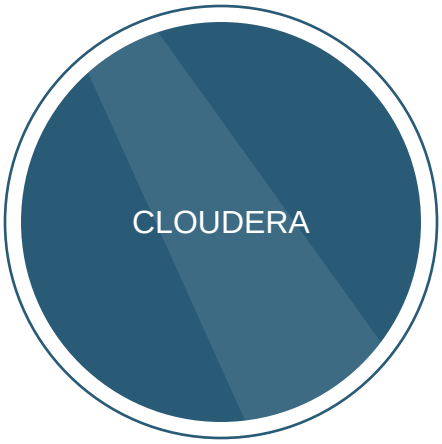
CLOUDERA

Cloudera Data Analyst

Formation officielle « Cloudera Certified Associate Data Analyst »

DESCRIPTION

Cloudera propose aux professionnels de la donnée les outils les plus performants pour accéder, manipuler, transformer et analyser des ensembles de données complexes, en utilisant SQL et les langages de script les plus courants.

Au cours de cette formation Data Analyst, vous apprendrez à appliquer vos compétences d'analyse de données et de business intelligence aux grands outils de données comme Apache Impala (en incubation) et Apache Hive.

Apache Hive, par l'intermédiaire de son langage HiveQL proche du SQL, permet la transformation et l'analyse de données complexes et multi-structurées évolutives dans Hadoop. Enfin, Cloudera Impala permet l'analyse interactive instantanée des données stockées dans Hadoop dans un environnement SQL natif.

Ensemble, Hive et Impala rendent les données multi-structurées accessibles aux analystes, aux administrateurs de base de données et à d'autres utilisateurs, sans nécessité de connaître la programmation Java.

OBJECTIFS PÉDAGOGIQUES

Acquérir, stocker et analyser des données à l'aide de Hive et Impala
Effectuer des tâches fondamentales d'ETL avec les outils Hadoop (extraire, transformer et charger) : ingestion et traitement avec Hadoop
Utiliser Hive et Impala pour améliorer la productivité sur les tâches d'analyse typiques
Relier des jeux de données de diverses provenances pour obtenir une meilleure connaissance commerciale
Effectuer des requêtes complexes sur les jeux de données

PUBLIC CIBLE

Analyste de données
Spécialiste de la business intelligence
Développeur
Architecte système
Administrateur de bases de données

PRÉ-REQUIS

- Connaissance de SQL
- Connaissance de base des lignes de commandes Linux
- Connaissance préalable d'Apache Hadoop non requise
- Connaissance d'un langage de script (comme Bash scripting, Perl, Python ou Ruby) est utile, mais pas indispensable.

MÉTHODE PÉDAGOGIQUE

Formation avec apports théoriques, échanges sur les contextes des

Stage pratique en présentiel
CLOUDERA

Code :
CLANA

Durée :
4 jours (28 heures)

Certification :
300 € HT

Exposés :
10%

Cas pratiques :
80%

Échanges d'expérience :
10%

Sessions à venir :

2 - 5 nov. 2020
Formation à distance / 2 695 eur

3 - 6 mai 2021
Paris / 2 695 eur

2 - 5 nov. 2021
Paris / 2 695 eur

Tarif & dates intra :
Sur demande

participants et retours d'expérience du formateur, complétés de travaux pratiques et de mises en situation.
Cette formation permet de préparer l'examen associé au titre de la certification « Cloudera Certified Associate Data Analyst » attestant des compétences acquises. La certification se déroule en dehors du temps de formation.

PROFILS DES INTERVENANTS

Toutes nos formations sont animées par des consultants-formateurs expérimentés et reconnus par leurs pairs.

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION

L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique. Une évaluation à chaud sur la satisfaction des stagiaires est réalisée systématiquement en fin de session et une attestation de formation est délivrée aux participants mentionnant les objectifs de la formation, la nature, le programme et la durée de l'action de formation ainsi que la formalisation des acquis.

