F.S.D……N° D'ordre :

Ministère de l'Enseignement Supérieur et de le Recherche Scientifique

Université Saad Dahlab de Blida

Faculté des sciences

Département d'informatique

Mémoire Présenté par

Chelfi Lyes

En vue d'obtenir le diplôme de Master

Domaine : Mathématique et Informatique. MI
Filière : Informatique.
Option : Ingénierie des logiciels.

Titre

Mise en œuvre d'un Framework pour la qualité des données décisionnelles

Promoteur : M. Bala Mahfoud, Maître Assistant

Organisme d'accueil :

Soutenue le: **devant le jury composé de :**

M. Ferfera Soufiane	**Maître Assistant**	**Président**
M. Hadj Yahia Ouahid	**Maître Assistant**	**Examinateur**
Mlle. Azzouz Mahdia	**Maître Assistant**	**Examinateur**

- promotion 2010/2011-

Remerciements

En tout premier lieu, je remercie le bon dieu, tout puissant, qui m'a éclairé le bon chemin et qui m'aide à réaliser dans les meilleures conditions, ainsi que Je tiens aussi à exprimer toute ma reconnaissance à mon promoteur M. Mahfoud BALA pour son encadrement, son conseil et son soutien tout au long de mon mémoire.

Je tiens à remercier sincèrement l'ensemble des membres du Jury qui me font grand honneur d'avoir accepté de juger mon travail.

Je voudrais remercier tous mes camarades de la promotion Master 2

Enfin je souhaite remercier tous les membres de ma famille qui m'ont beaucoup aidé et encouragé à la réalisation de ce travail.

Résumé

La gestion de la qualité des données est un problème clé au sein de toutes les organisations qu'elles soient privées ou publiques. Une large collection d'outils commerciaux et open source sont proposées pour gérer les problèmes de qualité des données dans les systèmes décisionnels (entrepôts des données) et les bases de données classiques. Cependant, chacun de ces outils a sa propre vision de la qualité des données. D'un côté, faire interagir ces outils entre eux, demeure un défi technique en raison de l'hétérogénéité de leurs modèles et méthodes d'accès. D'un autre côté, les utilisateurs de la qualité exigent de plus en plus des outils uniformes et qui contient plusieurs techniques pour identifier et corriger des problèmes complexes de la qualité des données. Ce document présente l'architecture d'un nouveau Framework **Data Quality** qui sert à évaluer et améliorer la qualité des données dans les systèmes décisionnels. Ce Framework contient trois principaux composants (data Auditing, data Cleansing et data reporting).

Mots clés : Qualité des données, dimension qualité des données, audit des données, nettoyage des données.

ملخص

إدارة جودة البيانات هي القضية الرئيسية في أي موسسة سواء كانت خاصة أو عامة, توجد مجموعة كبيرة من الأدوات التجارية ومفتوحة المصدر لإدارة مشاكل جودة البيانات في أنظمة صنع القرار (مستودع البيانات)أو قواعد البيانات التقليدية. ومع ذلك ، كل أداة لها رؤيتها الخاصة لجودة البيانات.وتفاعل هذه الأدوات مع بعضها البعض، لا يزال يشكل تحديا تقنيا بسبب عدم تجانس نماذجها وأساليب الوصول إلى البيانات. هذا من ناحية, ومن ناحية أخرى المستعملون يطالبون بنوعية أدوات أكثر فاعلية في تحديد وتصحيح مشاكل البيانات المعقدة.

تعرض هذه الورقة بنية أداة يتم استخدامها لتقييم وتحسين نوعية البيانات في أنظمة اتخاذ القرارات. هذه الأداة تحتوي على ثلاثة عناصر رئيسية هي (تدقيق البيانات، تطهير البيانات و تقارير البيانات).

الكلمات الرئيسية:نوعية البيانات، أبعاد البيانات الجيدة، وبيانات الحسابات، والبيانات التنظيف.

Table des matières

Table des figures

Introduction générale

Contexte

Aujourd'hui, les systèmes d'information ont connu une évolution importante est dans tous les domaines, le rôle de l'information est ainsi développée dans des différents domaines, c'est-à-dire a l'époque, les données sont utilisées dans les systèmes de gestion classique comme (OLTP), dans les dernières années les données de plus en plus sont utilisées dans les applications d'aide à la décision exemple (OLAP, Data Mining …), et pour cela les données ayant un impact très élevé.

Face à cette évolution, les utilisateurs sont intéressés plutôt par la qualité des données exploitées, car une qualité des données médiocre expose les entreprises et les organisations au risque de compromettre leurs décision et influence l'exécution des systèmes, il est donc nécessaire de déterminer la qualité de ces données, le problème qu'il n'existe pas d'une définition consensuelles sur la qualité des données, car chaque utilisateur a sa propre définition selon leurs domaine de recherche ou d'application.

Il existe pas mal des recherche centrées sur la qualité des données dans des divers domaines tel que [3], [11], [12], [13], qui parlent de la définition de la qualité des données, les approches et les techniques d'évaluation et d'amélioration de la qualité des données.

Dans le domaine des systèmes décisionnels, la qualité des données joue un rôle stratégique, les conséquences de la non ou mauvaise qualité des données quant prend décision et coûte les utilisateurs très cher, donc il est nécessaire de connaitre la qualité de leur données, puis l'engagement dans une initiative d'amélioration de la qualité de ces données pour faire des analyse fiable.

Motivations et problèmes

Les applications engendrées dans les systèmes décisionnels exploitent les données qui sont extraites de divers sources des données hétérogènes, ainsi que ces données souffrent plusieurs problèmes d'inexactitude, valeurs manquant, d'incohérence, dédoublons…etc.

Actuellement il existe beaucoup de recherches sur la qualité des données selon plusieurs approches et techniques d'évaluation et d'amélioration de la qualité des données dans les différents domaines, et même le domaine des systèmes décisionnels, ainsi que des solutions commerciales et open source qui focalisent sur la garantie de la qualité des données dans les systèmes décisionnels.

En effet, il y a des limites dans les outils de qualité des données existantes quand la détections et la correction des problèmes de la qualité des données telle que l'élimination des doublons et le traitement des valeurs manquantes.

Ces limites mettent en évidence l'importance d'avoir une plate-forme de gestion de la qualité qui gère les problèmes énumérés précédemment, et contient des techniques d'évaluations et d'amélioration de la qualité des données dans les systèmes décisionnels (ODS) mais aussi dans les bases de données classiques.

Objectifs

Dans ce mémoire de master nous nous intéressons à la qualité des données dans le système décisionnel, dont l'objectif peut se résumer comme suit:

- proposer une modélisation pour l'évaluation de la qualité dans les systèmes décisionnels
- proposer une architecture d'un Framework qui assure la qualité

Organisation du mémoire

Ce mémoire est organisé en quatre chapitres de la façon suivante :

- ✓ chapitre 1 : ce chapitre présente une étude bibliographique sur les systèmes décisionnels, et la technique d'extraction des données ETL (Extraction, Transformation, et Loading), puis une étude bibliographique approfondie sur la qualité des données, il présente aussi des outils traitants sur la qualité des données avec une comparaison entre eux.
- ✓ chapitre 2 : dans ce chapitre nous présentons en détail les deux techniques de détection et de correction des données (audit et nettoyage des données).
- ✓ chapitre3 : présente la modélisation de notre travail, puis l'architecture de notre Framework et ses composants.
- ✓ chapitre 4 : nous présentons le diagramme de déploiement de notre Framework ainsi que des exemples d'exécution.

Chapitre I Etat de l'art

Introduction

Le développement technologique et la généralisation de l'informatique ont révolutionné notre société et l'entreprise en particulier.

Les données constituant la mémoire de l'entreprise se trouvent concernées dans cette évolution technologique puisqu'elles ont été stockées, au fur et à mesure, dans des environnements et avec des formats différents. L'entreprise dispose donc de quantités énormes de données d'une valeur inestimable touchant ses différents processus métiers.

Dans un contexte caractérisé par la mondialisation et une concurrence accrue, l'entreprise se doit d'exploiter ces quantités de données stockées pendant des années et ce pour une capitalisation de son expérience et un retour sur investissement. Pour ce faire, un processus de préparation de données devra être entrepris afin de disposer de données ayant les qualités requises pour une exploitation réelle et rationnelle.

En se plaçant dans un contexte décisionnel, d'analyse et d'évaluation, la qualité des données devient un challenge. Comme tout système, le processus de préparation de données doit lui aussi se soumettre à des normes de qualités pour assurer à l'entreprise des données fiables, significatives, cohérentes, etc.

Afin de comprendre ce qu'est la qualité des données dans toutes ses dimensions, ce chapitre présente un état de l'art sur les systèmes décisionnels, les processus ETL et situe la qualité des données dans ce contexte.

1. Systèmes décisionnels

L'informatique décisionnelle s'insère dans l'architecture plus large des systèmes d'informations. Elle désigne l'ensemble des méthodes et outils permettant à une entreprise de mettre en place son projet décisionnel. Ces outils informatiques collectent, modélisent et restituent les données de sources internes ou externes afin de permettre, aux responsables, l'obtention d'indicateurs pertinents. Ils peuvent ainsi mesurer la performance de l'entreprise et prendre la meilleure.

1.1Définition

Un Système d'Information Décisionnel(SID) et parfois appelé « ***le décisionnel*** » désigne les moyens, les outils et les méthodes qui permettent de collecter, consolider, modéliser et restituer les données d'une entreprise en vue d'offrir une aide à la décision. Ce terme est l'équivalent français de « **Business intelligence** » [1]

Le Système d'Information Décisionnel (SID) correspond à :

L'exploitation coordonnée et cohérente des informations de l'établissement et de son environnement dans le but de faciliter la prise de décision par les décideurs

Vision transversale de l'établissement grâce à des informations en provenance de différents métiers

Entièrement dédié au pilotage de la performance, il met en œuvre une grande richesse de fonctions.

1.2. L'évolution des systèmes décisionnels

Les origines du système informatique d'aide à la décision remontent au début de l'informatique et des systèmes d'information. Ces systèmes ont connu une grande et complexe évolution liée notamment à la technologie. Cette évaluation se poursuit jusqu'à aujourd'hui.

Au début des années **1960**, le côté logiciel de l'informatique consistait à la création d'applications individuelles qui ont été effectuées en utilisant le recueil principal (dossier

principal d'un programme contenant des paramètres et des définitions essentielles à son fonctionnement).

Vers le milieu des années **1960**, la taille des fichiers « recueils principal » explose et produit d'énormes quantités de données redondantes. Ce qui représente de sérieux problèmes : le besoin de synchroniser les données, la complexité de gestion, la complexité d'ajout de nouveaux programmes et le besoin de matériel adéquat pour gérer cette situation.

Au début des années **1970**, l'apparition des disques à accès direct a permis l'accès direct aux données, d'où la réduction du temps d'accès à l'ordre de millisecondes. Du côté de l'offre logicielle, un nouveau type de logiciel. Il s'agit de systèmes de gestion de base de données dont est de faciliter le stockage et l'accès aux données [1].

Au milieu des années **1970**, l'apparition des systèmes OLTP rend encore plus rapide l'accès aux données. Cette nouvelle technologie a donné la possibilité d'utiliser l'ordinateur pour de nouveaux systèmes tel que : système de réservation, système bancaire, …etc.

C'est dans ce contexte, au milieu des années **80**, qu'apparaît l'ébauche d'un autre système d'information spécialement dédié à l'aide à la décision : l'informatique décisionnelle, en anglais **Business Intelligence**.

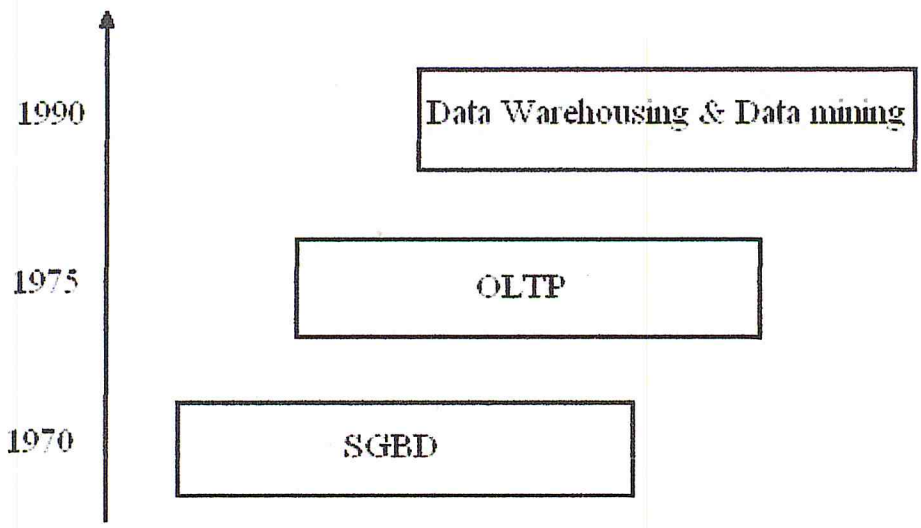

Fig. I.1 : L'évolution des systèmes décisionnels

1.3. Architecture générale d'un SID

Les outils de la plate-forme décisionnelle sont organisés en 4 classes, selon l'activité que chacun d'entre eux permet de réaliser dans la suite décisionnelle :

- La collecte des données au moyen des outils **ETL**,
- Le stockage des données dans l'entrepôt de données (DataWarehouse),
- Les analyses multidimensionnelles avec les outils OLAP,
- Les explorations de données avec les outils de Datamining, et présentation des résultats.

