INGENSI
Big Data Lab by CYRÈS

FORMATION HADOOP

Administrateur pour Hadoop (Apache)

CYRÈS Groupe

Direction commerciale et marketing : 87, avenue du Maine 75014 Paris - Tél. : 01 72 50 01 26
Centre de services : 19 rue Edouard Vaillant – 37000 Tours - Tel : 02 47 68 48 50 - Fax : 02 47 68 48 59 - www.cyres.fr
SAS au capital de 300 000 Euros - R.C.S. Tours B 442 155 818 - Code NAF: 6201Z

Sommaire

I. Objectifs

Encadrée par un formateur qualifié, cette formation vous permettra d'exploiter et de gérer un cluster Hadoop. De l'installation à la configuration en passant par l'optimisation, toutes les étapes seront traitées pour que vous soyez apte à administrer Hadoop. Les thématiques abordées seront les suivantes :

- Introduction au big data
- Hadoop, qu'est-ce que c'est ?
- Le système de fichiers distribués HDFS et l'algorithme MapReduce
- Bâtir une architecture Hadoop CDH5
- Déployer et configurer Hadoop, choix de l'infrastructure
- Les commandes d'exploitation et d'administration
- Hadoop avancé
- Comment alimenter un cluster Hadoop 1ère partie (Flume)
- Comment alimenter un cluster Hadoop 2ème partie (Hive)
- Comment alimenter un cluster Hadoop 3ème partie (Sqoop)
- Comment alimenter un cluster Hadoop 4ème partie (Impala)
- Visualisation des données (Hue)
- Diagnostics, problèmes et résolutions
- L'optimisation des configurations et les techniques d'améliorations des performances

II. Public concerné

Cette formation convient :

- aux administrateurs système qui ont déjà une expérience avec Linux.

- aux DBA, administrateurs BI

- aux chefs/directeurs de projets

III. Pré-requis

Connaissances en système d'exploitation Linux.

IV. Conditions Générales

Formation	Administrateur pour Hadoop (Apache)
Référence	HADADM1
Durée	4 jours (28 heures)
Tarif	A partir de 1 500 € H.T. / personne

V. Contenu de la formation

1) Introduction au big data

- Quelques chiffres
- Données structurées vs non-structurées
- Problématiques des solutions actuelles
- Exemples concrets d'applications big data
- Les 4 V
- Le cas Google

2) Hadoop, qu'est-ce que c'est ?

- Historique
- Concepts
- Les différentes versions
- Architecture d'un cluster hadoop
- Définition des différents services hadoop
- Présentation de l'écosystème hadoop

3) Le système de fichiers distribués HDFS et l'algorithme MapReduce

- HDFS : principes et techniques
- Mécanismes de lecture
- Mécanismes d'écriture
- Sécurité dans HDFS
- MapReduce v1 : qu'est-ce que c'est ?
- MapReduce v1 : fonctionnalités
- MapReduce v1 : concepts
- MapReduce v1 : architecture
- MapReduce v1 : failure recovery
- MapReduce v2 / YARN : qu'est-ce que c'est ?
- MapReduce v2 / YARN : concepts
- MapReduce v2 / YARN : architecture
- MRv1 vs MRv2
- Fair scheduler et capacity scheduler

4) Bâtir une architecture Hadoop CDH5

- Offre CDH5
- Produits inclus dans CDH5
- Cloudera express et Cloudera entreprise

5) **Déployer et configurer Hadoop, choix de l'infrastructure**

- Considérations générales
- Choix du matériel pour les slave nodes
- Choix du matériel pour les master nodes
- Choix du système d'exploitation
- Configuration de l'OS

6) **Les paramètres de configuration Hadoop**

- Hdfs-site.xml
- Core-site.xml
- Mapred-site.xml
- Yarn-site.xml
- Autres fichiers de configuration

7) **Travaux pratiques : Installer un cluster hadoop**

8) **Les commandes d'exploitation et d'administration :**

- Commandes hadoop
- Commandes hdfs
- Commandes yarn

9) **Travaux pratiques : HDS/MapReduce**

- Manipulation du système de fichier HDFS
- Lancement de jobs mapreduce (wordcount)
- Lancement de jobs mapreduce (calcul de PI)
- Lancement de jobs mapreduce (teragen/terasort)

