Système d'information et qualité des données

Sylvie Damy

Master 1 Informatique I2A

Master DVL Master ITVL

- Finance
- Histoire
- Géographie
- Informatique
- Mathématiques
- Sciences Pour l'Ingénieur
- Français Langue Étrangère
- Administration Économique et Sociale
- Diplôme d'accès aux études universitaires

Master mention Informatique

Parcours Informatique Avancée et Applications (I2A)

CTU
Centre de Télé-enseignement Universitaire
http://ctu.univ-fcomte.fr

Filière Informatique

VVI7MBDA

Bases de données avancées - Qualité des données

Mme DAMY - SYLVIE
sylvie.damy@univ-fcomte.fr

UNIVERSITÉ DE FRANCHE-COMTÉ
UBFC
UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Cette brochure a été réalisée en LaTeX 2_ε.

Préface

Les bases de données correspondent à un type d'outil très largement utilisé dans les entreprises. Toute entreprise a besoin de stocker et surtout d'accéder à une masse d'informations de plus en plus importante. En fonction de la nature de ces informations, des traitements que l'on souhaite pouvoir réaliser, on peut proposer différents types de base de données.
Actuellement :

- les bases de données relationnelles représentent les supports les plus employés ;
- les bases de données objet qui intègrent la notion d'objet dans les bases de données ;
- les bases de données réparties ;
- les bases de données image, multimédia, spatio-temporelle, ... permettent de gérer des informations de nature particulière ;
- les bases de données déductives, logiques proposent une approche spécifique pour les traitements ;
- on parle aussi de fouille de données, de data-mining, ...
- et dans la dernière décennie sont apparues les bases de données NoSQL.

Dans le cadre de ce cours, une première partie a présentée l'approche NoSQL, dans cette seconde partie nous nous intéresserons aux données et à leur qualité. En effet quelque soit le support utilisé pour gérer et stocker des données, la donnée doit répondre aux besoins de ses utilisateurs. Ainsi la qualité de la donnée dépend de l'utilisation que l'on fait de la donnée. J'ai pu à travers mon expérience de développement d'outils s'appuyant sur des bases de données dans le domaine de la recherche appréhender en particulier l'intérêt des données de référence que ce soit pour des problèmes de qualité de données ou d'interopérabilité.

Après avoir présenté les notions de données et d'information dans le chapitre 1, nous verrons les grands types d'applications informatiques de l'entreprise. Dans une seconde partie je présenterai la qualité des données et les causes et conséquences de la non-qualité des données. Enfin dans la troisième partie de ce cours nous verrons une proposition pour améliorer la qualité des données avec le MDM ou *Master Data Management*.

Ce cours est une compilation de différentes lectures : livres, articles et livres blancs ; certaines parties bien rédigées ont été reportées dans le document en italiques, voire ajoutées directement dans le cours (livre blanc Talend, chapitre 3). Suite au bilan de l'année dernière, la partie sur les systèmes d'information : "Les grands types d'applications informatiques dans l'entreprise" n'est présente dans ce support qu'à titre informatif. Il n'y aura pas de question sur cette partie à l'examen.

BDA : Parcours pédagogique

Sem.	Date	Chapitre	Exercice	Devoir	Correction
1	29 Jan	NoSQL - Chap 1-2			
2	05 Fév	NoSQL - Chap 3	Map/Reduce		
3	12 Fév	NoSQL - Chap 3	Map/Reduce		
4	19 Fév	NoSQL - Chap 4	Etudes de cas		
	26 Fév	Vacances			
5	04 Mars	NoSQL - Chap 4	Etudes de cas		
6	11 Mars		TP NoSQL		
7	18 Mars	NoSQL	TP NoSQL		
8	25 Mars			Devoir 1	
9	01 Avr	SI			
10	08 Avr	SI			
	15 Avr	Vacances			
	22 Avr	Vacances			
11	29 Avr	SI		Devoir 2	
12	06 Mai	Révision			Devoir 2
	16 Mai		Examen		

Contents

Part I

Systèmes d'information et données

Chapter 1

Données et information

Les entreprises doivent répondre de plus en plus rapidement à des sollicitations de tout ordre. Dans ce contexte la maitrise de l'information est devenue indispensable et le système d'information joue un rôle essentiel dans l'entreprise.
Nous présentons dans ce chapitre les rôles des systèmes d'information dans les entreprises et de leurs données.

1.1 L'évolution des systèmes informatiques des entreprises

Il y a une quarantaine d'années, les entreprises ont commencé à utiliser des systèmes client-serveur pour leurs applications, remplaçant les systèmes centralisés des années 1960. Dans les années 1990, les ERP ou *Entreprise Resource Planning* qui sont des progiciels de gestion intégrés s'appuyant sur des bases de données relationnelles, ont été adoptés par une grande partie des entreprises. Ces progiciels permettent de gérer l'ensemble des fonctions de l'entreprise (comptabilité, production, commercial, ...) de façon intégrative.
L'avènement d'internet a permis ensuite le partage et la diffusion de l'information. Les progrès réalisés en terme de stockage des données et l'augmentation de la puissance des processeurs ont participé à une large diffusion des systèmes d'information dans les entreprises.
Un système d'information d'entreprise correspond à un réseau complexe de relations entre des hommes, des machines, des processus et procédures plus ou moins formalisés. Toutes ces interactions engendrent des flux d'informations utiles et pertinentes à l'intérieur de l'entreprise et en interaction avec son environnement. Ces informations servent de base aux prises de décisions.
Les systèmes d'informations d'entreprise actuels peuvent être vus comme se composant de 3 sous-systèmes qui correspondent à leur organisation hiérarchique comme le montre la figure 1.1 . A la base de la pyramide, il y a le *système opérant* dit aussi système opérationnel. Il comprend l'ensemble des ressources consacrées à la réalisation de l'activité de l'entreprise. En haut de la pyramide il y a le *système décisionnel* qui commande et contrôle les actions du système opérant afin d'atteindre les objectif fixés. Le système d'information de gestion quant à lui, fait l'interface entre les 2 sous-systèmes précédents.
Le système d'information représente la mémoire de l'entreprise aussi bien pour les données, les flux que les traitements. Il répond à qui fait quoi, où, quand et comment. Il se différencie du système informatique car il n'est pas dépendant d'une plate-forme matérielle.

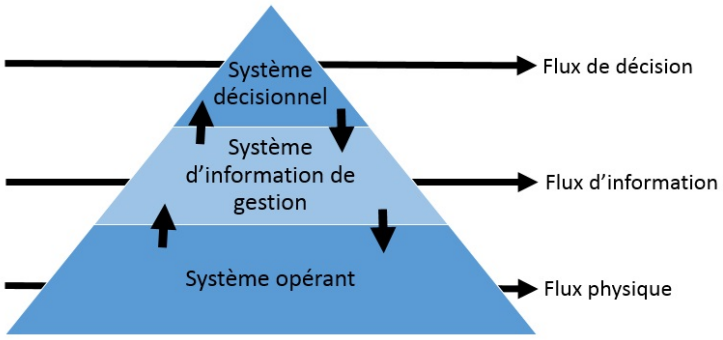

Figure 1.1 : Pyramide des systèmes d'entreprise

1.2 Donnée et information

Définir clairement les notions de donnée et information n'est pas une tâche aisée. On trouve ainsi les définitions suivantes.

1.2.1 Donnée : Définition

Le Larousse propose les définitions suivantes :

- *Ce qui est connu ou admis comme tel, sur lequel on peut fonder un raisonnement, qui sert de point de départ pour une recherche (surtout pluriel) : Les données actuelles de la biologie.*
- *Résultats d'observations ou d'expériences faites délibérément ou à l'occasion d'autres tâches et soumis aux méthodes statistiques.*

Dans wikipédia on trouve la définition suivante : *Une donnée est une description élémentaire d'une réalité. C'est par exemple une observation ou une mesure.*
Cette notion est précisée de la façon suivante :

- *La donnée est dépourvue de tout raisonnement, supposition, constatation, probabilité. Étant indiscutable ou indiscutée, elle sert de base à une recherche, à un examen quelconque.*
- *Les données sont généralement le résultat d'un travail préalable sur les données brutes qui permettra de leur donner un sens et ainsi, d'obtenir une information. Les données sont un ensemble de valeurs mesurables en fonction d'un étalon de référence. La référence utilisée et la manière de traiter les données (brutes) sont autant d'interprétations implicites qui peuvent biaiser l'interprétation finale (limites des sondages).*
- *Par exemple, des données dans un graphique permettront à un être humain d'y associer un sens (une interprétation) et ainsi créer une nouvelle information.*

Dans le livre "MDM : enjeux et méthodes de la gestion des données" [17], la donnée est définie par : *"Description élémentaire de nature numérique ou alphanumérique, représentée sous forme*

codée en vue d'être enregistrée, traitée, conservée et communiquée. Considérée individuellement, elle ne présente que peu d'intérêt humain et n'est donc utile que pour la machine."

