

Nettoyage des données en présence de données de référence

Elisa ABIDH, Julien LEVESY, Armand ROSSIUS

Table des matières

1	Introduction	3
1.1	Enjeux de la qualité de données	3
2	Démarche	4
3	Une réponse : le master data management	4
3.1	Présentation du procédé et objectifs	4
3.2	Données, de quoi parle-t-on ?	4
3.2.1	Donnée Transactionnelle	4
3.2.2	Donnée Analytique	5
3.2.3	Données de références, késako ?	5
3.2.4	Pourquoi définir un parc de données de références de qualité ?	6
4	Fonctionnement du M.D.M. dans un système d'information d'entreprise	6
4.1	Principe de fonctionnement	6
4.1.1	Fonctionnement de base d'un système de Master Data Management	6
4.1.2	Les conditions nécessaires au bon fonctionnement du système de M.D.M.	6
4.1.3	Les principaux modules d'un système de M.D.M.	7
4.1.4	Les processus du Master Data Management	7
4.2	Gestion spécifique selon le type de la donnée de référence et impact	8
4.2.1	Tiers	8
4.2.2	Produit	8
4.2.3	Financière	8
5	Positionnement au sein du SI de l'entreprise	8
5.0.4	Un peu d'histoire...	8
5.0.5	EAI : Intégration d'application opérationnelles dans le SI d'entreprise	8
5.0.6	Limites du modèle EAI & SOA	9
6	Présentation des offres du marché	9
6.1	Oracle	9
6.2	IBM	9
7	Les challenges	9
8	Conclusion	9
9	Bibliographie	9

1 Introduction

Le manque de qualité des données coûte environ 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 17%, alors que le reste du marché de l'informatique est "seulement" en hausse de 7%. Une brève définition de ce qu'est le data cleaning s'impose : l'objectif du data cleaning est de supprimer les erreurs et les incohérences d'une 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.

1.1 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 entreprises 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 maîtrise 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 gérées 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 [1] 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 contenu

- Justesse de l'information : en phase avec la réalité
- Adéquation aux besoins : réponds aux besoins réels
- Facilité d'interprétation : pas d'ambiguïté (abréviation, unités), compréhensible.

2. Accessibilité

- Disponibilité : disponible quand on 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.

2 Démarche

N'étant pas, par notre formation et nos affinités, des personnes ayant une grande connaissance des problématiques de gestion de données d'entreprise au sein d'un SI, nous avons décidé d'adopter une démarche d'ingénieur "ingénu" vis à vis du problème posé.

Par cela, nous entendons avoir une démarche progressive et critique vis à vis du vaste problème qu'est la qualité des données au sein d'un SI. Ce sujet étant extrêmement vaste et mal défini, nous ne souhaitons pas nous fermer de portes dans notre raisonnement.

De plus, nous souhaitons réellement avoir une approche qui permette de distinguer la réalité des intérêts technologiques des arguments commerciaux de bas étage, ces derniers étant plus que mis à contribution compte tenu du battage médiatique effectué autour de la problématique de la qualité des données. Nous tâcherons de passer outre ces aspects.

Ainsi, dans un premier temps, nous allons nous atteler à définir le domaine de la qualité des données, pour ensuite essayer d'effectuer un aperçu des approches possibles de la problématique de qualité des données, pour enfin déboucher sur le management des données de référence, son fonctionnement, son implémentation dans un SI d'entreprise et les différents challenges que ce type de service pose au sein d'une entreprise.

De cette manière nous espérons fournir un aperçu complet et critique de cette technologie.

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 comme fiables.

Outre cela, le Master Data Management englobe 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 côté opérationnel 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, considérées comme sûres, et limite ainsi l'entropie des données intégrées aux entrepôts de données, mais n'effectue pas à proprement parler de nettoyage des données, thème qui sera abordé dans la suite de la synthèse

A la différence d'un "simple" nettoyage instancié une ou plusieurs fois dans le parc de données de l'entreprise, le Master Data Management inscrit une démarche de qualité de données sur le long terme.

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'entrepôt de données"*

3.2 Données, de quoi parle-t-on ?

3.2.1 Donnée Transactionnelle

Chaque opération effectuée dans l'entreprise génère des données. Par exemple lors d'un achat les données générées sont (date de l'achat, quantité de produit acheté, montant transaction et cie...) Des

bases de données basées sur de l'OLTP sont utilisées pour la gestion de ce genre de transactions. Des techniques de gestion de ces données existent, pour gérer ces données, notamment via le nettoyage par données de références, que nous aborderons dans la suite de cette synthèse.

3.2.2 Donnée Analytique

Les données analytiques sont générées à partir des bases de données transactionnelles et des bases de données de références.

Il s'agit ici de traiter des données transactionnelles sur le plus ou moins long terme, les traitements et l'exploitation étant orientés en fonction de grands axes.

