Big Data : Pig, Hive et Impala avec Hadoop

Lien : https://innov-systems.com/formation/big-data-pig-hive-et-impala-avec-hadoop

DURÉE
5 jours (35h)

RÉFÉRENCE
BSI31

CATÉGORIE
Big Data : développement, langages et NOSQL

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Utiliser Hive et Impala pour améliorer la productivité sur les tâches d'analyse typiques
- Automatiser le transfert des données dans le stockage Hadoop avec Flume et Sqoop
- Stocker et analyser des données à l'aide de Hive et Impala
- Interroger plusieurs ensembles de données pour une analyse avec Pig et Hive
- Filtrer les données avec les opérations Extract-Transform-Load (ETL) avec Pig

POUR QUI ?

- Architectes techniques
- Développeurs
- DSI

Programme détaillé

1 / L’écosystème Hadoop

- Big Data : problématiques
- L’architecture MapReduce
- Le processus ETL
- Solutions apportées et manques d'Hadoop
- L’environnement d’Hadoop

2 / Stocker les données dans HDFS

- Réaliser un stockage fiable et sécurisé
- Surveiller les mesures du stockage
- Contrôler HDFS à partir de la ligne de commande

3 / Traitement parallèle avec MapReduce

- Détailler l'approche MapReduce
- Transférer les algorythmes et non les données
- Décomposer les étapes clés d'une tâche MapReduce

4 / Automatiser le transfert des données

- Faciliter l'entrée et la sortie des données
- Agréger les données avec Flume
- Configurer le fan in et le fan out des données
- Déplacer les données relationnelles avec Sqoop

5 / Explorer l’outil apache Pig

- Définition, caractéristiques et rayon d’action

- Les cas d’utilisation de Pig
- Le langage Pig Latin : caractéristiques et mise en œuvre
- Démarrer avec Pig

6 / Traiter des données basiques avec Pig

- Connaître les types et les caractéristiques de données simples
- Charger les données et définir les champs
- Gérer la sortie des données
- Filtrer les données avec Pig
- Utiliser les principales fonctions de traitement

7 / Traiter des données complexes avec Pig

- 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

8 / 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

9 / 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

10 / Explorer 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

11 / Analyse de données relationnelles avec Hive

- 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

12 / 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

13 / Repousser les limites de HIVEQL

- Trier, répartir et regrouper des données
- Réduire la complexité des requêtes avec les vues
- Améliorer la performance des requêtes avec les index

14 / Déployer Hive en production

- Concevoir les schémas de Hive
- Établir la compression des données
- Déboguer les scripts de Hive

15 / Rationaliser la gestion du stockage avec HCatalog

- Unifier la vue des données avec HCatalog
- Exploiter HCatalog pour accéder au metastore Hive
- Communiquer via les interfaces HCatalog
- Remplir une table Hive à partir de Pig

16 / Analyse de données textuelles et études sémantiques

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

17 / Optimisation et utilisation avancée

- Mettre en œuvre 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 œuvre des fonctions utilisateurs (UDF)

18 / Explorer le 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

19 / Analyse de données avec Impala

- 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

20 / Lancer le framework Spark

- Réduire le temps d'accès aux données avec Shark
- Interroger les données Hive avec Shark

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

17 au 21 Août 2026	Distanciel
12 au 16 Oct. 2026	Distanciel

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com