On dit aussi que *la donnée est une ressource, un actif de l'entreprise, au même titre qu'un bien matériel, et à ce titre :*

- *elle peut être valorisée,*
- *elle a un cycle de vie.*

1.2.2 Information : Définition

Le Larousse propose la définition : *"Élément de connaissance susceptible d'être représenté à l'aide de conventions pour être conservé, traité ou communiqué."*
Dans wikipédia on trouve la définition suivante : *"L'information est un concept. Au sens étymologique, l'information est ce qui donne une forme à l'esprit. Elle vient du verbe latin* **informare**, *qui signifie "***donner forme à** *" ou "* **se former une idée de** *".*

Dans [17] on propose la définition : *"Données agrégées en vue d'une utilisation par l'homme (par exemple, le résultat d'une requête décisionnelle qui somme des données individuelles est une information). On parle aussi d'élément de connaissance susceptible d'être représenté à l'aide de conventions pour être conservé, traité ou communiqué (image, texte, donnée structurée)."*

Dans le domaine de la gestion, l'information est vue comme [4]: *"Un ensemble d'éléments reflétant une réalité économique ou physique, susceptibles d'apporter de la connaissance utile à l'exercice de l'activité de l'entreprise".*
L'information représente un élément essentiel de la cohérence organisationnelle de l'entreprise.

1.2.3 Donnée versus information

Comme le montre la figure 1.2 les notions de données et d'information sont liées, ainsi la donnée sert à constituer l'information. De plus l'information est elle-même un élément de la connaissance.

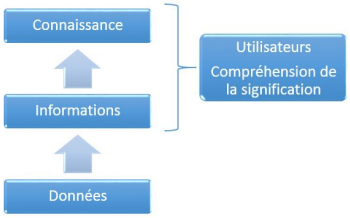

Figure 1.2 : Données, informations et connaissance

D'un point de vue pratique l'information est déduite d'un ensemble de données.
Les termes *donnée* et *information* sont donc assez proches mais pas équivalents. Les données peuvent être définies par des individus mais aussi par le système d'information. Par exemple, la date de création d'une commande, la quantité en stock d'un article, ... sont des données.
Une donnée peut être vue comme un signal, alors que l'information est issue de la transformation de cette donnée par un processus cognitif qui lui donne du sens [23]. Les données sont stockées

dans un système d'information et deviennent des informations dès lors qu'elles sont interprétées dans leur contexte d'utilisation.
Considérons les exemples ci-dessous : Le mois dernier, les données suivantes ont été relevées :

1. 3210 baguettes ont été vendues
2. 600 croissants on été fabriqués
3. 1 intérimaire a été embauché

Ces données peuvent, après interprétation, fournir les informations suivantes :

1. augmentation du nombre de baguettes vendues de 10% par rapport au mois précédent
2. augmentation du nombre de croissants fabriqués de 50% par rapport au mois précédent
3. l'emploi d'un intérimaire est lié à une période de vacances scolaires (périodes durant lesquelles les ventes augmentent).

L'interprétation des données pour proposer des informations suppose la connaissance d'un contexte.

1.2.4 A qui et/ou à quoi servent les données ?

Les données sont une constante au sein des systèmes d'information. L'information dans l'entreprise est variée et dépend des différentes fonctions. Les métiers de l'entreprise ne peuvent être bien exercés que si l'information adéquate est disponible au bon moment, avec le niveau de précision adapté. De plus certaines données doivent être disponibles pour des raisons légales ou fiscales, ...
On peut ainsi citer :

- les contraintes légales. Dans certains pays les entreprises doivent publier leurs comptes, réaliser leur bilan, réaliser des flux de trésorerie, ... Et les entreprises doivent se conformer à des réglementations de plus en plus nombreuses (CNIL, RGPD depuis mai 2018, ...),
- les besoins opérationnels. Ils supposent généralement de saisir, consulter des données, d'éditer des documents. Ici la notion de donnée partagée est essentielle et il faut éviter de dupliquer les données.
- les besoins d'analyse et d'aide à la décision. Pour gérer l'entreprise les managers ont besoin de données synthétiques et pertinentes appelées *indicateurs*. Les données nécessaires à l'analyse sont stockées dans des bases de données dédiées appelées entrepôts de données ou *datawarehouse*.
- l'apparition d'applications capables de traiter spécifiquement de la gestion des données
- les nouvelles orientations technologiques au sein des systèmes d'information (SOA, BPM),
- la valorisation et la protection de la donnée au service de l'entreprise,
- l'émergence des notions de gouvernance des données associées à l'ensemble de ces mouvements.

Historiquement, le système d'information d'une entreprise s'est structuré autour des applications opérationnelles et des applications décisionnelles. On trouve dans le système opérationnel de production toutes les applications transactionnelles de l'entreprise, que ce soit en front, back-office ou support. Toutes ces applications génèrent la grande masse des données qui sont ensuite utilisées, notamment dans les applications décisionnelles, lesquelles ont besoin de données fiables et à jour.
Or dans ce type de système les données sont souvent dispersées dans les différentes applications.

Il est alors difficile de savoir quelle application fournie *"la bonne référence"* pour telle donnée. Or certaines données sont utilisées par plusieurs applications du système et il faut alors savoir quelle donnée est utilisée. D'où l'intérêt d'un système d'information tel que nous l'avons présenté dans le premier paragraphe de ce chapitre, avec 3 sous-systèmes et en particulier le système d'information de gestion.

1.2.5 La structuration des données

Le volume de données ne cesse d'augmenter, et cette augmentation oblige à organiser le stockage et l'accès aux données en utilisant des bases de données.
La technologie a permis une très forte augmentation des capacités de stockage, ce qui a entrainé, avec une diminution des coûts, une augmentation très forte du volume d'informations stockées chaque année. Ce phénomène a été amplifié par l'utilisation d'internet avec les mails, les sites web, ... comme le montre par exemple la figure 1.3
En 2018, une étude d'IDC a estimé que le volume de données en 2025 sera de 175 Zo, soit 5,3

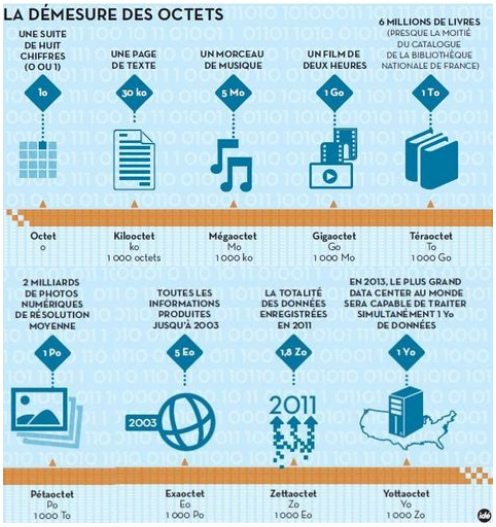

Figure 1.3 : Données stockées - Libération, Déc. 2012 [19]

fois plus quen 2018. L'internet des objets ou IoT ayant une grande part dans cette augmentation. Par ailleurs les sources d'information ont explosé au cours des dernières décennies :

- la mise à disposition d'ordinateurs dans les entreprises a constitué une étape importante dans la prolifération des données stockées et échangées. Chaque utilisateur d'un PC a ainsi pu créer, stocker, diffuser des données très facilement ; il consomme et produit beaucoup d'informations à tort ou à raison.
- l'arrivée d'internet et la généralisation des intranets a aussi accru les sources de données, participant là encore à une multiplication de celles-ci. Citons par exemple l'explosion du courrier électronique : en 2004, près de huit mille milliards de mails professionnels ont été échangés, nécessitant une capacité de stockage globale de cent cinquante mille Tera octets ! En 2018 on estime à 281 milliards le nombre de mail envoyés et reçus chaque jour dans le monde.
- la technologie participe aussi à ce mouvement : sans la contrainte d'un poste fixe, il est désormais possible de consulter, modifier, échanger des informations n'importe où et n'importe quand.

Cette évolution de la technologie a ainsi permis une utilisation sans contrainte qui explique la prolifération des données.
Les bases de données permettent de stocker une bonne partie de ces informations et de faciliter leur accès. Dans de nombreux systèmes d'information d'entreprise on utilise des bases de données relationnelles, cependant les bases de données NoSQL se développent de plus en plus. Dans le cadre qui nous intéresse nous resterons sur des systèmes dits de gestion et plutôt relationnels :

- les SGBDR ne stockent une donnée qu'une seule fois (non-redondance des données),
- les données sont indépendantes des applications qui les utilisent,
- les SGBD assurent la sécurité des données, à savoir leur confidentialité, leur fiabilité, leur traçabilité et leur intégrité.

1.3 Résumé

Nous avons vu dans ce chapitre le rôle essentiel de la donnée et de l'information dans les entreprises. La qualité et la maitrise des données sont un enjeu majeur pour les entreprises. L'explosion des données et de leurs sources a conduit à organiser ces données dans des bases de données.
En bref [9] : *"Dans un contexte où les défis des entreprises et des administrations sont de plus en plus nombreux, disposer d'un capital de données de qualité devient une nécessité incontournable. Déferlement d'informations sans précédent, pressions réglementaires, exigences de contrôle interne, cohérence des échanges avec les partenaires, satisfaction des clients sont autant de défis à relever par les entreprises.* **La maitrise de la qualité des données est désormais un enjeu important.** *Il s'agit de fournir des données correctes, complètes, à jour et cohérentes tout en mettant en place des indicateurs compréhensibles, faciles à communiquer, peu coûteux et simples à calculer. La direction générale et ses directions métiers doivent disposer d'une vision unifiée et exploitable des informations, afin de prendre les bonnes décisions au moment opportun."*

Chapter 2

Les grands types d'applications informatiques dans l'entreprise

Nous présentons dans ce chapitre les grandes applications informatiques d'entreprise, à savoir les ERP ou *Enterprise Resource Planning*, les CRM ou *Customer Relationship Management*, l'e-commerce et l'informatique décisionnelle ou *Business Intelligence*.