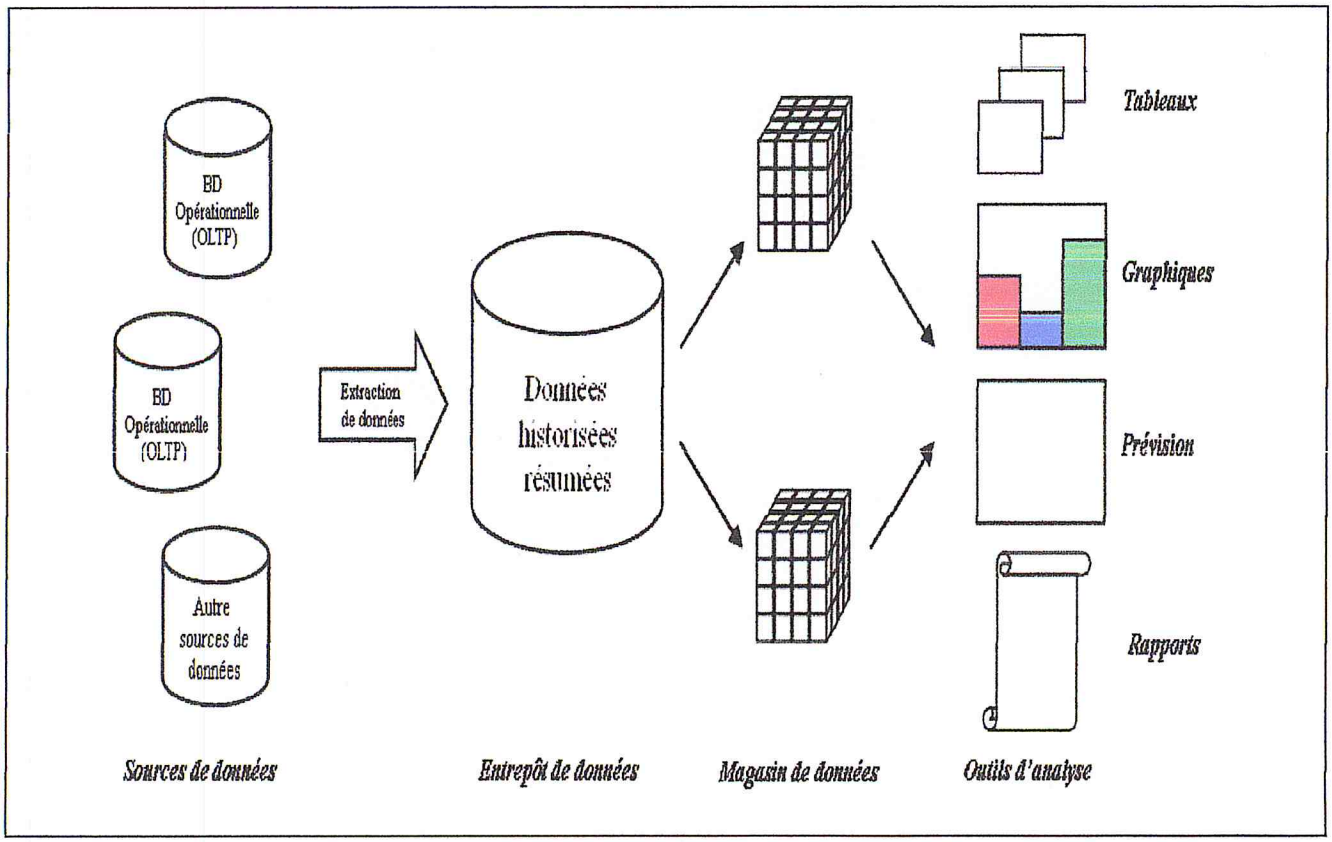

Fig. I.2 : Architecture générale d'un SID [1]

2. Aperçu général sur l'Extraction, Transformation et Chargement des données (ETL)

L'architecture du système décisionnel (fig. I.2) montre trois grandes phases à savoir acquisition, stockage, et analyse des données. L'acquisition des données elle-même se décompose en trois étapes principales qui sont l'Extraction, Transformation et Chargement des données [2].

Ce sont des processus amont (back-end) qui couvrent l'extraction des données à partir des systèmes source. Ensuite, ils comprennent toutes les fonctions et les procédures de nettoyage, filtrage, transformations vers des formats et des structures appropriés pour le stockage dans l'entrepôt de données. Les transformations traitant, entre autres, les agrégations consistant à calculer de nouvelles informations et donc obtenir des informations d'un niveau d'agrégation plus élevé. Les données sources sont très détaillées et donc extraites avec un niveau de granularité très fin. Ce niveau de détail pour certaines donner peut s'avérer inutile pour l'analyse et l'évaluation dans un système décisionnel.

Après la préparation des données, la dernière étape du processus consiste à déverser physiquement les données préparées dans l'entrepôt de données.

2.1 L'extraction de données

C'est la première étape dans le processus ETL, elle permet la collection des données sélectionnées à partir des diverses sources et ce en utilisant des outils appelés Wrapper travaillant de façon native avec les SGBD ou par la création de programmes « extracteurs »,

L'extraction efficace des données est la clé de réussite de l'entrepôt de données. Par conséquent, il est nécessaire de formuler une stratégie d'extraction de données pour l'entrepôt

de données et d'accorder une attention particulière aux questions ci-dessous, Voici une liste de questions liées à l'extraction de données [2] :

- Identification de sources : identifier les applications sources et les structures sources.
- Méthode d'extraction : définir pour chaque source de données, si le processus d'extraction est manuel ou automatique.
- Fréquence d'extraction : pour chaque source de données, établir la fréquence d'extraction des données : quotidienne, hebdomadaire, trimestrielle, et ainsi de suite.
- Intervalle de temps : pour chaque source de données, on précise l'intervalle de temps pour le processus d'extraction

2.2Transformation des données

Indépendamment de la variété et la complexité des sources des systèmes opérationnels, et quelle que soit l'étendu de l'entrepôt de données, les données extraites sont des données brutes. En d'autres termes, ces données ne remplissent pas les conditions exigées par l'entrepôt de données et ne disposent pas de la qualité requise pour être exploitées à des fins d'analyse et pour la prise de décisions. Il faudra donc les enrichir et améliorer leur qualité avant de les charger dans l'entrepôt de données, puis les transformer et les intégrer selon les normes puisqu'elles proviennent à partir de plusieurs systèmes sources disparates [2].

2.3 Chargement des données

Le chargement des données dans l'entrepôt est la tâche la moins complexe par rapport aux deux autres. Dans cette étape, il s'agit de charger les données préparées dans l'entrepôt de données. Pour ce faire, il faut créer des mappages de chargement entre les données sources et les données ciblent puis définir une stratégie de chargement pour assurer le bon chargement de données.

L'ensemble du processus de déplacement des données dans l'entrepôt de données référentiel est fait avec plusieurs façons. Le chargement des données, et rafraîchissante des données, le processus de déplacement est comme indiqué ci-dessous :

- Chargement initiale : peupler toutes les tables d'entrepôt de données pour la première fois -Application des modifications en cours si nécessaire de manière périodique - Actualisation : en effaçant complètement le contenu d'une ou plusieurs tables et effectuer le rechargement.

Le chargement des données doit se faire en mettant l'entrepôt de données en mode déconnecté pour éviter de perturber l'utilisateur [2].

3. Opérationnel Data Store

Introduction

Dans le processus de création de l'entrepôt de données, la premier phase consiste en la mise en place de la récupération des données du système de production, ensuite vient le nettoyage des données récupérées pour enfin on arrive à la construction de l'entrepôt de données.

La récupération des données venant du système de production est une étape cruciale. En effet, pour permettre le nettoyage des données et la construction de l'entrepôt de données, nous devons manipuler les données de production sans pour autant influencer les processus existants qui les utilisent, c'est pourquoi nous avons besoin d'un endroit dans lequel nous pouvons manipuler et nettoyage ces données. Cet endroit est l'Opérationnel Data Store (ou O.D.S.) [4].

3.1Définition :

Un ODS est un environnement où les données sont découlées de différentes bases de données opérationnelles. Le principale but est de fournir un lieu intermédiaire permet à l'utilisateur de l'entrepôt de données à prépare les données intégré a partir de divers source de données et l'assurance de la qualité des données avant le chargement de ces données dans l'entrepôt de données [20].

3.2Caractéristiques d'un ODS

Nous pouvons qualifier un O.D.S. comme étant une structure de données

- ✓ Orientée sujet: les données collectées doivent être orienté métier.

utilisateurs. Les données sont nettoyées avant leur chargement dans l'entrepôt de données, la fig. I.3 montre le positionnement de l'ODS dans l'architecture du système décisionnel.

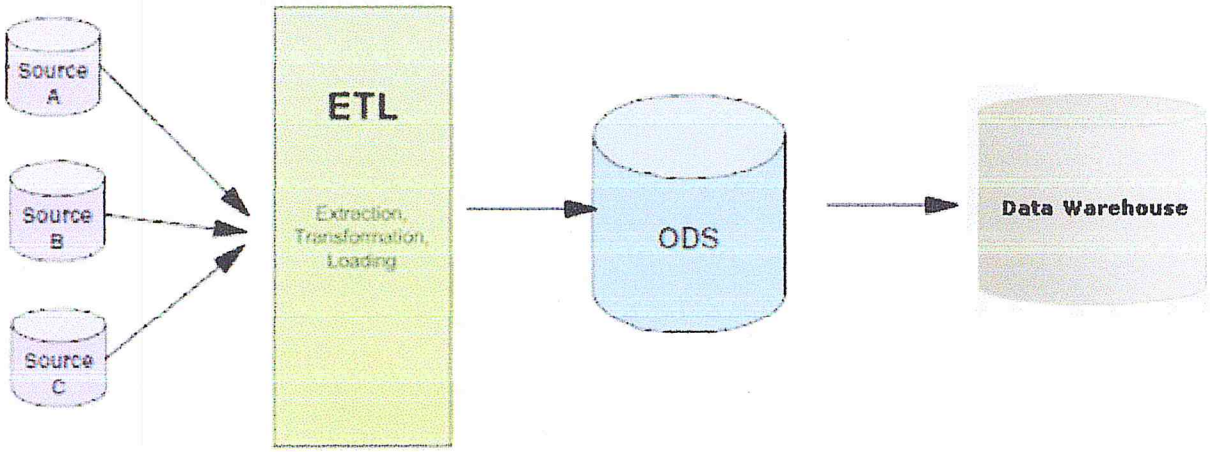

Fig. I.3 : Positionnement de l'ODS dans l'architecture le système décisionnels[20].

4. La qualité des données

Introduction

Durant ces dernières années, les sociétés et les administrations créent et stockent de grandes quantités des données. Il est évident qu'une mauvaise qualité des données affecte très négativement l'efficacité de ces entreprises. Ces impacts négatifs se manifestent par d'énormes coûts directs ou indirects. D'après une étude du **Gartner Group**, la plupart des initiatives de réingénierie de l'information n'aboutissent pas à cause d'un manque de qualité au niveau des données [4], Pour y remédier, nous devons avoir une approche structurée de la qualité des données. Nous étudions dans ce qui suit une approche hiérarchique de l'amélioration de la qualité des données et les techniques les plus courantes pour arriver à un nettoyage optimal des données.

4.1 Définition

Dans la littérature actuelle, il n'existe pas de définition standard de ce qu'est la qualité des données. Toutefois, l'ISO[1] nous fournit une définition acceptable de la qualité des données en utilisant la terminologie acceptée dans le domaine de la qualité. Parmi les nombreux standards ISO, nous trouvons l'ISO 8402: Quality Management and Quality Assurance Vocabulary.

Cette norme donne une définition formelle de la qualité :
« L'ensemble des caractéristiques d'une entité qui conduisent à une capacité de satisfaire des besoins explicites et implicites » [4].

Nous pouvons donc dire que des données sont de qualité si elles satisfont aux exigences définies dans les spécifications et que celles-ci reflètent les besoins implicites des utilisateurs. Mais comment pouvoir définir précisément la qualité des données?

[1]L'ISO est une fédération mondiale d'organismes nationaux de standardisation. Les travaux de l'ISO aboutissent à des accords internationaux qui sont publiés sous la forme de Normes internationales, Les normes sont des accords documentés contenant des spécifications techniques ou autres critères précis destinés à être utilisés systématiquement en tant que règles, lignes directrices ou définitions de caractéristiques pour assurer que des matériaux, produits, processus et services sont aptes à leur emploi[4].

