Traitement des jointures par similarité sur des architectures distribuées

Keywords: Jointure par similarité, Big Data, Parallélisme, Extensibilité, modèle MapReduce.

Personnes à contacter

- **Mostafa BAMHA / Sophie ROBERT / Sébastien LIMET**
  LIFO, Université d'Orléans.
  Email: {Mostafa.Bamha,Sophie.Robert,Sebastien.Limet}@univ-orleans.fr

Description du sujet de stage (4 à 6 mois)

La mise en correspondance de paires d'objets similaires, appelée également jointure par similarité, est une fonctionnalité fondamentale dans l'analyse des données et dans la prise de décisions. La jointure par similarité se retrouve au centre d'intérêt dans de nombreux domaines d'application (Recherche de plagiat/redondance de données, Détection de fraudes, Systèmes de recommandation, Filtrage Collaboratif, Bioinformatique, ...). Tous ces domaines d'application font face à des masses de données de plus en plus importantes qui ne peuvent être traitées sur une seule machine. Le but du stage est de proposer des approches extensibles permettant de traiter de manières efficaces (en termes de coûts de traitement et de communication, Coûts I/O, tailles des résultats intermédiaires ...), la jointure par similarité, sur des architectures distribuées, en utilisant le modèle MapReduce et des mesures de similarité (distance Euclidienne, Jaccard, Cosinus, Fréchet, ...) adaptées à la taille et à la nature des données à traiter. Le travail du stage consiste à

- Explorer, dans un premier, les différentes solutions proposées dans la littérature sur la jointure par similarité sur des architectures distribuées,
- Proposer et mettre en place des solutions efficaces pour les jointures par similarité basées sur le modèle MapReduce permettant de reduire les coûts de traitement, de communication et de replication des données,
- Évaluer les performances des solutions proposées sur des jeux de données synthétiques ou données réelles.

References

[1] Fier, Fabian and Augsten, Nikolaus and Bouros, Panagiotis and Leser, Ulf and Freytag, Johann-Christoph. Similarity Joins on Mapreduce: An Experimental Survey. of the VLDB Endowment, Volume 11, June 2018

[2] R. Vernica, M. J. Carey, and C. Li. Efficient parallel set-similarity joins using mapreduce. In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pages 495–506. 2010.