Hadoop - Architecture et administration de clusters

3 j (21 heures) Ref : BDHA

Public

Architectes et urbanistes SI, administrateurs systèmes

Pré-requis

Java, langages de script

Moyens pédagogiques

Formation réalisée en présentiel ou à distance selon la formule retenue
Exposés, cas pratiques, synthèse, assistance post-formation pendant trois mois
Un poste par stagiaire, vidéoprojecteur, support de cours fourni à chaque stagiaire

Modalités de suivi et d'évaluation

Feuille de présence émargée par demi-journée par les stagiaires et le formateur
Exercices de mise en pratique ou quiz de connaissances tout au long de la formation permettant de mesurer la progression des stagiaires
Questionnaire d'évaluation de la satisfaction en fin de stage
Auto-évaluation des acquis de la formation par les stagiaires
Attestation de fin de formation

A l'issue de ce cours, vous connaitrez les différents composants d'un cluster Hadoop et saurez dimensionner une solution répondant aux besoins de vos services. Vous saurez mettre en oeuvre les différentes façons de déployer Hadoop, et opérer les outils d'administration et de surveillance pour maintenir un cluster en condition opérationnelle.

Objectifs

Identifier les différentes distributions Hadoop et leurs composants
Concevoir un cluster à partir des besoins et spécifications utilisateurs
Dimensionner un cluster
Identifier une configuration matérielle adaptée
Déployer manuellement et via des outils automatiques
Gérer les jobs utilisateurs
Utiliser des outils de monitoring et de diagnostic
Utiliser des outils d'optimisation de performances

Programme détaillé

INTRODUCTION AUX TECHNOLOGIES BIG DATA

Stockage et traitement de données massives : problèmes et solutions

Panorama des technologies NoSQL, bases de données distribuées et en colonnes

PRESENTATION DE L'ECOSYSTEME HADOOP

Coût, performance et évolutivité : promesses et gains effectifs

Les composants logiciels majeurs : Zookeeper, HDFS, HBase

Les différentes utilisations : comme solution de stockage, d'ETL, de traitement batch

HADOOP PAR LA PRATIQUE

Notions de base : HDFS et programmation Map/Reduce

TP : Prise en main d'une installation en mode pseudo distribué, commandes de base et manipulation de fichiers sous HDFS

DESIGN D'UN CLUSTER HADOOP

Topologie : les différents serveurs et leurs rôles

Configuration matérielle

Les différentes distributions Hadoop

Choix des briques logicielles selon l'usage du cluster

Dimensionnement

PROVISIONNEMENT ET DEPLOIEMENT

Déploiement manuel

Outils d'automatisation (Chef/Puppet, Whirr/Pallet)

Installation d'Hive, Pig, Hbase

Configuration et paramétrage

TP : Provisionnement et déploiement d'un cluster de taille moyenne

ADMINISTRATION ET OPERATION

Gestion des données (backup, localisation et réplication)

Gestion des jobs et schedulers

Monitoring du cluster

Ajout et décommission de nœuds

Benchmarks, tuning et optimisation

Résolution de problèmes, login et debugging

INTEGRATION AU SI

Stratégies et étapes d'intégration

Les différentes couches d'abstraction selon le public utilisateur

Connection aux bases de données relationnelles via Sqoop et JDBC

Ingestion de données via Flume

Interfacer avec les services avals

HADOOP ET SES COMPLEMENTS

Forces et faiblesses de la plateforme selon les cas d'utilisation

Alternatives et compléments

Comment intégrer Hadoop à Storm, Cassandra, Mongo, Giraph ...

DEPLOIEMENT A GRANDE ECHELLE

Hadoop sur le cloud : l'offre d'Amazon, Elastic MapReduce

Hadoop chez Facebook, LinkedIn, Orbitz...