Programme de formation

Big Data : Pig, Hive et Impala avec Hadoop

Objectifs

Cette formation vous apportera une grande expertise dans l'utilisation d'outils de traitement de données issues du Big Data. Apprenez à combiner et à mettre en oeuvre Pig, Hive et Impala dans votre système Hadoop pour accroître votre potentiel BI. Vous serez en mesure d'exploiter ces outils et adapter leurs utilisations pour un traitement optimal des données : requêtes, transformations, combinaisons, interprétations, stockage, et plus encore !

Pré requis

Bases sur Hadoop et le Big Data, connaissances en gestion de données et SQL

Durée

4 jours

Public

Architectes-techniques, Developpeurs, DSI

Plan de formation

Chapitre introductif
Les problématiques du Big Data
Retour sur l'architecture MapReduce
Le processus ETL
Hadoop : solutions apportées et manques
Retour sur le système de fichiers distribués Hadoop (HDFS)
L'environnement d'Hadoop

Exploration de l'outil Apache Pig
Pig : définition, caractéristiques et rayon d'action
Les cas d'utilisation de Pig
Le langage Pig Latin : caractéristiques et mise en oeuvre
Démarrer avec Pig

Utilisation de Pig pour traiter des données basiques
Connaître les types et les caractéristiques de données simples
Charger les données et définir les champs
Gérer la sortie des données
Techniques de tri et de filtrage des données récoltées
Utiliser les principales fonctions de traitement

Utilisation de Pig pour traiter des données complexes
Les différents formats de stockage
Connaître les types et les caractéristiques des données complexes et emboîtées
Grouper les données et utiliser la fonction built-in
Programmer des itérations de traitement de données groupées

Utilisation avancée de Pig
Effectuer des combinaisons d'ensembles de données
Exécuter des opérations sur des groupes de données
Paramètres avancés
Utiliser des macros et des fonctions utilisateurs (UDF)
Utiliser Pig avec d'autres langages

Résolution de problèmes et optimisation
Méthodes de résolution de problèmes
Utiliser l'UI web d'Hadoop pour le trouble shooting
Méthodes de débogage par échantillonnage de données
Monitoring des performances

Exploration de l'outil Apache Hive
Hive : définition, caractéristiques et rayon d'action

Le modèle de stockage de données de Hive
Hive et Pig : concurrence et complémentarités
Le langage de requête HiveQL
Démarrer avec Hive

Utilisation de Hive pour l'analyse de données relationnelles
Les bases et tableaux de données sous Hive
Connaître les types de données et leurs caractéristiques
Les formats de données dans Hive
Méthodes d'assemblage de données et fonctions de built-in

Gestion des données avec Hive
Construire des bases de données et tableaux de gestion Hive
Utiliser des tableaux autogérés
Stocker le résultat des requêtes
Sécuriser l'accès aux données

Analyse de données textuelles et études sémantiques
Les principes du traitement de données textuelles
Utiliser les fonctions String
Principes et applications du « Opinion Mining »

Optimisation et utilisation avancée
Mettre en oeuvre les bonnes pratiques pour la performance des requêtes
Paramétrer les requêtes
Contrôler l'exécution des tâches
Partitionnement des données, bucketing et indexation
Utiliser des scripts pour transformer les données
Mettre en oeuvre des fonctions utilisateurs (UDF)

Exploration du moteur de requêtes Impala
Impala : définition, caractéristiques et rayon d'action
Impala, Pig et Hive : concurrence et complémentarités
Impala dans le monde des bases de données relationnelles
Exemples d'utilisations du Shell Impala

Utilisation d'Impala pour l'analyse de données
Utiliser la syntaxe Impala
Connaître les types de données et leurs caractéristiques
Techniques de tri et de filtrage des données récoltées
Méthodes d'assemblage de données
Optimiser les performances

Conclusion