

## Optimiser Apache Spark™ sur Databricks

### Formation officielle Optimizing Apache Spark™ on 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 tels que Netflix, Yahoo et eBay l'ont déployé à très grande échelle, traitant ensemble plusieurs péta-octets de données sur des clusters de plus de 8 000 nœuds.

Dans ce cours, les apprenants exploreront les 5 problèmes majeurs de performance rencontrés dans une application Apache Spark™ : skew, spill, shuffle, stockage et serialization.

Au travers d'exemples basés sur des datasets de 100Go à 1To, le focus sera mis sur investigation et la réalisation de diagnostic des différentes sources de goulets d'étranglement avec Spark UI, ainsi que sur l'appropriation de stratégies de résolution efficaces.

Enfin, un temps sera également consacré à la découverte des nouvelles fonctionnalités proposées par Spark 3.x qui adressent automatiquement ces problèmes de performance communs.

#### OBJECTIFS PEDAGOGIQUES

- S'approprier les 5 problématiques de performances les plus communes dans une application Spark et leurs principales méthodes de résolution
- Investiguer, identifier et traiter des problèmes de performances les plus communs associés à l'ingestion de données.
- Appréhender les nouvelles fonctionnalités de Spark 3.x permettant de traiter les problèmes de performance dans vos applications Spark.
- Configurer des clusters Spark pour une performance maximale pour des besoins métier spécifiques.

#### PUBLIC CIBLE

- Développeurs Spark™
- Data Engineers

#### PRE-REQUIS

Une expérience de développement sur Apache Spark™.

Une expérience de développement avec Python ou Scala.

**Stage pratique**  
Data Engineering

Code :  
**ASPOP**

Durée :  
**2 jour(s) (14,00 heures)**

Exposés : **70.00 %**  
Cas pratiques : **20.00 %**  
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.

Il est fortement recommandé d'avoir suivi la formation "Programmer avec Apache Spark de Databricks" (ASPWD) au préalable.

#### **METHODE PEDAGOGIQUE**

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

#### **PROFIL DES INTERVENANTS**

Cette formation est dispensée par un·e ou plusieurs consultant·es d'OCTO Technology ou de son réseau de partenaires, expert·es reconnus des sujets traités.

Le processus de sélection de nos formateurs et formatrices est exigeant et repose sur une évaluation rigoureuse leurs capacités techniques, de leur expérience professionnelle et de leurs compétences pédagogiques.

Par ailleurs, pour animer cette formation, nos intervenant·es doivent également avoir suivi un parcours d'habilitation imposé par Databricks, Inc.

#### **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.

En l'absence de réponse d'un ou plusieurs participants, un temps sera consacré en ouverture de session pour prendre connaissance du positionnement de chaque stagiaire sur les objectifs pédagogiques évalués.

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**

##### **RAPPEL DU FONCTIONNEMENT DE SPARK™**

- Revue de l'architecture de Spark et de Spark UI
- Skew
- Spill
- Shuffle

- Storage
- Serialization

## Jour 2

### MÉTHODES D'OPTIMISATION

- Les bases de l'ingestion
- Prédire et anticiper les goulots d'étranglement
- Partitionnement de disque
- Z-ordering
- Bucketing
- Optimisation avec Adaptive Query Execution (AQE)
- Concevoir et configurer des clusters à haute performance

### BILAN ET CLÔTURE DE SESSION

- Revue des concepts clés présentés lors de la formation
- Temps d'échange sur les questions et réponses additionnels
- Retour à chaud et clôture

---

#### Accessibilité

L'inclusion est sujet important pour OCTO Academy.  
Nos référent-es sont à votre disposition pour faciliter l'adaptation de votre formation à vos besoins spécifiques.  
Pour les contacter : [academy.accessibilite@octo.com](mailto:academy.accessibilite@octo.com)