Pour savoir intuitivement si un élément de données est de haute qualité ou non en l'examinant, nous devons faire référence à certains critères concrets de reconnaissance de la qualité des données dans l'entrepôt de données ou dans la base de données.

4 .2. Dimensions de la qualité des données

Aujourd'hui, la qualité des données est décrite et caractérisée par plusieurs attributs, facteurs, critères ou dimensions, lesquels représentent les différents aspects qui définissent une caractéristique particulière ou qui forment la qualité des données elle-même.

Les dimensions de la qualité peuvent être considérées comme une extension de la données, et sont définies de façon qualitative, et sont donc associées à une ou plusieurs métriques, pour chaque métrique une ou plusieurs méthodes de mesure [20], ces dimensions sont associées en trios niveaux :

1. Qualité du modèle conceptuel des données : ce niveau centré sur la qualité du modèle conceptuel des données. Beaucoup de recherches accordé a ce niveau qui sont définis plusieurs dimensions nous citons (*Lisibilité*, *Expressivité*, *Complétude*, *Simplicité*, *Traçabilité*, etc.) [3].

2. Qualité du processus manipulant des données : dans ce niveau nous intéressons à la qualité des processus de traitement des données, divers dimensions définis pour ce telle que *Temps de réponse*, *fiabilité*, *sécurité*, etc. [12]

3. Qualité des données livrées à l'utilisateur (Qualité des instances des données) : dans ce niveau nous intéressons à la qualité des instances des données ou des attributs et pour cela les chercheurs sont définis plusieurs dimensions telles que le *Fraîcheur*, *exactitude*, *complétude*, *Cohérence*, etc.

Dans le cadre de ce mémoire de PFE, nous nous intéressons à la qualité des instances des données.

Beaucoup de listes de dimensions de qualité ont été proposées, quelques-unes contiennent un bon nombre de définitions. Par exemple Redman a identifié quatre dimensions de qualité : l'exactitude, la complétude, la crédibilité et la consistance [13] alors que Wang et Strong ont

analysé les attributs de qualité du point de vue de l'utilisateur [11]. Ci-dessous quelques dimensions de la qualité des données,

Exactitude : La plupart des études sur la qualité des données classé l'exactitude des données comme une dimension clé, on peut dire que les données exacte c'est la valeur stockée dans le système pour un élément de données est la bonne valeur, c'est à dire les données doivent être correctes, fiables et sans erreurs.

Cohérence : La forme et le contenu d'un élément de données doit être le même dans les multiples systèmes sources [7].

Complétude : les données sont complétés c'est-à-dire Il ne doit pas y avoir de valeurs manquantes pour un élément de données dans le système.

Duplication : Le nombre des enregistrements dupliqués dans un système, c'est-à-dire ne trouve pas le même enregistrement deux ou trois fois dans une table de données avec des identifiants différents.

Représentation : Les données doivent être représentées de façon compacte sans perdre de leur signification [4].

La conformité aux règles métier : Les valeurs de chaque élément de données doit se conformer aux exigences et aux règles métier.

Exemple : Dans un système de prêt bancaire, le solde du prêt doit toujours être positif et supérieure à zéro

Clarté : Un élément de données peut respecter toutes les caractéristiques de qualité ci-dessus, mais si les utilisateurs ne comprennent pas clairement le sens de la donnée, celle-ci sera sans valeur pour eux. Pour assurer une clarté pour les données vis-à-vis des utilisateurs, une convention de nommage correct doit être respectée [2].

Respect des règles d'intégrité des données : Les données stockées dans les bases de données relationnelles dans un système source doivent respecter les règles d'intégrité référentielles et l'intégrité d'entité.

- Toute table qui contient des valeurs NULL dans la colonne clé primaire ne respecte pas l'intégrité d'entité.

- Par contre, l'intégrité référentielle implique l'établissement correct des relations parent-enfant. C'est l'intégrité référentielle qui assure l'existence d'un client pour chaque commande dans la base de données commerciale [4].

4.3 Exemple de problèmes de qualité des données au niveau instance

Exemple de problèmes de qualité des données au niveau instance		
Problème	**Données**	**Descriptif**
inexacte	Tél = 000-00-00-00	Valeur non correct pour un numéro de téléphone
Valeur manquant	Ville = Null Adresse = " "	Valeurs null pour un champ de données
doublons	Nom = ahmed lakhale Nom = ahmed L	Le Même Nom entrée deux fois avec deux manières différentes dans une même source
Valeur imbriquées	Nom = Khalid 12 /04/1990	valeurs multiples saisies dans un attribut (au format texte Multiple)
incohérence	Code postale = Alger	Valeur incohérente par rapport au champ
Représentation	Ahmed Dj, Khalid L	Deux noms en même champ
Conformité à la règle	Prix = - 4,00 DZ	Valeurs non conforme à la règle de métier

Table I.2problèmes de qualité des données au niveau instance

4.4 Source des problèmes de la qualité des données

Les données et dans tout leur durée de vie ont connu des tâches ou des actions qui on peut les considère que sont les sources des problèmes de la qualité de ces données.

Ci-dessous et dans la table suivant on a vue quelque source des problèmes de qualité des données, et dans les différentes étapes de traitement des données.

ÉTAPES DE TRAITEMENT	**SOURCES DE PROBLEMES DE QUALITE DES DONNEES**
Création des données	• Entrée manuelle : absence de vérifications systématiques des formulaires de saisie • Entrée automatique : problèmes de capture OCR, de reconnaissance de la parole • Incomplétude, absence de normalisation ou inadéquation de la modélisation conceptuelle des données : attributs peu structurés, absence de contraintes d'intégrité pour maintenir la cohérence des données • Entrée de doublons • Contraintes matérielles ou logicielles
Collecte / import des données	• Destruction ou mutilation d'information par des prétraitements inappropriés • Perte de données : *buffer overflows*, problèmes de transmission et l'absence de vérification dans les procédures d'import massif • Introduction d'erreurs par les programmes de conversion des données

Stockage des données	• Absence de méta-données • Absence de mise à jour et de rafraîchissement des données obsolètes ou répliquées • Modèles et structures de données inappropriés, spécifications incomplètes ou évolution des besoins dans l'analyse et conception du système • Modifications *ad hoc* • Contraintes matérielles ou logicielles
Intégration des données	• Problèmes d'intégration de multiples sources de données ayant des niveaux de qualité et d'agrégation divers • Problèmes de synchronisation temporelle • Systèmes de données non conventionnels • Facteurs sociologiques conduisant à des problèmes d'interprétations et d'intégration des données • Jointures *ad hoc*
Recherche et analyse des données	• Erreur humaine • Contraintes liées à la complexité de calcul • Contraintes logicielles, incompatibilité • Problèmes de passage à l'échelle, de performances et de confiance dans les résultats • Utilisation de boîtes noires pour l'analyse • Expertise insuffisante d'un domaine

Table I.3 source des problèmes de qualité des données [3].

4 .5 Coûts de la mauvaise qualité des données

Les problèmes de qualité des données stockées dans les bases et les entrepôts des données se propagent de façon endémique à tous les types de données (structurées ou non) et dans tous les domaines d'application, données gouvernementales, commerciales, industrielles ou scientifiques. Les conséquences de la non qualité des données (ou de leur qualité médiocre)

sur les prises de décision et les coûts financiers qu'elle engendre sont considérables, ci-dessous une liste des conséquences de la mauvaise qualité des données.

- Des mauvaises décisions
- Perdit l'occasion d'affaires en raison de sale des données
- La souche et les frais généraux sur les systèmes source en raison de l'origine des données corrompues.
- Amendes des agences gouvernementales pour non-respect ou violation des règlements
- Les données redondantes
- Temps et d'efforts pour corriger les données

4.6 Approches générales pour détecter et corriger les problèmes de qualité des données

Comme le représente la Figure I.3, on peut classer la plupart des travaux abordant la problématique de la qualité des données selon quatre grands types d'approches complémentaires [3].

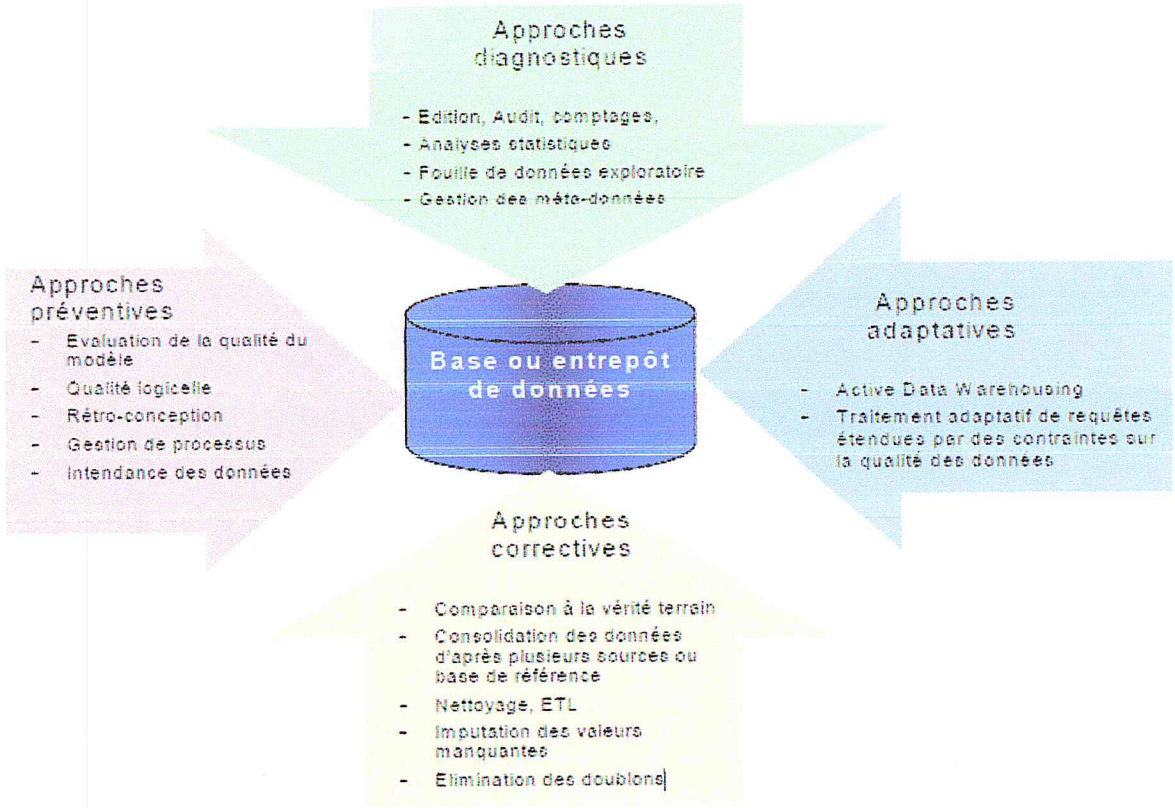

Fig. I.3 Panorama des approches pour l'évaluation et le contrôle de la qualité des données [3].

1. Les approches préventives centrée sur l'ingénierie des systèmes d'information et le contrôle des processus avec des techniques permettant d'évaluer la qualité des modèles conceptuels, la qualité des développements logiciels et celle des processus employés pour le traitement des données.

2. Les approches diagnostiques centrées sur des méthodes statistiques, d'analyse et de fouille de données exploratoire permettant de détecter des anomalies sur les données

3. Les approches correctives centrées sur des techniques de nettoyage et de consolidation de données et utilisant des langages de manipulation des données étendus et des outils d'extraction et de transformation de données (*ETL – Extraction-Transformation-Loading*)

4. Les approches adaptatives ou actives appliquées généralement lors de la médiation ou de l'intégration des données : elles sont centrées sur l'adaptation des traitements (requêtes ou opérations de nettoyage sur les données) de telle façon que ceux-ci incluent à l'exécution en temps-réel la vérification de contraintes sur la qualité des données [3].

4.7 Outils traitant sur la qualité des données

1. Oracle data Quality

Oracle Data Quality est une solution éprouvée d'amélioration de la qualité des données et reconnue par le marché pour répondre aux exigences les plus complexes dans de domaine. Son puissant moteur basé sur les règles fonctionnelles, son architecture robuste et évolutive le placent au cœur d'une stratégie d'intégration de données d'entreprise. Oracle Data Quality répond aux attentes des entreprises souhaitant améliorer la qualité des données, notamment sur les projets de Business Intelligence, de référentiels de type métier ou d'entreprise. Oracle Data Quality possède la meilleure et la plus vaste prise en charge des langues et des règles fonctionnelles propres à chaque pays.

Principale fonctionnalité d'Oracle Data Quality

- Standardisation et nettoyage de noms, d'adresses et d'autres données complexes, Oracle Data Quality offre des fonctions optimisées d'analyse et d'association pour le nettoyage des noms et d'adresses. Les règles intégrées propres à chaque pays (basées sur la détection de mots-clés et de masque de recherche) peuvent être personnalisées afin de répondre aux attentes spécifiques

- Validation, réparation et enrichissement de vos données Vous pouvez valider, corriger et enrichir de façon optionnelle les données de type nom et adresse à l'aide d'un répertoire postal. Vous pouvez utiliser des sources supplémentaires pour enrichir vos adresses avec des informations géographiques (longitude, latitude, etc.) ou des règles d'enrichissement personnalisées.

