

Programmer avec Apache Spark de Databricks

Formation officielle Databricks «Apache Spark™ Programming with Databricks »

DESCRIPTION

Apache Spark est un moteur d'analyses unifiées ultra-rapide pour le big data et le machine learning. Depuis sa sortie, il a connu une adoption rapide par les entreprises de secteurs très divers. Des acteurs majeurs du monde de l'internet tels que Netflix, Yahoo et eBay l'ont déployé à très grande échelle, traitant ensemble plusieurs peta-octets de données sur des clusters de plus de 8 000 nœuds.

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

A l'issue de la session, 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.

Ce cours officiel prépare à la certification "Databricks Certified Associate Developer for Apache Spark 3.0". La certification se passe après la formation et n'est pas obligatoire.

OBJECTIFS PEDAGOGIQUES

- 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
- Découvrir les bases du fonctionnement interne de Spark
- Découvrir le pattern Deltalake

PUBLIC CIBLE

Stage pratique
Data Engineering

Code :
ASPWD

Durée :
2 jour(s) (14,00 heures)

Exposés : **40.00 %**
Cas pratiques : **50.00 %**
Echanges d'expérience : **10.00 %**

Inter-entreprises :
Prochaines sessions disponibles [sur notre site web](#).
Tarif : 1 780,00 € HT / participant

Intra-entreprise :
Tarifs et dates sur demande.

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.

PRE-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

METHODE PEDAGOGIQUE

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. La formation mélange les langages Python et Scala.

Chaque sujet abordé 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.

PROFIL DES INTERVENANTS

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.

Afin de valider les compétences acquises lors de la formation, un formulaire d'auto-positionnement est envoyé en amont et en aval de celle-ci.

Une évaluation à chaud est également effectuée en fin de session pour

mesurer la satisfaction des stagiaires et un certificat de réalisation leur est adressé individuellement.

PROGRAMME PEDAGOGIQUE DETAILLE

Jour 1

APERÇU DE SPARK ET DATAFRAMES

- Introduction
- L'écosystème Databricks
- Spark SQL
- Lecture et écriture de données
- Dataframe et colonnes

TRANSFORMATIONS ET MANIPULATIONS DE DONNÉES

- Agrégations
- Datetimes
- Types complexes
- Fonctions additionnelles
- UDF : User Defined Functions

Jour 2

OPTIMISATION DE SPARK

- Architecture
- Shuffle et Cache
- Optimisation des requêtes
- Spark UI
- Gestion des partitions

STRUCTURED STREAMING

- Exposés
- Streaming et requêtes
- Processing streaming
- Agrégations
- Deltalake

CONCLUSION

- Évaluation de la session
- Partage sur la formation
- Questions/réponses additionnelles

