Cloudera Data Analyst Training: Using Pig, Hive and Impala with Hadoop

Course Summary

Description

Cloudera University's four-day data analyst training course focusing on Apache Pig and Hive and Cloudera Impala will teach you to apply traditional data analytics and business intelligence skills to big data. Cloudera presents the tools data professionals need to access, manipulate, transform and analyze complex data sets using SQL and familiar scripting languages.

Objectives

By the end of this course, students will learn such topics 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
- 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

Topics

- Hadoop Fundamentals
- Basic Data Analysis with Pig
- Processing Complex Data with Pig
- Multi-Dataset Operations with Pig
- Pig Troubleshooting and Optimization
- Introduction to Hive and Impala
- Querying with Hive and Impala
- Data Management
- Data Storage and Performance
- Relational Data Analysis with Hive and Impala
- Working with Impala
- Analyzing Text and Complex Data with Hive
- Hive Optimization
- Extending Hive
- Choosing the Best Tool for the Job

Prerequisites

- 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.
- Prior knowledge of Apache Hadoop is not required.

Duration

Four days

ProTech *"Charting the Course ...*

protechtraining.com

... to Your Success!"

Cloudera Data Analyst Training: Using Pig, Hive and Impala with Hadoop

Course Outline

I. Hadoop Fundamentals
- A. The Motivation for Hadoop
- B. Hadoop Overview
- C. Data Storage: HDFS
- D. Distributed Data Processing: YARN, MapReduce and Spark
- E. Data Processing and Analysis: Pig, Hive and Impala
- F. Data Integration: Sqoop
- G. Other Hadoop Data Tools
- H. Exercise Scenarios Explanation

II. Introduction to Pig
- A. What Is Pig?
- B. Pig's Features
- C. Pig Use Cases
- D. Interacting with Pig

III. Basic Data Analysis with Pig
- A. Pig Latin Syntax
- B. Loading Data
- C. Simple Data Types
- D. Field Definitions
- E. Data Output
- F. Viewing the Schema
- G. Filtering and Sorting Data
- H. Commonly-Used Functions

IV. Processing Complex Data with Pig
- A. Storage Formats
- B. Complex/Nested Data Types
- C. Grouping
- D. Built-In Functions for Complex Data
- E. Iterating Grouped Data

V. Multi-Dataset Operations with Pig
- A. Techniques for Combining Data Sets
- B. Joining Data Sets in Pig
- C. Set Operations
- D. Splitting Data Sets

VI. Pig Troubleshooting and Optimization
- A. Troubleshooting Pig
- B. Logging
- C. Using Hadoop's Web UI
- D. Data Sampling and Debugging
- E. Performance Overview
- F. Understanding the Execution Plan
- G. Tips for Improving the Performance of Your Pig Jobs

VII. Introduction to Hive and Impala
- A. What Is Hive?
- B. What Is Impala?
- C. Schema and Data Storage
- D. Comparing Hive to Traditional Databases
- E. Hive Use Cases

VIII. Querying with Hive and Impala
- A. Databases and Tables
- B. Basic Hive and Impala Query Language Syntax
- C. Data Types
- D. Differences Between Hive and Impala Query Syntax
- E. Using Hue to Execute Queries
- F. Using the Impala Shell

IX. Data Management
- A. Data Storage
- B. Creating Databases and Tables
- C. Loading Data
- D. Altering Databases and Tables
- E. Simplifying Queries with Views
- F. Storing Query Results

X. Data Storage and Performance
- A. Partitioning Tables
- B. Choosing a File Format
- C. Managing Metadata
- D. Controlling Access to Data

Cloudera Data Analyst Training: Using Pig, Hive and Impala with Hadoop

Course Outline (con't)

XI. Relational Data Analysis with Hive and Impala
- A. Joining Datasets
- B. Common Built-In Functions
- C. Aggregation and Windowing

XII. Working with Impala
- A. How Impala Executes Queries
- B. Extending Impala with User-Defined Functions
- C. Improving Impala Performance

XIII. Analyzing Text and Complex Data with Hive
- A. Complex Values in Hive
- B. Using Regular Expressions in Hive
- C. Sentiment Analysis and N-Grams
- D. Conclusion

XIV. Hive Optimization
- A. Understanding Query Performance
- B. Controlling Job Execution Plan
- C. Bucketing
- D. Indexing Data

XV. Extending Hive
- A. SerDes
- B. Data Transformation with Custom Scripts
- C. User-Defined Functions
- D. Parameterized Queries

XVI. Choosing the Best Tool for the Job
- A. Comparing MapReduce, Pig, Hive, Impala and Relational Databases
- B. Which to Choose?