- Oracle Data Quality peut identifier et lier des enregistrements similaires ou en double, Il permet également de rapprocher des enregistrements sur une notion commune.

- Des règles prédéfinies propres à un pays sont fournies pour les projets de nettoyage de noms et d'adresses. Les règles de comparaison ainsi que les règles communes peuvent être entièrement personnalisées.

- L'interface utilisateur Oracle Data Quality permet la création de projets sur la base de modèles prédéfinis et d'en personnaliser les règles en vue de refléter au mieux les attentes d'amélioration de la qualité. Cette interface intuitive permet la personnalisation des projets et des règles sans connaissances techniques particulières.

2 Informatica Data Quality

Informatica Data Quality propose une qualité globale de données à tous les intervenants, sur tous les projets, dans tous les domaines et toutes les applications métiers, sur site ou dans les applications bureau, au moyen d'une plate-forme unique et unifiée. Informatica est le leader des fournisseurs de logiciels et de services d'intégration de données. Les produits Informatica permettent aux entreprises de tirer davantage de valeur de leurs informations en intégrant toutes leurs sources de données. Plus de 3 350 entreprises dans le monde s'appuient sur Informatica pour réduire les coûts et les délais de réponse à leurs besoins d'intégration de données, quelles qu'en soient l'échelle et la complexité.

Principales Fonctionnalités d'Informatica Data Quality

- Jeu d'outils basés sur les rôles, l'entreprise peut participer aux processus de qualité de données et réduire la dépendance vis-à-vis de ressources informatiques limitées.

- Support complet de toutes les données et de tous les objectifs ,vous pouvez appliquer des règles de qualité de données aux données concernant les clients, les produits, les

finances et les ressources, et réutiliser ces règles sur tous les types de projets d'intégration de données et de qualité de données

- Ouverture à toutes les applications, vous pouvez accéder à toutes les sources de données, à tout moment (qu'elles soient sur site, dans les systèmes des partenaires ou dans des applications normal) et déployer des règles de qualité de données centralisées afin d'améliorer la qualité des données dans l'ensemble des applications

- Identification, résolution et prévention des problèmes de qualité de données

- Nettoyer les données pour toutes les applications de manière proactive et les garder exploitables

- Fonctionnement plus efficace grâce à la participation possible des équipes métiers aux processus de qualité de données

3 Talend Data Quality

Talend Data Quality est un environnement de gestion de la qualité des données qui permet aux utilisateurs de déposer les composants traitement de données directement dans l'éditeur graphique. Ces composants traitent des données diverses telles que les adresses, les numéros de téléphone, les synonymes, les abréviations …, Ses fonctionnalités étendues de Profiling des données et sa structure intégrée de gestion des rapports permettent une mise en place plus efficace et plus rapide des projets d'analyse de données. Talend Data Quality est la première solution open source de qualité de données offrant des fonctionnalités et du support technique d'entreprise. Il relève le défi de la qualité de données grâce à plusieurs modules interconnectés :

Principal fonctionnalité de Talend Data Quality

- La première fonctionnalité c'est le profilage de données. Le data profiler est une application sophistiquée mais simple d'utilisation qui n'exige pas de connaissance particulière. Les utilisateurs métier ou les équipes en charge de la gestion des données

peuvent ainsi effectuer toutes sortes d'analyses à l'aide d'un ensemble d'indicateurs, de motifs et de règles pour chaque élément de données à analyser ou superviser.

- Talend Data Quality dispose d'outils puissants pour redresser et réparer toutes les données non conformes à vos normes. Elle vous permet d'utiliser des données de référence pour paramétrer les normes de valeur, des expressions régulières pour les normes de format et de taille.

- Des algorithmes de correspondance pour les doublons et quasi-doublons contenues dans vos données.

- Enrichissement des données par compléter les données manquantes de vos données pour atteindre vos objectifs de qualité.

- Data Quality Portal est un outil de reporting de la qualité de données, personnalisable et basé web, permettant aux entreprises de suivre de près les métriques de qualité de données qui peuvent impacter les processus métier importants

- Talend Data Quality offrant à la fois une vue graphique et fonctionnelle des processus d'intégration, il permet un approche accessible et non technique des données. Les systèmes, connexions, étapes et pré-requis d'un workflow sont représentés par des symboles graphiques intuitifs.

4 Etude comparative

Pour faire la comparaison entre ces outils on a proposé les critères ci-dessous, ces critères sont des fonctionnalités de qualité des données offres par ces outils, le rôle de ces services c'est l'assurance de la qualité des données dans les entrepôts des données ou dans les bases des données, ces critères sont présentés ci-dessous avec des définitions.

- **Data Profiling(Profilage)** : analyse de la qualité des données afin de déterminé les données d'amélioration

- **Data Cleansing(Nettoyage)** : détection et correction des données corrompus ou inexactes
- **Data monitoring (surveillance):** suivi la qualité des données dans le temps et production des rapports de qualité

- **Standardisation des noms et des adresses :** moteur de règle pour standardiser les noms et les adresse, et pour assuré que les données sont conforme aux règles de qualité
- **Validation, réparations et enrichissement des données :** par utilisation des sources externe pour améliore la complétude des données
- **Identifier les doublons :** analyse les données afin de déterminer les enregistrements on double
- **Portail web :** pour suivre la qualité des données avec des métriques personnalités et via le web

Outils / critère	Oracle data Quality	Informatica data Quality	Talend Data Quality(open source)
Data Profiling	Oui	oui	oui
Data Cleansing	Oui	oui	oui
Data monitoring	**Non**	oui	**non**
Standardisation des noms et des adresses	Oui	oui	oui
Validation réparations enrichissement des données	Oui	oui	oui
Identifier les doublons	Oui	oui	oui
Portail web	**Non**	**non**	oui
Intégrité à l'outil data Intégration	Oui	oui	oui

Table I. 3 Comparaisons entre les outils qualité des données

à partir de cette comparaison on a vus que les trois Framework offre des bonnes fonctionnalité pour garantir la qualité des données dans les entrepôts des données ou dans les bases des données, telle que l'audit et le nettoyage des données, avec une avantage pour les deux outils Informatica Data Quality et Talend data Quality car elles sont fournies par des entreprises spécialistes seulement sur l'intégration des données et la qualité des données et tous ce qui concerne la gestion des données, ainsi que Informatica est le leader des fournisseurs de logiciels et de services d'intégration de données.

On ne conclut que ces outils parmi les meilleurs outils existent dans le marché qui propose des solutions sérieuses pour les différents problèmes de la qualité des données dans les bases des données et les systèmes d'intégration de données.

Conclusion

Le but de ce chapitre est de donner un aperçu global sur la qualité des données, et avant cela un résumé sur les systèmes décisionnels et l'ETL (Extraction, Transformation, Loading), puis les sources des problèmes liés à la qualité des données, l'impact d'une mauvaise qualité des données, et comment les données erronées se propagent dans les systèmes d'information. On a vu des outils traitant sur la qualité des données et les différentes fonctionnalités fournis par celle-ci, avec une étude comparative.

En conclusion, la qualité des données a une grande importance pour les entreprises, les établissements ou les administrations, car une mauvaise qualité des données engendre à ces derniers des coûts très élevés. Enfin, l'amélioration de la qualité des données passe par la mise en place d'une initiative très sérieuse.

Chapitre II Technique de détection / correction des problèmes de la qualité des données

Introduction

Face à la croissance augmentée des volumes des données dans les bases et l'entrepôt des données, l'évaluation de leur qualité est devenue une tâche d'autant plus difficile que les experts ont recherché au fil du temps à des nouvelles moyen et techniques pour l'évaluation et l'amélioration de la qualité des données.

Parmi ces techniques nous présentons dans ce chapitre les deux techniques les plus connu, qui sont l'audit des données et le nettoyage des données, ainsi que nous essayons de montré les différentes fonctionnalités de ces technique par des exemples.

1. Technique de détection / correction des problèmes de la qualité des données

Aujourd'hui, ils existent beaucoup des solutions ou des technique pour l'évaluation et l'amélioration de la qualité dans les base et les entrepôts des données, dans le cadre de ce mémoire nous nous intéressons sur les deux techniques les plus connu sont l'audit et le nettoyage des données.

1.1 Audit de données

Les outils de l'audit des données permet de vérifier les valeurs des données satisfont des contraint de plusieurs type : cohérence par rapport à des règles logique, ou contraint spécifique à l'application ou d'intégrité [3]. Ainsi que nous donne des statistiques sur les données : non précise, null, zéro ....c'est à dire elles ne permettent pas l'amélioration de la qualité des donnés.

L'audit des données vise l'intégrité c'est-à-dire la conformité à des règles préalablement définis, mais elles ne garantie pas l'exactitude de données [2] [5] [10].Donc on peut dire que l'audit de données c'est une technique ou moyen pour identifier ou détecter les diverses erreurs existe dans les base ou les entrepôts de données, ci-dessous on a présenté les fonctionnalités de cette outils.

1.1.1Fonctionnalités de découverte l'erreur

La liste suivante décrit quelques fonctions typiques de découverte les erreurs que les outils de audit des données sont capables d'accomplir :

- Identifier les enregistrements en double
- Identifier les éléments de données dont les valeurs sont en dehors des règles préalables.
- Trouver des données incohérentes
- Détecter les valeurs manquant et les valeurs null
- Détecter les incohérences entre les éléments de données provenant de différentes sources

Ci-dessous la marche de l'audit de données : c'est-à-dire comment on a appliqué les outils d'audit de données sur nos données :

- définition du périmètre de l'audit de données, selon les dimensions de qualité à considérer
- Identification des segments de données à analyser (par exemple, données client, ville, … etc.)
- Choix d'un ensemble représentatif de données (par exemple, par zone géographique)
- Analyse du dictionnaire de données (par exemple, le nom des attributs, type, domaine, taux de remplissage, etc.)
- Enumération des contraintes : par exemple, unicité de clés pour les n-uplets d'une table, respect des contraintes d'intégrité, respect de règles syntaxiques dans les valeurs de certains attributs (tel que le numéro de Sécurité Sociale), respect du zonage géographique (défini par exemple comme une règle de cohérence entre la ville et le code postal), etc.
- Multiples comptages : par exemple, taux d'informations non renseignées, taux d'anomalies de zonage, taux des données ne respectent pas chaque contrainte, détection de doublons), normalisation des adresses, vérification syntaxique du numéro de téléphone, taux de faux téléphones, taux de fax erronés, etc.
- Calculs croisés : par exemple, taux d'individus avec même email, même nom, même adresse, même téléphone, etc. [3].

1.1.2 Exemple d'un audit des données

Ci-dessous des exemples des requêtes pour analyser les données, afin de détecter les données manquant et les enregistrements en double, nous utilisons comme source de données la base de données (pubs) qui fourni par Microsoft.

Exemple 1 :

On va analyse une colonne pour but de voire quels est les champs null, est pour cela on utiliser la requête suivant :

```
Select NomdeTable, NomdeColonne, Nombre d'enregistrement, nbNull,
    (NbNull*100) /Nombred'enregistrement as 'Pourcentage%'
    From (select "Table" as NomdeTable) as t,
```

(Select "Colonne" as NomdeColonne) as t1,

(Select count (*) as Nombred'enregistrement FROM "Table") as t2,

(Select count (*) as nbNull

From "Table" Where "Colonne" is null) as t3

C'est par exemple on a analysé la colonne " State " qui existe dans la table Publisher, on va obtenir le résultat qui s'afficher dans la table suivant :

Nom de Table	Nom de Colonne	Nombre enregistrement	Nombre valeur Null	% des valeurs Null	% des valeurs non Null
Publishers	State	18	2	11	89

Table II.1 : % des champs null dans une colonne

Exemple 2 :

Dans cet exemple on a analysé la colonne mail dans la table Publishers afin de voir quels est les enregistrements qui ont la même adresse mail, pour cela on a utilisé la requête suivant et le résultat s'afficher dans la table II.2.

Select NomdeTable, NomdeColonne, Nombre d'enregistrement, mail_noncorrect,

*(mail_noncorrect *100) /Nombred'enregistrement as 'Pourcentage%'*

From (select "Table" as NomdeTable) as t,

(Select "Colonne" as NomdeColonne) as t1,

(Select count (*) as Nombred'enregistrement FROM "Table") as t2,

(Select count (*) as mail_noncorrect

From "Table" Where "Colonne" not like('%@%.%')) as t3

Nom de Table	Nom de Colonne	Nombre enregistrement	Nombre des @ mail non correcte	% des@ mail non correcte	% @ mail correcte
Publishers	mail	18	4	22	78

Table II.1 : % des adresses mail non correcte

1.2 Nettoyage de données

L'étude de la qualité des données et le nettoyage des données sont souvent les premiers pas cruciaux dans les processus de découverte de connaissance (knowledge discovery en anglais), de fouille de données (data Mining en anglais) et de data Warehousing [4]. Le nettoyage des données est une étape très importante car les fichiers ou les bases de données utilisées en tant que sources de données peuvent contenir beaucoup d'inexactitudes et d'inconsistances.... Ces erreurs résultent d'une multitude de différents facteurs, voire table I.2.

