CLOUDERA DATA ANALYST TRAINING(DATA-ANALYST-OD)

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
- Exercise Scenarios Explanation

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 of Your Pig Jobs

Introduction to Hive and Impala

- What Is Hive?
- What Is Impala?
- Schema and Data Storage
- Comparing Hive to Traditional Databases
- Hive Use Cases

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 Databases and Tables
- Loading Data
- Altering Databases and Tables
- 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

Analyzing Text and Complex Data with Hive

- Complex Values in Hive
- Using Regular Expressions in Hive
- Sentiment Analysis and N-Grams
- Conclusion

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 for the Job

- Comparing MapReduce, Pig, Hive, Impala, and
- Relational Databases
- Which to Choose?