UNIVERSITE PARIS 13, Sorbonne Paris Cité
Laboratoire d'Informatique de Paris-Nord, UMR CNRS 7030
*99, avenue Jean-Baptiste Clément - 93430 Villetaneuse – **France***
Tél : (33) 01 49 40 40 71 ; Fax : (33) 01 48 26 07 12
www.univ-paris13.fr

Sujet de stage de Master 2 Informatique

Qualité des données, Elimination des doubles et des similaires

Responsable du stage : M. Faouzi Boufarès
Maître de Conférences HDR
Université Paris 13, Sorbonne Paris Cité
Laboratoire LIPN UMR CNRS 7030
99 avenue Jean-Baptiste Clément
F-93430 Villetaneuse, France
Bureau A109
Email: boufares@lipn.univ-paris13.fr
Web: http://lipn.univ-paris13.fr/~boufares/
----------+++----------

Contexte :

Les données contribuent au succès de l'activité d'une organisation. Leur qualité représente un enjeu très important. Le coût de la non-qualité peut s'avérer très élevé : prendre une décision à partir de mauvaises informations peut nuire à l'organisation ou à ses clients et partenaires. La gouvernance des données est un sujet qui prend de plus en plus d'importance dans les entreprises et administrations. L'importance des données et de leur qualité est de plus en plus reconnue. Une bonne gouvernance des données permet l'amélioration des interactions entre les différents collaborateurs d'une ou plusieurs organisations concernées.

Le problème de l'élimination des doublons (et des similaires) et de la fusion/intégration de données est très complexe. En effet, il s'agit de localiser, fusionner et enrichir des entités qui représentent le même monde réel. Beaucoup de questions se posent, lors du traitement d'une volumétrie très importante, notamment celles de l'optimisation du temps de réponse. On peut les classer à deux niveaux :

- L'identification des valeurs proches ou égales (les fonctions Match & Merge).

- Le processus d'élimination des doubles et des similaires en utilisant un traitement massivement parallèle.

Mots clés : Données Volumineuses, BigData, Elimination des doubles et des similaires, Architectures distribuée et parallèle, Cloud computing.

Objectifs du stage :

- **Concevoir** de nouveaux algorithmes massivement parallèles qui permettent l'élimination des doublons et des similaires pour de très gros volumes de données.

- **Implémenter** ces algorithmes dans un environnement distribué.

- **Comparer** les performances de chaque algorithme.

Lieu du stage :

Laboratoire d'Informatique de Paris Nord (LIPN)- UMR CNRS 7030
Université Paris 13 Sorbonne Paris Cité

Période de stage :

Avril/ Septembre 2014

Environnement de travail :

- Langages : JAVA, SQL, PL/SQL
- Hadoop (Cluster Paris13)
- BD : Oracle
- Talend Open Studio, Talend Data Quality

Profil recherché :

Etudiants en Master 2 informatique.