Les outils de nettoyage de données sont des outils de correction des données qui d'aide à fixer les données corrompues [10]. Ces outils ont des caractéristiques et des fonctions qui découverte et éliminer des données polluées. Dans la section suivante, on va voire les différentes fonctionnalités de correction des données.

2.2.1 Fonctionnalités de la correction des données

Le nettoyage des données n'est pas un processus en une phase mais un processus itératif. A chaque étape, nous raffinerons la phase précédente pour essayer d'atteindre les objectifs suivants:

- Normaliser les données incohérentes
- Eliminer les données en doubles
- traitement des valeurs manquant
- l'établissement des standards et la standardisation des différents types de données
- Valider les valeurs admissibles.

Maintenait on va détailler par des exemples sur ces fonctionnalité :

a) Elimination des doublons :

Pour détecter les doublons, on a utilisé plusieurs technique ou algorithme, parmi ces technique nous expliquent la technique de jointure approximative, c'est on a par exemple un enregistrement décri en deux manier différent dans la même source « Khalid lakhale » ou « Khalid. L », il peut être difficile de détecter que les deux enregistrements représentent la même personne.

Exemple :

	Matricule	**Nom et Prénom**	**Téléphone**	**Mail**
R1	10/1000	lakhale Khalid	0554302034	khalid@yahoo.fr
R2	10/1001	la Khalid	0554302034	khalid@yahoo.fr
R3	10/1002	chelfi Ahmed	0773234521	Ahmed_ch @yahoo.fr
R4	10/1003	Nabi Ali	0663321290	Ali_nabi@gmail.com
R5	10/1003	Ch. Ahmed		Ahmed_ch @yahoo.fr

Table II.2: Enregistrements avec doublons

À partir de la table Table II.2 ci-dessus on remarque que les enregistrements (R1, R2) représentent la même adresse e-mail et le même numéro de téléphone, donc (R1, R2) représente la même personne (lakhale Khalid), et aussi les enregistrements (R3, R5) représentent la même personne (chelfi Ahmed) avec une matricule et manière différent, donc ce n'est pas facile pour détecter que les enregistrements (R1, R2) représentent la même personne.

En conséquence, pour détecter les doublons, on a utilisé la technique de jointure approximative : une même adresse mail ou un même N° de téléphone implique qu'il s'agit d'une même personne.

La technique de jointure approximative consiste à regrouper et trier les enregistrements par groupes selon une fonction de hachage sur les valeurs d'un ou plusieurs attributs(par exemple, utilisant les premières lettres ou les consonnes des noms propres).Les enregistrements qui se trouvent dans les mêmes groupes sont candidats à l'appariement et, pour chaque paire de candidats, une distance de similarité est calculée, seules les paires de plus haut score sont effectivement appariées (ou assimilées à des doublons). La méthode classique de jointure approximative et de détection de doublons est présentée ci-après.

Méthode générique pour la recherche des doublons

1. Pré-traitement des données (standardisation des attributs, des abréviations, structuration des adresses, etc.)
2. Choix d'une fonction permettant de réduire l'espace de recherche par :
- Tri ou hachage selon une clé
3. Choix d'une fonction de comparaison permettant d'exprimer la distance entre lespaires telle que :
- Identité stricte, distance simple ou complexe
- Distance pondérée par la fréquence ou dirigée par des règles
- Distance d'édition, distance de Jaro, Jaro-Winkler, etc.

4. Choix d'un modèle de décision
- Méthodes probabilistes : avec/sans ensemble d'apprentissage
- Méthodes basées sur des règles et connaissances du domaine
5. Vérification de l'efficacité de la méthode

Pour des domaines d'attributs textuels, l'appariement des chaînes des caractères (*string matching*) peut être calculée par une distance comptabilisant le nombre d'opérations d'édition (telles que l'ajout, la suppression d'un caractère ou le changement de lettre) nécessaires pour transformer une chaîne de caractères en une autre. Par exemple, « SRH » et « RH » ont une distance d'édition de 1. Les chaînes de caractères dont la distance d'édition est inférieure à un seuil fixé seront alors appariées. Le Tableau II.6 présente les principales mesures de similarité pouvant être employées pour apparier les chaînes de caractères.

Calcul de similarité	**Définition et principales caractéristiques**
Distance de Hamming	Applicable à des champs numériques fixes (N°Sécu, Code Postal) mais ne prend pas en compte les ajout/suppressions de caractères
Distance d'édition	Soient s1 et s2 deux chaînes de caractères à apparier, le calcul du coût minimal de conversion de s1 en s2 en cumulant le coût unitaire des opérations d'ajout (A), suppression (S) ou remplacement de caractères (R) est tel que : $Edit(s1,s2) = \min(\Sigma A(s1,s2) + S(s1,s2) + R(s1,s2))$

	Calcul par programmation dynamique mais complexité quadratique
Mesure TF-IDF	Soit un terme s1 et un document d dans un ensemble de documents D, TF le nombre d'occurrences du terme s1 dans le document d et IDF la fraction du nombre de documents dans D sur le nombre de documents contenant s1 TFIDF(s1, d, D)=log(TF(s1,d)+1)*log(IDF(s1, D)) Usage traditionnel en recherche d'information, les termes rares sont rendus plus importants
Indice de Jaccard	Soient deux ensembles de termes S et T *Jaccard(S, T)* = $\|S \cap T\| / \|S \cup T\|$
Distance de Jaro	Soient s1 et s2, deux chaînes de caractères de longueur respective L1 et L2, ayant C caractères communs et T transpositions de caractères : Jaro (s1, s2)=(C/L1 +C/L2 +(2C-T)/2C)/3 Utilisé pour les chaînes de caractères courtes

Table II.3Distances de similarité pour comparer les chaînes de caractères et identifier les doublons potentiels

Pour compléter, dans l'étape précédente nous avons calculé la distance de similarité entre les valeurs, mais pour détecter les enregistrements en double on a besoin de comparé toutes les valeurs de cette enregistrement, et pour cela il y a des nombreux modèles de décision ont été proposés pour confirmer ou informer les hypothèses d'appariement entre les enregistrements candidats.qui sont classées en trois catégories (modèle probabiliste, empirique ou basé sur des connaissances),

b) Traitement des valeurs manquant :

Les données incomplètes sont un problème fréquent dans la plupart des recherches centrées sur la qualité des données [18], il existe plusieurs méthodes pour traiter les valeurs manquantes.

1) **catégories des données manquant :** on a deux catégories des données manquant

- les données manquant totale, lorsque aucune information n'est recueillie sur une unité échantillonnée

- les données manquant partielle, lorsque le manque d'information est limité à certaines variables [19].

Puis, il existe trios mécanismes pour distinct les données manquant :

a. données manquant entièrement au hasard (MCAR, *Missing Completely at random)* : les enregistrements ayant une donnée manquante ne peuvent pas être distingués de ceux ayant une donnée renseignée. La probabilité qu'une donnée soit manquante ne dépend ni des valeurs des variables observées ni de la valeur non observée.
b. données manquant au hasard (MAR, *Missing at random*) : le fait d'avoir une donnée manquante dépend d'autres caractéristiques observées, mais pas de la valeur manquante (qui aurait pu être renseignée). La probabilité qu'une donnée soit manquante dépend des valeurs des variables observées mais non de sa vraie valeur.

c. données ne manquant pas au hasard (NMAR, *Non missing at random)*. le fait d'avoir une donnée manquante n'est pas aléatoire, ne peut pas être déduit des autres variables et dépend de la valeur manquante (qui aurait pu être renseignée). La probabilité qu'une donnée soit manquante dépend de sa vraie valeur (non observée).

2) Les différents traitements des données manquantes :

- Ne rien faire : Cela oblige à travailler avec un fichier de données incomplet. Si les valeurs manquantes sont peu nombreuses, on peut les oublier sans aucun scrupule.

- utiliser uniquement les enregistrements complets : Si les données sont présentées sous forme de tableau, cela revient à oublier une ligne dès qu'il manque une valeur dans cette ligne : on oublie donc aussi les autres valeurs de cette ligne, qui sont effectivement présentes. Bien que cette option soit simple et permette d'utiliser un fichier complet, elle présente certains risques. En effet :

L'échantillon de ceux qui ont répondu à toutes les questions peut être

-soit trop réduit pour être significatif,

- soit non représentatif de la population globale.

- imputer une valeur : L'imputation consiste à produire une « valeur artificielle » pour remplacer la valeur manquante, pour objectif de produire des estimations approximativement sans biais. il existe certains méthodes d'imputation, dans la table II.7 on va montre quelques technique d'imputation, il y a des explications on détaillé sur les méthodes d'imputation en [19].

méthode	**Description**	**Valeur imputée**
L'imputation par règle	on applique à une valeur Manquante une valeur déterminée suivant une réglementation	Exemple : Calcul montant TTC à partir du montant HT
imputation par la moyenne	On remplace chacune des valeurs manquantes par la valeur moyenne de l'ensemble de réponses obtenues	$y_i^* = \frac{1}{r}\sum_{i \in S_r} y_i = \bar{y}_r$
imputation par le ratio	chaque valeur manquante y_i est remplacée par la valeur prévue y_i^* obtenue par régression de y sur x.	$y_i^* = \frac{\bar{y}_r}{\bar{x}_r} x_i$
imputation par la méthode par le plus proche voisin	on attribue à l'enregistrement pour lequel la réponse à une question manque la valeur figurant pour cette question dans l'enregistrement obtenu pour le répondant le plus proche, où l'expression « le plus proche » est habituellement définie par une fonction de distance basée sur une ou plusieurs variables auxiliaires.	$y_i^* = y_j$ pour certains j $\in S_r$ tels que $dist(x_i, x_j)$ soit minimal

Table II.4 quelques méthodes d'imputation [19]

Conclusion

L'évaluation et l'amélioration de la qualité des données n'est pas une étape simple, mais ils nécessitées plusieurs technique qui fourni des certains fonctionnalité pour arriver a assuré la qualité des données puis a corriger les données corrompus.

Le but de ce chapitre et de montre les deux technique les plus connu pour l'évaluation et l'amélioration de la qualité des données dans les base et l'entrepôt de données.

Le chapitre suivant concerne la conception et l'architecture de notre Framework.

Chapitre III Modélisation du Framework data Quality

Introduction

Dans ce chapitre on va voire une modélisation de la qualité des données dans les base et les entrepôts des données, après on va voire la conception de notre Framework par l'utilisation du langage de modélisation UML, on va voire aussi un diagramme de cas d'utilisation est après on a proposé un diagramme de classe.

Enfin, on a proposé une architecture pour notre Framework qui consiste a amélioré la qualité des données dans les systèmes décisionnels.

2. la conception

D'une façon générale, nous modélisons la qualité des données selon le formalisme UML comme le représente la Figure III.1 et Figure III.2. On a commencé par le diagramme de cas d'utilisation pour définir le besoin d'utilisateur de la qualité des données, et après cela on à modéliser le diagramme de classe.