POUR ALLER PLUS LOIN :

- Site internet de Cloudera
- Formation "Spark avec Hadoop pour développeurs de Cloudera"
- Formation "Administrer la plateforme Hadoop Cloudera"
- Livre blanc "Hadoop, feuille de route"
- Toutes nos formations HADOOP HORTONWORKS
- Toutes nos formations sur la Data Science

Programme pédagogique détaillé par journée

JOUR 1

LES FONDAMENTAUX D´HADOOP

- Pourquoi choisir Hadoop ?
- Présentation d'Hadoop
- Stockage de données : HDFS
- Traitement des données distribuées : YARN, MapReduce et Spark
- Traitement et analyse des données : Hive et Impala
- Intégration de base de données : Sqoop
- Les autres outils Hadoop
- Présentation des exercices

INTRODUCTION À HIVE ET À IMPALA

- Présentation de Hive
- Présentation d'Impala
- Pourquoi utiliser Hive et Impala?
- Schéma et stockage de données
- Comparaison de Hive et Impala avec les bases de données traditionnelles
- Cas d'utilisation

REQUÊTES AVEC HIVE ET IMPALA

- Bases de données et tables
- Syntaxe de base des langages de requête Hive et Impala
- Types de données
- Utilisation de Hue pour exécuter des requêtes
- Utilisation de Beeline (Shell Hive)
- Utilisation de Impala Shell

Jour 2

LES OPÉRATEURS COMMUNS ET FONCTIONS BUILT-IN

- Opérateurs
- Fonctions scalaires
- Fonctions d'agrégation

GESTION DES DONNÉES AVEC HIVE ET IMPALA

- Stockage de données
- Création de bases de données et de tables
- Chargement des données
- Modification des bases de données et des tables
- Simplification des requêtes au moyen de vues
- Enregistrement des résultats de requêtes

STOCKAGE DE DONNÉES ET PERFORMANCES

- Tables partitionnées

- Chargement des données dans des tables partitionnées
- Quand utiliser le partitionnement
- Choisir un format de fichier
- Utilisation des formats de fichier Avro et Parquet

Jour 3

ANALYSE RELATIONNELLE DE DONNÉES AVEC HIVE ET IMPALA

- Jointure de jeux de données
- Fonctions communes intégrées
- Agrégation et fenêtrage

LES FONCTIONS ANALYTIQUES ET LE FENÊTRAGE

- Utiliser des fonctions analytiques
- Autres fonctions analytiques
- Fenêtres glissantes

DONNÉES COMPLEXES AVEC HIVE ET IMPALA

- Données complexes avec Hive
- Données complexes avec Impala

ANALYSE DE TEXTE AVEC HIVE ET IMPALA

- Utilisation d'expressions régulières avec Hive et Impala
- Traitement des données textuelles dans Hive avec des SerDes
- Analyse de sentiment et n-grams

Jour 4

OPTIMISATION DE HIVE

- Comprendre les performances des requêtes
- Bucketing
- Indexation des données
  - 4Hive sur Spark

OPTIMISATION D´IMPALA

- Exécution de requête avec Impala
- Améliorer la performance d'Impala

EXTENSION DE HIVE ET D´IMPALA

- SerDes et formats de fichier personnalisés dans Hive
- Transformation de données avec des scripts personnalisés dans Hive
- Fonctions définies par l'utilisateur
- Requêtes paramétrées

CHOISIR LE MEILLEUR OUTIL

- Comparaison de Pig, Hive, Impala et des bases de données relationnelles
- Critères de choix

MODULE OPTIONNEL (EN FONCTION DE L´AVANCEMENT) : APACHE KUDU

- Qu'est-ce que Kudu
- Les tables Kudu
- Utiliser Impala avec Kudu

CLÔTURE DE LA SESSION

BEST

Développer des applications pour Spark avec Hadoop Cloudera

Formation officielle "Cloudera Developer Training for Spark and Hadoop"

DESCRIPTION

Apache Spark s'est imposé ces dernières années comme le framework big data de référence, et comme un outil central de l'écosystème hadoop. Cette formation intensive emmène le participant de la découverte de Spark jusqu'à l'utilisation de ses fonctionnalités avancées.