2.1 Progiciels de gestion intégrés : ERP

Un ERP ou *Enterprise Resource Planning* encore appelé en français PGI ou Progiciel de gestion intégré est un système d'information qui permet de gérer et suivre au quotidien, l'ensemble des informations et des services opérationnels d'une entreprise. Il répond aux caractéristiques suivantes :

- il émane d'un concepteur unique,
- il garantit à l'utilisateur l'unicité d'information assurée par la disponibilité de l'intégralité de la structure de la base de données à partir de chacun des modules, même pris individuellement,
- il repose sur une mise à jour en temps réel des informations modifiées dans tous les modules affectés,
- il fournit des pistes d'audit basées sur la garantie d'une totale traçabilité des opérations de gestion,
- il couvre soit une fonction (ou filière) de gestion, soit la totalité du système d'information de l'entreprise.

En d'autres termes un ERP est une solution logicielle visant à unifier le système d'information d'une entreprise en intégrant les différentes composantes fonctionnelles autour notamment d'une base de données unique, comme le montre la figure 2.1 .
On trouve sur le marché un certain nombre de produits propriétaires ou libres tels que :

- Propriétaires : SAP, Oracle, GEAC, SAGE, SSA Global, ...
- Libres : Aria, Compiere, ERP5, Fisterra, OFBiz, OpenBravo, PGI Suite, Tiny ERP/Open ERP, TiOlive, Value Entreprise, ...

Figure 2.1 : l'ERP et les composantes fonctionnelles de l'entreprise

2.2 Systèmes de gestion de la relation Client : CRM

Un CRM ou *Customer Relationship Management* encore appelé en français GRC ou Gestion de la relation client, permet de mieux maitriser les interactions avec les clients. C'est un système d'acquisition de données clients qui est orienté en fonction du domaine d'activité de l'entreprise. Le CRM permet de partager une base de données des clients (contacts, prospects, ...) unique consolidée et surtout mise à jour en temps réel. Il permet à l'ensemble des personnes de l'entreprise de disposer d'une vision complète des clients (historique, besoins, commandes, ...).
Ce progiciel permet d'optimiser la relation *Entreprise - Client* avec une approche globale du client (beaucoup d'entreprises sont encore organisées autour des produits). Il permet de gérer l'organisation marketing de l'entreprise.

On trouve sur le marché un certain nombre de produits propriétaires ou libres tels que :

- Propriétaires : SALES Force, INFOCOB, C-FIRST, UPDATE, AKOBA-Solution, ...
- Libres : CRM Open Source, CiviCRM, CremeCRM, Dolibarr, Odoo, Vtiger, ...

2.3 E-commerce

Ces dix dernières années ont vu le e-commerce exploser.

2.3.1 E-commerce : définitions

"Le commerce électronique fait référence à des activités dans lesquelles des entreprises - ou des consommateurs - utilisent Internet pour identifier des fournitures, sélectionner des produits et des services, effectuer des transactions financières et / ou obtenir des services. La livraison peut avoir lieu en ligne ou en dehors d'Internet. La couverture de la recherche comprend des plates-formes et des technologies de commerce électronique ainsi que le commerce électronique mobile." [26]
Wikipédia donne la définition suivante :

"Le commerce électronique (ou e-commerce, commerce en ligne, vente en ligne ou à distance, parfois cybercommerce) est l'échange pécuniaire de biens, de services et d'informations par l'intermédiaire des réseaux informatiques, notamment Internet."

2.3.2 Quelques chiffres

1. Le site Suisse, Bilan ([24]) a affiché pour l'année 2017 les résultats suivant :
   "Les ventes en ligne ont crû de 10% l'an dernier par rapport à 2016, pour totaliser 8,6 milliards de francs. La part du tourisme d'achat sur Internet a explosé de 23% à 1,6 milliard."
2. En France, la FEVAD ou "Fédération e-commerce et vente à distance" ([25]) a publié le bilan e-commerce de la France pour l'année 2017 :
   Le chiffre d'affaires global a progressé de 14% entre 2016 et 2017.
   Les principaux chiffres à retenir sont :
   - Les statistiques du e-commerce (janvier-mars 2017)
     - 81,7 milliards d'euros de chiffre d'affaires, +14,3% par rapport à 2016
     - 1,2 milliard de transactions, +20,5% par rapport à l'année 2016
   - Depuis deux ans, la croissance du e-commerce est tirée par la hausse du nombre de transactions (et non par le panier moyen, en baisse) : +19% depuis deux ans.
   - Les ventes en ligne se banalisent
     - 36 millions de Français ont effectué au moins un achat sur Internet (+1 million)
     - Fréquence d'achat en hausse : +13%
     - On réalise en moyenne 33 transactions par an (28 en 2016)
     - 763 dépensés en moyenne par acheteur lors de l'année 2017
     - Panier moyen en baisse (-5%) : 65,5 vs 69 en 2016
   - La FEVAD ([25]) indique que le panier moyen sur Internet se rapproche progressivement du panier moyen en magasin : l'achat sur Internet se banalise.

Pour avoir les derniers chiffres (2018, ...) concernant la France, vous pouvez consulter le site de la FEVAD.

2.3.3 E-commerce : pour faire quoi ?

On vend tout ou presque sur Internet, mais tout ne se vend pas de la même manière ([20]). Ainsi on peut faire une première distinction entre les différents sites de e-commerce par rapport à ce qu'ils vendent :

- les produits physiques, qu'il faut stocker, manipuler, livrer,
- les services, qui sont achetés en ligne pour être servis à terme: prestation de tourisme, entrée dans un musée, ...
- les produits et services numériques, qui peuvent être livrés par le web : dépôts d'annonces, abonnements à un journal, achats de musique, ...

Ces produits possèdent leurs propres contraintes et spécificités. Tous les outils de e-commerce ne sont pas forcément capables de prendre en compte des notions telles que un calendrier de réservation, la gestion fine de stocks, la sécurisation des contenus pour les produits numériques, ...

Une solution e-commerce ne se réduit pas à un outil de vente en ligne. On peut avoir des sites marchands qui ne proposent pas d'acte d'achat en ligne, mais proposent un catalogue.

2.4 Informatique décisionnelle : BI

Wikipédia propose la définition suivante de l'informatique décisionnelle :
"L'informatique décisionnelle (en anglais business intelligence *(BI) ou* decision support system *(DSS)) est l'informatique à l'usage des décideurs et des dirigeants d'entreprises. Elle désigne les moyens, les outils et les méthodes qui permettent de collecter, consolider, modéliser et restituer les données, matérielles ou immatérielles, d'une entreprise en vue d'offrir une aide à la décision et de permettre à un décideur d'avoir une vue d'ensemble de l'activité traitée."*

On parle aussi d'aide au pilotage ou en anglais de *Business Intelligence*. Ces systèmes permettent d'analyser et de synthétiser les données de l'entreprise pour guider la prise de décision au sein de l'entreprise. Les informations synthétiques sont appelées **indicateurs**.
Ces systèmes sont composés d'un ensemble de solutions informatiques qui permettent de les analyser. Le processus de l'informatique décisionnelle est composé de quatre phases (cf. figure

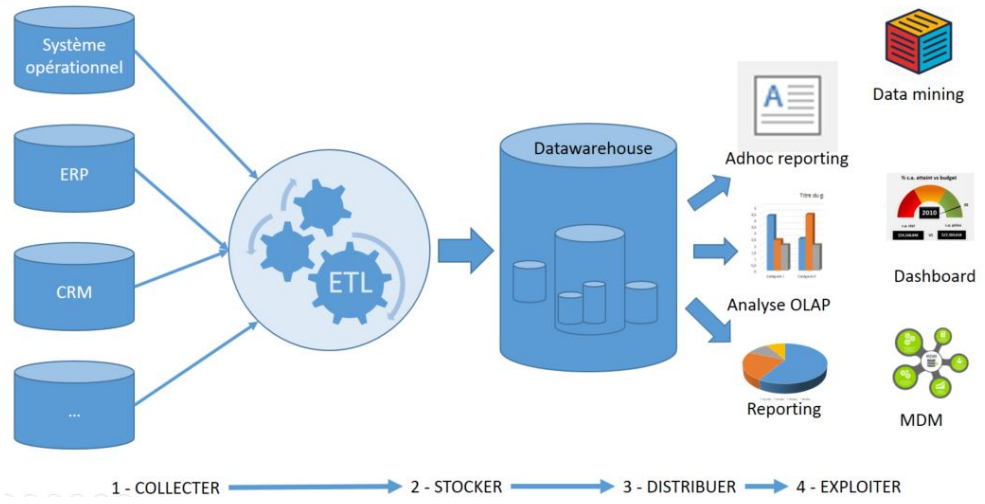

Figure 2.2 : Le flux de données en BI

2.2) : de la donnée à l'information.