2.1 Diagramme de cas d'utilisation

Tout d'd'abord dans notre diagramme de cas d'utilisateur on a un seul acteur intervenant sur le système (Utilisateur), les objectif qui doit accomplir le système sont les suivant :

- Audit de données
- Nettoyage de données
- Rapport de qualité

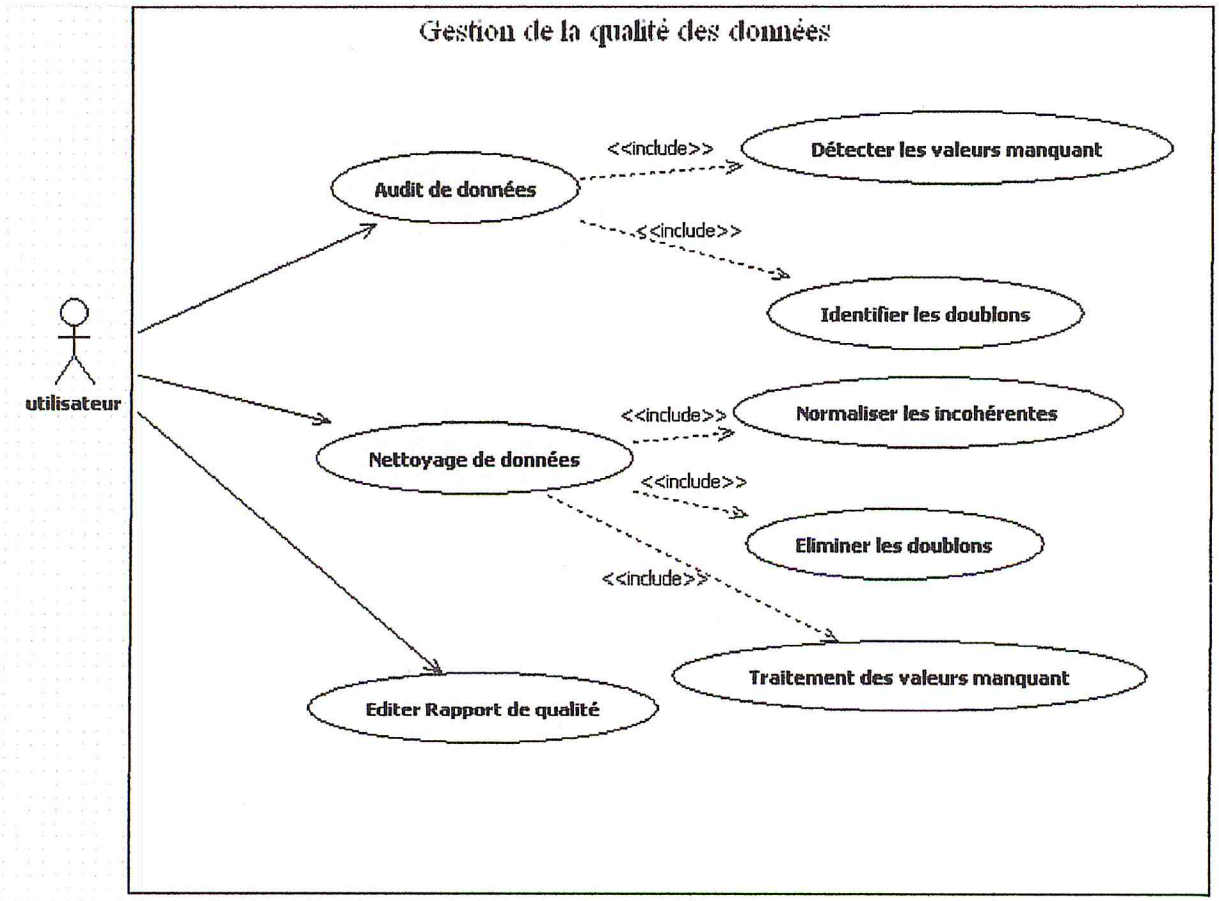

Fig.III.1 : diagramme cas d'utilisation qualité des données

Ces objectifs vont se transformer on cas d'utilisation intervenant par l'acteur (Utilisateur), qui inclut les fonctionnalités suivant:

- ✓ le premier cas d'utilisation offre les fonctionnalités de détecter et identifier les problèmes ou les erreurs des données suivant :
  _ trouver les incohérent
  _ détecter les valeurs manquant
  _ identifier les doublons.

- ✓ le deuxième cas d'utilisation inclut les différentes fonctions de correction les problèmes de données :
  _ normaliser les incohérent
  _ Traitement des valeurs manquant
  _ Détection et élimination des doublons.

- ✓ le troisième cas d'utilisation est pour faire des rapports sur la qualité des données

2.2 Diagramme de classe

Le diagramme de classe représente la structure des modules de notre système qui composé de dix classes d'objets, La base de données ou l'ODS est une classe abstraite représentant les données à traiter pour laquelle les données sont représentées selon un modèle conceptuel et gérées par des processus de traitement.

La qualité de celles-ci peut être évaluée selon plusieurs dimensions (représentant les différentes facettes de la qualité).Chaque dimension peut être mesurée par une ou plusieurs métriques à un instant donné.

Et la classe besoin d'utilisateur qui est une classe abstraite pour auquel l'utilisateur choisit les services qui assure la qualité des données, on a aussi les classe audit et nettoyage de données qui sont héritées de la classe besoin utilisateur, ces classe (audit et nettoyage de données) contient tous les méthodes qui évaluer et améliorer la qualité des données dans les entrepôts et les bases de données.

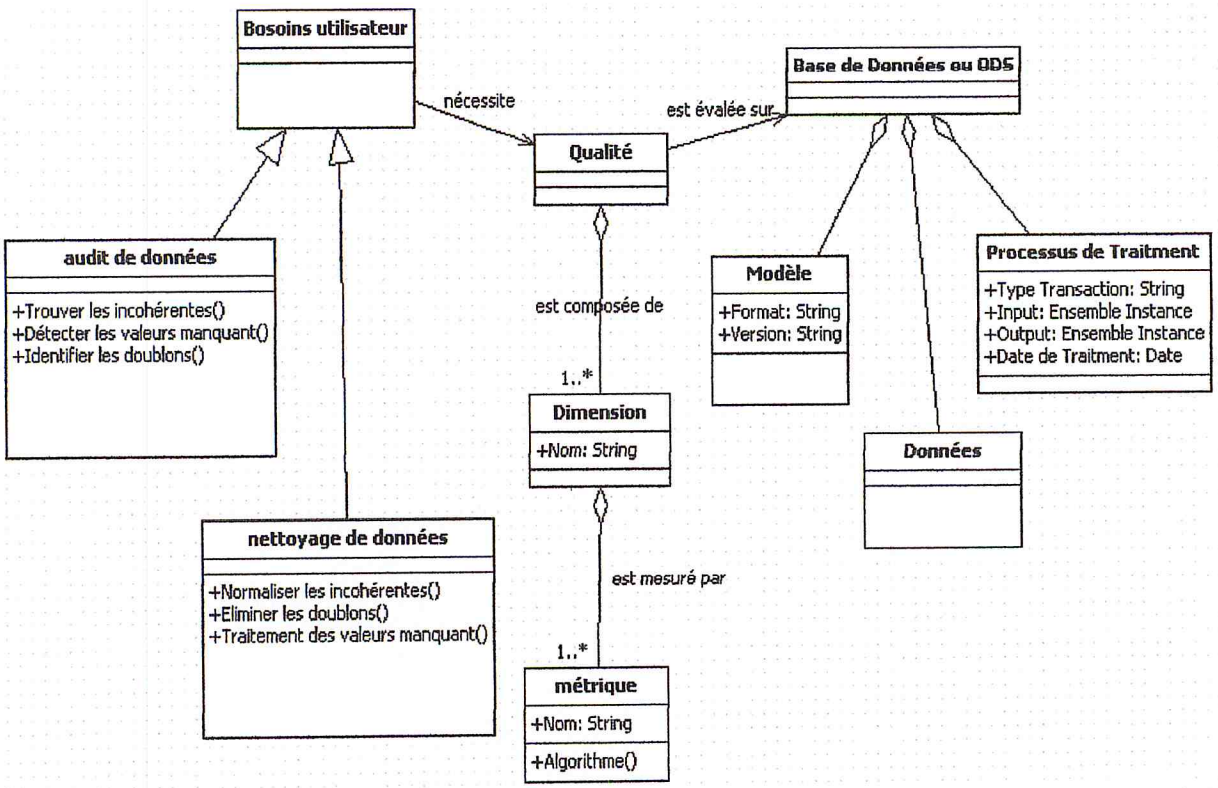

Fig.III.2 : Diagramme de classe Qualité des données

La table ci-dessous contient la description détailles des classe de notre diagramme

Classe	Description	Les méthodes
Besoin utilisateur	Est une classe avec deux classes filles audit et nettoyage des données dans laquelle sont représentés les besoins de l'utilisateur.	Get besoin utilisateur () ;
Audit de données	C'est une classe d'où on a trouvéles méthodes d'évaluation des données	Identifier les doublons () ; Détecter les valeurs manquent () ; Trouver les incohérent () ;

Nettoyage des données	C'est une classe d'où on a trouvé les méthodes d'amélioration des données.	Eliminer les doublons () ; Normaliser les incohérent () ; Traitement des données manquant () ;
qualité	C'est une classe abstraite	
dimension	Dans cette classe on a trouvé les différentes dimensions de la qualité.	Get data Dimension () ;
métrique	C'est là on a mesuré les dimensions par des algorithmes	Chaque dimension a des métriques précisées.
ODS ou Base de données	est une classe abstraite pour laquelle les données sont représentées selon un modèle conceptuel et gérées par des processus de traitement, et contient des données.	

III. 1 description de diagramme de classe.

3. Architecture de notre Framework

Nous allons maintenant décrire l'architecture de notre plateforme. On a commencé par montrer les principaux composants de l'architecture et la manière dont ils sont reliés entre eux.

La figure ci-dessous représente l'architecture de notre Framework.

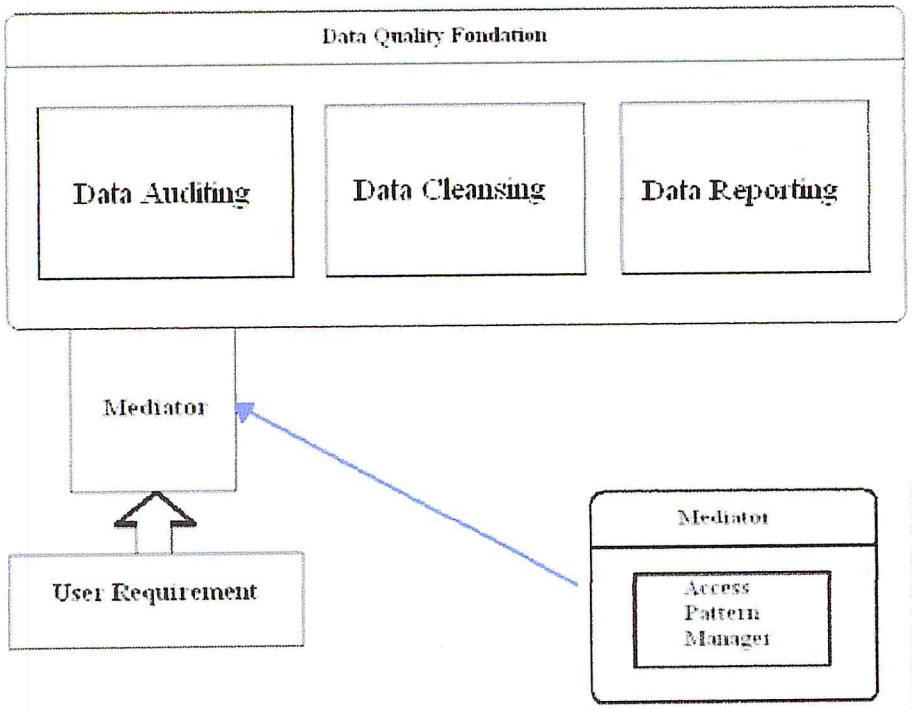

Fig III.3: Architecture Data Quality Framework

Notre plateforme composée de deux principales composantes (Data Quality Fondation, et Médiator).

1. **Data Quality Fondation** : ce composant contient avec son tour trois composant (Data Auditing, Data Cleansing, et Data reporting), Ce composant offre aussi une interface graphique à l'utilisateur pour choisir les indicateur et la source de donnée, ainsi que exécuter les requêtes, et afficher les résultats obtenus et les analyser sous forme tabellaire, et sous forme graphique.
   - Data Auditing: ce composant fournir une moyenne qui permet l'analyse des données afin de déterminer les incohérent, les valeurs manquant et identifier les doublons.

- Data Cleansing : ce composant fait le nettoyage de données par la normalisation des incohérent et des données corrompus ou inexact, et aussi les différents traitements des données manquant.

- Data reporting: ce dernier composant fournir une moyenne qui fait des rapports sur la qualité des données.

2. **Mediator :** fournit la fonctionnalité qui nous permet de choisir le type du SGBD (SQL server, ORACLE, MySQL, etc.).Et le type de driver utiliser pour cette connections, Qui contient le composant suivant :
   - Access Patterns Manager ou Gestionnaire des patterns d'accès est un composant du Mediator, il choisir:

     _ Les Drivers utiliser pour la connections

     _ Le type du SGBD ou est stocké l'objet (SQL sevrer, ORACLE, MySQL, etc.).

Conclusion

Ce chapitre est dédie à la conception et la modélisation de la qualité des données dans les base et l'entrepôt de données, et après cela on a proposé une architecture pour notre Framework, ainsi que les fonctionnalités qui doit fournir cette Framework.

Le chapitre suivant décrit en détail le prototype de notre Framework Data Quality, ainsi que les technologies que nous avons utilisées pour le développer, puis une étude de cas pour tester notre Framework sur des données réelles.

Chapitre VI Le Prototype

Introduction

Ce chapitre sera entièrement consacré à la description de notre Framework conçu, d'une manière générale. Nous allons entre autres décrire les principales composants qui constituent notre Framework et les technologies qui nous ont permis de les développer.

Puis nous allons présenter les différents services et les méthodes fournissent par les composants de notre Framework.

1. l'environnement et outils de développement

Nous utilisons pour implémenter ce Framework comme langage de programmation Visual Studio.net et pour le SGBD SQL server 2000 le travail se fait sur la plateforme Windows.

1.1 visual studio 2008

Microsoft Visual Studio est une suite de logiciels de développement pour Windows conçue par Microsoft.

Visual Studio est un ensemble complet d'outils de développement permettant de générer des applications WebASP.NET, des Services WebXML, des applications bureautiques et des applications mobiles. Visual Basic, Visual C++, Visual C# et Visual J# utilisent tous le même environnement de développement intégré (IDE, Integrated Development Environment), qui leur permet de partager des outils et facilite la création de solutions faisant appel à plusieurs langages. Par ailleurs, ces langages permettent de mieux tirer parti des fonctionnalités du Framework .NET, qui fournit un accès à des technologies clés simplifiant le développement d'applications Web ASP et de Services Web XML grâce à Visual Web Developer.

