# Nettoyage des données en présence de données de référence

Elisa ABIDH, Julien LEVESY,Armand ROSSIUS

# Table des matières

# 1 Enjeux de la qualité de données

Depuis plusieurs années la gestion des données est devenue cruciale pour les entreprises, le volume de données stockées et échangées augmentent, ce qui confronte les entreprise à la problématique de comment stocker les données et pouvoir y accéder facilement, on parle alors de "Data Management" ou gestion des données.
Que ce soit pour des raisons légales, des besoins opérationnelles ou pour des choix stratégiques la gestion de l'information est importante. Au sein d'une entreprise beaucoup d'activités et fonctions sont concernées :

- La gestion d'activité optimale pour répondre à la demande : demande une maitrise de l'information.
- Toutes les fonctions des entreprises sont gérées par le SI
- Les données sont un flux présent dans toutes les entreprises
- Les dirigeants : parce que les décisions, le plan stratégique nécessite de l'information
- Les responsables opérationnels : ils traitent de l'information pour pouvoir gérer au mieux les problèmes.
- Marketing : données sur les fournisseurs, les clients, les concurrents, les marchés
- Les collaborateurs opérationnels : approvisionner le stock, lister des interventions sur une machine, nom des pièces changées

On comprends que la qualité des données gérer par le SI est importante pour répondre aux attentes du client mais aussi pour gérer de manière optimale l'entreprise. Cependant l'entreprise va rencontrer plusieurs difficultés pour répondre à cette problématique de qualité de données :

- Détecter la mauvaise qualité des données.
- Trop de données car beaucoup de données inutiles.

Avant de répondre à cette problématique il est légitime de se poser la question : "Qu'est ce qu'une donnée de qualité ? " D'après le livre de Christophe Brasseur "Data Management : qualité des données et compétitivité" la qualité ne se résume pas à une donnée juste, c'est une condition nécessaire mais non suffisante. Il est difficile de donner une définition précise de cette notion, cependant on peut dégager plusieurs axes pour juger de la qualité des données : qualité du contenu, accessibilité, flexibilité, sécurité.

1. Qualité du contenue
   - Justesse de l'information : en phase avec la réalité
   - Adéquation aux besoins : réponds aux besoins réels
   - Facilité d'interprétation : pas d'ambiguïté(abrévation,unités), compréhensible.
2. Accessibilité
   - Disponibilité : disponible quand on en a besoin
   - Facilité d'accès : ergonomie des applications.
3. Flexibilité
   - Evolutivité : définition et codification de la donnée (pas de remise en cause)
   - Cohérence avec d'autres sources (identifier les données partagées),
   - Possibilité de traduction.
4. Sécurité Protéger l'information des menaces accidentelles et des attaques malveillantes
   - Confidentialité
   - Fiabilité
   - Traçabilité
   - Intégrité des données.

C'est un enjeu d'actualité qui

# 2 Les solutions actuelles

## 2.1 Solution spécifique

## 2.2 ERP

## 2.3 Le MDM

# 3 Une réponse : le master data management

## 3.1 Présentation du procédé et objectifs

Le Master Data Management, traduit en français par Gestion des données de références, est une discipline des technologies de l'information ayant pour objectif de définir des concepts et méthodes visant à établir au sein d'un système d'information un schéma de base de données de références considérées commes fiables.
Outre cela, le Master Data Management engloble aussi les disciplines d'intégration, d'exposition et d'utilisation de ces données de références au sein d'un système d'information d'entreprise, autant du coté opérationel que analytique.
Ce procédé, permet de répondre en partie à la problématique de la qualité des données, en définissant un cadre de données dites de références, sures, et limite ainsi l'entropie des données intégrées au DataWarehouse, mais n'effectue pas à proprement parler de nettoyage des données, thème qui sera abordé dans la suite de la synthèse

## 3.2 Principes

L'hypothèse de base est la suivante : *"En assurant la qualité sur les données de références, on limite les erreurs lors de l'alimentation et l'exploitation de l'entrepot de données"*

### 3.2.1 Données de références, késako ?

Les données de références sont un sous enssemble des données opérationelles, qui ont la praticularité de ne pas êtres issues d'opération de transactions. Ainsi elle possèdent une certaine constance dans le temps, qui n'est cependant pas une invariance, ces données pouvant être modifiées, complétées voire étendues. Ce sont ces mêmes données qui vont définir les axes d'exploration, d'exploitation et d'analyse.
On différencie trois grandes catégories de données de références.

- Produit : Chaque entreprise possède une quantité de réféence produits, qui peuvent êtres transersaux à plusieurs secteurs de l'entreprise. Typiquement, un produit pourra être référencé par une documentation technique issue d'un bureau d'étude, une opération de vente ou encore un référenciel fournisseur. L'unicité devra donc être assurée sur l'enssemble des entrées dans ce domaine.
- Tiers : De façon similaires, les "tiers" d'entreprises sont aussi considérés comme données de références. Par tiers nous entendons toutes personne ou entité ayant une intéraction possible avec le système d'information, typiquement un collaborateur, un client ou encore un fournisseur.
- Finance : Les données de finances sont des informations critiques pour le fonctionnement de l'entreprise, obligatoire en ce qui concerne n'importe quel aspect légal et primoridal en ce qui concerne le pilotage des activités. Ces deux approches sont intégrées aux données de références.