La démarche pédagogique équilibre apports théoriques sur les structures fondamentales Spark (RDD, DataFrame, DataSets) et de nombreux travaux pratiques. Les participants manipulent la console interactive pour prototyper. Ensuite, ils codent, déploient et monitorent des applications sur un cluster. Le programme intègre les évolutions majeures de la nouvelle version Spark 2, et des cas d'usages complexes de traitement en flux (streaming).

Au cours de la formation, un panorama de l'écosystème hadoop est dressé, en insistant sur les concepts essentiels des environnements distribués : stockage sur HDFS, calcul avec Map-Reduce et gestion des ressources via YARN.

Des compléments sur l'ingestion de données avec Sqoop et Kafka sont proposés, afin que les participants maitrisent l'ensemble des outils nécessaires pour développer des applications Spark. Ils disposent ainsi d'une expertise complète pour préparer des données massives et les analyser sur un cluster hadoop.

OBJECTIFS PÉDAGOGIQUES

Identifier et utiliser les outils appropriés à chaque situation dans un écosystème hadoop
Utiliser Apache Spark et l'intégrer dans l'écosystème hadoop
Utiliser Sqoop, Kafka, Flume, Hive et Impala

PUBLIC CIBLE

Développeur
Analyste

PRÉ-REQUIS

- Être à l'aise pour programmer dans l'un de ces langages : Scala et/ou Python
- Connaissance de base des lignes de commande Linux requise
- La connaissance de base du SQL est un plus
- Aucune expérience préalable avec hadoop n'est nécessaire

MÉTHODE PÉDAGOGIQUE

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation. Les exemples Apache Spark et les

Stage pratique en présentiel
CLOUDERA

Code :
CLSPH

Durée :
4 jours (28 heures)

Certification :
300 € HT

Exposés :
40%

Cas pratiques :
50%

Échanges d'expérience :
10%

Sessions à venir :

14 - 17 déc. 2020
Formation à distance / 2 695 eur

18 - 21 jan. 2021
Paris / 2 695 eur

15 - 18 mar. 2021
Paris / 2 695 eur

25 - 28 mai 2021
Paris / 2 695 eur

24 - 27 août 2021
Paris / 2 695 eur

Tarif & dates intra :
Sur demande

exercices de "hands-on" sont présentés avec Scala et Python.

A la suite de la formation, les stagiaires auront la possibilité de passer l'examen Certification « CCA Spark and Hadoop Developer » de Cloudera.

PROFILS DES INTERVENANTS

Toutes nos formations sont animées par des consultants-formateurs expérimentés et reconnus par leurs pairs.

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION

L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique. Une évaluation à chaud sur la satisfaction des stagiaires est réalisée systématiquement en fin de session et une attestation de formation est délivrée aux participants mentionnant les objectifs de la formation, la nature, le programme et la durée de l'action de formation ainsi que la formalisation des acquis.

POUR ALLER PLUS LOIN :

- Site internet Cloudera
- Livre blanc "Hadoop, feuille de route"
- Toutes nos formations Hadoop Cloudera
- Toutes nos formations sur la Data Science

Programme pédagogique détaillé par journée

Jour 1

INTRODUCTION À HADOOP ET À SON ÉCOSYSTÈME

- Introduction générale à hadoop
- Traitement de données
- Introduction aux exercices pratiques

HDFS : LE SYSTÈME DE FICHIERS HADOOP

- Les composants d'un cluster hadoop
- L'architecture d'HDFS
- Utiliser HDFS

LE TRAITEMENT DISTRIBUÉ SUR UN CLUSTER HADDOP

- L'architecture de YARN
- Travailler avec YARN

LES BASES DE SPARK

- Introduction à Spark
- Démarrer et utiliser la console Spark
- Introduction aux Datasets et DataFrames Spark
- Les opérations sur les DataFrames

