Updated 07/27/2023

Sign up

Hadoop Training: Development

3 days (21 hours)

Presentation

Hadoop is an open source framework developed by Google for data storage and application execution Hadoop applies MapReduce (MR) algorithms in which data is processed in parallel with other datasets. It offers massive storage for all types of data, enormous processing power and the ability to handle an unlimited number of simultaneous tasks or jobs. Our Hadoop: Development course will teach you the techniques for processing large volumes of data. During this course, you'll learn about the Hadoop ecosystem, the principles of the framework and how to develop parallel algorithms with MapReduce. You'll also see how to use Hadoop tasks to extract relevant elements from the dataset and load unstructured data from HBase and HDFS. At the end of this course, you'll be able to develop applications compatible with the Apache Hadoop platform for processing Big Data. As with all our training courses,
it will introduce you to the latest version of Apache Hadoop 3.3.

Objectives

- Master the Cloudera/Hortonwork Hadoop ecosystem
- Implement Hadoop framework functionalities
- Extract relevant elements from large and varied data sets using Hadoop tasks
- Developing efficient parallel algorithms with MapReduce
- Load unstructured data from HDFS and HBase systems

Target audience

- Developers
- Project managers
- Data scientists
- Architects

Prerequisites

Knowledge of an object programming language such as Java and of scripting.

Our Hadoop training program: Development

Introduction to the Hadoop framework

- Installing Hadoop
- Objective of the Hadoop project
- Basic principles of the framework
- Essential features
- Applications in various fields
- Coudera and Hortonworks platform

MapReduce

- MapReduce implementation using the Hadoop framework
- MapReduce programming principles
- Map() and Reduce() functions
- Using MapReduce technologies
- Developing efficient parallel algorithms
- Create, customize and deploy tasks
- Synthesizing data with MapReduce
- Best practices for developing MapReduce applications

The Hadoop ecosystem

- Ecosystem overview
- Hadoop features at a glance
- Hadoop architecture
  - HDFS MapReduce FIL
- Name node
- Data node
- Secondary name node
- Blocks
- The difference between RDBMS and Hadoop

Hadoop YARN

- Using MapReduce through Yarn
- Using a cluster
- Cloud cluster management
- Different applications on the same cluster

- YARN components

Relational database with Hadoop

- What is Hive
- Basic syntax
- Integrating MySQL with Hadoop
- Simplify queries
- HiveQL extension
- User-Defined-Functions (UDF)
- Using Sqoop to import data from MySQL to HFDS/Hive
- Using Sqoop to export data from Hadoop to MySQL

Programming Hadoop with Pig

- Definition and use
- Best practices map/reduce
- Java development and integration
- Extension with UDF

Hadoop with Spark

- Why choose Spark?
- Spark architecture
- Essential components
- Resilient distributed datasets (RDD)
    - Operations
    - Persistence
    - Shared Variables
- Integrated functions

Data storage on HDFS

- Hadoop Distributed File System
- Loading unstructured data from HDFS
- XML data types
- Parallelize calculations on large volumes of data
- Distributed mode operation

Data storage with HBase

- Load unstructured data from HBase
- HBase cluster operation
- Independent operation
  - HRegionServer
  - HMaster
  - ZooKeeper
- Security mechanisms in Hadoop
- Authentication management

Hadoop Streaming

- Configuring Hadoop
- Defining MapReduce to Streaming
- Python language with Hadoop Streaming
- Creating a MapReduce job in Python
- Followed by a MapReduce streaming job

Further information

Companies concerned

This course is aimed at both individuals and companies, large or small, wishing to train their teams in a new advanced computer technology, or to acquire specific business knowledge or modern methods.

Teaching methods

Practical course: 60% Practical, 40% Theory. Training material distributed in digital format to all participants.

Organization

The course alternates theoretical input from the trainer, supported by examples, with brainstorming sessions and group work.

Validation

At the end of the session, a multiple-choice questionnaire verifies the correct acquisition of skills.

Sanction

A certificate will be issued to each trainee who completes the course.