Cloudera Data Analyst Training for Apache Hadoop

Duration: **4 Days**

Prerequisites:

To ensure you gain the maximum benefit from the Cloudera Data Analyst Training for Apache Hadoop course, the following are the minimum required prerequisites:

- Basic understanding of SQL: Familiarity with the SQL query language will help you grasp Hive and Impala query syntax more easily.
- Basic knowledge of Linux command line: As Hadoop runs on Linux, being comfortable with Linux commands will be beneficial for interacting with the Hadoop ecosystem.
- Fundamental understanding of databases: Knowing how traditional databases work will help you understand the motivations behind Hadoop and the use cases for Hive and Impala.
- Analytical skills: The ability to think critically and analytically will assist you in understanding data processing and analysis techniques.
- (Optional) Experience with traditional data warehousing concepts: While not mandatory, prior exposure to data warehousing can provide useful context for learning Hadoop's approach to data analysis.

While these are the minimum prerequisites, remember that the course is designed to guide you through each concept step-by-step, building your knowledge as you progress through the modules.

Course Description:

The Cloudera Data Analyst Training for Apache Hadoop is a comprehensive course designed for analysts who want to leverage the power of Hadoop to work with big data. It provides hands-on experience with tools like Hive and Impala, key components of the Hadoop ecosystem. Learners will gain insights into Hadoop's motivation, its architecture, and how to perform data processing and analysis with various Hadoop tools. By the end of the course, participants will be well-versed in querying, managing data, and optimizing performance within the Hadoop ecosystem. This knowledge is critical for earning the Cloudera Data Analyst Certification. The certification demonstrates proficiency in data analysis techniques and the use of Hadoop tools, making individuals stand out in their professional field. The Cloudera Data Analyst Training equips learners with the skills necessary to make data-driven decisions and to choose the right tool for any data analysis task.

Target Audience:

Cloudera Data Analyst Training for Apache Hadoop equips participants with essential skills for big data analytics using Hadoop tools.

- Data Analysts interested in big data and Hadoop
- Business Intelligence Professionals seeking to understand Hadoop ecosystems
- Database Administrators looking to expand into Hadoop-based systems
- Data Engineers who require proficiency in Hive and Impala
- IT Professionals aiming to specialize in big data analytics
- Software Developers who need to understand data processing on Hadoop
- System Architects planning to design big data solutions
- Technical Managers overseeing data analytics projects
- Data Scientists seeking to enhance their data processing capabilities
- Hadoop Developers and Engineers looking to deepen their expertise in Hive and Impala

Course Outlines:

Module 1: Apache 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 Scenario Explanation

Module 2: 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

Module 3: 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

Module 4: Data Management

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

Module 5: 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

Module 6: Working with Multiple Datasets

- UNION and Joins
- Handling NULL Values in Joins
- Advanced Joins

Module 7: Analytic Functions and Windowing

- Using Common Analytic Functions
- Other Analytic Functions
- Sliding Windows
- Complex Data
- Complex Data with Hive
- Complex Data with Impala

Module 8: Analyzing Text

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

Module 9: Apache Hive Optimization

- Understanding Query Performance
- Bucketing
- Hive on Spark

Module 10: Apache Impala Optimization

- How Impala Executes Queries
- Improving Impala Performance

Module 11: Extending Apache Hive and Impala

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

Module 12: Choosing the Best Tool for the Job

- Comparing Hive, Impala, and Relational Database