MANIPULATION DES DATAFRAMES ET DES SCHEMAS

- Créer des DataFrames depuis diverses sources de données
- Sauvegarder des DataFrames
- Les schémas des DataFrames
- Exécution gloutonne et paresseuse de Spark

Jour 2

ANALYSER DES DONNÉES AVEC DES REQUÊTES SUR DATAFRAMES

- Requêter des DataFrames avec des expressions sur les colonnes nommées
- Les requêtes de groupement et d'aggrégation
- Les jointures

LES RDD – STRUCTURE FONDAMENTALE DE SPARK

- Introduction aux RDD
- Les sources de données de RDD
- Créer et sauvegarder des RDD
- Les opérations sur les RDD

TRANSFORMER LES DONNÉES AVEC DES RDD

- Écrire et passer des fonctions de transformation
- Fonctionnement des transformations de Spark

- Conversion entre RDD et DataFrames

AGRÉGATION DE DONNÉES AVEC LES RDD DE PAIRES

- Les RDD clé-valeur
- Map-Reduce : principe et usage dans Spark
- Autres opérations sur les RDD de paires

Jour 3

REQUÊTAGE DE TABLES ET DE VUES AVEC SPARK SQL

- Requêter des tables en Spark en utilisant SQL
- Requêter des fichiers et des vues
- L'API catalogue de Spark

TRAVAILLER AVEC LES DATASETS SPARK EN SCALA

- Les différences entre Datasets et DataFrames
- Créer des Datasets
- Charger et sauvegarder des Datasets
- Les opérations sur les Datasets

ÉCRIRE, CONFIGURER ET LANCER DES APPLICATIONS SPARK

- Écrire une application Spark
- Compiler et lancer une application
- Le mode de déploiement d'une application
- L'interface utilisateur web des applications Spark
- Configurer les propriétés d'une application

LE TRAITEMENT DISTRIBUÉ AVEC SPARK

- Rappels sur le fonctionnement de Spark avec YARN
- Le partitionnement des données dans les RDD
- Exemple : le partitionnement dans les requêtes
- Jobs, étapes et tâches
- Exemple : le plan d'exécution de Catalyst
- Exemple : le plan d'exécution de RDD

PERSISTANCE DE LA DONNÉE DISTRIBUÉE

- La persistance des DataFrames et des Datasets
- Les niveaux de persistances
- Voir les RDD persistés

LES ALGORITHMES ITÉRATIFS AVEC SPARK

- D'autres cas d'usages courants de Spark
- Les algorithmes itératifs en Spark
- Machine Learning avec Spark
- Exemple : K-means

Jour 4

INTRODUCTION À SPARK STRUCTURED STREAMING

- Introduction à Spark Streaming
- Créer des streaming DataFrames
- Transformer des DataFrames
- Exécuter des requêtes de streaming

STRUCTURED STREAMING AVEC KAFKA

- Introduction
- Recevoir des messages Kafka
- Envoyer des messages Kafka

AGGREGATION ET JOINTURES SUR DES STREAMING DATAFRAMES

- Aggregation sur des streaming DataFrames
- Jointure sur des streaming DataFrames

Suppléments

LE TRAITEMENT DE MESSAGES AVEC KAFKA

- Introduction à Kafka
- Passer à l'échelle avec Kafka
- L'architecture d'un cluster Kafka
- La ligne de commande Kafka

Administrer la plateforme Hadoop 2.X Hortonworks : fondamentaux

Formation officielle Hortonworks "ADM 221 – HDP Operations: Administration Foundations"

DESCRIPTION

Cette session prépare au rôle d'administrateur au sein d'un contexte technologique innovant et en particulier au cours d'un projet Big Data. A travers des exercices concrets, vous apprendrez à concevoir, installer, configurer et maintenir un cluster Hadoop.

