Cloudera Data Analyst

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 Databases