AVIS DE SOUTENANCE

THESE DE DOCTORAT

Présentée par

Mr: MOHAMMED BERGUI

Discipline : Informatique

Spécialité : Informatique

Sujet de la thèse : Modélisation des performances des travaux MapReduce géo-distribués à l'aide du deep learning.
Formation Doctorale : Sciences de l'ingénieur, Sciences Physiques, Mathématiques et Informatique.

Thèse présentée et soutenue le samedi 18 mars 2023 à 15h au Centre de conférences de la faculté des Sciences et Techniques devant le jury composé de :

Nom Prénom	Titre	Etablissement	
Fatiha MRABTI	PES	Faculté des Sciences et Techniques de Fès	Président
Zouhair ABDELHAMID	PH	Faculté des Sciences et Techniques de Tanger	Rapporteur
Nourddine EN-NAHNAHI	PH	Faculté des Sciences Dhar El Mehraz de Fès	Rapporteur
Med Chaouki ABOUNAIMA	PH	Faculté des Sciences et Techniques de Fès	Rapporteur
Azeddine ZAHI	PES	Faculté des Sciences et Techniques de Fès	Examinateur
Said NAJAH	PH	Faculté des Sciences et Techniques de Fès	Directeur de thèse

Laboratoire d'accueil : Laboratoire des Systèmes Intelligents et Applications.

Etablissement : Faculté des Sciences et Technique de Fès

Résumé de la thèse

L'analyse de données géo-distribuées a suscité un grand intérêt ces dernières années en raison du besoin croissant de dériver des connaissances à partir de données géo-distribuées. Bien que les applications de traitement en cluster, telles que MapReduce et Spark, aient été largement déployées dans les centres de données pour prendre en charge les applications commerciales et la recherche scientifique, elles ne parviennent pas à répondre aux exigences de performance de la plupart des applications. De plus, elles entraînent un gaspillage des ressources en raison des différences inhérentes entre les environnements, notamment la nature éparse, hautement hétérogène et dynamique des ressources distribuées : puissance de calcul et bande passante du réseau.

Ce type d'analyse géo-distribuée implique le transfert de données sur les liens du réseau étendu (WAN) entre différents centres de données. La nature de ces liens est très limitée et hétérogène, ce qui rend le transfert de données lent et coûteux.

Dans cette thèse, nous passons en revue les systèmes de traitement de big data géodistribués qui prennent en compte la bande passante du WAN et nous fournissons les avantages et les inconvénients de la plupart de ces systèmes. En outre, nous les classons en fonction de la technique de traitement et du framework sur lesquels ils sont basés et nous les comparons en fonction de plusieurs caractéristiques. Nous identifions les travaux futurs, tels que l'utilisation de l'apprentissage automatique pour améliorer la planification des tâches.

Puisque l'une des fonctions les plus critiques de Hadoop est la gestion des ressources. Une gestion plus efficace sera obtenue si l'estimation du temps d'exécution des travaux est faite avec précision. Ainsi, dans cette thèse, nous proposons d'utiliser l'apprentissage automatique pour prédire le temps d'exécution des travaux MapReduce géo-distribué. Pour cela, nous proposons une nouvelle base de données de traces de travaux MapReduce version dans le cloud avec une bande passante réseau limitée. Nous présentons la méthodologie et le dispositif expérimental que nous avons adoptés pour créer la base de données, puis nous décrivons le processus de génération et de collecte de cette base. Par la suite, nous étudions l'anatomie de MapReduce ainsi que les caractéristiques qui impactent les performances. Nous proposons ensuite un réseau de neurones profond pour prédire le temps d'exécution des travaux MapReduce. Le résultat estprometteur car il montre que l'apprentissage profond peut être appliqué comme solution à ce problème et, avec suffisamment de données, peut atteindre une grande précision.