Deux principales approches se distinguent dans le monde de la business intelligence :

- Modélisation OLAP : Transformation des entrepôts de données d'entreprises en "hypercubes", permettant une exploitation facilitée (création de rapports, tableaux d'indicateurs...)
- Datamining : Recherche de regroupements de tuples en fonction de leurs attributs, dans les bases de données d'entreprise, de façon à effectuer de l'analyse prédictive entre-autres. Un exemple concret est le "pattern" de navigation internet d'un utilisateur moyen, permettant au final de "prédire" quelle sera sa prochaine étape de navigation

3.2.3 Données de références, késako ?

Les données de références sont un sous ensemble des données opérationnelles, considérées comme données de support pour les différentes opération d'alimentation ou d'exploitation des données du SI. Elles 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 référence produits, qui peuvent être transversaux 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 référentiel fournisseur. L'unicité devra donc être assurée sur l'ensemble des entrées dans ce domaine.
- Tiers : De façon similaires, les "tiers" d'entreprises sont aussi considérés comme données de références. Par tiers nous entendons toute personne ou entité ayant une interaction 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 primordial en ce qui concerne le pilotage des activités. Ces deux approches sont intégrées à la gestion de données de références.

Une question demeure en suspens... Quand une donnée normale peut être considérée comme donnée de référence ?

En effet, la simple différenciation basée sur le fait qu'une donnée est transactionnelle ou non, peut dans certains cas se révéler mise en défaut.

Prenons pour exemple une opération de vente du produit P1 à Mr M enregistrée par une application opérationnelle, cela se traduit par une entrée dans la table vente de l'entrepôt en relation avec les données de références produit...

Maintenant, est-il possible de considérer le fait que l'acheteur constitue en lui même une donnée de référence ? Mieux encore, la quantité de produit P1 vendue à ce client peut à son tour être considérée comme une donnée de référence...

Cet exemple met en évidence que la frontière entre une donnée de référence ou une donnée standard n'est pas clairement définie aux yeux de tous. Clairement, en suivant la logique précédente, la totalité de l'entrepôt de données de l'entreprise peut être considéré comme donnée de référence.

Plusieurs pistes pour répondre à cette question :

- Donnée de support pour l'alimentation et l'exploitation de l'entrepôt de données de l'entreprise. Dimensions d'analyse lors de l'exploitation en OLAP par exemple.
- Objets métiers partagés entre plusieurs applications de l'entreprise

3.2.4 Pourquoi définir un parc de données de références de qualité ?

Comme expliqué auparavant, les données de références au sein d'un système d'information servent d'axes d'exploitation et d'analyse. Ainsi chaque opération effectuée au sein de ce dernier est obligatoirement rattachée à une ou plusieurs données de référence. Si celles-ci sont corrompues, fausses, non unifiées, le risque d'augmentation de l'entropie au sein de l'entrepôt s'en trouve décuplé.

Par exemple, dans un repère orthogonal de dimension 3, caractérisé par les axes (x,y,z), positionner un point aux coordonnées (1,2,3) est quelque chose de relativement aisé, si et seulement si les valeurs 1,2 et 3 des axes sont garanties comme unique. Émettons alors l'hypothèse que non, les valeurs présentes ne sont pas uniques... La question de l'insertion d'un point aux coordonnées (1,2,3) se révèle alors beaucoup plus complexe, quelle valeur choisir ?

Nous nous retrouverions automatiquement en face d'un parc de n valeurs possédant toutes les caractéristiques (1,2,3) ... différentes !

Autre chose, recherchons maintenant tous les points résolvant la condition $y = 2$, ce qui peut s'apparenter à la recherche de caractéristique commune pour des entrées produit dans le cas réel. La encore, si "2 possède plusieurs valeurs" dans le repère, la tâche de regroupement s'avère encore plus complexe.

Imaginez les risques, sur une base de données opérationnelles, avec un nombre de tuples extrêmement grand ?

4 Fonctionnement du M.D.M. dans un système d'information d'entreprise

4.1 Principe de fonctionnement

Nous l'avons vu, il est indispensable d'avoir un parc de données de références de qualité si l'on souhaite pouvoir exploiter correctement ces données dans les applications opérationnelles du S.I., et ainsi maintenir une certaine cohérence. Cette cohérence est également indispensable pour l'exploitation de ces données par le système de Business Intelligence de l'entreprise.

4.1.1 Fonctionnement de base d'un système de Master Data Management

Afin d'expliquer clairement le fonctionnement d'un système de Master Data Management, il convient de présenter le fonctionnement d'un S.I. sans la présence de M.D.M. Sans système de M.D.M, chaque application du système d'information possède une base de données qui lui est propre, et y accède de manière informelle. Les modifications, créations et suppression de données ne sont donc pas tracées. De nombreux problèmes de cohérence et de corruption de données sont présents de par la présence de sources de données multiples, de données incomplètes, erronées ou incohérentes.