A l'issue de cette formation, vous aurez grâce aux mises en pratique une solide compréhension d'Apache Ambari et vous appréhenderez son utilisation comme outil de gestion de la plateforme Hortonworks.

OBJECTIFS PÉDAGOGIQUES

Dimensionner un cluster Hadoop
Installer un cluster Hadoop
Configurer un cluster Hadoop
Sécuriser un cluster Hadoop
Maintenir un cluster Hadoop

PUBLIC CIBLE

Architecte
Administrateur

PRÉ-REQUIS

- Connaissances de l'environnement Linux.
- Capacité à lire et exécuter des scripts shell Linux simples.
- Il est recommandé d'avoir des connaissances de base autour des requêtes SQL et de l'expérience sur les sujets opérationnels tels que la gestion des incidents et la gestion des versions.

MÉTHODE PÉDAGOGIQUE

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.
Cette formation prépare à la certification éditeur Hortonworks.

PROFILS DES INTERVENANTS

Toutes nos formations sont animées par des consultants-formateurs expérimentés et reconnus par leurs pairs.

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION

L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique. Une évaluation à chaud sur la satisfaction des stagiaires est réalisée systématiquement en fin de session et une attestation de formation est délivrée aux participants mentionnant les objectifs

Stage pratique en présentiel
CLOUDERA

Code :
HWADM

Durée :
4 jours (28 heures)

Certification :
300 € HT

Exposés :
40%

Cas pratiques :
50%

Échanges d'expérience :
10%

Sessions à venir :

1 - 4 sept. 2021
Paris / 2 695 eur

Tarif & dates intra :
Sur demande

de la formation, la nature, le programme et la durée de l'action de formation ainsi que la formalisation des acquis.

POUR ALLER PLUS LOIN :

- Toutes nos formations HADOOP HORTONWORKS
- Formation officielle Hortonworks "Analyse de données pour Hadoop 2.X Hortonworks avec Pig, Hive et Spark" (HDP Developer: Apache Pig and Hive) (HWAPH)
- Formation officielle Hortonworks "Développer des applications pour Apache Spark avec Python ou Scala" ("DEV 343 – Spark Developer") (HWSPK)
- Livre blanc "Hadoop, feuille de route"
- Parcours de formation "Intelligence artificielle par la pratique : des fondamentaux à l'industrialisation"
- Formation "Fondamentaux de la Data Science" (DSFDX)
- Formation "Data Science : niveau avancé" (DSNVA)
- Formation "Cadrage et pilotage d'un projet de Data Science" (DSGDP)
- Formation "Industrialisation d'un projet de Data Science" (DSIND)
- Formation "Architecture des données : stockage et accès" (DSARC)

Programme pédagogique détaillé par journée

Jour 1:

BIG DATA, HADOOP ET LA PLATEFORME HORTONWORKS : LES BASES DU BIG DATA

- Les produits de la HDP
- Qu'est-ce que Hadoop ?
- Introduction à Ambari

INSTALLER LA HDP

- Identifier les options de déploiement de cluster
- Planifier un déploiement de cluster
- Faire une installation avec Ambari
- Mise en pratique : « Installer la HDP »

GESTION DES UTILISATEURS AVEC AMBARI

- Gérer les utilisateurs et les groupes
- Gérer les permissions
- Mise en pratique : « Gestion des utilisateurs avec Ambari »

GESTION DES SERVICES HADOOP VIA AMBARI

- Configuration des services
- Surveillance des services
- Maintenance des services
- Mise en pratique : « Gestion des services Hadoop »

UTILISER LE STOCKAGE HDFS

- Accéder aux données
- Gestion des fichiers
- Mise en pratique : « Utiliser le stockage HDFS »
- Les web services d'HDFS
- Mise en pratique : « Utiliser WebHDFS »
- Protéger les accès
- Mise en pratique : « Utiliser les ACLs HDFS »

