Publication date : 09/26/2025

Course : Hadoop, installation and administration

Practical course - 4d - 28h00 - Ref. HOD
Price : 2960 CHF E.T.

★★★★★ 5 / 5

The Apache Hadoop platform was the first solution to really enable (distributed) processing of huge quantities of data. This course will show you how to install, configure and administer a Hadoop cluster and other components of the ecosystem (Hive, Pig, HBase, Flume...).

Teaching objectives

At the end of the training, the participant will be able to:

- Discover Hadoop concepts and challenges
- Understand how the platform and its components work
- Install and manage the platform
- Optimizing the platform

Intended audience

Hadoop cluster administrators, developers.

Prerequisites

Good knowledge of Linux administration. Experience required.

Practical details

Hands-on work

Hadoop cluster installation and configuration.

Teaching methods

Magisterial" teaching method with practical exercises after each notion or group of notions.

Course schedule

PARTICIPANTS

Hadoop cluster administrators, developers.

PREREQUISITES

Good knowledge of Linux administration. Experience required.

TRAINER QUALIFICATIONS

The experts leading the training are specialists in the covered subjects. They have been approved by our instructional teams for both their professional knowledge and their teaching ability, for each course they teach. They have at least five to ten years of experience in their field and hold (or have held) decision-making positions in companies.

ASSESSMENT TERMS

The trainer evaluates each participant's academic progress throughout the training using multiple choice, scenarios, hands-on work and more.
Participants also complete a placement test before and after the course to measure the skills they've developed.

1 Introduction to the Apache Hadoop framework

- The challenges of big data and the contributions of the Hadoop framework.
- Introduction to the Hadoop architecture.
- Description of the main components of the Hadoop platform.
- Overview of the main on-premise and on-cloud distributions, and the hybrid approach.
- Advantages/disadvantages of the platform versus alternative solutions.
- Overview and comparison of native and complementary components (Storm, Flink, Spark, etc.).

2 Preparing and configuring the Hadoop cluster

- Hadoop Distributed File System (HDFS) operating principles.
- MapReduce operating principles.
- Design "type" of the cluster.
- Equipment selection criteria.

Hands-on work

Hadoop cluster configuration.

3 Installing a Hadoop platform

- Type of deployment.
- Hadoop installation.
- Installation of other components (Hive, Pig, HBase, Nifi...).
- Overview and comparison of historical (HDP, HDF, CDH) and current (CDP/CDSW...) software stacks.
- Kappa, Lambda, SMACK architectures (Spark, Mesos, Akka, Cassandra, Kafka).

Hands-on work

Installation of a Hadoop platform and its main components.

4 Managing a Hadoop cluster

- Manage Hadoop cluster nodes.
- MapReduce V2 (Yarn, Resource Manager, Application Master, Node Manager, etc.).
- Resource managers (Yarn versus Mesos).
- Task management via schedulers.
- Log management.
- Process scheduling (Oozie).
- Use a manager.

Hands-on work

List jobs, queue status, job status, task management, WebUI access.

TEACHING AIDS AND TECHNICAL RESOURCES

• The main teaching aids and instructional methods used in the training are audiovisual aids, documentation and course material, hands-on application exercises and corrected exercises for practical training courses, case studies and coverage of real cases for training seminars.
• At the end of each course or seminar, ORSYS provides participants with a course evaluation questionnaire that is analysed by our instructional teams.
• A check-in sheet for each half-day of attendance is provided at the end of the training, along with a course completion certificate if the trainee attended the entire session.

TERMS AND DEADLINES

Registration must be completed 24 hours before the start of the training.

ACCESSIBILITY FOR PEOPLE WITH DISABILITIES

Do you need special accessibility accommodations? Contact Mrs. Fosse, Disability Manager, at psh-accueil@orsys.fr to review your request and its feasibility.

5 Data management in HDFS

- Import external data (files, relational databases) into HDFS.
- Handling HDFS files.
- File formats (SequenceFile, ORC, Parquet...), their uses and relationship to performance.
- Database storage (structured or unstructured): NoSQL Hbase, SQL with Impala, Hive, Hive LLAP...

Hands-on work
Import external data with Flume or Nifi, import data from relational databases with Sqoop.

6 Advanced configuration

- Authorizations and security: administration, authentication, authorizations, auditing, data protection.
- Components involved in security: Ranger, Knox, Kerberos, KMS...
- NameNode high availability (MRV2/YARN).

Hands-on work
Configuration of service-level authentication (SLA) and Access Control List (ACL).

7 Monitoring and optimization/tuning

- Monitoring (Ambari, Cloudera Manager, Ganglia...).
- Benchmark types (DFSIO, Teragen/TeraSort/TeraValidate) and results available online (TPCx-HS...)
- Comparing MapReduce and TEZ.
- Examples of optimization and optimization tools (CDP advisor...).
- Choice of block size.
- Other tuning options (use of compression, memory configuration, etc.).

Hands-on work
Parameterize, launch and analyze Bench, understand cluster monitoring and optimization commands.

8 The benefits of Hadoop v3

- Object storage approaches (Ozone).
- Erasure coding.
- Yarn Federation.
- Migration scenarios, aspects to consider, and a few examples (Hortonworks to Cloudera...).

Dates and locations

REMOTE CLASS
2026 : 23 June, 22 Sep.