1.2 Microsoft SQL Server 2000

Microsoft SQL Server est un système de gestion de base de données (abrégé en SGBD ou SGBDR pour « Système de gestion de base de données relationnelles ») développé et commercialisé par la société Microsoft.

2. Le Framework DATA QUALITY

Nous allons présenter dans cette section le Framework que nous avons développé, et qui implémente l'architecture décrite précédemment dans le chapitre III. Ce Framework est composé de deux principales composants : le composants *data qualité fondation* et le composants de médiator (médiateur).

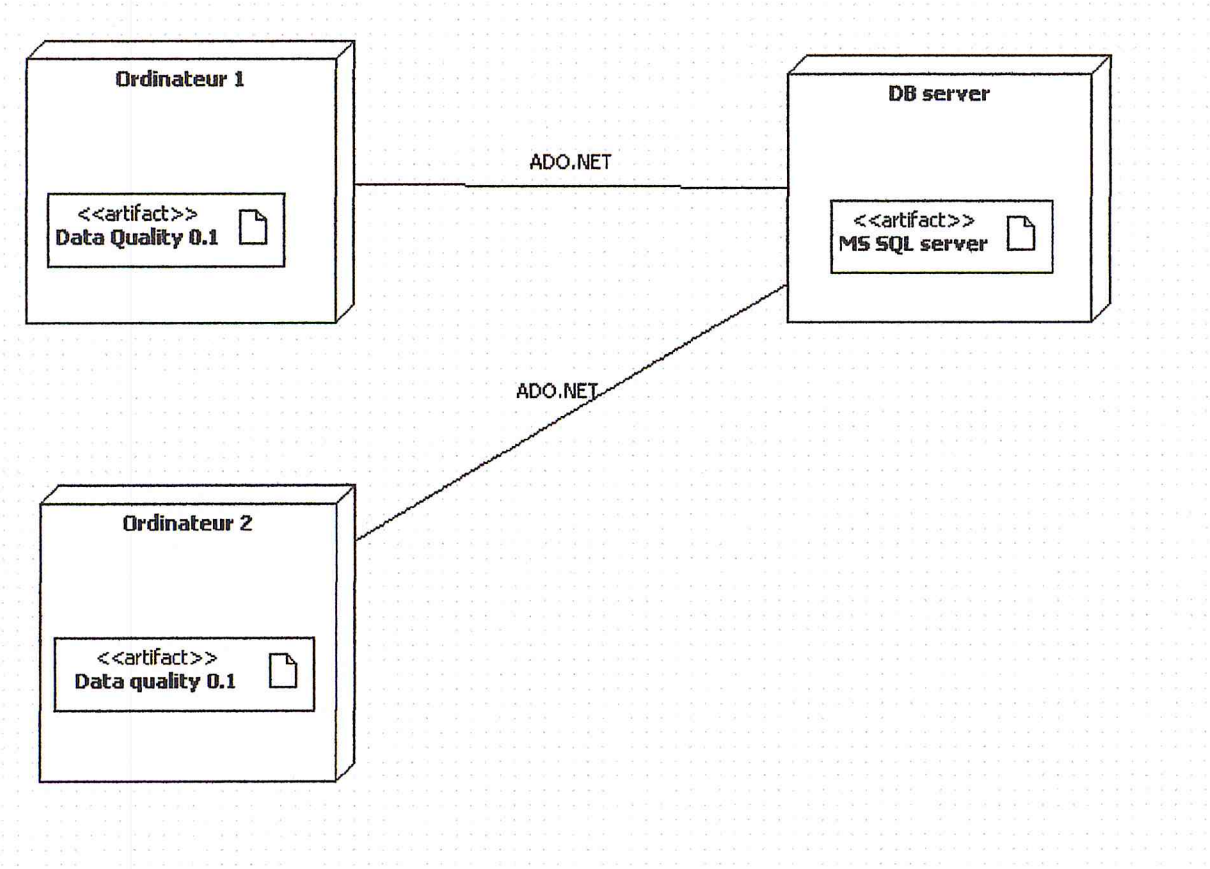

Fig. IV.1Diagramme de déploiement du Framework data Quality

Notre système fonctionne en client/serveur (2/3) puisque l'application doit être installée sur chaque poste client alors que les bases de données à traiter sont regroupées sur un seul serveur de données.

3. Descriptions des différents composants de notre Framework

Notre Framework rassemblé deux composants principaux data Quality fondation et le médiator (médiateur).

3 .1 Data Quality

Ce composant inclut des services qui permet la évaluation et l'amélioration de la qualité des données, ces services sont data Auditing, data Cleansing, et data reporting.

Ci-dessous il y a une description détaillée de ces trios service.

3.1.1 Data Auditing (audit de données)

Data Auditing est un outil qui fournit des méthodes d'évaluation de la qualité à partir de différentes sources de données, afin de déterminé les données d'amélioration, l'analyse fait selon des indicateur définit par l'utilisateur, et le résultat de cette analyse s'affiche se forme tabulaire et se forme graphique.

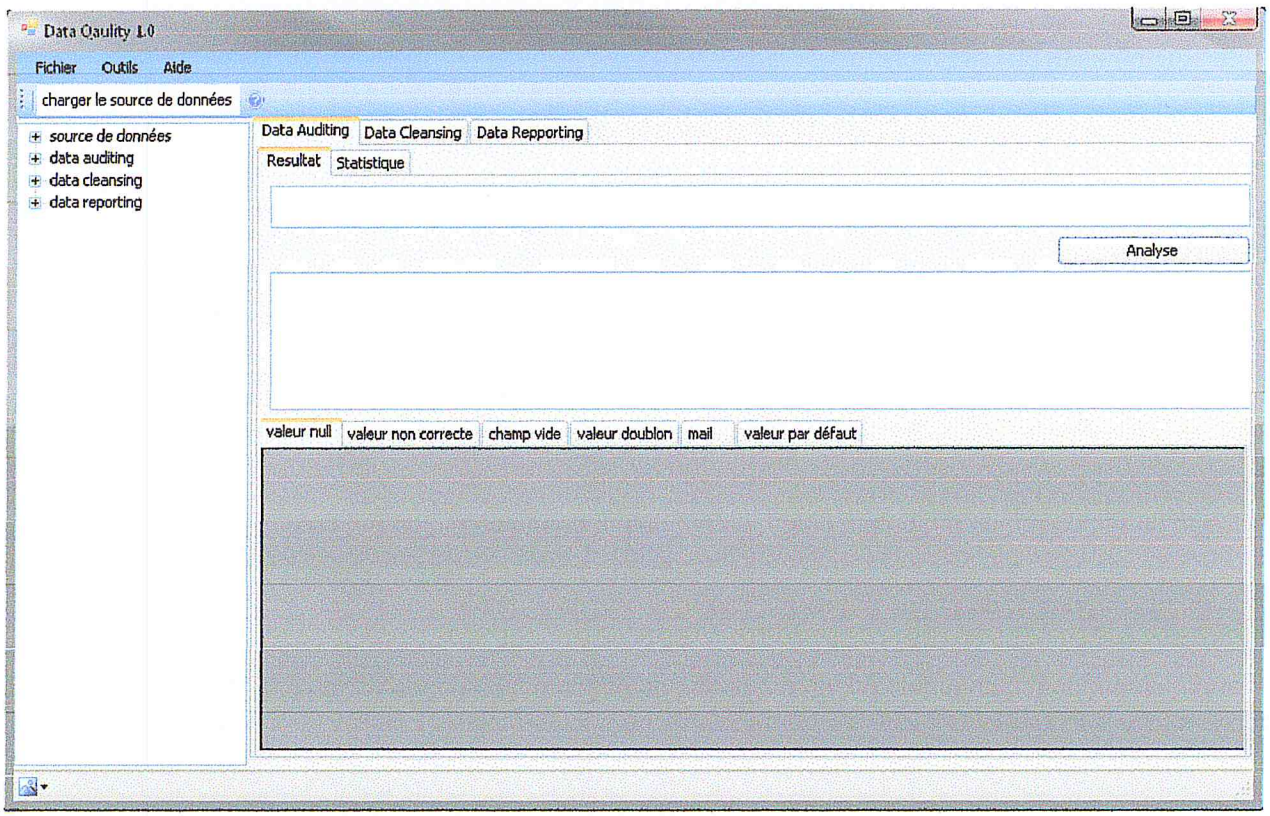

Fig. IV.2 l'interface du composant data Auditing

3.1.2 Data Cleansing (nettoyage des données)

Ce composant sert à améliorer la qualité des données dans les entrepôts des données et dans les bases de données, dans ce composant nous traitons deux problèmes complexes, qui sont le traitement des valeurs manquants et l'limitation des doublons.

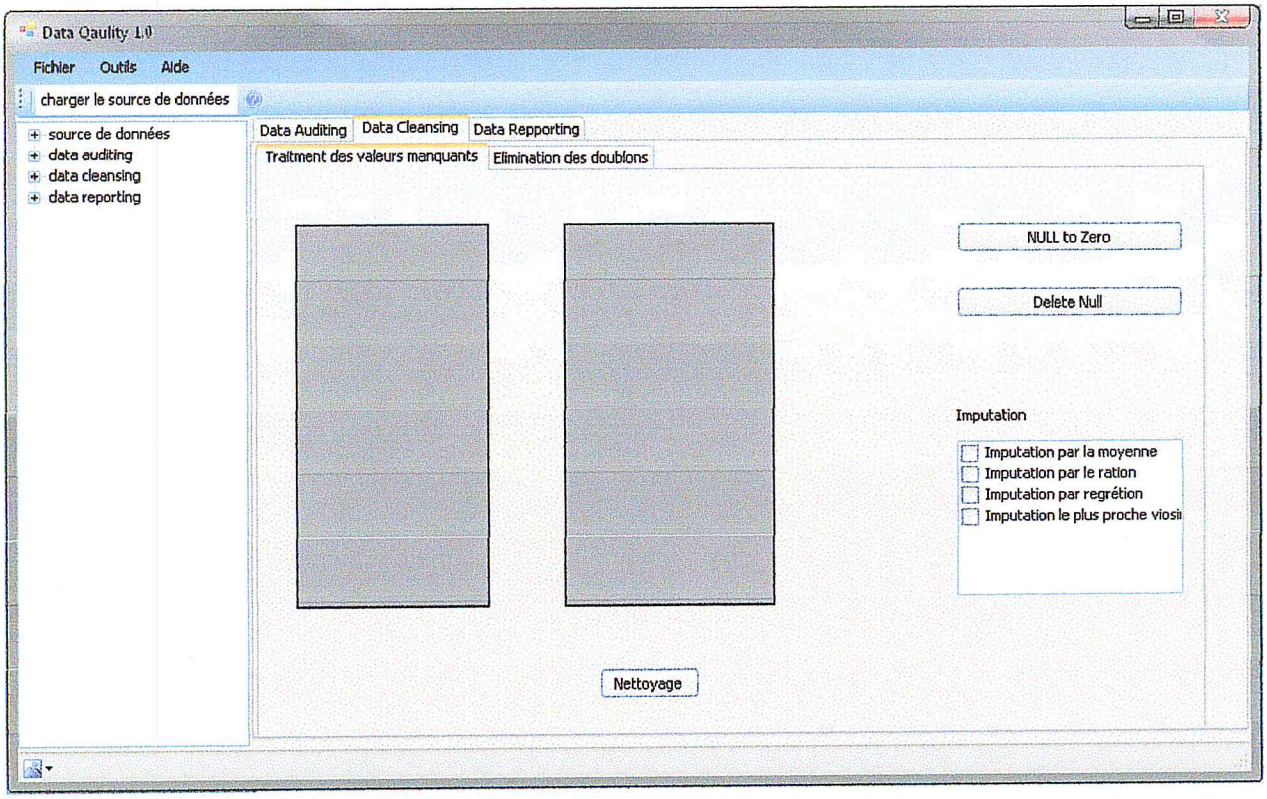

Fig. IV.3 l'interface du composant data Cleansing

3.2 Médiator (médiateur)

Le médiateur représente la logique d'accès à l'importe quels base de données de Microsoft SQL server et l'import quels ODS.

