Big-Data and Hadoop Analyst
Analyst Training for Pig, Hive and Impala

COURSE BROCHURE & SYLLABUS

Course Overview

Apache Hive makes multi-structured data accessible to analysts, database administrators, and others without Java programming expertise. **Apache Pig** applies the fundamentals of familiar scripting languages to the Hadoop cluster. **Impala** enables real-time, interactive analysis of the data stored in Hadoop via a native SQL environment.

This data analyst training course focusing on Apache Pig, Hive and Impala 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.

Audience Profile

This course is designed for data analysts, business intelligence specialists, developers, system architects, and database administrators.

Knowledge of SQL is assumed, as is basic Linux command-line familiarity. Knowledge of at least one scripting language (e.g., Bash scripting, Perl, Python, Ruby) would be helpful but is not essential.

At Course Completion

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

- The features that Pig, Hive, and Impala offer for data acquisition, storage, and analysis
- The fundamentals of Apache Hadoop and data ETL (extract, transform, load), ingestion, and processing with Hadoop tools
- How Pig, Hive, and Impala improve productivity for typical analysis tasks
- Joining diverse datasets to gain valuable business insight
- Performing real-time, complex queries on datasets

Analyst Training for Pig, Hive and Impala

Course Duration: 30 Hours

Course Outline

- **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
  - Data Integration: Sqoop
  - Other Hadoop Data Tools
- **Introduction to Pig**
  - What Is Pig?
  - Pig's Features
  - Pig Use Cases
  - Interacting with Pig
- **Basic Data Analysis with Pig**
  - Pig Latin Syntax
  - Loading Data
  - Simple Data Types
  - Field Definitions
  - Data Output
  - Viewing the Schema
  - Filtering and Sorting Data
  - Commonly-Used Functions
- **Processing Complex Data with Pig**
  - Storage Formats
  - Complex/Nested Data Types
  - Grouping
  - Built-In Functions for Complex Data
  - Iterating Grouped Data
- **Multi-Dataset Operations with Pig**
  - Techniques for Combining Data Sets
  - Joining Data Sets in Pig
  - Set Operations
  - Splitting Data Sets

- **Pig Troubleshooting and Optimization**
  - Troubleshooting Pig
  - Logging
  - Using Hadoop's Web UI
  - Data Sampling and Debugging
  - Performance Overview
  - Understanding the Execution Plan
  - Tips for Improving the Performance
- **Introduction to Hive and Impala**
  - What Is Hive?
  - What Is Impala?
  - Schema and Data Storage
  - Comparing Hive to Traditional Databases
- **Querying with Hive and Impala**
  - Databases and Tables
  - Basic Hive and Impala Query Language Syntax
  - Data Types
  - Differences Between Hive and Impala Query Syntax
  - Using Hue to Execute Queries
  - Using the Impala Shell
- **Data Management**
  - Data Storage
  - Creating an altering Databases and Tables
  - Loading Data
  - Simplifying Queries with Views
  - Storing Query Results
- **Data Storage and Performance**
  - Partitioning Tables
  - Choosing a File Format
  - Managing Metadata
  - Controlling Access to Data
- **Relational Data Analysis with Hive and Impala**
  - Joining Datasets
  - Common Built-In Functions
  - Aggregation and Windowing
- **Working with Impala**
  - How Impala Executes Queries
  - Extending Impala with User-Defined
  - Functions
  - Improving Impala Performance

- **Analysing Text and Complex Data with Hive**
  - Complex Values in Hive
  - Using Regular Expressions in Hive
  - Sentiment Analysis and N-Grams
- **Hive Optimization**
  - Understanding Query Performance
  - Controlling Job Execution Plan
  - Bucketing
  - Indexing Data
- **Extending Hive**
  - SerDes
  - Data Transformation with Custom Scripts
  - User-Defined Functions
  - Parameterized Queries
- **Choosing the Best Tool**
  - Comparing MapReduce, Pig, Hive, Impala, and Relational Databases