1. collecter, nettoyer et consolider les données. Extraire les données des systèmes de production et les adapter à un usage décisionnel ;
2. stocker, centraliser les données structurées et traitées afin qu'elles soient disponibles pour un usage décisionnel ;
3. distribuer ou plutôt faciliter l'accessibilité des informations selon les fonctions et les types d'utilisation ;
4. exploiter ou assister du mieux possible l'utilisateur afin qu'il puisse extraire la substance de l'information des données stockées à cet usage.

2.4.1 ETL

En général les données nécessaires à l'informatique décisionnelle proviennent du système d'information global de l'entreprise qui génère d'immense volumes de données. On doit ainsi collecter les données sources de l'entreprise, ce qui est fait en général avec un ETL ou *Extract Transform and Load*.
L'ETL est un outil qui permet d'extraire, de transformer et d'homogénéiser des données provenant de sources très différentes et hétérogènes. Les données fournies par l'ETL, dans un format adapté, sont ensuite stockées dans un entrepôt de données. Les données sont souvent gardées plusieurs années afin de permettre une analyse et une comparaison des données sur des périodes assez longues.
On trouve sur le marché un certain nombre de produits propriétaires ou libres plus ou moins complets. Notons en particulier des outils tels que Talend ETL , Pentaho, Informatica PowerCenter, SAS ou un outil très simple tel que OpenRefine.

2.4.2 L'entrepôt de données

Un entrepôt de données, ou *datawarehouse*, est une vision centralisée et universelle de toutes les informations de l'entreprise. C'est une une base de données qui regroupe les données de l'entreprise à des fins analytiques et pour aider à la décision. On peut le voir comme un tas d'informations épurées, organisées, historisées et provenant de plusieurs sources de données, servant aux analyses et à l'aide à la décision.
L'entrepôt de données repose souvent sur un modèle de données dit dimensionnel (cf. figure 2.3). Dans un tel modèle les données sont représentées comme des faits (mesures) et des dimensions (contexte : qui, quoi, où, quand, comment) et ces dernières ne sont pas normalisées ; on obtient ainsi des données regroupées selon des catégories qui ont un sens pour l'utilisateur et la dénormalisation accroit les performances en évitant des jointures très couteuses. On parle de schéma en étoile, une table de faits est liées à plusieurs tables de dimensions.
Ce type de modèle est adapté à la définition de plusieurs axes d'analyse des données, on pourra

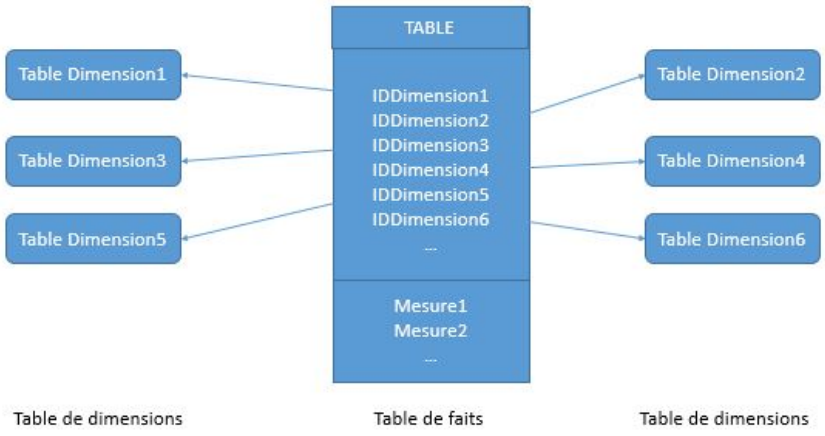

Figure 2.3 : Schéma en étoile

ainsi analyser les données selon des axes géographique, temporel, client, produit, ...

2.4.3 Générateur de rapports ou *Reporting*

Un rapport restitue des informations de façon lisible et synthétique. Il est généralement imprimé. Le générateur de rapports est quand à lui une application qui permet de concevoir un rapport et de le générer. A chaque génération, l'utilisateur peut définir les valeurs des différents paramètres du rapport. La phase de conception reste dédiée à des experts alors que la phase de génération peut être réalisée par tout utilisateur.

2.4.4 Analyse à la demande ou Ad hoc

Dans les années 2000-2010, le besoin de conception de rapports simples pour les utilisateurs finaux a conduit à la proposition d'outils de reporting dits *"Ad hoc"*. On propose deux modes de travail :

- le mode rapport statique : il permet de trouver une information récurrente qui est définie en amont par des experts. Une fois construit ce type de rapport peut être généré de façon quasi-automatique.
- le mode interactif : il permet de chercher une information. Dans ce cas c'est l'utilisateur final qui construit la requête. Les outils permettant ce type de requêtage doivent donc être simples d'utilisation et fournir des résultats avec de bons temps de réponse.

2.4.5 Analyse multidimensionnelle ou OLAP

L'analyse multidimensionnelle ou OLAP (*OnLine Analytical Processing*) est un mode d'analyse courant dans l'informatique décisionnelle. On part de jeux d'informations élémentaires, en grand nombre. Chaque information représente un événement caractérisé par un identifiant unique, des attributs qualifiant l'information et des grandeurs portant une information quantitative.

- Les informations sont *"recomposées"* avant d'être analysées (on dénormalise la base). On obtient ainsi un tableau de données avec beaucoup de redondances.
- Ensuite, ces informations sont agrégées en fonction de certaines caractéristiques, et les valeurs de certains attributs seront additionnées, ...
- Enfin, l'analyse multidimensionnelle sélectionne des axes d'analyse et l'ordre dans lesquels on les utilise et définit des grandeurs qui seront étudiées.

OLAP et les entrepôts de données sont complémentaires. Un entrepôt de données stocke et gère les données tandis que OLAP transforme les données de l'entrepôt en informations stratégiques. OLAP peut réaliser des calculs ou des analyses comme les séries temporelles ou la modélisation complexe.
On peut définir OLAP comme étant l'ensemble des technologies qui, se basant sur une représentation multi-dimensionnelle des données, permettent aux analystes et décideurs de traiter leurs données de façon analytique, interactive et rapide et de voir les données de l'entreprise sous plusieurs angles (dimensions).
Exemple : Ci-dessous voici un exemple tiré du document [21].
Considérons une entité élémentaire : *la ligne de facture de vente*. C'est souvent l'information la plus fine dont on dispose par rapport aux processus de vente.
La ligne de facture porte sur la vente d'un produit à un client à une date (*axes*), dans une quantité, à un prix unitaire (*mesures*).

Sur le client lui-même, on possède d'autres informations : pays, région, type de client, secteur de métier, etc. Par ailleurs, le client est peut-être affecté à un commercial.
L'information des axes peut être hiérarchisée de la façon suivante :

- jour –> mois –> trimestre –> année
- produit –> catégorie de produit
- client –> secteur de métier.

Une première étape est donc l'identification des informations nécessaires aux analyses. Ce qui donne par exemple :

- Date : année, mois, jour, ...
- Produit : SKU, catégorie, ...
- Client : secteur de métier, pays, commercial attitré, ...
- Ligne de commande : Quantité, Prix Unitaire, Chiffre d'affaires.

Les premières informations constituent les axes d'analyse potentiels, la dernière, les grandeurs ou mesures à analyser.
Dans l'analyse multidimensionnelle, la modélisation relationnelle applicative des sources opérationnelles n'est pas la plus pertinente, ni la plus efficace. On préfère généralement une modélisation en étoile où l'on dénormalise les axes, c'est à dire que l'on travaille sur des tables dans lesquelles ont été rassemblées toutes les informations utiles.
Dans l'exemple on obtient :

Date	Prod	Segment	Famille	Client	Pays	Comm.	Qté	CA
10/03/17	0991	Tondeuse	Jardin	Castorama	France	Lepaul	50	50000
10/03/17	0952	Perceuse	Outil	Castorama	France	Lepaul	120	11000
30/04/17	0991	Tondeuse	Jardin	LeroyMerlin	France	Legrand	250	25000
...								

Table 2.1 : Exemple OLAP : Données dénormalisées

Le tableau dénormalisé contient des redondances, mais l'objectif ici n'est pas de gérer les problèmes d'intégrité ou de cohérence de données mais bien de pouvoir simplement d'analyser l'information.
L'étape suivante consiste à réaliser un premier niveau d'agrégation, c'est à dire à réunir certaines lignes. On suppose ici que que les données ne seront pas utilisées au niveau de la référence produit, mais uniquement par segment. On réunit toutes les lignes identiques pour la clé (date, segment, famille, client, pays, commercial), et on cumule les grandeurs quantité et CA.
La dernière étape est celle de l'analyse multidimensionnelle proprement dite, qui consiste à sélectionner des axes d'analyse. Parmi ces axes, on peut distinguer :

- Des axes à valeurs discrètes, ou discontinues, c'est à dire qui portent un nombre fini de valeurs, par exemple un code postal, un segment.
- Des axes à valeurs continues, typiquement une date, un prix. On peut les ramener à un nombre discret de valeurs en définissant des tranches : tranches de prix, tranches d'âges.

On distingue également :

- des grandeurs cumulables, par exemple un montant, un nombre d'items,

– des grandeurs non cumulables, par exemple l'âge ou la date.