Jour 2 :

GESTION DU STOCKAGE HDFS

- Architecture HDFS
- Gestion d'HDFS à travers l'interface Ambari Web
- Gestion d'HDFS en ligne de commande
- Mise en pratique : « Gestion du stockage sur HDFS »
- Les quotas HDFS
- Mise en pratique : « Gestion des quotas sur HDFS »

GESTION DES RACKS SUR HADOOP

- Les bénéfices de la « rack awareness »

- Configurer la « rack awareness »
- Mise en pratique : « Configurer la rack awareness »

PROTÉGER SES DONNÉES

- De l'importance des backups
- Les snapshots HDFS
- Utiliser DistCP
- Mise en pratique : « Gestion des snapshots HDFS »
- Mise en pratique : « Utiliser DistCP »

CONFIGURER LE STOCKAGE HÉTÉROGÈNE HDFS

- Les principes du stockage hétérogène
- Mise en pratique "Configuration des règles de stockage HDFS"

CONFIGURER LE CACHE CENTRALISÉ HDFS

- De l'utilité d'un cache HDFS centralisé
- Définir et gérer des groupes et instructions de cache
- Mise en pratique « Configuration du cache centralisé HDFS »

GATEWAY NFS SUR HDFS

- Les cas d'utilisations d'une gateway NFS sur HDFS
- Architecture et opération de la gateway NFS
- Installer et configurer la gateway NFS
- Mise en pratique « configurer une gateway NFS sur HDFS »

Jour 3 :

GESTION DES RESSOURCES AVEC YARN

- Architecture et Opération de YARN
- Les différentes façons de gérer YARN
- La gestion YARN des échecs de composants
- Mise en pratique : « Configurer et gérer YARN »
- Mise en pratique : « Gestion de YARN sans Ambari »

DÉCOUVERTE DES APPLICATIONS YARN

- Les bases d'une application YARN
- Mise en pratique : « Démarrer une application YARN »

LE CAPACITY SCHEDULER DE YARN

- Contrôler la répartition des ressources grâce aux queues YARN
- Configuration et gestion des queues YARN
- Contrôler les accès sur les queues YARN
- Mise en pratique : « Configurer le capacity scheduler »
- Mise en pratique : « Gérer les ressources et queues YARN »
- Mise en pratique : « Gérer les autorisations et les limites utilisateurs pour YARN »

LES LABELS SUR LES NŒUDS YARN

- Principes de base et application
- Activer et configurer les labels
- Gestion des labels (ajout, suppression et modification)
- Configurer les queues pour accéder aux ressources des labels
- Tester les labels pour valider leur comportement
- Mise en pratique : « Configurer les labels de nœuds YARN »

Jour 4 :

ACTIVER LA HAUTE DISPONIBILITÉ AVEC HDFS ET YARN

- Les principes de la haute disponibilité
- Haute disponibilité du Namenode
- Haute disponibilité du Resource manager
- Mise en pratique : « Configurer la haute disponibilité du namenode »
- Mise en pratique : « Configurer la haute disponibilité du resource manager »

GESTION DES NŒUDS DANS UN CLUSTER

- Ajouter, enlever un nœud du cluster
- Déplacer des composants
- Mise en pratique : « Ajouter, décommissionner et recommissionner un nœud »

SURVEILLANCE DE CLUSTER

- Surveillance avec Ambari
- Lever des alertes avec Ambari
- Mise en pratique : « Configurer les alertes avec Ambari »

LES BLUEPRINTS AMBARI

- Déploiement de cluster à la volée grâce aux blueprints
- Mise en pratique : "Déploiement de cluster avec les blueprints Ambari"

MONTÉE DE VERSION HDP

- Comprendre la stack HDP et sa version
- Les types et méthodes de montée de version avec HDP
- Le processus de montée de version, restrictions et prérequis
- Mise en pratique : "Faire une montée de version HDP"

BEST

Administrer la plateforme Hadoop Cloudera

Formation officielle « Cloudera Administrator Training for Apache Hadoop »

DESCRIPTION

Vous souhaitez exploiter le potentiel de vos données pour créer de la valeur et développer votre activité. Avec Hadoop et son architecture flexible et évolutive, vous pouvez stocker, traiter et analyser vos données à partir d'une plateforme unique fonctionnant sur du matériel standard.