Le manque de qualité des données coûte 600 milliards de dollars à l'économie américaine chaque année. (Interaction between Record Matching and Data Repairing, Wenfei Fa et ali., 2011).
Ce constat alarmant montre la nécessité de s'intéresser au problème de la qualité des données, afin de le corriger en amont (faire la prévention, par le biais de M.D.M. par exemple), mais aussi en aval (de la correction, par le biais de "data cleaning"). Le data cleaning est un sujet d'étude finalement assez récent, mais qui semble prometteur, puisque le marché du data cleaning est en hausse de 17Une vrève définition de ce qu'est le data cleaning s'impose : l'objectif du data cleaning est de supprimer les erreurs et les incohérences d'un base de données afin d'améliorer la qualité des données.
Cependant, avant d'aborder plus avant le sujet du data cleaning à proprement parler, il est nécessaire d'aborder le sujet de la qualité de données. Comprendre l'origine et la diversité des problèmes de la qualité des données est nécessaire pour correctement aborder le sujet du data cleaning.
Nous verrons donc dans un premier temps quels peuvent être les différentes origines de la mauvaise qualité de données.
Les données fournies au système de nettoyage de données sont la plupart du temps d'origines diverses, et elle proviennent notamment souvent de bases de données différentes.
Ces origines diverses sont à l'origine de plusieurs problèmes concernant la qualité des données.
Voici un schéma récapitulant brièvement les différents problèmes concernant la qualité des données (source : Fig. 2, Data Cleaning : Problems and Current Approaches, Erhard Rahm et al.)

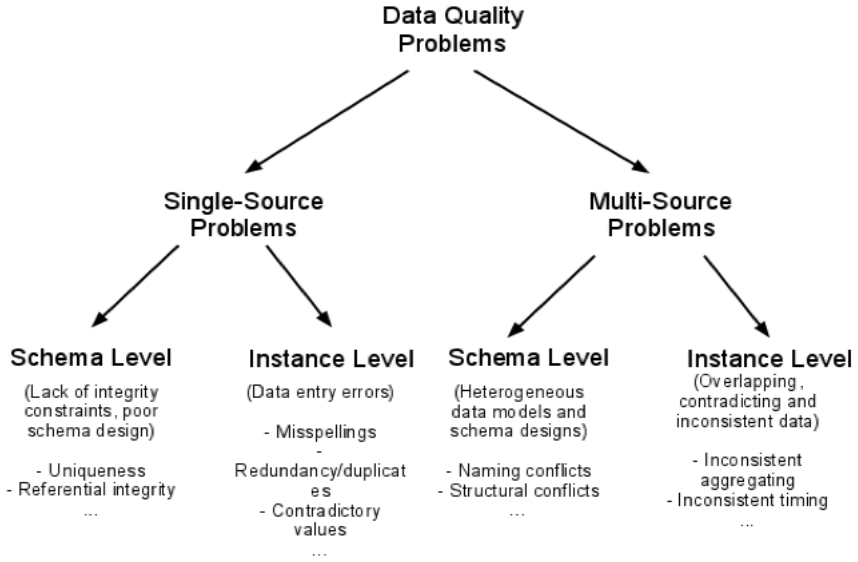


FIGURE 1 –

Nous voyons donc qu'il est possible de séparer les problèmes de qualité de données selon deux origines principales :

- les problèmes internes à une source de données unique
- les problèmes liés à des données ayant des sources multiples

Il est aussi possible de distinguer deux autres sous-catégories de problèmes liés à la qualité des données :

- les problèmes liés au modèle de données (le schéma de données)
- les problèmes liés aux données elles-mêmes (incohérences au niveau des données entre autres)

Nous aborderons dans un premier temps les problèmes liés aux sources de données uniques.
Les problèmes liés aux modèles de données sont principalement des problèmes liés à de mauvaises définitions du modèle données (violation de contraintes d'intégrité, unicité non respectée, valeurs illégales, etc....).
Les problèmes liés aux données elles-même sont quant-à-eux un peu plus variés. Il peut s'agir de problèmes liés à des valeurs manquantes ou erronées (fautes de frappes, d'orthographe, abréviations, erreurs de champs, etc...), à des incohérences entre plusieurs valeurs d'une même données (exemple : ville Paris, et code postal 42000), ou encore à des incohérences entre différentes données (données enregistrées plusieurs fois - et éventuellement de manière légèrement différentes, données incohérentes entre-elles).
Il est évident que ces problèmes se retrouvent souvent combinés entre-eux, et ainsi créer des problèmes bien plus complexes.
Les meilleurs moyens de résoudre (la plupart) de ces problèmes consiste à introduire un modèle de données fiable et cohérent laissant le minimum de libertés aux données, évitant ainsi un maximum d'erreurs.
Nous verrons ensuite les problèmes liés aux sources de données multiples. Ceux si sont sensiblement aggravés en comparaison des problèmes liés aux sources de données uniques, et sont bien plus nombreux.
Les problèmes liés aux modèles de données sont assez nombreux. Tout d'abord, l'un des problèmes provient du fait que les schémas de données des multiples sources sont différents.
Il faut donc dans premier temps convertir toutes les sources de données vers un schéma unique. Il faut à ce moment là définir deux autres types de problèmes :