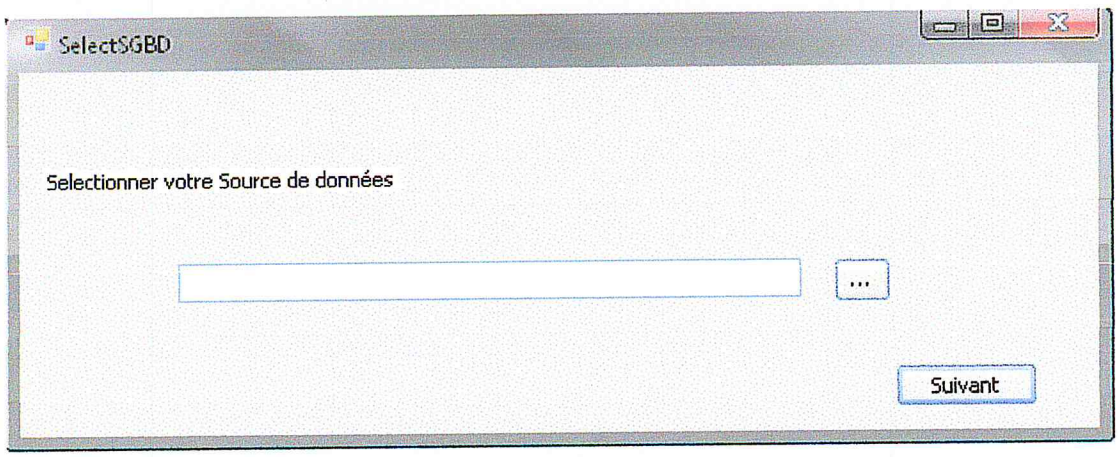

Fig. IV.4 l'interface du composant Médiator

4. Etude de cas

Dans cette section nous présentons une étude de cas, qui se fait sur des données réelles, nous utilisons pour celle-ci les données de la base de données (pubs) fournit par Microsoft, et une ODS d'un entrepôt de données métiers.

4.1 Data Auditing (audit de données)

Alors, nous voulons analyser la colonne (e-mail) dans la table autours et la base de données pubs, afin de connaître quels sont las autours qui n'ont pas un e-mail ou bien leur e-mail non correct. Donc, nous allons à l'interface Indicateur et nous sélectionnons les cases (valeurs null, champs vide, mail) comme représente la figure ci-dessous.

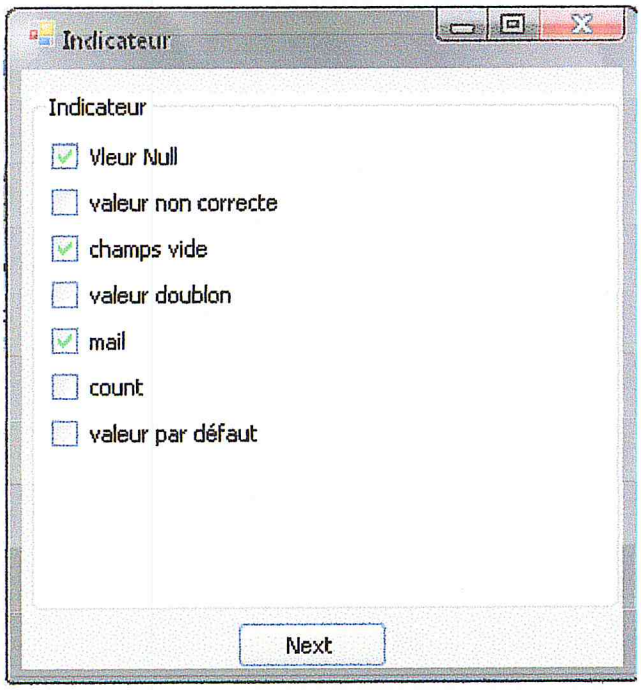

Fig. IV.5 interface pour sélectionner les indicateurs

Ensuite, nous allons charger la source de données (dans ce cas la base de données pubs), puis nous sélectionnons la colonne mail dans la table autours, le résultat de cette analyse s'affiche ci-dessous.

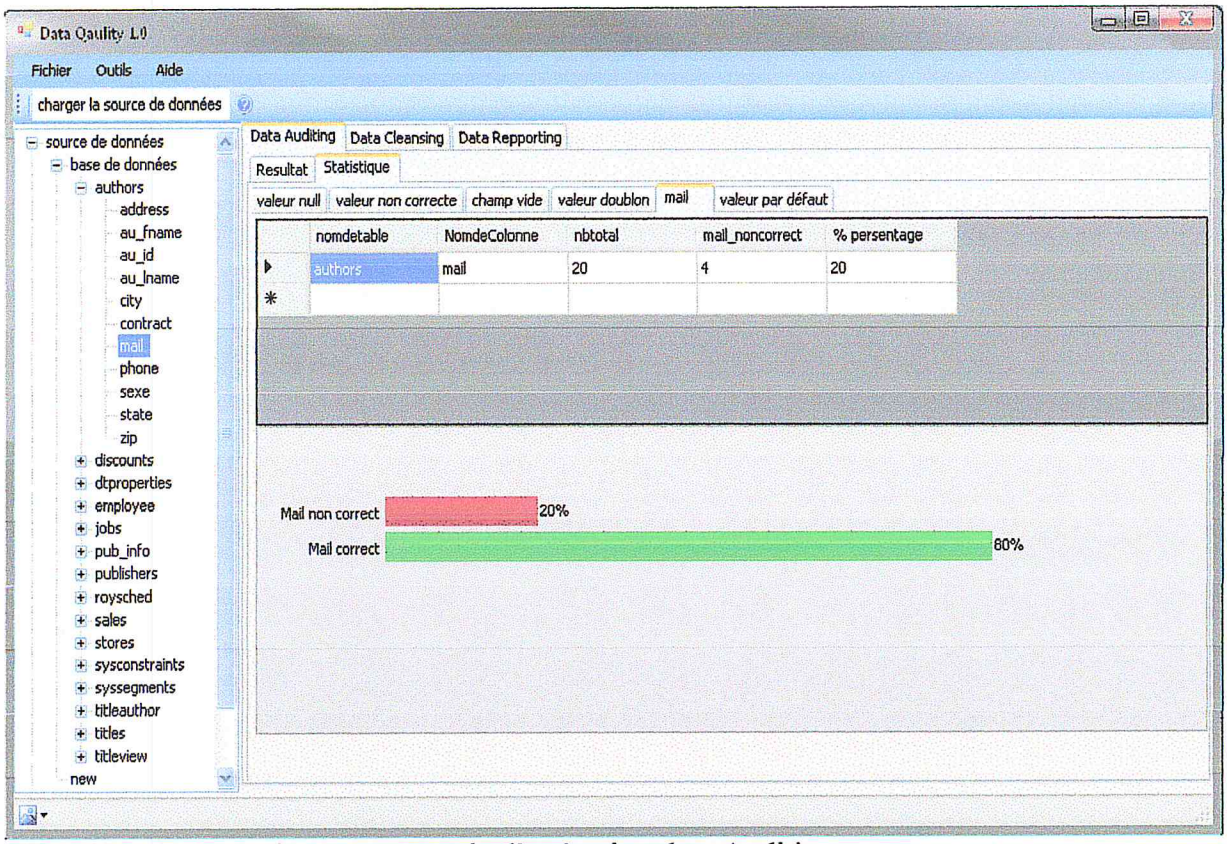

Fig. IV.6 exemple d'exécution data Auditing

4.2 Data Cleansing (nettoyage de données)

La table suivant représente les enregistrements de la table autours (dans ce cas la table autours et de l'ODS de l'entrepôt de données chiffre d'affaire).

Si nous voyons dans cette table, l'enregistrement 1 et 2 ont la même state et le même numéro de téléphone, mais le nom et le prénom syntaxiquement différent.

	au_lname	prenom	state	télé
	Benne	Hells	USA	415 548-7723
	Bennet	Halls	USA	415 548-7723
	Carson	mark	IR	
	DeFrance	alexonder	CA	408 286-2428
	del Castillo	alberto	CA	707 938-6445
	Dull	jake		415 585-4620
	Greene	maickel	CA	503 745-6402

Fig. IV.7 table contient des enregistrements en double

Si nous appliquons l'algorithme de détection et d'élimination des doublons, l'algorithme se trouve que les deux enregistrements représente le même autour, donc data Cleansing supprimer un enregistrement, le résultat de nettoyage de la table autours s'affiche ci-dessous.

au_lname	prenom	state	télé
Bennet	Halls	USA	415 548-7723
Carson	mark	IR	
DeFrance	alexonder	CA	408 286-2428
del Castillo	alberto	CA	707 938-6445
Dull	jake		415 585-4620
Greene	maickel	CA	503 745-6402

Fig. IV.8 résultat de nettoyage de données

Conclusion

Dans ce chapitre, nous avons présenté le prototype développé et les principales technologies que nous avons utilisés. Nous avons notamment présenté les composants de cette Framework et les services de qualité fournis par ces composants (audit ou l'analyse des données, puis le nettoyage des données.

Enfin, nous présentons une étude de cas qui va permettre de mettre en application ce prototype sur des données réelles, ces données sont des bases de données ou d'entrepôt de données.

Conclusion et perspectives

La qualité des données dans les systèmes décisionnels tels que les entrepôts des données ou les bases de données reste un problème d'actualité. Au long de ce mémoire nous avons identifié plusieurs problèmes liés à la qualité des données, nous avons identifié ces problèmes suite à une étude détaillée sur les approches et les méthodologies existantes concernant l'évaluation et l'amélioration de la qualité des données dans les systèmes décisionnels.

Nous avons fait une étude de quelques Framework existants sur le marché traitant de la qualité des données sur les entrepôts et les bases de données, ce qui nous a permis de déduire ce qui suit :

_ L'absence d'un Framework unique qui contient les deux techniques les plus connues (l'audit et le nettoyage des données).

_ Manque d'une interface pour traiter des problèmes spécifiques tels que les valeurs manquantes et les doublons.

Le Framework DATA QUALITY que nous avons développé fournit à l'utilisateur une large collection de services de qualité d'analyse (audit de données) et de Cleansing (nettoyage de données).

Une piste possible dans la continuité de nos travaux, est d'intégrer d'autres techniques et approches qui traitent la qualité des données dans les systèmes décisionnels, ainsi que d'étudier d'autre nouvelles dimensions de la qualité des données, avec l'évaluation des techniques d'analyse et de nettoyage de données par l'intégration de la technologie Web Ontology Language.

Bibliographie :

[1] Marcel P. *Manipulation de données Multidimensionnelles et Langages Règles*, Thèse de Doctorat de l'Institut des Sciences Appliquées de Lyon, 1998.

[2]*Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals.* Paulraj PonniahCopyright © 2001 John Wiley & Sons, Inc.
ISBNs: 0-471-41254-6 (Hardback); 0-471-22162-7 (Electronic)

[3] HABILITATION À DIRIGER DES RECHERCHES présentée devant L'Université de Rennes 1Spécialité : Informatique par Laure Berti-Équille
*Quality Awareness for Managing and Mining Data.*soutenue le 25 Juin 2007.

[4] Amélioration de la qualité des données dans les entrepôts de données et son impact dans les Pratiques organisationnelles. Par Christian Clemmen Année 2000-2001

[5] Un livre blanc de JEMM research janvier 2008

[6] Sixième atelier qualité des données et des connaissances 26 janvier 2010, Hammamet, Tunisie

[7]*Évaluation de la qualité des données et informations* : la prise en compte de l'utilisateurLaure Berti-Équille University of Rennes 1, France Visiting Researcher@ AT&T Labs-Research, NJ

[8] Data Quality Framework March 2010

[9] 7e AtelierQualité des Données et des Connaissances
Evaluation des méthodes d'Extraction des Connaissances dans les Données
En conjonction avec EGC 201125 Janvier 2011Brest, France.

[10]–A Small Sample Survey *Data Quality Tools for Data Warehousing* Using Information in Government Program© 1998 Center for Technology in Governmentthe Center grants permission to reprint this document provided that it is printed in its entirety.

[11] Richard Y. Wang and Diane M. Strong. Beyond accuracy : What data quality means to data consumers. *Journal on Management of Information Systems*, 12 :29, 1996.

[12]Verónika PERALTA. *Data Quality Evaluation in Data Integration Systems*.PhD thesis, Université de Versailles Saint-Quentin-en-Yvelines (France) and Universidad de la República (Uruguay), 2006.

[13] Thomas C. Redman. *Data quality for the information age*.Artech House, Inc, 1997.

[14]Master II Recherche en Informatique Concept aux Systèmes Université de Versailles Saint-Quentin 2008-2009 *Système de médiation à base de services pour l'évaluation de la qualité des données* Réalisé par Mohamed Reda BOUADJENEK

[15]M. Hernandez et S. Stolfo : *Real-world data is dirty: Data cleansing and the merge/purge problem* ; Data Mining and Knowledge Discovery, 2(1), pp. 9-37, 1998.

[16]N. Koudas et D. Srivastava : *Approximate joins: Concepts andtechniques* ; tutorial donné à International Conference on Very Large Databases (VLDB), 1363, 2005.

[17]G. Navarro: *A guided tour to approximate string matching* ; ACM Computer Surveys, 33(1), pp. 31-88, 2001.

[18]Trivellore E. Raghunathan, James M. Lepkowski, John Van Hoewyk et Peter Solenberger*Une technique multidimensionnelle d'imputation multiple des valeurs manquantes à l'aide d'une séquence de modèles de régression* juin 2001

[19] Florence NICOLAU *Traitement des valeurs manquantes et des valeurs aberrantes 2005* – 2006

[20] thèse *qualité des données capteur pour les systèmes de surveillance de phénomènes environnementaux* par Claudia Catalina soutenue le 4 juin 2010

Site web:

www.talend.com

www.informatica.com