Résumé et points forts
Statut : Traduit automatiquement de Anglais
Traduit automatiquement de Anglais
Nous vous félicitons ! Vous avez terminé ce module. Dans ce module, vous avez appris

- Spark est un système informatique distribué qui traite efficacement des ensembles de données à grande échelle.
- Les ingénieurs de données bénéficient de diverses fonctionnalités fournies par Spark, qui les aident à créer des pipelines de traitement de données évolutifs et efficaces.
- Grâce à l'approche informatique distribuée de Spark, à ses capacités de traitement en mémoire et à ses API, les ingénieurs de données peuvent s'attaquer à des structures de données complexes et travailler avec des ensembles de données massifs.
- La classification à l'aide de Spark ML consiste à prédire des étiquettes ou des classes catégoriques pour un ensemble donné de caractéristiques.
- Les algorithmes de classification peuvent être appliqués dans divers domaines, tels que l'analyse des sentiments, la détection des spams, la détection des fraudes et la classification des images.
- La régression, à l'aide de Spark ML, implique la construction et l'entraînement de modèles de régression en utilisant les bibliothèques d'apprentissage automatique offertes par Spark.
- Les algorithmes de régression permettent de créer des modèles qui prédisent des variables cibles continues à l'aide de caractéristiques d'entrée.
- Le clustering, réalisé à l'aide de Spark ML, consiste à regrouper des points de données similaires en clusters sur la base de caractéristiques ou de schémas partagés.
- Les algorithmes de classification sont bien adaptés à des applications telles que la segmentation des clients, la détection des anomalies, la segmentation des images et les systèmes de recommandation.
- L'instance de l'algorithme K-means spécifie le nombre de grappes (k) et une graine aléatoire pour la reproductibilité.
- GraphFrames, intégré à Apache Spark, permet le traitement des graphes à l'aide de DataFrames.
- Il fournit des DataFrames distincts pour les sommets et les arêtes des graphes, qui peuvent être analysés avec SparkSQL.
- GraphFrames inclut également des algorithmes graphiques intégrés populaires qui peuvent être appliqués aux DataFrames des arêtes et des sommets pour diverses tâches d'analyse.