Dès sa création en 2008, Cloudera a lié son histoire à celle de l'écosystème Hadoop. Avec ses fondations composées à 100 % de logiciels open source et de standards ouverts, la plate-forme Cloudera vous assure un meilleur contrôle des coûts, plus de souplesse et des résultats plus performants pour votre organisation. CDH, la plate-forme open source de Cloudera, est ainsi devenue la distribution la plus populaire de Hadoop.

De l'installation à la configuration en passant par l'équilibrage de charge et le réglage, cette formation de quatre jours fournit aux participants une compréhension complète de toutes les étapes nécessaires pour opérer et maintenir un cluster Hadoop à l'aide de Cloudera Manager.

OBJECTIFS PÉDAGOGIQUES

Reposer les bases de l'environnement Hadoop, MapReduce, Spark et HDFS
Gérer un cluster avec les fonctionnalités de Cloudera Manager
Déterminer le matériel et l'infrastructure appropriés pour son cluster
Configurer et déployer correctement le cluster pour l'intégration avec le système d'information
Charger des données dans le cluster à partir de fichiers générés dynamiquement à l'aide de Flume, ou à partir de SGBDR en utilisant Sqoop
Configurer FairScheduler pour répartir les ressources entre plusieurs utilisateurs d'un cluster
Préparer et maintenir Apache Hadoop en production en utilisant les bonnes pratiques
Dépanner, diagnostiquer, mettre au point et résoudre les problèmes sur Hadoop

PUBLIC CIBLE

Administrateur système
Responsable informatique
Architecte système
Développeur
Analyste de données
Administrateur de bases de données

PRÉ-REQUIS

- Connaissance de base de la ligne de commande Linux

MÉTHODE PÉDAGOGIQUE

Formation avec apports théoriques, échanges sur les contextes des

Stage pratique en présentiel
CLOUDERA

Code :
CLADM

Durée :
4 jours (28 heures)

Certification :
300 € HT

Exposés :
40%

Cas pratiques :
50%

Échanges d'expérience :
10%

Sessions à venir :

23 - 26 nov. 2020
Formation à distance / 2 695 eur

8 - 11 mar. 2021
Paris / 2 695 eur

20 - 23 sept. 2021
Paris / 2 695 eur

Tarif & dates intra :
Sur demande

participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

Cette formation permet de préparer l'examen associé au titre de la certification « Cloudera Certified Associate Administrator ».

PROFILS DES INTERVENANTS

Toutes nos formations sont animées par des consultants-formateurs expérimentés et reconnus par leurs pairs.

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION

L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique. Une évaluation à chaud sur la satisfaction des stagiaires est réalisée systématiquement en fin de session et une attestation de formation est délivrée aux participants mentionnant les objectifs de la formation, la nature, le programme et la durée de l'action de formation ainsi que la formalisation des acquis.

POUR ALLER PLUS LOIN :

- Site internet de Cloudera
- Formation "Développer des applications pour Spark avec Hadoop Cloudera"
- Livre blanc "Hadoop, feuille de route"
- Formation "Cloudera Data Analyst"
- Toutes nos formations Big Data
- Toutes nos formations sur la Data Science et la Data visualisation

Programme pédagogique détaillé par journée

Jour 1

INTRODUCTION

CLOUDERA ENTERPRISE DATA HUB

- Cloudera Enterprise Data Hub
- Introduction au CDH
- Introduction à Cloudera Manager
- Les responsabilités d'un administrateur Hadoop

