EXCLU Programmer avec Apache Spark de Databricks

Formation officielle Databricks « Apache® Spark™ Programming SPARK 105 »

DESCRIPTION

Cette formation de 3 jours propose un panorama pratique de la solution Apache Spark en alternant des présentations et des exercices pratiques. Elle couvre les APIs de base de Spark, les fondamentaux et les mécanismes du framework, mais aussi les outils plus haut-niveau dont SQL, ainsi que ses capacités de traitement en streaming et l'API de machine learning.

Chaque sujet couvert comprend une partie d'exposé couplée à une mise en pratique de Spark au travers d'un environnement type notebook web. Inspiré d'outils tels IPython/Jupyter, les notebooks permettent aux participants de développer des jobs, des requêtes d'analyse et des représentations visuelles s'appuyant sur leur propre cluster Spark, le tout depuis leur navigateur web.

A l'issue du cours, les notebooks peuvent être conservés et être réutilisés dans le service cloud gratuit Databricks Community Edition, pour lequel la compatibilité est garantie. Il est également possible d'exporter le notebook sous forme de code source pour exécution sur n'importe quel environnement Spark.

OBJECTIFS PÉDAGOGIQUES

Décrire les fondamentaux de Spark
Exploiter les APIs de base de Spark pour manipuler des données
Concevoir et implémenter des cas d'usage typiques de Spark
Construire des pipelines de données et requêter de larges jeux de données grâce à Spark SQL et aux DataFrames
Analyser les jobs Sparks à l'aide des interfaces d'administration et des logs au sein des environnements Databricks
Créer des jobs de type Structured Streaming et Machine Learning
Découvrir les bases du fonctionnement interne de Spark

PUBLIC CIBLE

Data engineers et data analysts ayant l'expérience des traitements Big Data, qui souhaitent apprendre à utiliser Apache Spark pour effectuer leurs traitements Big Data, construire des jobs Spark à destination de la production et comprendre mais aussi déboguer des applications Spark.

PRÉ-REQUIS

- Une première expérience avec Apache Spark est conseillée
- Avoir utilisé les Spark DataFrames dans des cas simples est souhaitable
- Une expérience de programmation en langage objet ou fonctionnel est nécessaire

MÉTHODE PÉDAGOGIQUE

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

Stage pratique en présentiel
DATABRICKS

Code :
SP105

Durée :
3 jours (21 heures)

Exposés :
40%

Cas pratiques :
50%

Échanges d'expérience :
10%

Sessions à venir :

16 - 18 nov. 2020
Paris / 2 390 eur

3 - 5 mar. 2021
Paris / 2 390 eur

17 - 19 mai 2021
Paris / 2 390 eur

15 - 17 sept. 2021
Paris / 2 390 eur

17 - 19 nov. 2021
Paris / 2 390 eur

Tarif & dates intra :
Sur demande

La formation mélange les langages Python et Scala.

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.

Programme pédagogique détaillé par journée

Jour 1

APERÇU DE SPARK

- Exposés
  - Aperçu de Databricks
  - Fonctionnalités de Spark et de son écosystème
  - Composants de base
- Exercices pratiques
  - Environnement Databricks
  - Travailler avec des Notebooks
  - Fichiers et clusters Spark

SPARK SQL ET DATAFRAMES

- Exposés
  - Cas d'usages de Spark SQL et des DataFrames
  - APIs DataFrames / SQL
  - Optimisation de requêtes Catalyst
  - ETL
- Exercices pratiques
  - Créer des DataFrames
  - Requêter à l'aide des DataFrames et de SQL
  - ETL avec des DataFrames
  - Cache
  - Visualisation

Jour 2

FONCTIONNEMENT INTERNE DE SPARK

- Exposés
  - Jobs, Stages et Tasks
  - Partitions et Shuffling
  - Localité de la donnée
  - Performance des Jobs
- Exercices pratiques
  - Visualisation des requêtes SQL
  - Observer l'exécution des Tasks
  - Comprendre la performance
  - Mesurer l'utilisation mémoire

STRUCTURED STREAMING

- Exposés
  - Streaming sources et Sinks
  - APIs Structured Streaming
  - Windowing
  - Checkpointing et Watermarking

  - Streaming des DataFrames
  - Robustesse et tolérance aux pannes
- Exercices pratiques
  - Lire depuis TCP
  - Lire depuis Kafka
  - Visualisation en continu

Jour 3

MACHINE LEARNING

- Exposés
  - API Spark MLlib Pipeline
  - Création de features et algorithmes fournis
- Exercices pratiques
  - K-Means
  - Régression logistique

TRAITEMENTS ORIENTÉS GRAPHES AVEC GRAPHFRAMES

- Exposés
  - Analyse simple de graphes
  - API GraphFrames
  - Trouver des motifs GraphFrames
  - Persistance des données de graphes
- Exercices pratiques
  - ETL avec GraphFrames
  - Analyse du Property Graph GraphFrames

BILAN ET CLÔTURE DE LA FORMATION