optimization
  - How Impala Executes Queries
  - Improving Impala Performance
- Extending Hive and Impala
  - User-Defined Functions
  - Parameterized Queries
- Choosing the Best Tool for the Job
  - Comparing Hive, Impala, and
  - Relational Databases
  - Which to Choose?
- CDP Public Cloud Data Warehouse



Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

- Data Warehouse Overview
- Auto-Scaling
- Managing Virtual Warehouses
- Querying Data Using CLI and Third-Party Integration

- Appendix: Apache Kudu
  - What Is Kudu?
  - Kudu Tables
  - Using Impala with Kudu

## REQUIREMENTS:

This course is designed for data analysts, business intelligence specialists, developers, system architects, and database administrators. Some knowledge of SQL is assumed, as is basic Linux command-line familiarity.

## Difficulty level



## CERTIFICATE:

The participants will obtain certificates signed by Cloudera (course completion).

Upon completion of the course, attendees are encouraged to continue their study and register for the CDP Data Analyst exam https://www.cloudera.com/about/training/certification/cdp-dataanalyst-exam-cdp-4001.html and/or CDP Data Engineer exam https://www.cloudera.com/about/training/certification/cdp-data-engineer-exam-guide-cdp-3002.html

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.

## TRAINER:

Certified Cloudera Instructor



Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007