Les grandeurs cumulables sont celles qu'il est pertinent d'agréger, c'est à dire dont on peut calculer la somme, (ou la moyenne ou d'autres fonctions mathématiques), pour un sous-ensemble de lignes, par exemple pour chaque thématique. On obtient ainsi le type de schéma (2.4) :

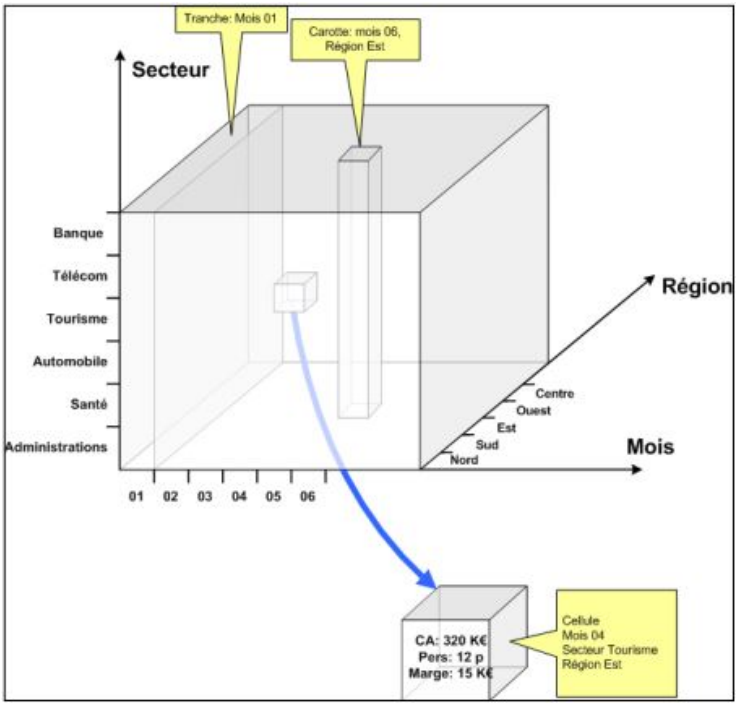

Figure 2.4 : Cube OLAP

2.4.6 Fouille des données ou *Data Mining*

Le Data Mining, appelé encore fouille de données ou forage de données (2.5), est une composante essentielle des technologies Big Data et des techniques d'analyse de données volumineuses.
En règle générale, le terme Data Mining désigne l'analyse de données depuis différentes perspectives et le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns. Le datamining recherche des informations statistiques (tendance, corrélation, similitude, ...) cachées dans de grands volumes de données et qui ne sont pas encore identifiées par l'utilisateur.
Citons par exemple l'analyse des achats sur un site de vente en ligne qui peut faire apparaitre des corrélations entre les achats de certains produits. Ou plus récemment citons les travaux de la société Cambridge Analytica (créée en 2013) qui a étudié des données concernant des citoyens lors de campagnes électorales en utilisant des outils de datamining.

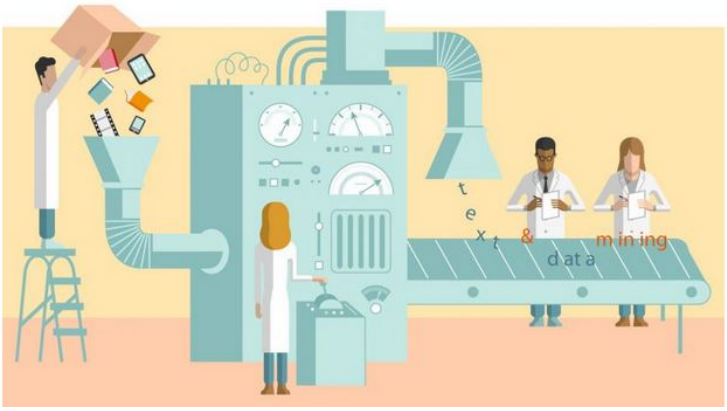

Figure 2.5 : Data Mining

Dis-moi qui tu « likes »,
je te dirai pour qui voter

Pour l'économiste Michael Wade, les méthodes utilisées par Cambridge Analytica apportent un changement majeur dans le domaine du marketing

groupes spécifiques d'électeurs selon des catégories telles que le sexe, l'âge, le revenu, le niveau d'éducation, le nombre de personnes dans le ménage, l'affiliation politique, les préférences d'achats... La machine d'analyse des données d'Hillary Clinton utilisait des

Figure 2.6 : Article du 26 mars 2018 - Le monde

2.4.7 Tableau de bord

C'est une forme particulière de reporting, particulièrement synthétique (tient sur une feuille A4), qui regroupe plusieurs indicateurs significatif de l'activité de l'entreprise et qui peut être personnalisé en fonction de son destinataire.

2.4.8 Master Data Management ou MDM

Le MDM ou gestion des données référentielles, a pour objectif d'assurer la cohérence, la qualité et la pérennité des données de référence dans un système d'information dont les données proviennent de sources de données hétérogènes. Nous reviendrons sur cette gestion des données dans la dernière partie du cours.

2.4.9 Les outils de l'informatique décisionnelle

Citons les outils :

- Libres ([21]): Pentaho, Talend, Birt, Jasper Reports, JPivot, Palo, Weka, ...
- propriétaires ([27]): Business Objects (SAP), SAS, IBM, Microsoft, Information Builders, Oracle, OpenText, MicroStrategy, ...

2.5 En résumé

Cette présentation assez rapide des grands outils du système d'information de l'entreprise montre

l'importance des données. Le coeur du système est la donnée et sans données de qualité aucun système ne saura produire des informations, voire de la connaissance de qualité.

Part II

Qualité des données

Chapter 3

Qu'est-ce que la qualité des données ?

Pour exploiter au mieux ses données, une entreprise doit gérer la qualité de ses données. Ce processus doit être permanent et nécessite la mise en place de processus de qualité des données. Ce chapitre présente la notion de qualité des données et ses différentes facettes. Si actuellement les différents acteurs des données ont pris conscience de l'importance de la qualité des données, il n'existe pas de définition *"universelle"* de la notion de qualité. Celle-ci est décomposée et présentée grâce à des dimensions, caractéristiques, critères, ...

3.1 La qualité

Avant de définir la notion de qualité des données, nous présenterons de façon plus générale la notion de qualité.

3.1.1 Qu'est-ce que la qualité ?

Depuis de nombreuses années des travaux ont été réalisés sur la qualité. D'après [16], W. E. Deming a affirmé que les améliorations de la qualité mènent à l'amélioration de la productivité et donc de la compétitivité. Il a aussi mis en avant que "le client est la partie la plus importante de la chaine de production".
Quelques années plus tard, J.M. Juran a proposé une définition simple de la qualité *"la meilleure adéquation au besoin"* ou *"aptitude à l'utilisation"*. Il identifie 3 raisons qui poussent une entreprise à s'intéresser à la qualité : la baisse des ventes, les coûts d'une mauvaise qualité et les menaces pour l'entreprise liées aux produits de mauvaise qualité. Pour gérer la qualité il propose une trilogie des processus de gestion de la qualité : la planification, le contrôle et l'amélioration de la qualité.
Pour finir P.B. Crosby est revenu sur le rôle essentiel du client : *"la seule caractéristique absolument essentielle dans la gestion du 21^{ième} siècle est celle d'acquérir la capacité à diriger une organisation qui donne à ses clients exactement ce qu'ils demandent et ce avec la plus grande efficacité"*.
Par ailleurs la série de normes ISO 9000 s'intéresse à la gestion de la qualité et la certification

ISO 9000 permet aux entreprises de montrer leurs capacités à fournir des produits ou des services de qualité.

3.1.2 Qu'est-ce que la qualité des données ?

Les problèmes de qualité des données stockées dans les systèmes d'information des entreprises se propagent et touchent tous les types de données et dans tous les domaines d'application (données gouvernementales, commerciales, industrielles, de recherche) [3]. Il s'agit souvent d'erreurs sur les données, des doublons, des valeurs manquantes, incomplètes, obsolètes,... Les conséquences de la non qualité des données sont considérables, d'après une étude présentée en 2015 *"la pauvre qualité des données coûte aux entreprises américaines 600 milliards de dollars par an"*.

On a autant de définitions de la qualité des données que d'articles ou livres parlant de cette notion.
Wikipédia propose la définition suivante en s'appuyant sur des propositions de J.M. Juran :
"La qualité des données, en informatique se réfère à la conformité des données aux usages prévus, dans les modes opératoires, les processus, les prises de décision, et la planification. De même, les données sont jugées de grande qualité si elles représentent correctement le mode de fabrication auquel elles se réfèrent. Ces deux points de vue peuvent souvent entrer en contradiction, y compris lorsqu'un même ensemble de données est utilisé avec un objectif commun".

"Les données sont de haute qualité si elles sont aptes à être utilisées dans le but qui a conduit à les recueillir, que ce soit pour l'aménagement, l'aide à la décision, ou la planification" d'après [22].
Certaines normes telles que les normes ISO 8402:1994, ISO 9000:2005 et enfin ISO 9000:2015 définissent de façon plus générale la qualité comme *"l'ensemble des caractéristiques d'une entité qui lui confèrent l'aptitude à satisfaire des besoins exprimés ou implicites"*. On parle ainsi du *"degré d'adéquation de l'entité à l'usage que l'on en fait"* [?].
La qualité des données est multidimensionnelle : elle implique la gestion des données, la modélisation et l'analyse, le contrôle et l'assurance qualité, le stockage et la présentation ([6]).

