Introduction to Cloudera

Course ID: ITG-BIG-200

Introduction to Cloudera ITG-BIG-200

Course Overview:

This 5 day course introduces the student to the Hadoop architecture, file system, and the Hadoop Ecosystem. This course will cover the basic concepts of processing unstructured files, as well as the Cloudera Ecosystem.

Prerequisites:

Basic Linux Knowledge and some knowledge of traditional database systems.

Who Should Attend this course?

Those developers who have an interest in learning the Cloudera framework.

Topics:

• Introduction to Hadoop	• HDFS
• MapReduce	• Clusters
• Developing a MapReduce App	• MapReduce APIs
• Unit testing of MapReduce App	• Hadoop API
• Development Techniques	• Partitioners & Reducers
• Data inputs & outputs	• Output formats
• Document Frequency	• Joining datasets in MapReduce tasks
• Hadoop Integration	• How to use Sqoop
• FuseDSF & HttpFS	• Oozie

DAY 01 MODULES: ITG-BIG-200

Module 01: Introduction to Hadoop

- Overview
- Hadoop & Conventional large-scale systems
- Introduction to Hadoop
- Hadoopable Issues

Module 02: HDFS

- Overview
- HDFS Basic Concepts
- The Hadoop Project & Components
- Understanding the Distributed File System

Module 03: MapReduce

- MapReduce Overview
- WordCount
- Using Mappers
- Using Reducers

Module 04: Clusters

- Clusters Overview
- Hadoop Tasks
- Hadoop Jobs
- Other elements of the Hadoop ecosystem

DAY 02 MODULES: ITG-BIG-200

Module 05: Developing a MapReduce App Using Java

- MapReduce API
- MapReduce Drivers, mappers, reducers
- Using Eclipse to make Hadoop development faster
- Old and New: Primary differences

Module 06: MapReduce APIs

- Overview
- Developing a MapReduce App with Streaming
- Building Mappers & Reducers
- Streaming APIs

Module 07: Unit Testing of MapReduce Apps

- Testing Units
- Working with JUnit and MRUnit testing frameworks
- Creating unit tests using MRUnit
- How to run unit tests

Module 08: Hadoop API

- Toolrunner
- How to set up and tear down mappers & reducers
- Less intermediate data via combiners
- How to access HDFS via Software
- Working with Distributed Cache
- Library of mappers, reducers, partitioners

Module 09: Development Techniques

- Debugging Map Reduce Code
- Local Job Runner
- Log Files: writing & reading
- Obtaining Job details using counters
- Reusing Objects
- Creating map-only MapReduce jobs

Module 10: Partitioners & Reducers

- Overview
- Partitioners & Reducers working together
- How to determine the best number of reducers
- Creating customer partitioners

Module 11: Data Inputs & Outputs

- Creating custom writable implementations
- Saving binary data with Sequence File and Avro data files
- Things to remember when using file compression
- Implementing custom Input Formats

Module 12: Output Formats & Common MapReduce Algorithms

- Overview
- Sorting & Searching big data sets
- Indexing data
- Calculating term frequency

Module 13: Document Frequency

- Overview
- How to calculate word co-occurrence
- Secondary sorting

Module 14: How to join Data Sets in MapReduce Tasks

- Map-Side Joins
- Reduce-Side Joins

Module 15: Integrating Hadoop into your Enterprise Workflow

- Integrating Hadoop into your Enterprise Workflow
- Loading data from RDBMS into HDFS

Module 16: How to use Sqoop

- Management of real data with Flume
- How to access HDFS from legacy systems

DAY 05 MODULES: ITG-BIG-200

Module 17: FuseDFS & HttpFS

- Hive, Impala, and Pig
- Motivation
- Overview of Hive
- Overview of Impala
- Overview of Pig
- B Hive, Impala, and Pig: Making a Choice

Module 18: Oozie

- Oozie Overview
- Creating Oozie Workflows