L'objectif du Master Data Management est de fournir une base de données unique et unifiée pour l'ensemble des applications opérationnelles du S.I.

Ainsi, l'ensemble des applications du S.I. partagent les mêmes informations, éliminant ainsi les problèmes de cohérence et de corruption des données.

La maintenance des données ne s'effectue donc qu'à un seul endroit, et est donc simplifiée. Le système de M.D.M. distribue ensuite les données aux applications qui en ont besoins de manière régulière (via un système de pull). Il est également possible de mettre en place un système de push où l'information est directement demandée par l'application en ayant besoin.

4.1.2 Les conditions nécessaires au bon fonctionnement du système de M.D.M.

Afin de fournir des données de qualité, il est indispensable que le système de M.D.M. respecte trois règles de base :

- L'ensemble des données de référence du S.I. doit être stocké en un unique endroit.

L'objectif de cette règle est tout d'abord d'empêcher la présence d'incohérence entre différentes

données due à l'existence de plusieurs sources de données. La présence d'une source unique de données permet également de simplifier le coût opérationnel dû à la maintenance des données.

- Le système est le maître des données. Grâce à cette règle, le système de Master Data Management permet un contrôle d'accès aux données. Il détermine quelle application peut accéder à quelle données, permettant de sécuriser les données, parfois sensibles, du système d'information.
- Les données du système de M.D.M sont des données de références. Cela permet d'empêcher des problèmes liés à la corruption de données, de données erronées, ou encore de données obsolètes. En effet, le système de M.D.M. est l'unique possesseur de l'information, et l'information qu'il possède est considérée comme étant sûre, de qualité. Ainsi, en cas de litige, doute sur la validité ou l'intégrité d'une données, une simple vérification auprès du système de M.D.M permet de s'assurer de la validité des données, puisqu'il est la base de référence des données.

4.1.3 Les principaux modules d'un système de M.D.M.

L'architecture basique d'un système de M.D.M. est composée de 6 principaux blocs :

La gestion du cycle de vie : Ce module permet définir et implémenter tous les processus, rôles et responsabilités liés à la modification, création ou suppression de données.

L'administration : Ce module se charge de la gestion des différents acteurs, et de leurs droits d'accès concernant les données.

Le stockage : Cette partie concerne la manière dont sont stockées les données et les références entre-elles.

La gestion des méta-données : Ce module se charge de la gestion de l'ensemble des données concernant les données de références (les méta-données), comme par exemple les dates de dernières modifications, etc...

La gestion de l'accès aux données : Ce module permet de définir tout ce qui concerne l'accès aux données. Cela comprend les interfaces d'accès aux données, mais aussi de création, suppression et modification de celles-ci. Il est aussi en charge des protocoles de transmission des données, ainsi que de la politique d'accès aux données (push, pull...).

Les règles directrices : Ces règles directrices assurent la conformité du système avec des règles de base (format des données, attribut d'une donnée indispensable ou facultatif). Ce sont des règles logiques permettant d'assurer la qualité des données selon un modèle adapté à l'entreprise. Ces règles sont implémentées par le biais de routines permettant de contrôler la conformité des informations.

4.1.4 Les processus du Master Data Management

Maintenant que nous avons vu les principaux modules de l'architecture d'une solution de Master Data Management, voyons quels sont les processus indispensables qu'un système de M.D.M. doit implémenter.

Profiler : Chaque source de données de références que l'on souhaite intégrer au système de Master Data Management doit être vérifiée, et notamment en terme de qualité. L'objectif du Master Data Management est d'améliorer la qualité des données en utilisant une base que l'on considère comme étant saine. Il est donc indispensable que la base que l'on utilise soit effectivement saine, et il faut donc s'assurer que chaque source utilisée pour créer la base de référence est de qualité. C'est l'objectif de ce processus, qui s'assure de la complétude des données, vérifie que les données sont bien dans la plage qui leur correspond, etc...

Consolider : La consolidation est la partie la plus importante du Master Data Management. En effet, l'objectif du Master Data Management étant de fournir une base de données uniques pour de multiples applications, la consolidation est donc l'élément central du M.D.M.

Gouverner et nettoyer : Le nettoyage de données a pour but la standardisation des données, la correction d'erreur, l'établissement de la correspondance de données, la suppression de doublons. Gouverner les données consiste à établir des règles sur les données, sur la qualité des données, et à définir des politiques d'accès aux données.

Partager : Il est indispensable de combiner le système de M.D.M. à une organisation du S.I. de type S.O.A. afin de faire profiter les processus métiers des données de référence.

Exploiter : L'objectif est d'exploiter les données collectées dans le M.D.M. L'objectif est de maintenir une "vue à 360°" et une référence croisée avec les bases du Datawarehouse.

