Republique Tunisienne
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Jendouba
Institut Supérieur d'Informatique du Kef

الجمهورية التونسية
وزارة التعليم العالي والبحث العلمي
جامعة جندوبة
المعهد العالي للإعلامية بالكاف

Optimisation des pipelines de traitement de Big Data en utilisant des modèles d'apprentissage automatique distribués

Encadrant académique: Dr. Anwer KALGHOUM

Email: anouar.kalghoum@isikef.u-jendouba.tn | anwer.kalghoum@gmail.com

Description:

Le traitement de Big Data implique des volumes massifs de données nécessitant une gestion et une analyse efficaces pour en extraire des informations précieuses. Les pipelines de traitement de Big Data sont utilisés pour orchestrer les différentes étapes de ce processus, allant de la collecte et du prétraitement des données à l'analyse et à l'extraction des connaissances. Cependant, avec l'augmentation exponentielle des données, l'optimisation de ces pipelines devient un défi crucial. L'objectif de ce projet est d'explorer l'utilisation de modèles d'apprentissage automatique distribués pour améliorer l'efficacité des pipelines de traitement de Big Data. En appliquant des techniques d'apprentissage distribué, telles que les réseaux neuronaux distribués et le machine learning parallèle, l'optimisation des pipelines vise à accélérer les processus tout en maintenant ou en améliorant la précision des résultats. Ce projet inclura également l'intégration de frameworks comme Apache Spark, Hadoop, et TensorFlow Distributed pour la mise en œuvre de modèles distribués.

Objectifs:

1. **Étudier l'état de l'art des pipelines de Big Data:** Analyser les défis liés au traitement de Big Data, notamment la gestion des données massives, leur prétraitement, et l'orchestration des processus. Étudier les méthodes existantes pour optimiser ces pipelines.
2. **Développement de modèles d'apprentissage automatique distribués:** Implémenter des modèles d'apprentissage automatique adaptés aux environnements distribués. Appliquer des techniques telles que le machine learning parallèle, les réseaux neuronaux distribués, et les algorithmes de consensus pour traiter de grandes quantités de données de manière efficace.
3. **Optimisation du prétraitement des données dans les pipelines de Big Data:** Concevoir des méthodes distribuées pour le prétraitement des données (nettoyage, transformation, normalisation) afin de réduire les coûts de calcul et d'améliorer la qualité des données en entrée.
4. **Évaluation de la performance des modèles et des pipelines optimisés:** Comparer l'efficacité des pipelines optimisés à celle des pipelines traditionnels en termes de temps d'exécution, de précision, et de capacité à traiter de grandes quantités de données. Évaluer la scalabilité et la résilience des modèles distribués.

5. **Intégration avec des frameworks Big Data:** Mettre en œuvre les modèles d'apprentissage automatique distribués à l'aide de frameworks tels qu'Apache Spark, Hadoop, ou TensorFlow Distributed. Garantir une intégration fluide dans un environnement de traitement de Big Data en temps réel.

Résultats attendus:

1. **Optimisation des pipelines de Big Data:** Un pipeline de traitement de Big Data plus rapide et plus efficace, grâce à l'application de modèles d'apprentissage automatique distribués.
2. **Amélioration des performances:** Réduction des temps de traitement tout en maintenant une haute précision dans l'analyse des données.
3. **Scalabilité des solutions proposées:** Des solutions capables de gérer des volumes massifs de données tout en maintenant des performances optimales même à grande échelle.
4. **Réduction des coûts de calcul:** Optimisation des ressources informatiques en réduisant les coûts de calcul tout en améliorant l'efficacité globale des pipelines.
5. **Adaptation aux environnements en temps réel:** La solution proposée pourra être utilisée pour des applications nécessitant un traitement en temps réel des données.

Références:

1. **Zaharia, M., et al. (2024).**
Optimizing Distributed Machine Learning Pipelines with Apache Spark. *IEEE Transactions on Big Data, 10*(1), 1-14.
2. **Zhao, Y., & Liu, J. (2024).**
A Survey of Machine Learning Algorithms for Big Data Processing in Distributed Systems. *Future Generation Computer Systems, 151*, 35-50.
3. **Wang, Y., & Xu, Z. (2024).**
TensorFlow Distributed: A Framework for Efficient Large-Scale Machine Learning. *Journal of Machine Learning Research, 25*(6), 122-145..
4. **Chen, S., & Yu, J. (2024).**
Parallel Machine Learning Models for Big Data Analytics. *Computational Intelligence and Neuroscience, 2024*, Article ID 9812345.
5. **Li, T., & Zhang, Q. (2024).**
Distributed Learning Algorithms for Big Data in Cloud Environments. *International Journal of Cloud Computing and Services Science, 12*(2), 103-116.

Mots-clés : Big Data; Machine Learning Distribué; Apache Spark; Hadoop; TensorFlow Distributed; Optimisation de pipelines; Scalabilité.