Bibliographie

- [ABB+2007] J. Akoka, L. Berti-Équille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué, Z. Kedad, S. Nugier V. Peralta, S. Si-Said-Cherfi, A Framework for Quality Evaluation in Data Integration Systems, Proceedings of the 9th International Conference on Enterprise Information Systems (ICEIS 2007), Madeira, Portugal, June 2007.
- [ABB+2008] J. Akoka, L. Berti-Équille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué, Z. Kedad, S. Nugier V. Peralta, M. Quafafou, S. Sisaïd-Cherfi, Évaluation de la qualité des systèmes multisources. Une approche par les patterns, Proceedings of the 2nd Workshop on Data and Knowledge Quality (QDC 2008) in conjunction with the French National Conf. on Extraction and Management of Knowledge (Extraction et Gestion des Connaissances - EGC), Nice, France, January 29, 2008.
- [B2004] L. Berti-Équille, La qualité des données comme condition à la qualité des connaissances : un état de l'art. Mesures de qualité pour la fouille de données, Revue Nationale des Technologies de l'Information (RNTI-E)-Cépaduès, France, 2004.
- [B2007a] L. Berti-Équille, Quality Awareness for Data Managing and Mining, Université de Rennes 1, France, Juin 2007.
- [B2007b] L. Berti-Équille, Data Quality Awareness: a Case Study for Cost Optimal Association Rule Mining, Special Issue of Knowledge Information Systems, London, UK, 2007, pp. 191-215.
- [BBC2012a] F. Boufarès, A. Bensalem and S. Correia, « Qualité de données dans les entrepôts de données : élimination des similaires », Revue des Nouvelles Technologies de l'Information, (RNTI), B-8, Entrepôts de Données et Analyse en ligne (EDA'2012). Juin 2012. Pages 22-31. Actes des 8èmes Journées francophones sur les Entrepôts de Données et l'Analyse en ligne, (EDA'2012), 12-13 Juin 2012, Bordeaux, France.
- [BBC2012B] F. Boufarès, A. Bensalem and S. Correia, « Deduplication algorithms for DataBases and Data warehouses », Proceedings of the Twenty First International Conference on Software Engineering and Data Engineering, (SEDE'2012), 27-29 June 2012, Los Angeles, California, USA. Pages 73-78.
- [BBH+ 2009] M. Badri, F. Boufarès, S. Hamdoun, V. Heiwy, K. Lellahi, Construction and Maintenance of Heterogeneous Data Warehouses, Information Sciences reference: Data Warehousing Design and Advanced Engineering Applications Methods for Complex Construction, Hershey, New York, 2009, pp. 189-204.
- [BD2009] L. Berti-Équille, T. Dasu, Data Quality Mining: New Directions, IEEE International Conference on Data Mining (ICDM), Miami, Florida, USA, 7 Décembre 2009.
- [BGMS2009] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. Euijong Whang, J. Widom**.** Swoosh: A Generic Approach to Entity Resolution. The VLDB Journal, January 2009.
- [BN2006] J. Bleiholder and F. Naumann, Conflict Handling Strategies in an Integrated Information System, Humboldt-Universitat zu Berlin Unter den Linden 6 Berlin, May 22–26, 2006, Edinburgh, UK, pp. 1-6.
- [DBS2009] X.-L. Dong, L. Berti-Equille, and D. Srivastava, Integrating conflicting data: the rôle of source dependence, Proceedings of the International Conference on

Very Large Databases (VLDB), Lyon, France, August 2009.

- [KPS+2005] D. Kostadinov, V. Peralta, A. Soukane, X. Xue, Intégration de données hétérogènes basée sur la qualité, Grenoble, France, 2005, pp. 1-16.
- [ORH+2005] P. Oliveira, F. Rodrigues, P. Henriques, H. Galhardas, A Taxonomy of Data Quality Problems, 2nd Int. Workshop on Data and Information Quality (DIQ' 05), Porto, Portugal, 2005, pp. 219-233.
- [PRB2007] V. Peralta, R. Ruggia., M. Bouzeghoub, Data Quality Evaluation in a Data Integration System, AMW'2007, Punta del Este, Uruguay, October 26th 2007, pp. 1-25.
- [T2008] C. Toulemonde, JEMM research_Informatica, Exploiter le capital de votre organisation, Un livre blanc de JEMM research - Des données de qualité, France, 2008, pp. 1-26.
- [JS2004] J. Dean, S. Ghemawat. MapReduce: Simplified data processing on large clusters. In Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI2004), pages 137–150, 2004.
- [CBZ2009] E.Y. Chang, H. Bai, K. Zhu, Parallel algorithms for mining large-scale richmedia data, In: Proceedings of the 17th ACM International Conference on Multimedia (MM '09), New York, NY, USA, 2009, pp. 917-918.
- [DG2008] J. Dean and S. Ghemawat, -MapReduce: simplified data processing on large clusters, Communications of the ACM, vol. 51, Jan. 2008, pp. 107–113.

Exemple :

Données de départ : Plusieurs types d'anomalies !

	Nom et Prénom	Téléphone	Mail
t1	Le Bon Adam	0666600007	lebon@yahoo.fr
t2	Le Bon A.	0666677777	
t3	Le B. Adam	0666677777	lebon@yahoo.fr
t4	Grande Clémence	0666688887	grande@yahoo.fr
t5	Adam LeBon	0666677777	
t6	Unique Eve	0666622227	unique@gmail.com
t7	Le Bon Adam	0666600007	lebon@yahoo.fr
t8	Le Bon	0666600007	lebon@yahoo.fr
t9	LeBon Adeline	0666611117	Alebon@yahoo.fr

Données résultats à l'arrivée (sans double ni similaire) !

	Nom et Prénom	Téléphone	Mail
t'1	Le Bon Adam	0666600007, 0666677777	lebon@yahoo.fr
t'2	Grande Clémence	0666688887	grande@yahoo.fr
t'3	Unique Eve	0666622227	unique@gmail.com
t'4	LeBon Adeline	0666611117	Alebon@yahoo.fr