Maîtriser l’optimisation avancée de Spark sur Databricks

Lien :
https://innov-systems.com/formation/matriser-loptimisation-avancee-de-spark-sur-databricks

DURÉE
3 jours (21h)

RÉFÉRENCE
BSI427

CATÉGORIE
Big Data : développement, langages et NOSQL

Innov Systems

OBJECTIFS DE LA FORMATION

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

- Comprendre en profondeur les mécanismes internes de Spark™ pour mieux diagnostiquer les problèmes de performance
- Identifier et corriger les goulots d’étranglement liés à l’ingestion, au traitement et au stockage des données
- Exploiter les fonctionnalités avancées de Spark 3.x pour améliorer la rapidité et la fiabilité des traitements
- Configurer et dimensionner des clusters Databricks adaptés à différents cas d’usage métier
- Mettre en œuvre de bonnes pratiques pour un code Spark optimisé et maintenable

POUR QUI ?

- Développeurs Spark™
- Data Engineers
- Architectes Big Data

Programme détaillé

1 / PRINCIPES ET ARCHITECTURE DE SPARK™

- Comprendre l'architecture distribuée et les composants clés de Spark
- Explorer Spark UI pour diagnostiquer les performances
- Identifier les points critiques : Skew, Spill, Shuffle
- Gestion de la mémoire et des formats de sérialisation
- Optimisation du stockage et des caches

2 / OPTIMISATION DE L'INGESTION ET DU TRAITEMENT DES DONNÉES

- Bonnes pratiques d'ingestion à grande échelle
- Gestion et optimisation des partitions
- Utilisation efficace du Z-ordering et du bucketing
- Prédiction et réduction des goulets d'étranglement
- Exploitation d'Adaptive Query Execution (AQE) pour ajuster les plans d'exécution

3 / CONFIGURATION ET DIMENSIONNEMENT DES CLUSTERS

- Choisir le type et la taille de cluster en fonction du workload
- Configurer la mémoire, les cores et le parallelisme
- Stratégies de scaling automatique pour optimiser les coûts
- Ajustements spécifiques pour Spark 3.x dans Databricks

4 / OUTILS ET MÉTHODES D'ANALYSE DES PERFORMANCES

- Analyse et interprétation des métriques dans Spark UI
- Utilisation de logs et d'outils intégrés à Databricks
- Cas pratiques de débogage et de tuning
- Détection proactive des problèmes avant mise en production

Innov Systems

5 / BONNES PRATIQUES DE CODAGE SPARK

- Structuration des jobs pour un traitement efficace
- Choix des API (RDD, DataFrame, Dataset) selon les cas
- Gestion des jointures et agrégations volumineuses
- Réduction de la latence et optimisation des pipelines ETL

6 / ÉTUDE DE CAS RÉELLE

- Analyse d’un projet Spark existant présentant des problèmes de performance
- Identification des causes racines
- Application des techniques de tuning
- Comparaison des performances avant/après optimisation

7 / ATELIERS PRATIQUES

- Exercices guidés sur l’optimisation de requêtes et transformations
- Tuning des clusters en fonction de scénarios métier
- Tests de charge et interprétation des résultats

8 / SYNTHÈSE ET PLAN D’ACTION

- Récapitulatif des techniques et outils abordés
- Élaboration d’une checklist d’optimisation pour vos projets Spark
- Échanges sur les problématiques rencontrées par les participants
- Conclusion et ressources complémentaires

Innov Systems

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

03 au 05 Fév. 2027	Présentiel - Casablanca
31 Mars au 02 Avr. 2027	Présentiel - Casablanca
26 au 28 Mai 2027	Présentiel - Casablanca
21 au 23 Juil. 2027	Présentiel - Casablanca
22 au 24 Sep. 2027	Présentiel - Casablanca

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

Innov Systems — Tous droits réservés