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 :

23 - 26 mar. 2020
Paris / 2 695 eur

7 - 10 juil. 2020
Paris / 2 695 eur

2 - 5 nov. 2020
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