Titre : Optimisation des requêtes dans des environnements parallèles : Application au Big Data

Mots clés : Big Data, Passage à l'échelle, performances, Optimisation de requêtes, Base de données distribuées.

Encadrants :
Amin Mesmoudi (amin.mesmoudi@univ-poitiers.fr)
Ladjel Bellatreche (ladjel.bellatreche@ensma.fr)

1. Contexte

Le Big Data représente un défi non seulement pour le monde socio-économique mais aussi pour la recherche scientifique (Zicari et al.). En effet, comme il a été souligné dans plusieurs articles scientifiques (*e.g., Wu et al.*) et rapports stratégiques (*e.g., Wang et al.*), les applications informatiques modernes sont confrontées à de nouveaux problèmes qui sont liés essentiellement au stockage et à l'exploitation de données générées par les instruments d'observation et de simulation. La gestion de telles données représente un véritable goulot d'étranglement qui a pour effet de ralentir la valorisation des différentes données collectées non seulement dans le cadre de programmes scientifiques internationaux mais aussi par des entreprises, ces dernières s'appuyant de plus en plus sur l'analyse de données massives.

La recherche scientifique, à l'ère des Big Data, est devenue multidisciplinaire. En effet, il est nécessaire de combiner des techniques issues de plusieurs disciplines (informatique, physique, mathématique, …) afin de faire avancer la science. D'ailleurs, à titre d'exemple, le projet LSST[1] ambitionne la construction du plus grand télescope au monde. Le défi ultime de LSST est de mettre à disposition des scientifiques une base de données commune à partir de laquelle seront conduites des recherches scientifiques qui s'intéressent, entre autres, à la recherche de petits objets dans le système solaire, à l'astrométrie de précision des régions extérieures à la Voie Lactée, à la surveillance des effets transitoires dans le ciel optique et à l'étude de l'Univers lointain. La communauté française utilisera ces données pour mener des études sur l'énergie noire responsable de l'accélération de l'expansion de l'univers, incomprise à ce jour. Le goulot d'étranglement lié à ces analyses repose en grande partie sur la méthodologie d'accès et de traitement des données retenues. LSST produira des images CDD de 3,2 Gigapixel toutes les 17 secondes (la nuit), pendant 10 ans. Il permettra à terme de générer 15 à 30 Téraoctets de données par nuit pour arriver à un volume d'environ 140 Pétaoctets d'images en fin de programme. Le catalogue de données est constitué de tables relationnelles ayant des tailles allant jusqu'à 5 Pétaoctets (*Ivezić et al.*). Par conséquent, de telles applications sont orientées par des questions telles que : comment stocker, organiser, indexer et distribuer des milliers de PetaOctets de données ? comment combiner l'indexation et la gestion de mémoire pour des bases de données extrêmement volumineuses, distribuées et multidimensionnelles ? comment évaluer des jointures entres des objets ayant plus de 100 milliards d'éléments, ce qui induit un

[1] https://www.lsst.org/

problème de passage à l'échelle ? Quels algorithmes utilisés pour évaluer des requêtes et des fonctions d'agrégations sur ce genre de base de données ?

2. Objectif du stage

Dans le cadre de ce stage, nous travaillerons sur le problème de choix du meilleur plan d'exécution pour des requêtes issues du projet LSST. En effet, plusieurs plans d'exécution peuvent être considérés pour la même requête. Nous avons déjà prouvé que ce problème est NP Hard. Nous devons considérer deux sous problèmes : 1) le choix du modèle du coût qui prend en compte les environnements parallèles d'exécution et 2) le développement d'un algorithme efficace permettant d'explorer l'espace de recherche. D'ailleurs, les techniques proposées seront intégrées dans le système massivement parallèle QDAG qui a pour objectif de garantir à la fois le passage à l'échelle et les performances lors du traitement des Big Data.

Technologies impliquées

- ***Frameworks Parallèles :*** Hadoop et Spark
- ***Environnement cloud de testes :*** machines virtuelles Ubuntu déployées avec OpenStack
- ***Langages de programmation*** **:** Python, Java et éventuellement C++

3. Références

Zicari, Roberto V. "Big data: Challenges and opportunities." *Big data computing* (2014): 564.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

Wang, S., Wang, H. J., Qin, X. P., & Zhou, X. (2011). Architecting big data: challenges, studies and forecasts. *Jisuanji Xuebao(Chinese Journal of Computers)*, *34*(10), 1741-1752.

Ivezić, Ž., Connolly, A. J., & Jurić, M. (2016). Everything we'd like to do with LSST data, but we don't know (yet) how. *arXiv preprint arXiv:1612.04772*.

Amin Mesmoudi, Mohand-Saïd Hacid, Farouk Toumani: Benchmarking SQL on MapReduce systems using large astronomy databases. Distributed and Parallel Databases 34(3): 347-378 (2016)

Ladjel Bellatreche, Pedro Furtado, Mukesh K. Mohania: Special Issue in Physical Design for Big Data Warehousing and Mining. Distributed and Parallel Databases 34(3): 289-292 (2016)


isae
Institut Supérieur de l'Aéronautique et de l'Espace
ENSMA

LIAS

1431
Université
de Poitiers