- les problèmes de nommage : un même nom d'attribut correspondant à plusieurs types d'attributs différents (homonymes, par exemple id) ou plusieurs noms d'attribut correspondant finalement à unique type d'attribut (synonymes, name et nom par exemple).
- les problèmes structurels : ce sont des problèmes dus à des représentations de la même donnée de manières différentes (pas le même types de données - bool, string -, contraintes d'intégrité différentes, etc...).

Les problèmes liés aux données elles-mêmes sont assez similaires aux mêmes problèmes liés aux sources de données uniques (duplicité des données, incohérences, etc...), combinés aux problèmes de représentation différentes des données. Même si ces problèmes de représentation ne semble pas toujours présents au premier abord (même nom d'attributs, même types de données) il faut rester prudent quant à l'exploitation des résultats (interprétation des données différentes par exemple - prix en dollar ou en euro -...).
La principale difficulté posée par la présence de sources de données multiples est en fait la difficulté de déterminer quelles sont les données se référant à une même entité réelle.
Si la plupart des problèmes liés aux modèles de données peuvent - et doivent dans la mesure du possible - être corrigés en amont, il n'est pas toujours possible de faire de même concernant les problèmes liées aux données elle-mêmes. C'est donc principalement sur ce problème que s'attardent les solutions de data cleaning.
Pour résumer, les données, une fois présentées via le même schéma de données, présentent encore des problèmes de cohérence entre-elles.
Il faut donc être capable de déterminer ces incohérences et des les corriger. Ce problème peut se décomposer en deux sous-problèmes principaux :

- déterminer les données se référant à une unique entité du monde réel (record matching, implémenté par la plupart des systèmes de data cleaning)
- corriger ces données incohérentes, et supprimer les données redondantes (data repairing, ou merge/purge, que seuls quelques systèmes de data cleaning intègrent)

Pour le record matching, dans le meilleur des cas, il y a un attribut unique (ou un groupe d'attributs unique) permettant d'identifier clairement les données, et donc de déterminer si deux entrées se réfèrent à la même entité réelle.
Cependant, si on ne dispose pas d'attribut permettant d'identifier les données, ou si ces données sont de mauvaise qualité, il est impossible de déterminer si deux entrées se réfèrent à la même entité réelle par de simple comparaison d'attributs. Il est donc nécessaire d'introduire un système de "fuzzy matching" (correspondance floue ?).
Ce système fait intervenir des "règles de correspondance" permettant de déterminer si deux entrées correspondent ou non à la même entité. Ces règles permettent de déterminer le degré de correspondance de deux entrées, souvent exprimé par un chiffre entre 1 et 0.
Pour chaque entrées, chaque attribut est pris en compte, avec un poids différent, pour le calcul du degré de correspondance.

### 5.1 Recherche sur le sujet

Les systèmes de data cleaning traitent le record matching et le data repairing en tant que deux processus indépendants et séparés.
Cependant, dans, certains cas, ces processus peuvent interagir entre eux et s'aider l'un l'autre. La réparation aide à la "concordance" et la "concordance" aide à la réparation.
L'article Interaction between Record Matching and Data Repairing, Wenfei Fa et ali., 2011 propose une solution permettant d'unifier ces deux processus. Chacune des règles utilisées pour le record matching et le data repairing (les Conditional Functional Dependencies, les Conditionals Inclusion Dependencies et les Matching Dependencies) sont unifiées et traitées en tant que processus unique.
Cela permet d'améliorer la qualité des données finales.

## 6 Conclusion

## 7 Bibliographie

**Enjeu de la qualité des données**
BRASSEUR, Christophe. Data Management : qualité des données et compétitivité. Paris : Hermès, 2005. 164 p. Management et Informatique. ISBN 2-7462-1210-2

**Data Cleaning**
`http://www.lifl.fr/~bonifati/teaching/dq/lucidi/pods08Tutorial.pdf` `http://homepages.inf.`

ed.ac.uk/sma1/pubs/sigmod2011.pdf http://wwwiti.cs.uni-magdeburg.de/iti_db/lehre/dw/paper/data_cleaning.pdf http://qdb2011.dia.uniroma3.it/participants/program/p47-GOLAB.PDF

**Le MDM**

http://www.oracle.com/us/products/applications/master-data-management/018876.pdf, White Paper, Master Data Management, Oracle
http://www-01.ibm.com/software/data/master-data-management/library.html#White%20papers , White paper, "How Master Data Management Serves the Business", IBM