10) **Hadoop avancé :**

- HDFS High Availability
- HDFS federation
- Capacity scheduler : concepts, paramétrage et mise en place

11) **Travaux pratiques : Mettre en place le capacity scheduler**

12) **Comment alimenter un cluster Hadoop 1ère partie :Flume**

- Flume : concepts, paramétrage et mise en place

13) Travaux pratiques : Flume

- Installation et utilisation de Flume
- Intégrer un fichier de log système dans hdfs

14) Comment alimenter un cluster Hadoop 2ème partie : Hive

- Hive historique
- Hive architecture
- Les différentes versions de Hive
- Hive data types
- Hive formats de stockage
- Hive DDL
- Hive DML
- Installation de Hive
- Les files d'attente avec Hive

15) Travaux pratiques : Hive

- Installation et utilisation de Hive
- Création d'une base de données
- Création d'une table et chargement d'un fichier csv
- Calcul des statistiques
- Ecriture de requêtes
- Création d'une table partitionnée et chargement des données
- Calcul des statistiques
- Ecriture de requêtes et comparaison des performances avec la table non-partitionnée

16) Comment alimenter un cluster Hadoop 3ème partie : Sqoop

- Sqoop historique
- Sqoop concepts et fonctionnalités
- Sqoop vs sqoop2
- Sqoop et sqoop2 architectures
- Les commandes sqoop
- Sqoop import
- Sqoop installation

17) Travaux pratiques : Sqoop

- Installation et utilisation de sqoop
- Importer une table depuis une base PostgreSQL vers Hive

18) Comment alimenter un cluster Hadoop 4ème partie : Impala

- Impala historique
- Impala historique

- Impala architecture
- Les différentes versions d'Impala
- Impala data types
- Impala formats de stockage
- Impala DDL
- Impala DML
- Installation d'Impala
- Paramétrage du cache hdfs
- Impala-shell
- Les files d'attente avec Impala

19) Travaux pratiques : Impala

- Installation et utilisation d'Impala
- Création d'une base de données
- Création d'une table au format textfile et chargement d'un fichier csv
- Calcul des statistiques
- Ecriture de requêtes
- Création d'une table au format parquet dans le cache hdfs et chargement des données
- Calcul des statistiques
- Ecriture de requêtes et comparaison des performances avec la table au format textfile
- Création d'une table partitionnée au format parquet dans le cache hdfs et chargement des données
- Calcul des statistiques
- Ecriture de requêtes et comparaison des performances avec la table non-partitionnée

20) Visualisation des données

- Hue concepts et architecture
- Installation et configuration de Hue

21) Travaux pratiques : Hue

- Installation et utilisation de Hue

22) Diagnostics, problèmes et résolutions

- Identifier et utiliser les fichiers de logs
- Hadoop safemode
- Perte du namenode
- Perte d'un ou plusieurs datanodes
- Block under replicated
- Hdfs balancer
- Ajout d'un datanode
- Décomissionner un datanode
- Distcp

23) Travaux pratiques : crash et incidents

- Simulation du crash d'un datanode
- Simulation du crash du namenode
- Ajouter un datanode
- Décomissionner un datanode

24) L'optimisation des configurations et les techniques d'améliorations des performances

- Les ressources de YARN
- Les étapes d'un mapreduce
- Paramétrage pour hdfs et mapreduce
- Paramétrage réplication et taille des blocs
- Optimisation des IO HDFS
- Paramètres liés à la configuration matérielle
- Optimisation de YARN
- Calcul du nombre de containers par serveur
- Règles à respecter

25) Travaux pratiques : construction d'un cluster Hadoop de A à Z

- Faire la construction d'un cluster pour abriter 400 To de données
    - Matériel (nombre de serveurs, cpu, ram, disques)
    - Configuration optimisée d'hadoop
    - Calculer le nombre de serveurs au bout de 2 ans