INSTALLATION DE CLOUDERA MANAGER ET DU CDH

- Introduction à l'installation du cluster
- Installation de Cloudera Manager Installation
- Installation du CDH
- Les services du cluster CDH

CONFIGURER UN CLUSTER CLOUDERA

- Introduction
- Paramètres de configuration
- Modifier la configuration des services
- Fichiers de configuration
- Gérer les instances de rôle
- Ajouter des nouveaux services
- Ajouter et supprimer des hôtes

HADOOP DISTRIBUTED FILE SYSTEM

- Introduction
- Topologie et rôles HDFS
- Modifier les logs et le checkpointing
- La performance HDFS et la tolérance à la panne
- Introduction à la sécurité de HDFS et de Hadoop
- Interfaces utilisateurs web pour HDFS
- Utiliser la ligne de commande HDFS
- Autres outils de ligne de commande

Jour 2

INGESTION DE DONNÉES SUR HDFS

- Introduction à l'ingestion de données
- Formats de fichiers
- Ingérer de la donnée en utilisant File Transfer ou les interfaces REST
- Ingérer de la donnée d'une base de donnée relationnel avec Sqoop
- Ingérer de la donnée d'une source externe avec Flume
- Les bonnes pratiques d'ingestion de donnée

HIVE ET IMPALA

- Apache Hive
- Apache Impala

YARN ET MAPREDUCE

- Introduction à YARN
- Exécuter des applications sur YARN
- Explorer les applications YARN
- Les logs d'application YARN
- Les applications Map Reduce
- Réglage mémoire et CPU pour YARN

APACHE SPARK

- Introduction à Spark
- Les applications Spark
- Comment les applications Spark s'exécutent sur YARN
- Monitorer les applications Spark

Jour 3

DIMENSIONNEMENT DE VOTRE CLUSTER HADOOP

- Considérations générales relatives au dimensionnement
- Choix du matériel
- Considérations sur le réseau
- Options de virtualisation
- Options de déploiement cloud
- Configuration des noeuds

CONFIGURATION AVANCÉ DU CLUSTER

- Configurer les ports de service
- Paramétrer HDFS et MapReduce
- Activer la Haute Disponibilité HDFS

GESTION DES RESSOURCES

- Configuration de cgroups avec des centres de services statiques
- Le Fair Scheduler
- Configurer la gestion dynamique des ressources
- Planification des requêtes Impala

MAINTENANCE DU CLUSTER

- Vérification du statut HDFS
- Copier les données entre clusters
- Rééquilibrage du cluster
- Snapshots de répertoires
- Mise à niveau du cluster

Jour 4

MONITORING DU CLUSTER

- Fonctionnalités de monitoring de Cloudera Manager
- Tests de santé
- Événements et alertes
- Graphiques et rapports
- Recommandation de monitoring

DIAGNOSTIC DU CLUSTER

- Introduction
- Outils de diagnostic
- Exemples de mauvaises configurations

INSTALLER ET GÉRER HUE

- Introduction
- Gérer et configurer Hue
- Authentification et autorisation Hue

SÉCURITÉ

- Les concepts de sécurité sur Hadoop
- Authentification sur Hadoop en utilisant Kerberos
- Authorisation sur Hadoop
- Chiffrement sur Hadoop
- Sécuriser un cluster Hadoop

CONCLUSION

Suppléments sous réserve de temps disponible

APACHE KUDU

- Introduction à Kudu
- Architecture
- Installation et Configuration
- Outils de monitoring et de gestion

APACHE KAFKA

- Qu'est-ce que Apache Kafka ?
- Introduction à Kafka
- Architecture de cluster Kafka
- Outils de ligne de commande Kafka
- Utiliser Kafka avec Flume

STOCKAGE D´OBJETS DANS LE CLOUD

- Système de stockage d'objet
- Connecter Hadoop et un système de stockage objet