CLOUDERA APACHE HADOOP DEVELOPER COURSE CONTENT

- **THE MOTIVATION FOR HADOOP:**
  - Problems with Traditional Large-Scale Systems
  - Requirements for a New Approach
  - Introducing Hadoop

- **HADOOP: BASIC CONCEPTS:**
  - The Hadoop Project and Hadoop Components
  - The Hadoop Distributed File System
  - Hands-On Exercise: Using HDFS
  - How MapReduce Works
  - Hands-On Exercise: Running a MapReduce Job
  - How a Hadoop Cluster Operates
  - Other Hadoop Ecosystem Projects

- **WRITING A MAPREDUCE PROGRAM:**
  - The MapReduce Flow
  - Basic MapReduce API Concepts
  - Writing MapReduce Drivers, Mappers and Reducers in Java
  - Writing Mappers and Reducers in Other Languages Using the Streaming API
  - Speeding Up Hadoop Development by Using Eclipse
  - Hands-On Exercise: Writing a MapReduce Program
  - Differences Between the Old and New MapReduce APIs

- **UNIT TESTING MAPREDUCE PROGRAMS:**
  - Unit Testing
  - The JUnit and MRUnit Testing Frameworks
  - Writing Unit Tests with MRUnit
  - Hands-On Exercise: Writing Unit Tests with the MRUnit Framework

- **DELVING DEEPER INTO THE HADOOP API:**
  - Using the ToolRunner Class
  - Hands-On Exercise: Writing and Implementing a Combiner
  - Setting Up and Tearing Down Mappers and Reducers by Using the Configure and Close Methods
  - Writing Custom Partitioners for Better Load Balancing
  - Optional Hands-On Exercise: Writing a Partitioner
  - Accessing HDFS Programmatically
  - Using The Distributed Cache
  - Using the Hadoop API's Library of Mappers, Reducers and Partitioners

- **PRACTICAL DEVELOPMENT TIPS AND TECHNIQUES:**
  - Strategies for Debugging Map Reduce Code
  - Testing Map Reduce Code Locally by Using Local Job Reducer
  - Writing and Viewing Log Files
  - Retrieving Job Information with Counters
  - Determining the Optimal Number of Reducers for a Job
  - Creating Map-Only Map Reduce Jobs
  - Hands-On Exercise: Using Counters and a Map-Only Job

- **DATA INPUT AND OUTPUT:**
  - Creating Custom Writable and Writable Comparable Implementations
  - Saving Binary Data Using Sequence File and Avro Data Files
  - Implementing Custom Input Formats and Output Formats
  - Issues to Consider When Using File Compression
  - Hands-On Exercise: Using Sequence Files and File Compression

- **COMMON MAPREDUCE ALGORITHMS:**
  - Sorting and Searching Large Data Sets
  - Performing a Secondary Sort
  - Indexing Data
  - Hands-On Exercise: Creating an Inverted Index
  - Computing Term Frequency — Inverse Document Frequency
  - Calculating Word Co-Occurrence
  - Hands-On Exercise: Calculating Word Co-Occurrence (Optional)
  - Hands-On Exercise: Implementing Word Co-Occurrence with a Customer Writable Comparable (Optional)

- **JOINING DATA SETS IN MAPREDUCE JOBS:**
  - Writing a Map-Side Join
  - Writing a Reduce-Side Join

- **INTEGRATING HADOOP INTO THE ENTERPRISE WORKFLOW:**
  - Integrating Hadoop into an Existing Enterprise
  - Loading Data from an RDBMS into HDFS by Using Sqoop
  - Hands-On Exercise: Importing Data with Sqoop
  - Managing Real-Time Data Using Flume
  - Accessing HDFS from Legacy Systems with FuseDFS and HttpFS

- **MACHINE LEARNING AND MAHOUT:**
  - Introduction to Machine Learning
  - Using Mahout
  - Hands-On Exercise: Using a Mahout Recommender

- **AN INTRODUCTION TO HIVE AND PIG:**
  - The Motivation for Hive and Pig
  - Hive Basics
  - Hands-On Exercise: Manipulating Data with Hive
  - Pig Basics
  - Hands-On Exercise: Using Pig to Retrieve Movie Names from Our Recommender
  - Choosing Between Hive and Pig

- **AN INTRODUCTION TO OOZIE:**
  - Introduction to Oozie
  - Creating Oozie Workflows
  - Hands-On Exercise: Running an Oozie Workflow

- **AN INTRODUCTION TO HBASE:**
  - Introduction to Hbase.
  - Understanding the Column based Database.
  - Practicing Hbase commands. Introduction about Zookeeper.