La première définition que nous proposons est la suivante : *Une donnée est dite de qualité si elle satisfait les besoins de ses utilisateurs.*
Il est immédiat que la qualité d'une donnée dépend de son utilisation et donc de ses utilisateurs. Une donnée peut ainsi être considérée comme étant de qualité par un utilisateur et être jugée de mauvaise qualité par un autre utilisateur.
Prenons l'exemple d'un fichier client dans lequel on dispose de la commune du client. Ce fichier sera tout à fait adapté pour réalisé des études, pour connaitre le nombre de clients dans une commune, par contre si l'objectif est de livrer le client, ce fichier sera considéré comme de mauvaise qualité. On met en avant ici le coté subjectif de cette notion.
Il n'y a pas de consensus sur la définition même de la qualité des données [2]. Si tout le monde s'accorde sur le fait que la qualité d'une donnée peut se décomposer en un certain nombre de *dimensions*, *critères*, *facteurs*, *éléments* ou *attributs* (les uns, subjectifs nécessitant un jugement et une expertise humaine et les autres, quantifiables et pouvant se mesurer par une grande variété de techniques et de métriques), aucune définition ne fait l'unanimité. Plus de deux cents dimensions ont été recensées dans la littérature.

3.2 Panorama des critères de qualité des données

Il existe différents critères de qualité des données, citons : la qualité du contenu des données avec la justesse, la pertinence, la compréhensibilité, l'accessibilité des données avec la disponibilité et la facilité d'accès aux données, la crédibilité des données, leur complétude, leur actualité, ...
En fonction des auteurs on trouve différents critères, classés de différentes façons.
Les différents travaux sur la qualité des données ont permis de définir tout un ensemble de critères regroupés en dimensions. L'article de Wang al. [14] a ainsi étudié plus 120 articles proposant une classification des différentes dimensions et critères de la qualité de l'information. Nous présentons ici quelques propositions mis en avant dans [2].

Auteurs	**Date**	**Nbre de dimensions**	**Dimensions**
Brodie [5]	1980	6 concepts	Intégrité, Maintenance des données Niveau d'abstraction du modèle conceptuel Expressivité sémantique Validité par rapport à des données de référence Efficacité dans l'utilisation des ressources
Delen, Rijsenbrij [7]	1992	4 dimensions 21 aspects 40 attributs	Développement et contrôle du SI Propriétés statistiques de maintenance Fonctionnement dynamique Importance de l'information : donnée correcte, complète, mise à jour, précise, vérifiable
Wang, Storey Firth	1995	4 catégories 179 attributs	Qualité intrinsèque, d'accessibilité Qualité contextuelle Qualité de la représentation
Redman	1996	4 dimensions pour les valeurs 8 dimensions pour le format de représentation	Précision, complétude, actualité, cohérence Donnée appropriée, interprétable, portable, précision et flexibilité du format, possibilité de représenter les valeurs nulles, utilisation efficace, cohérence
Calabretto,Pinon, Pouillet, Richez	1998	3 critères	Disponibilité Fiabilité, Adaptabilité
Aebi, Perrochon	1998	3 composantes	Donnée correcte, complète et minimale
ICIS [8]	2009	5 dimensions	Exactitude, Actualité, Comparabilité Facilité d'utilisation, Pertinence
Régnier-Pécastaing Gabassi, Finet [17]	2008	4 types de critères	Relativité, Critères intrinsèques Critères de services, de sécurité

Table 3.1 : Propositions de dimensions pour décrire la qualité des données

3.3 Présentation de quelques critères de qualité des données

Nous proposons ici de présenter les critères de qualité de données en partant de la classification proposée dans [17] (dernière proposition du tableau 3.1). Nous verrons ainsi trois grandes catégories de critères de qualité : les critères intrinsèques aux données, les critères de services et les critères de sécurité. En règle générale, les critères d'évaluation de la qualité des données d'un système sont définis en fonction des objectifs et des priorités de l'organisation qui utilise le système d'information.

3.3.1 Critères intrinsèques aux données

3.3.1.1 L'unicité

L'unicité permet de savoir si il existe de multiples et redondantes représentations des mêmes instances de données dans le système.
On souhaite ici que chaque entité du monde réel représentée dans l'entreprise ne le soit qu'une unique fois. L'utilisation dans l'entreprise de différents outils renforce le risque de duplication des données. Ce qui provoque des difficultés de mise à jour des données et la vue de l'entité n'est plus unifiée.
Chaque entité doit être identifiée par un **identifiant unique**. Ce type d'identifiant permet d'éviter les doublons et évite les confusions entre entités distinctes.
Il existe deux façons de définir l'unicité de données :

1. la façon *déterministe* : repose sur des règles définissant quels champs sont caractéristiques d'une entité. Par exemple deux entités ayant le même identifiant correspondent à la même entité, ou deux clients ayant le même téléphone correspondent au même client.
2. la façon *probabiliste* : repose comme la première méthode sur des champs caractéristiques mais aussi sur des fréquences ou approximations portant sur l'ensemble des données correspondant aux entités comparées.

3.3.1.2 La complétude

La complétude permet de dire si toutes les informations requises sont disponibles.
Il est important d'avoir des données complètes, ceci signifie que certains champs de données doivent être renseignés. Il est aussi important d'avoir des groupes avec des données complètes pour pouvoir réaliser certains traitements, la question ici est donc : est-ce que les données nécessaires sont disponibles ? Il faut de plus définir des règles par rapport aux données manquantes et définir des seuils de données manquantes.
Au niveau d'une entité la complétude peut s'exprimer par exemple, pour une entité représentant une entreprise, par le fait que le numéro SIRET soit rempli. Au niveau d'un jeu de données, la complétude peut signifier que beaucoup d'entreprises du CAC40 sont dans la table `Entreprise`.
Ainsi la complétude concerne à la fois un concept manquant, une donnée manquante dans un champ mais aussi des valeurs parasites (numéro SIRET rempli avec la valeur *"0000000"*) et un taux de couverture de base (nombre de données manquantes par rapport à une *"population de base"*) .

3.3.1.3 L'exactitude

L'exactitude permet de savoir si les données représentent bien la réalité.
Une donnée est *exacte* si la valeur des attributs de l'entité concernée est égale ou peut être considérée comme égale à la grandeur qu'elle est censée représenter dans le monde réel. Les notions d'exactitude et de précision sont souvent confondues. Plus formellement on peut définir l'exactitude comme étant la distance entre une valeur V et une valeur V' qui sont respectivement la représentation d'une réalité dans le système et la représentation exacte de la réalité. On fait référence à la proximité des valeurs mesurées, observées ou estimées avec la valeur réelle.

- **Exactitude :** On considère qu'une donnée est exacte lorsqu'elle représente la réalité. Par exemple le code postal *25000* correspond à *"Besançon"*.

- **Précision :** On parle aussi de résolution. La précision statistique correspond à la proximité de valeurs d'observation répétées. On peut avoir une bonne précision statistique mais une mauvaise exactitude.

L'exactitude d'une donnée dépend aussi de son niveau de granularité qui doit correspondre à l'usage de la donnée. Une donnée *"relativement précise"* pourra être utilisée dans certains cas. Considérons une donnée du type *le nombre d'étudiants à l'UFC est supérieur à 20000*, cette donnée est exacte et utilisable pour un traitement travaillant sur des classements d'universités selon leur nombre d'étudiants avec des paliers. Par contre si l'objectif est de pouvoir imprimer des plaquettes pour les fournir ensuite à chaque étudiant, cette donnée n'est pas suffisamment précise.
La figure 3.1 présente l'exemple d'un point sur une carte qui correspond à la réalité et aux mesures qui sont proposées.

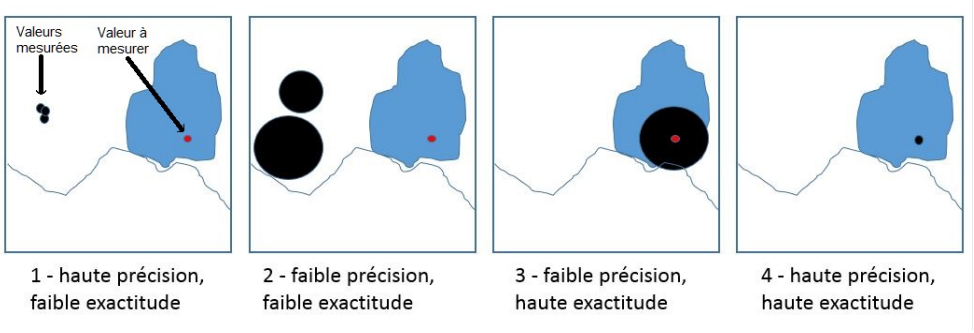

Figure 3.1 : Précision et exactitude

3.3.1.4 La conformité

L'information est-elle dans un format non prévu ?
La conformité d'un ensemble de données est le respect par celles-ci d'un ensemble de contraintes. Considérons par exemple le code INSEE d'une personne ; celui-ci doit être composé de 13 chiffres et commencer par 1 ou 2.
La conformité peut être vue comme une sous-propriété de l'exactitude : des données exactes sont conformes, le contraire n'étant pas vrai. Ce critère est cependant important d'un point de vue pratique car il permet une mise en place simple de contrôles de données dits contrôles de

conformité. Ainsi vérifier qu'un numéro Insee commence bien par 1 ou 2 est simple et permet d'éliminer des premières erreurs de façon simple.
Les contrôles de conformité permettent d'éliminer dès la saisie des données inexactes. La conformité par rapport à un ensemble de contraintes ne constitue cependant pas une garantie d'exactitude.

