

Hadoop, installation et administration

**SII-315** **4 Days (28 Hours)**

Description

The Apache Hadoop platform makes it easier to create distributed applications. This internship will allow you to understand its architecture and give you the knowledge necessary to install, configure and administer a Hadoop cluster. You will also learn how to optimize it and maintain it over time.

Who is this training for ?

For whom

Hadoop cluster administrators, developers.

Prerequisites

Aucune

Training objectives

- Discover the concepts and issues related to Hadoop
- Optimize the platform
- Understand how the platform and its components work
- Install the platform and manage it

Training program

Présentation du Framework Apache Hadoop

- Big Data challenges and contributions of the Hadoop framework.
- Presentation of the Hadoop architecture.
- Description of the main components of the Hadoop platform.
- Presentation of the main market distributions and complementary tools (Cloudera, MapR, Dataiku).
-).
- Advantages/disadvantages of the platform.

Préparations et configuration du cluster Hadoop

- Hadoop Distributed File System (HDFS) working principles.
- MapReduce working principles.
- Cluster "type" design.
- Hardware selection criteria.
- Practical work Configuration of the Hadoop cluster.

Installation d'une plateforme Hadoop

- Deployment type.
- Installation of Hadoop.
- Installation of other components (Hive, Pig, HBase, Flume.
-).
- Practical work Installation of a Hadoop platform and main components.

Gestion d'un cluster Hadoop

- Management of Hadoop cluster nodes.
- TaskTracker, JobTracker for MapReduce.
- Management of tasks via schedulers.
- Management of logs.
- Using a manager.
- Practical work List jobs, queue status, job status, task management, access to the web UI.

Gestion des données dans HDFS

- Import of external data (files, relational databases) to HDFS.
- Handling of HDFS files.
- Practical work Import external data with Flume, consult relational databases with Sqoop.

Configuration avancée

- Authorization and security management.
- Recovery from name node failure (MRV1).
- NameNode high availability (MRV2/YARN).
- Practical work Configuration of a service-level authentication (SLA) and an Access Control List (ACL).

Monitoring et optimisation Tuning

- Monitoring (Ambari, Ganglia.
-).
- Benchmarking/profiling of a cluster.
- Apache GridMix tools, Vaaidya .
- Choose block size.
- Other tuning options (use of compression, memory configuration.
-).
- Practical work Understand cluster monitoring and optimization commands as they come.