Hadoop, installation et administration

Cours Pratique de 4 jours - 28h
Réf : HOD - Prix 2024 : 2 860€ HT

La plateforme Apache Hadoop est la première solution a avoir réellement rendu possibles des traitements (distribués) sur d'énormes quantités de données. Ce cours vous montrera comment installer, configurer et administrer un cluster Hadoop ainsi que d'autres composants de l'écosystème (Hive, Pig, HBase, Flume...).

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Découvrir les concepts et les enjeux liés à Hadoop

Comprendre le fonctionnement de la plateforme et de ses composants

Installer la plateforme et la gérer

Optimiser la plateforme

MÉTHODES PÉDAGOGIQUES

Méthode pédagogique de type "magistral" avec des exercices pratiques à l'appui, après chaque notion ou groupe de notions exposées.

TRAVAUX PRATIQUES

Installation du cluster Hadoop et paramétrage.

LE PROGRAMME

dernière mise à jour : 01/2022

1) Présentation du framework Apache Hadoop

- Enjeux du Big Data et apports du framework Hadoop.
- Présentation de l'architecture Hadoop.
- Description des principaux composants de la plateforme Hadoop.
- Présentation des distributions principales du marché on-premise et on-Cloud, et l'approche hybride.
- Avantages/inconvénients de la plateforme vs les solutions alternatives.
- Synthèse des différents composants natifs, complémentaires, et comparatif (Storm, Flink, Spark…).

2) Préparations et configuration du cluster Hadoop

- Principes de fonctionnement de Hadoop Distributed File System (HDFS).
- Principes de fonctionnement de MapReduce.
- Design "type" du cluster.
- Critères de choix du matériel.

Travaux pratiques : Configuration du cluster Hadoop.

3) Installation d'une plateforme Hadoop

- Type de déploiement.
- Installation d'Hadoop.
- Installation d'autres composants (Hive, Pig, HBase, Nifi...).
- Présentation et comparatif des piles logicielles historiques (HDP, HDF, CDH) et actuelles (CDP/CDSW…).
- Architectures Kappa, Lambda, SMACK (Spark, Mesos, Akka, Cassandra, Kafka).

Travaux pratiques : Installation d'une plateforme Hadoop et des composants principaux.

PARTICIPANTS

Administrateurs de cluster Hadoop, développeurs.

PRÉREQUIS

Bonnes connaissances de l'administration Linux. Expérience requise.

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

4) Gestion d'un cluster Hadoop

- Gestion des nœuds du cluster Hadoop.
- MapReduce V2 (Yarn, Resource Manager, Application Master, Node Manager…).
- Gestionnaires de ressources (Yarn vs Mesos).
- Gestion des tâches via les schedulers.
- Gestion des logs.
- Ordonnancement des traitements (Oozie...).
- Utiliser un manager.

Travaux pratiques : Lister les jobs, statut des queues, statut des jobs, gestion des tâches, accès à la Web UI.

5) Gestion des données dans HDFS

- Import de données externes (fichiers, bases de données relationnelles) vers HDFS.
- Manipulation des fichiers HDFS.
- Les formats de fichiers (SequenceFile, ORC, Parquet…), leurs usages et leurs relations avec les performances.
- Le stockage sous forme de base de données (structurée ou non) : NoSQL Hbase, SQL avec Impala, Hive, Hive LLAP…

Travaux pratiques : Importer des données externes avec Flume ou Nifi, importer des données des bases de données relationnelles avec Sqoop.

6) Configuration avancée

- Autorisations et sécurité : administration, authentification, autorisations, audit, protection des données.
- Les composants impliqués dans la sécurité : Ranger, Knox, Kerberos, KMS...
- NameNode high availability (MRV2/YARN).

Travaux pratiques : Configuration d'un service-level authentication (SLA) et d'un Access Control List (ACL).

7) Monitoring et optimisation/Tuning

- Monitoring (Ambari, Cloudera Manager, Ganglia...).
- Les types de benchmark (DFSIO, Teragen/TeraSort/TeraValidate) et les résultats disponibles en ligne (TPCx-HS, …)
- Comparatif entre MapReduce et TEZ.
- Exemples d'optimisation et d'outils d'aide à l'optimisation (CDP advisor…).
- Choix de la taille des blocs.
- Autres options de tuning (utilisation de la compression, configuration mémoire...).

Travaux pratiques : Paramétrer, lancer et analyser des Bench, Appréhender les commandes au fil de l'eau de monitoring et d'optimisation de cluster.

8) Les apports de Hadoop v3

- Les approches de type stockage Objet (Ozone).
- Erasure coding.
- Yarn Federation.
- Scénarios de migration, les aspects à prendre en compte, et quelques exemples (Hortonworks vers Cloudera...).

LES DATES

CLASSE À DISTANCE
2025 : 25 mars, 17 juin, 23 sept., 04 nov.

PARIS
2025 : 18 mars, 10 juin, 16 sept., 28 oct.