3.3.1.5 La cohérence

Deux instances distinctes de données d'un même objet produisent-elles de l'information conflictuelle ?
Ce critère est à la fois un critère intrinsèque et de service.
Au niveau intrinsèque, la cohérence signifie l'absence d'informations conflictuelles concernant une même entité. Considérons par exemple un étudiant et l'historique de ses notes dans une unité d'étude. La note conservée par l'étudiant est normalement la plus haute note qu'il a obtenu à ses différents examens. Si la note conservée est inférieure à l'une des notes qu'il a obtenu, il y a un problème car ces deux valeurs sont en conflit.
Au niveau service, la cohérence signifie l'absence de conflit avec les valeurs d'une autre entité. Reprenons notre étudiant, il ne peut pas avoir une note supérieure dans une UE au nombre de points associé à cette UE.
On parle aussi de cohérence avec d'autres sources de données. Dans un système d'information on a de nombreuses données partagées entre les différentes applications informatiques qui le constituent. Il est important de les identifier et des les traiter de façon cohérente. De plus dans la mesure du possible il faut voir si il existe des "normes" de codification de ces données.

3.3.1.6 L'intégrité

Les relations importantes entre objets sont-elles toutes présentes ?
Les données ne sont pas indépendantes les unes des autres, ainsi une note fait référence à un étudiant et une UE. Si la note ne renvoie pas vers une étudiant et une UE, c'est un problème d'intégrité.

3.3.2 Critères de services

3.3.2.1 L'actualité

Les données sont-elles suffisamment à jour au moment de leur utilisation ?
Dans [8] l'actualité est présentée de la façon suivante : *"L'actualité désigne principalement le caractère courant ou à jour des données au moment de leur diffusion selon l'écart entre la fin de la période de référence à laquelle les données se rapportent et la date à laquelle les données deviennent accessibles aux utilisateurs."* On parle aussi d'opportunité. Il existe un rapport assez fort entre les données et le temps.
On parle en particulier d'obsolescence des données. Une valeur de donnée qui a été à un moment exacte peut devenir incorrecte suite à un changement de l'objet observé ou devient périmée à une date donnée.
Considérons les exemples suivants :

- l'âge d'une personne (enregistrée dans un système) devient obsolète à l'anniversaire de la personne.

– considérons un fichier client avec des adresses, un certain pourcentage de ces adresses deviennent obsolètes chaque année. Ces données doivent donc être actualisées régulièrement

3.3.2.2 L'accessibilité

Les données sont-elles facilement accessibles ?
L'accessibilité correspond à la disponibilité de l'information et la facilité avec laquelle on peut y accéder.

– **Disponibilité :** L'information doit être disponible au moment où l'on en a besoin. Ainsi le bilan mensuel d'une entreprise ne peut pas être connu le 31 janvier au soir, car ce calcul suppose de traiter un ensemble conséquent de données mises à jour en toute fin de mois ; cependant si les résultats de l'entreprise doivent être publiés dès le 02 février, ces données devront être disponibles .
– **Facilité d'accès :** Des données disponibles ne signifient pas pour autant qu'elles soient facilement accessibles. Les données-clés et celles régulièrement consultées doivent pouvoir être accessibles en quelques clics de souris et non par un enchaînement fastidieux d'écrans. On touche là à l'ergonomie des applications, élément essentiel pour l'efficacité opérationnelle.

Il est important de pouvoir déterminer l'existence de données. On parle ainsi de données facilement trouvables ou découvrables.

3.3.2.3 La pertinence

Les données sont-elles utiles ?
La pertinence décrit de quelle façon des données répondent aux besoins actuels et potentiels des utilisateurs, elle définit l'utilité de la donnée.
Pour assurer la pertinence, il faut rester en contact avec les utilisateurs. La donnée doit être en adéquation avec son usage, en particulier la granularité de l'information doit correspondre aux besoins.

3.3.2.4 La compréhensibilité

La donnée est-elle compréhensible ?
Les données doivent être compréhensibles par l'utilisateur (humain ou informatique) et ne laisser aucune ambiguïté quand à leur signification et interprétation. Elles doivent aussi être dans un format qui aide l'utilisateur à interpréter les valeurs et respecter les standards (normes, dictionnaires de données) lorsqu'ils existent.
Ainsi on évitera de mettre des initiales pour le nom d'un collecteur de données par exemple : *SD* signifie-t-il *Sylvie Damy* ou *Serge Dupont* ?

3.3.3 Critères de sécurité

Si la sécurité physique et/ou logique d'un système laisse à désirer, les données peuvent être corrompues accidentellement ou volontairement. La sécurité reste l'une des premières règles de bonne gestion des données. Elle recouvre la confidentialité, l'intégrité, et la traçabilité.

3.3.3.1 L'intégrité

Il est nécessaire de protéger l'information, en interdisant des modifications non autorisées, non prévues ou non intentionnelles dans le but de prévenir la corruption ou la falsification des données.
L'intégrité référentielle quand à elle s'applique lorsque qu'une donnée est référencée dans une autre table (clé étrangère).

3.3.3.2 La confidentialité

L'accès aux données peut être filtré en fonction des utilisateurs. Par exemple, les données comptables d'un client pourront être consultées par l'ensemble des utilisateurs de l'entreprise, mais seul le service comptabilité clients pourra créer, modifier ou supprimer ces informations.

3.3.3.3 La traçabilité

Il s'agit ici de "suivre" la donnée. Celle-ci doit être bien documentée, vérifiable et sa source doit être identifiable. On conserve ainsi les opérations effectuées sur la donnée et leurs auteurs.
Les gestionnaires de données doivent garder la trace des opérations de vérification : quelles données ont été vérifiées et quand. Ceci permet d'éviter les redondances et d'empêcher que des données ne disparaissent (journal des opérations).

3.4 En résumé

La qualité des données peut être vue de différentes façons et les critères pour la définir sont nombreux. C'est en fonction de l'utilisation de la donnée que l'on définit les principaux critères de qualité à contrôler.

Chapter 4

Causes et conséquences de la non-qualité des données

A l'origine ce sont des constats sur la non-qualité des données et leurs conséquences qui ont été mis en avant dans les entreprises. Des cas ont été décrits dans des domaines commerciaux, médicaux, domaine public, ...
Nous présentons dans ce chapitre quelques exemples célèbres de non-qualité de données, les principales causes de non-qualité et quelques exemples tirés d'un livre blanc.

4.1 Quelques exemples célèbres de non-qualité de données

4.1.1 NASA

En 1999, la NASA perd un satellite lors de sa mise en orbite autour de Mars [28]. *"Le 23 septembre 1999, le satellite Mars Climate Orbiter doit effectuer sa manoeuvre d'insertion en orbite autour de Mars. Peu avant que la sonde ne survole la planète, la propulsion principale doit fonctionner en continu un peu moins de 17 minutes afin de réduire suffisamment sa vitesse pour qu'elle soit capturée par le champ gravitationnel de Mars. La procédure est entièrement automatique. ... aucun signal n'est reçu à l'heure où la sonde doit réapparaître (11 h 27). L'engin, n'ayant pas repris de contact radio par la suite, est considéré le lendemain comme perdu. Très rapidement, les ingénieurs de la NASA se rendent compte que la trajectoire suivie par la sonde la faisait passer à une altitude beaucoup trop faible au-dessus de la surface de Mars. Au lieu de survoler le pôle au moment de son freinage à 193 km de hauteur, elle est en fait passée à 57 km. À cette altitude, l'atmosphère est beaucoup trop dense pour que la sonde, qui circule à plus de 20 000 km/h, survive. Celle-ci a dû se transformer en une boule de feu au fur et à mesure de son approche de Mars."*
Le rapport de la commission chargée de détecter l'origine de l'anomalie, publié en février 2000, montre que deux logiciels travaillaient avec une même donnée mais exprimée dans des unités différentes. Les calculs de trajectoire effectués sur la base de calculs erronés ont ainsi trop rapproché la sonde de la surface de Mars et entraîné finalement sa destruction.
Cette erreur a couté 125 millions de dollars aux contribuables américains.

4.1.2 Airbus

[29], *"le plus gros client du nouvel avion A380 du constructeur européen Airbus, la compagnie aérienne de Dubaï, Emirates, serait mécontente des premiers appareils qui lui ont été livrés. C'est le magazine allemand Der Spiegel qui l'affirme.*
Selon ce magazine, une réunion de crise aurait eu lieu mi-février sur le site d'Airbus à Toulouse, où Emirates aurait présenté un rapport de 46 pages. Ce rapport aurait fait état, photos à l'appui, de défauts de fabrication sur plusieurs A380 déjà livrés. Ces défauts auraient contraint la compagnie à annuler des vols.
Parmi les défauts mis en évidence figurent des câbles électriques brûlés, des tôles d'habillage arrachées et des problèmes affectant certains éléments des moteurs, selon le magazine allemand. Ces défauts de câblage de la cabine passagers avaient retardé, on s'en souvient, le lancement du gros porteur d'Airbus. La différence des différences des outils de CAO, conception assistée par ordinateur, entre les usines allemandes de Hambourg et les usines de Toulouse étaient à l'origine du problème de câblage."
Ces problèmes de câblage ont retardé la première livraison de deux ans. Il a fallu recâbler 26 avions à la main, et revoir entièrement la conception électrique des appareils suivants. Ce qui a fait exploser les coûts. Ils s'élevaient déjà à 10,2 milliards fin 2006, date à laquelle Airbus a pudiquement cessé d'en publier le décompte. Entre-temps, l'avionneur a dû provisionner 4,9 milliards pour l'A380. En ajoutant diverses dépenses non détaillées, l'ardoise finale atteindrait, selon de bons connaisseurs du dossier, au moins 18 milliards.

