Avis de Soutenance

Monsieur Sébastien RIVAULT

Informatique

Soutiendra publiquement ses travaux de thèse intitulés

Parallélisme, équilibrage de charges et extensibilité dans le traitement des mégadonnées sur des systèmes à grande échelle

dirigés par Monsieur SEBASTIEN LIMET
Ecole doctorale : Mathématiques, Informatique, Physique Théorique et Ingénierie des Systèmes - MIPTIS
Unité de recherche : LIFO - Laboratoire d'Informatique Fondamentale d'Orléans

Soutenance prévue le ***mardi 02 juillet 2024*** à 14h00
Lieu : 6 Rue Léonard de Vinci, 45067 Orléans
Salle : Amphi H

Composition du jury proposé

M. SEBASTIEN LIMET	LIFO, Université d'Orléans	Directeur de thèse
M. Laurent D'ORAZIO	Université Rennes 1, CNRS, IRISA	Rapporteur
M. Gaétan HAINS	Université Paris Est Créteil	Rapporteur
M. Mostafa BAMHA	LIFO, Université d'Orléans	Co-encadrant de thèse
Mme Sophie ROBERT	LIFO, Université d'Orléans	Co-encadrante de thèse
M. Sofian MAABOUT	LaBRI, Université de Bordeaux	Examinateur

Mots-clés : MapReduce,Parallélisme,Extensibilité,Traitement de Mégadonnées,

Résumé :

Durant les deux dernières décennies, grâce à la réduction des coûts de stockage, d'échange et de traitement de l'information, le volume de données générées chaque année ne cesse d'exploser. Les enjeux liés au traitement de ces mégadonnées sont souvent décrits par la règle des 3V : le volume, la variété et la vitesse de création, de collecte, d'analyse et de partage des données. Pour stocker et analyser ces ensembles de données volumineux, il est essentiel d'utiliser des grappes de machines et des algorithmes extensibles et insensibles aux déséquilibres pouvant se produire pour répartir équitablement la charge sur l'ensemble des machines de traitement. Des applications telles que le filtrage collaboratif, la déduplication et la résolution d'entités sont nécessaires pour identifier des relations dans des mégadonnées en se reposant sur une notion de similarité entre les enregistrements. Ces applications permettent entre autres de retrouver des utilisateurs ayant des goûts similaires, de nettoyer les données et de détecter des fraudes. Dans ces cas, les opérations de jointure et de recherche par similarité sont utilisées pour retrouver tous les enregistrements similaires dans une ou plusieurs collections de données en utilisant une distance et un seuil défini par l'utilisateur. Bien que les équi-jointures parallèles aient été largement étudiées et mises en œuvre avec succès sur des architectures parallèles et distribuées, ces algorithmes ne sont pas adaptés à l'opération de jointure par similarité, car il n'y aucune technique de hachage ou de tri dans la littérature permettant de retrouver tous les couples d'enregistrements potentiellement similaires en une seule étape. Des techniques existent dans la littérature pour réduire l'espace de recherche tout en garantissant la complétude du résultat et en évitant le calcul d'un produit Cartésien et la comparaison de tous les couples d'enregistrements. Cependant, l'extensibilité de ces techniques est limitée et ne permet pas le traitement efficace de mégadonnées sur des systèmes à grandes échelles. Des méthodes approximatives ont été proposées pour traiter la jointure par similarité en renonçant à une petite partie du résultat et en fournissant une garantie probabiliste sur l'exhaustivité des résultats et permettant de réduire l'espace de recherche. Ces méthodes reposent sur des fonctions de hachages dont les probabilités de collisions sont sensibles à la similarité des objets. Nous nous reposons sur ces techniques pour proposer des solutions efficaces, pour le traitement de la jointure et de la recherche par similarité, basées sur l'utilisation de LSH (Locality Sensitive Hashing), d'histogrammes distribués et de schémas de communication randomisés dans le but de réduire les coûts de traitement, de communication et de lecture/écriture de données sur disques aux seules données pertinentes pour diverses distances et sur divers objets, et ce, dans le but de proposer un framework générique basé sur le modèle de programmation MapReduce répondant aux enjeux de volume, de variété et de vitesse d'analyse des données. L'efficacité et l'extensibilité des solutions proposées ont été étudiées en utilisant un modèle de coût et ont été confirmées par une série d'expériences mesurant également l'exhaustivité du résultat et la réduction de l'espace de recherche, garantissant ainsi une solution efficace quels que soient la taille, le déséquilibre des données en entrée, la distance et les seuils fournis par l'utilisateur.