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 Projet de Fin d'Etudes

Sujet de stage de Master 2 Informatique

Le MDM de L'ENISO

Le Master Data Management de L'Ecole Nationale d'Ingénieurs de Sousse

Qualité des données, Intégration des données Elimination des doubles et des similaires

Responsable du stage : M. Faouzi Boufarès
Université Paris 13, Sorbonne Paris Cité
Laboratoire LIPN UMR CNRS 7030
99 avenue Jean-Baptiste Clément
F-93430 Villetaneuse, France
Email: boufares@lipn.univ-paris13.fr
Web: http://lipn.univ-paris13.fr/~boufares/
----------+++----------
Collaboration avec : M. Aref Meddeb directeur de l'ENISO
----------+++----------

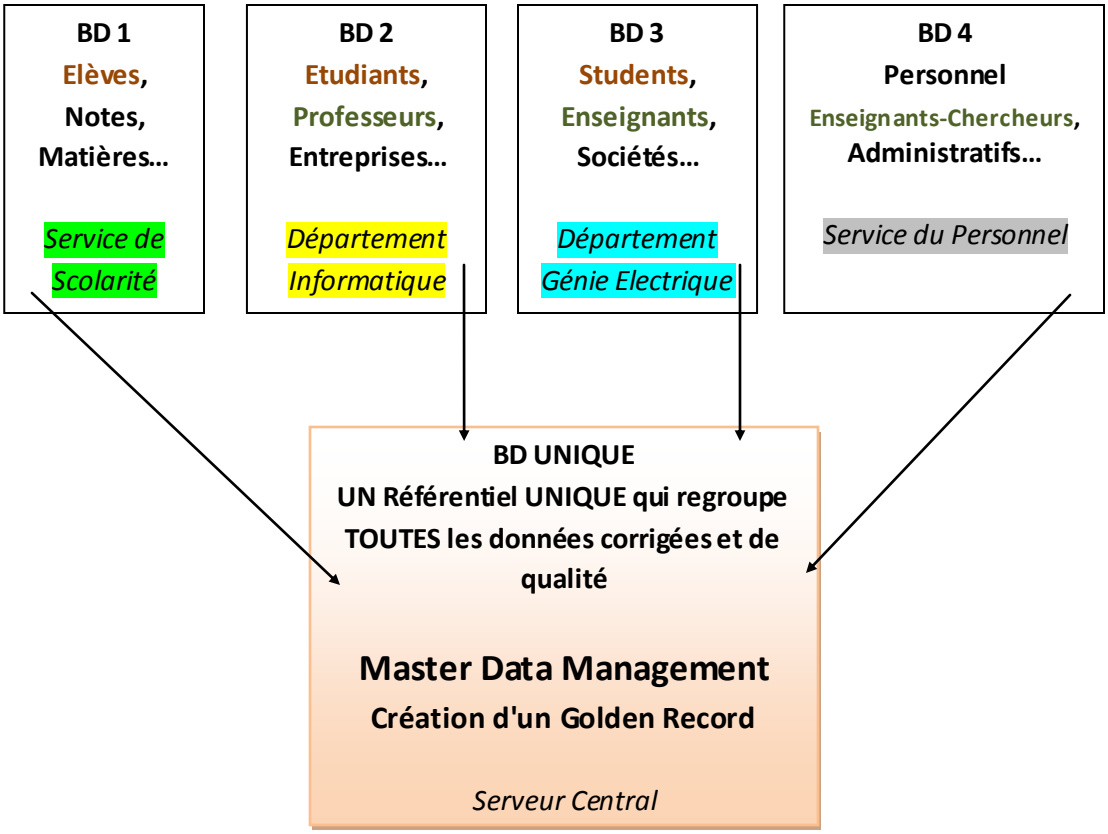
Contexte :

Le **Big Data** s'est imposé récemment comme une innovation majeure pour toutes les organisations et les entreprises qui cherchent à construire un avantage concurrentiel grâce à l'exploitation de leur **données** (données internes et données externes).

En effet, les données contribuent au succès de l'activité de toute organisation et toute entreprise. **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 les 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.

De plus en plus d'entreprises tentent de capitaliser sur leurs **données métier** les plus importantes en **construisant des référentiels de type MDM « Master Data Management »** offrant **une vue centrale et unique de ces dernières**. La qualité des données est un pré-requis essentiel pour ce type de projets, plus encore que pour les projets BI « Business Intelligence ».

Exemple : Création d'un MDM pour un établissement de formation (Ecole, Université…)



Le problème de la **fusion/intégration** des données est complexe. En effet, il s'agit de localiser, **fusionner et enrichir des entités qui représentent le même monde réel**. Parmi les problèmes qui se posent, On peut citer l'identification des valeurs proches ou égales et l'élimination des doubles et des similaires.

Mots clés : Intégration des données hétérogènes, Données Volumineuses, Big Data, Gouvernance des Données, Master Data Management, Qualité des Données, Elimination des doubles et des similaires, Architectures distribuée et parallèle.

Objectifs : Créer les données de référence

- Concevoir la Base de Données de référence
- Intégrer les différentes données existantes (CSV, Excel, Access, Oracle, MySQL…)
- Unifier la codification et Corriger les données
- Créer de nouvelles données en respectant des normes
- Définir les modalités d'accès et de mises à jour des données

Lieu du Stage/Travail/Projet :

L'Ecole Nationale d'Ingénieurs de Sousse (ENISO)
+
Laboratoire d'Informatique de Paris Nord (LIPN)- UMR CNRS 7030
Université Paris 13 Sorbonne Paris Cité

Période de Stage/Travail/Projet :

2016-2017

Environnement de travail (Exemple) :

- Langages : JAVA, SQL, PL/SQL, PHP MySQL…
- BD : Oracle
- ETL : TALEND OPEN STUDIO
- MapReduce, Hadoop, Spark

Profil recherché : Le travail peut être effectué par plusieurs

Etudiants sérieux d'un niveau >= Bac + 4, Bac + 5 en informatique.

Bibliographie

- [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.
- [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.
- [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.
- [ABS2015] A. BenSalem. Thèse : Qualité contextuelle des données : Détection et nettoyage guidés par la sémantique des données ; 31 mars 2015, Université Sorbonne Paris cité, Paris, France. Directeur : M. Faouzi BOUFARES.
- [HZ2017] H. Zaidi, Thèse : Amélioration de la qualité des données ; Correction sémantique des anomalies inter-colonnes ; 1er février 2017, Cnam Paris, Université Sorbonne Pris Cité, Paris 13, France. Directeur : M. Faouzi BOUFARES.