4.2 Gestion spécifique selon le type de la donnée de référence et impact

Les données métiers pouvant se révéler complexes, notamment à exploiter, la majeure partie des systèmes de Master Data Management se spécialise dans la gestion d'un type particulier de données de références.

4.2.1 Tiers

Les systèmes de M.D.M. spécialisés dans la gestion des données référentielles "Tiers" sont les M.D.M. traitant plus particulièrement les données liées à des personnes physiques/morales (personnel, client, fournisseur, etc...). La principale difficulté liée à ces systèmes de M.D.M. est la consolidation. Il est extrêmement important d'être capable de rapprocher plusieurs informations provenant de sources diverses, de les enrichir, et de contrôler leur validité.

4.2.2 Produit

Les systèmes de M.D.M. spécialisés dans la gestion des données référentielles "Produit" sont principalement utilisés dans la grande distribution (pour les processus de référencement) et dans l'industrie (pour les processus de développement de produit). La principale difficulté liée à ces systèmes de M.D.M. est la notion de gestion collaborative des données de références. En effet, la fiche produit est mise à jour à partir de nombreux acteurs et de nombreux métiers. Cela nécessite une coordination fine qu'il n'est pas à négliger.

4.2.3 Financière

Les systèmes de M.D.M. spécialisés dans la gestion des données référentielles "Financières" sont utilisés pour deux raisons principales :

- l'établissement des comptes annuels
- les rapports de gestion

Les données financières sont souvent modélisées sous forme hiérarchique, c'est pourquoi de nombreuses solutions se sont spécialisées dans ce domaine en utilisant des modèles de données hiérarchiques plutôt que les modèles de données traditionnels.

5 Positionnement au sein du SI de l'entreprise

5.0.4 Un peu d'histoire...

Historiquement, à l'âge (pas tant) de pierre (que ça) du système d'information, chaque application opérationnelle possédait son propre SGBD dédié à l'application... Celle-ci ne possédait que les données qui lui étaient utiles, que ce soit de référence, ou de simple transaction.

Le problème de la propagation des mises à jour des données est alors posé, car laissé à la responsabilité de l'opérateur, et comme le dit l'adage " *La seule source d'erreur possible dans un ordinateur se trouve entre la chaise et le clavier !*".

La continuité logique des choses est donc d'essayer de "faire communiquer" les différents SGBDs entre eux... Vient donc la problématique de l'intégration n-carrée : chaque application est raccordée directement aux multiples bases de données qu'elle utilise, sans réel moyen de contrôle de la mise à jour de ces dernières... Fort risque de corruption lors de la propagation de données, de création de doublons sur certaines entrées et aucune trace des modifications portées.

Ce système s'est donc révélé catastrophique en terme de maintenance et de qualité des données, mais il avait au moins le mérite d'avoir permis d'identifier une solution possible à la propagation des données au sein d'un SI : il faut contrôler et uniformiser les modes de communication entre les différentes bases de données

5.0.5 EAI : Intégration d'application opérationnelles dans le SI d'entreprise

Compte-tenu des expériences décrites précédemment, les développeurs ont orienté la démarche vers la création d'un bus commun de communication entre les différentes entités du système d'information. Ainsi

ce service fourni sera en charge de l'archivage et du transit des données de l'entreprise le tout de façon générique, moyennant le développement de services "connecteurs" entre les applications et le système de communication, appelé Entreprise Service Bus, ou ESB.

- Applications Opérationnelles : Applications métier de l'entreprise, raccordées à l'ESB,
- Synchronisation des données basées sur les méta données. Toutes les informations traitant des opérations à effectuer sur les différentes bases sont stockées à part. Ainsi la tâche de synchronisation du contenu est externalisée. Cela est aussi appelé ESB.
- Les fonctionnalités clés des applications métiers sont maintenant exposées comme "services" dans un système d'orchestration. Ainsi, il est possible de définir plusieurs processus d'orchestration de services. En ce qui concerne le MDM, il ne s'agit pas de seulement être en A2A (Application to application), mais aussi d'exposer les données de références à la couche d'orchestration.

5.0.6 Limites du modèle EAI & SOA

Les architectures de gestion de systèmes d'informations présentées précédemment présentent l'avantage de faciliter l'intégration de nouveaux services à un système d'information d'entreprise, en s'occupant principalement de la problématique de communication / synchronisation des données entre plusieurs applications opérationnelles. En d'autres termes, ils sont conçus pour gérer et limiter les problèmes de fragmentation, mais ils ne les éliminent pas.

6 Présentation des offres du marché

6.1 Oracle

6.2 IBM

7 Les challenges

8 Conclusion

9 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

Références

- [1] BRASSEUR Christophe. *Data Management : qualité des données et compétitivité*. Paris : Hermès, 2005.