Course duration

- 4 days

Course Benefits

- How the open source ecosystem of big data tools addresses challenges not met by traditional RDBMSs
- How Apache Hive and Apache Impala are used to provide SQL access to data
- How Hive and Impala syntax and data formats, including functions and subqueries, help answer questions about data
- How to create, modify, and delete tables, views, and databases; load data; and store results of queries
- How to create and use partitions and different file formats
- How to combine two or more datasets using JOIN or UNION, as appropriate
- What analytic and windowing functions are, and how to use them
- How to store and query complex or nested data structures
- How to process and analyze semi-structured and unstructured data
- Different techniques for optimizing Hive and Impala queries
- How to extend 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

Course Outline

1. Apache Hadoop Fundamentals
    1. The Motivation for Hadoop
    2. Hadoop Overview
    3. Data Storage: HDFS
    4. Distributed Data Processing: YARN, MapReduce, and Spark
    5. Data Processing and Analysis: Hive and Impala
    6. Database Integration: Sqoop
    7. Other Hadoop Data Tools
    8. Exercise Scenario Explanation
2. Introduction to Apache Hive and Impala
    1. What Is Hive?
    2. What Is Impala?
    3. Why Use Hive and Impala?
    4. Schema and Data Storage
    5. Comparing Hive and Impala to Traditional Databases
    6. Use Cases

3. Querying with Apache Hive and Impala
    1. Databases and Tables
    2. Basic Hive and Impala Query Language Syntax
    3. Data Types
    4. Using Hue to Execute Queries
    5. Using Beeline (Hive's Shell)
    6. Using the Impala Shell
4. Common Operators and Built-In Functions
    1. Operators
    2. Scalar Functions
    3. Aggregate Functions
5. Data Management
    1. Data Storage
    2. Creating Databases and Tables
    3. Loading Data
    4. Altering Databases and Tables
    5. Simplifying Queries with Views
    6. Storing Query Results
6. Data Storage and Performance
    1. Partitioning Tables
    2. Loading Data into Partitioned Tables
    3. When to Use Partitioning
    4. Choosing a File Format
    5. Using Avro and Parquet File Formats
7. Working with Multiple Datasets
    1. UNION and Joins
    2. Handling NULL Values in Joins
    3. Advanced Joins
8. Analytic Functions and Windowing
    1. Using Analytic Functions
    2. Other Analytic Functions
    3. Sliding Windows
9. Complex Data
    1. Complex Data with Hive
    2. Complex Data with Impala
10. Analyzing Text
    1. Using Regular Expressions with Hive and Impala
    2. Processing Text Data with SerDes in Hive
    3. Sentiment Analysis and n-grams in Hive
11. Apache Hive Optimization
    1. Understanding Query Performance
    2. Bucketing
    3. Hive on Spark
    4. Apache Impala Optimization
    5. How Impala Executes Queries
    6. Improving Impala Performance
12. Extending Apache Hive and Impala

    1. Custom SerDes and File Formats in Hive
    2. Data Transformation with Custom Scripts in Hive
    3. User-Defined Functions
    4. Parameterized Queries
13. Choosing the Best Tool for the Job
    1. Comparing Hive, Impala, and Relational Databases
    2. Which to Choose?
14. Conclusion
    1. Apache Kudu
    2. What Is Kudu?
    3. Kudu Tables
    4. Using Impala with Kudu

Class Materials

Each student will receive a comprehensive set of materials, including course notes and all the class examples.

Class Prerequisites

Experience in the following *is required* for this Hadoop class:

- Some knowledge of SQL.
- Basic Linux command-line familiarity. .