Titre : Big Data et passage à l'échelle : vers une nouvelle approche de gestion intelligente et efficace

(English version below)

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

Encadrants : Ladjel Bellatreche (ladjel.bellatreche@ensma.fr) et Amin Mesmoudi (amin.mesmoudi@univ-poitiers.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. Défis scientifiques : évaluation et optimisation de requêtes

Les systèmes classiques de gestion des bases de données et les techniques associées doivent être revisités *(Mesmoudi et al., Bellatreche et al.*) afin de faire face aux nouveaux défis engendrés par les Big Data. D'ailleurs, les techniques permettant une utilisation efficace de nouvelles plateformes matérielles et logicielles représentent une étape importante pour le développement du "Big Data". Dans cette thèse, les contributions scientifiques attendues sont liées principalement à 1) l'identification des bonnes abstractions pour capturer les nouveaux environnements d'exécution via une étude expérimentale, 2) le développement de nouvelles techniques qui supportent la parallélisation massive des traitements sur des grandes masses de données, et 3) la définition formelle des modèles de coûts pour évaluer l'efficacité des algorithmes utilisés dans les plateformes technologiques modernes.

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)

Title: Big Data and scalability: Towards a new smart and effective management Approach

Keywords: Big Data, Scalability, Performances, Query Optimization, Parallel databases.

Supervisors: Ladjel BELLATRECHE (ladjel.bellatreche@ensma.fr) and Amin Mesmoudi (amin.mesmoudi@univ-poitiers.fr)

1. Context

Big Data represents a challenge not only for the socio-economic world but also for scientific research (*e.g., Zicari et al.*). Indeed, as witnessed in several scientific articles (*e.g., Wu et al.*) and strategic reports (*e.g., Wang et al.*), modern computer applications are confronted with new problems that are essentially related to storage and use of data generated by observation and simulation instruments. The management of such data represents a real bottleneck which has the effect of slowing down the valorization of different data collected not only by companies but also within the framework of international scientific programs, the latter relying more and more on massive data analysis.

Scientific research, in the Big Data era, has become multidisciplinary. Indeed, it is necessary to combine techniques coming from several disciplines (computer science, physics, mathematics, ...) in order to make advances in science. By the way, for example, the LSST project aims to build the largest telescope in the world. The ultimate challenge of LSST is to provide scientists with a common database from which will conducted scientific research that focuses, inter alia, on search of small objects in the solar system, precision astrometry of regions outside the Milky Way, monitoring of transient effects in the optical sky and the study of the distant Universe.
The French community will use this data to conduct studies on the dark energy responsible for the acceleration of universe expansion, misunderstood to this day. The bottleneck associated with these analyzes is largely based on the methodology used to access and process data. LSST will produce 3.2 Gigapixel CDD images every 17 seconds (at night) for 10 years. It will eventually generate 15 to 30 terabytes of data per night to reach a volume of about 140 petabytes of image at the end of the program. The data catalog consists of relational tables with sizes up to 5 Petabytes (*Ivezić et al.*). Therefore, such applications are guided by questions such as: how to store, organize, index and distribute thousands of Petabytes data? How to combine indexing and memory management for very large, distributed, and multidimensional databases? How to evaluate joins between objects with more than 100 billion elements, which leads a scalability problem? Which algorithms to evaluate queries and aggregation functions on this kind of datasets?

2. Scientific challenges: Query evaluation and Optimization

Traditional database management systems and related techniques must be revisited (*Mesmoudi et al., Bellatreche et al.*) in order to face the new challenges generated by the Big Data. Moreover, techniques allowing the efficient use of new hardware and software platforms represent an important step for the development of Big Data. In this thesis, the expected scientific contributions are mainly related to 1) the identification of good abstractions to capture the new execution environments via an experimental study, 2) the development of new techniques that support massive parallel processing on large databases, and 3) formal definition of cost models to evaluate the effectiveness of algorithms used in modern technology platforms.

3. References

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)