Module Title : CLOUDERA DATA ANALYST TRAINING

Duration : 4 days

Overview

Cloudera University's four-day Data Analyst Training 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.

Advance Your Ecosystem Expertise

Apache Hive makes transformation and analysis of complex, multi-structured data scalable in Cloudera environments. Apache Impala enables real-time interactive analysis of the data stored in Hadoop using a native SQL environment. Together, they make multi-structured data accessible to analysts, database administrators, and others without Java programming expertise.

What to Expect

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

- How the open source ecosystem of big data tools addresses challenges not met by traditional RDBMSs
- Using Apache Hive and Apache Impala to provide SQL access to data
- Hive and Impala syntax and data formats, including functions and subqueries
- Create, modify, and delete tables, views, and databases; load data; and store results of queries
- Create and use partitions and different file formats
- Combining two or more datasets using JOIN or UNION, as appropriate
- What analytic and windowing functions are, and how to use them
- Store and query complex or nested data structures
- Process and analyze semi-structured and unstructured data
- Techniques for optimizing Hive and Impala queries
- Extending the capabilities of Hive and Impala using parameters, custom file formats and SerDes, and external scripts
- How to determine whether Hive, Impala, an RDBMS, or a mix of these is best for a given task

Audience & Prerequisites

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. Prior knowledge of Apache Hadoop is not required.

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 Outline

Introduction

Apache Hadoop Fundamentals

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

- 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 Common 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

Apache Hive Optimization

- Understanding Query Performance
- Bucketing
- Hive on Spark

Apache Impala Optimization

- How Impala Executes Queries
- Improving Impala Performance

Extending Apache 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 Hive, Impala, and Relational Databases
- Which to Choose?

Conclusion