ANALYSE DES DONNEES (ECOSYSTEM SPARK)

Durée et lieu de la formation : 3 jours à Nanterre (92) – ou à distance le temps de la crise sanitaire (matériel de formation fournis par Devenez)

Pour connaitre les dates de la prochaine formation :
https://www.devenez.fr/espace-candidats/les-offres-en-cours/formation-modules-boosters/

PRÉREQUIS

- Identifier et utiliser les outils appropriés à chaque situation dans unécosystème hadoopUtiliser Apache Spark et l'intégrer dans l'écosystème hadoopUtiliser Sqoop, Kafka, Flume, Hive et Impala

OBJECTIFS :

- Être à l'aise pour programmer dans l'un de ces langages : Scala et/ouPython
- 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

PROGRAMME

1. LES BASES DE SPARK

Introduction à Spark

Démarrer et utiliser la console Spark

- Introduction aux Datasets et DataFrames Spark
- Les opérations sur les DataFrames

devenez.fr
GROUPE FITEC

ANALYSE DES DONNEES (ECOSYSTEM SPARK)

2. 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

3. 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

4. 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

5. TRANSFORMER LES DONNÉES AVEC DES RDD

- Écrire et passer des fonctions de transformation
- Fonctionnement des transformations de Spark
- Conversion entre RDD et DataFrames

6. 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

7. 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

8. 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

9. 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

ANALYSE DES DONNEES (ECOSYSTEM SPARK)

10. 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

11. PERSISTANCE DE LA DONNÉE DISTRIBUÉE
- La persistance des DataFrames et des Datasets
- Les niveaux de persistances
- Voir les RDD persistés

12. 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

13. INTRODUCTION À SPARK STRUCTURED STREAMING
- Introduction à Spark Streaming
- Créer des streaming DataFrames
- Transformer des DataFrames
- Exécuter des requêtes de streaming

14. STRUCTURED STREAMING AVEC KAFKA
- Introduction
- Recevoir des messages Kafka
- Envoyer des messages Kafka

15. AGGREGATION ET JOINTURES SUR DES STREAMING DATAFRAMES
- Aggregation sur des streaming DataFrames
- Jointure sur des streaming DataFrames

Suppléments

16. 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

ANALYSE DES DONNEES (ECOSYSTEM SPARK)

APPROCHE PEDAGOGIQUE ET MODALITES D'EVALUATION

- **En temps normal** : formation présentielle alternant présentation théoriques et TP – sur le site de FITEC NANTERRE
- **En période de crise sanitaire** : Idem qu'en temps normal, mais à distance via VISIO CONFERENCE en SYNCHRONE
- **A la sortie de la crise sanitaire :** La formation reste accessible en visio-conférence.

FORMATEUR ET PÉDAGOGIE

Chaque formateur à FITEC doit avoir une double casquette celle d'expert et de pédagogue. C'est cette double compétence qui garantit une acquisition de compétences optimum pour chaque stagiaire.
Tous les formateurs ont suivi une formation et maîtrisent l'outil Teams. Ils ont également été formés quant à l'organisation de la formation à distance et cela dès sa mise en place.

QUALITÉ DES CONDITIONS TECHNIQUES

Un formateur anime le cours sur un **groupe Teams**, partage sa **vidéo** et son **écran**, diffuse ses **supports**, met une **base documentaire** à disposition, **interagit** avec les stagiaires, lance les TP sur des outils et plateformes cloud, peut **prendre la main** sur les postes.
Enregistrement de la session possible pour consultation hors créneaux horaires.

RYTHME ET SÉQUENÇAGE D'APPRENTISSAGE

Les formations à distance sont donc des cours **synchrones** qui requièrent la présence de tous comme une salle de formation en virtuel.
Le **rythme** est de 9h à 13h et 14h à 18h par jour.

INTERACTIVITÉ

A tout moment, les stagiaires peuvent **interagir** avec le formateur, poser des **questions orales** grâce au micro ou par écrit sur le **chat** de Teams.
Les micros des stagiaires sont coupés par défaut et chacun peut l'activer pour intervenir.

ACCOMPAGNEMENT

Pour chaque session de formation, en plus du formateur, un **modérateur** membre de l'équipe pédagogique veille à la bonne marche de la formation à distance. Il reste connecté à la formation, vérifie les connexions, répond aux questions ou difficultés des stagiaires, enregistre la session.
Le formateur peut ainsi se concentrer sur l'animation pédagogique de son cours.
Un **questionnaire en ligne de satisfaction** est rempli par les stagiaires.

MODALITES D'EVALUATION - VALIDATION DE LA MISE EN OEUVRE DE LA COMPETENCE

20% du temps est consacré au cours théorique et 80% du temps est consacré aux travaux pratiques, exercices et projets.
Les cursus complets de Reskilling sont validés par une **soutenance en ligne** face un jury de professionnels du métier.
Toute notre approche pédagogique est basée sur le respect des critères de la norme ISO 9001.

ANALYSE DES DONNEES (ECOSYSTEM SPARK)

TARIFS ET FINANCEMENTS (possibilité de prise en charge partielle ou intégrale)

Tarif public de cette formation : 950€ HT soit 1140€ TTC (tarif indicatif)

Votre formation peut être prise en charge par un organisme de financement. Dans le cadre d'un FNE, CPF, AIF, CSP, n'hésitez pas à nous contacter à recrutement@devenez.fr

En fonction de votre statut au moment de l'entrée en formation (salarié ou demandeur d'emploi), cette prise en charge pourra être **partielle ou intégrale**.
Plus d'informations sur les financements : https://www.devenez.fr/financements/

MODALITES ET DELAIS D'ACCES

Délai d'inscription variable selon le mode de financement.

INSCRIPTION SUR NOTRE SITE INTERNET :
https://www.devenez.fr/espace-candidats/les-offres-en-cours/formation-modules-boosters/

ACCESSIBILITE :

Public en situation de handicap, contactez notre Référent : M. Olivier BENANOU – o_benanou@devenez.fr

POUR TOUTE DEMANDE DE RENSEIGNEMENT :

Contactez-nous à recrutement@devenez.fr