4.2 Principales causes de la non-qualité

Les problèmes de qualité des données peuvent survenir à différents moments [2] :

1. Lors de la modélisation des données : les attributs peuvent être insuffisamment structurés ou normalisés, le modèle n'est pas validé ou il manque des contraintes d'intégrité et des procédures pour maintenir la cohérence des données.
2. L'interprétation de données peut être incohérente, suite à l'utilisation de codes ou symboles différents (notamment dans différents pays).
3. Des erreurs peuvent être introduites lors du développement logiciel.
4. Des erreurs humaines peuvent être facilitées par une méthode de saisie des données mal conçue et qui manque de procédures de contrôle.
5. Si la sécurité du système n'est pas suffisante, des données peuvent être corrompues volontairement.
6. Si les entreprises ne gèrent pas l'actualité de leurs données, ne les mettent pas à jour ou ne les enrichissent pas.
7. Lors d'intégration de données à partir de sources hétérogènes, les données peuvent être contradictoires ou incohérentes entre les différents systèmes. de plus la qualité des données de ces différents systèmes peut ne pas être homogène.
8. Lors de la migration de systèmes, on peut voir apparaitre de nouvelles erreurs liées à la perte du contexte de définition, de production ou d'usage de la donnée.

4.3 Exemples de non-qualité de données : Livre blanc Talend

Ci-joint un livre blanc Talend "*les 10 causes principales des problèmes de qualité des données*", qui présente des causes de problèmes de qualité des données.

Les 10 causes principales des problèmes de qualité de données

White Paper

Sommaire

Nous reconnaissons tous les problèmes de qualité de données lorsque nous en voyons. Ils peuvent ébranler la capacité de votre entreprise à travailler efficacement, à respecter les réglementations gouvernementales et à faire des bénéfices. Les problèmes techniques spécifiques comprennent les données manquantes, les attributs saisis dans le mauvais champ, les enregistrements en doublon et les modèles de données cassés, pour n'en citer qu'une partie.

Mais plutôt que de rafistoler de mauvaises données, la plupart des experts pensent que la meilleure stratégie pour lutter contre les problèmes de qualité de données est de comprendre les causes à la racine et d'implémenter de nouveaux processus afin de les éviter. Ce livre blanc traite des dix causes principales des problèmes de qualité de données et suggère des implémentations à effectuer dans votre entreprise afin de les prévenir.

Erreurs typographiques et données non conformes

Malgré l'automatisation quasi systématique de l'architecture des données, des informations sont toujours saisies dans des formulaires Web et d'autres interfaces utilisées par les clients. Les erreurs de saisie sont une source fréquente d'imprécision des données. Les gens font parfois des fautes de frappe. Ils choisissent la mauvaise entrée dans une liste. Ils saisissent la bonne donnée au mauvais endroit.

Etant donnée la complète liberté laissée pour renseigner un champ, les personnes saisissent des données de mémoire. Le nom du vendeur est-il Grainger, WW Granger, ou W. W. Grainger ? Idéalement, il devrait y avoir un ensemble de données de référence d'entreprise (métadonnées) afin que les formulaires permettent aux utilisateurs de trouver les bons vendeurs, noms de clients, villes, références, etc.

Stratégie de l'entreprise

- Formation – Assurer que les personnes saisissant les données connaissent leur impact sur les applications en aval.
- Définition des métadonnées – En restreignant ce que les gens peuvent saisir dans un champ via une liste exhaustive, de nombreux problèmes peuvent être évités. Ces métadonnées (pour le nom des vendeurs, les références, etc.) peuvent devenir une partie de la qualité des données dans l'intégration de données, les applications métier et d'autres solutions.
- Monitoring – Rendre publics le résultat des données mal saisies et féliciter les personnes ayant saisi des données correctement. Vous pouvez en garder un suivi grâce à un logiciel de monitoring de données, tel que Talend Data Quality Portal.
- Validation en temps réel – En plus des formulaires, des outils de validation de qualité de données peuvent être implémentés afin de valider des adresses, des adresses e-mail ainsi que d'autres informations importantes lors de la saisie. Assurez-vous que votre solution de qualité de données puisse diffuser la qualité de données dans des environnements de serveurs d'application, dans le cloud ou dans un Enterprise Service Bus (ESB).

Obfuscation d'informations

Les erreurs lors de la saisie de données ne sont pas toujours faites par erreur. Combien de fois les gens ne donnent-ils pas des informations incomplètes ou incorrectes afin de protéger leur vie privée ? Si cela n'a aucune incidence sur les personnes saisissant les données, elles ont tendance à les falsifier.

Même si les personnes saisissant les données souhaitent le faire correctement, c'est parfois impossible. Si un champ n'est pas disponible, un autre champ est souvent utilisé. Cela peut causer des problèmes de qualité de données, tels que des numéros de TVA dans le champ du nom, ou des coordonnées dans le champ des commentaires.

Stratégie de l'entreprise

- Récompense – Offrir une prime aux gens saisissant des données personnelles correctement. Cela doit être centré sur les gens saisissant des données de l'extérieur, par exemple via les formulaires Web. Les employés ne doivent pas avoir besoin d'une récompense pour faire leur travail. Le type de récompense dépendra de l'importance d'avoir les informations correctes.
- Accessibilité – En tant qu'expert en charge de l'arbitrage de données, soyez ouvert et acceptez les critiques des utilisateurs. Soyez à l'écoute lorsque les changements de processus nécessitent un changement de technologie. Si vous n'êtes pas accessible, les utilisateurs chercheront tous les moyens de valider leur formulaire, même incorrect.
- Validation en temps réel – En plus des formulaires, des outils de validation de qualité de données peuvent être implémentés afin de valider des adresses, des adresses e-mail ainsi que d'autres informations importantes lors de la saisie.

Informatique traîtresse et fichiers Excel

Un renégat est une personne qui déserte et trahit un ensemble de principes d'une organisation. C'est exactement ce que font sans le savoir certains chefs d'entreprise impatients en déplaçant leurs données dans et hors de leurs solutions métier, bases de données, etc. Plutôt que d'attendre de l'aide d'équipes informatiques professionnelles, des équipes métier zélées peuvent décider de créer leur propre ensemble d'applications locales, sans connaissance informatique particulière. Si l'application doit répondre aux besoins immédiats du département, il est peu probable qu'elle soit conforme aux standards de données, de modèles de données ou d'interfaces. La base de données doit commencer par faire une copie d'une base de données approuvée vers une application locale sur le bureau des membres de l'équipe. D'importants morceaux de données stockés dans des feuilles de calcul Excel, dénommés « spreadmarts », sont facilement duplicables sur le bureau des membres de l'équipe. Dans ce scénario, vous perdez le contrôle des versions et des standards. Il n'y a pas de sauvegarde, de versionnement ou de règle métier.

Stratégie de l'entreprise

- Culture d'entreprise – Il devrait y avoir des conséquences pour les personnes propageant ces spreadmarts, les dissuadant de créer des applications de données locales.
- Communication – Eduquez et formez vos employés sur les conséquences négatives des spreadmarts.
- Gestion de « Small data » – Il est crucial d'avoir des outils permettant à des utilisateurs métier et aux professionnels informatique de gérer les données. Les solutions comme Talend Master Data Management (MDM) permettent de combler les

lacunes entre les applications informatiques coûteuses et limitées et la gestion métier effective des données.

- Verrouiller les données – Le but est d'atteindre une culture d'entreprise où créer des spreadmarts sans sanction n'est pas possible. Certaines entreprises ont trouvé le succès en verrouillant les données afin de les rendre plus difficiles à exporter.

Après la fusion

Les fusions d'entreprises augmentent le risque d'erreurs de qualité de données car elles se déroulent en général rapidement et ne sont pas prévues par les départements informatiques. Presque immédiatement, il y a une certaine pression pour consolider et raccourcir le planning. La consolidation inclura probablement le besoin de partager des données dans un ensemble varié d'applications disjointes. De nombreux raccourcis sont pris pour rendre cela possible, ce qui comprend souvent des risques connus ou inconnus pour la qualité de données.

Parmi les priorités du planning raccourci, la fusion des départements informatiques peut provoquer un conflit culturel et différentes versions de la vérité. De plus, des fusions peuvent créer une perte d'expertise lorsque des personnes clés partent au milieu du projet pour chercher de nouvelles aventures.

Stratégie de l'entreprise

- Conscience d'entreprise – Lorsque c'est possible, la direction doit ordonner la division du travail afin d'éviter des conflits de culture et la rétention de l'information par les gens assoiffés de pouvoir.

- Document – Votre projet informatique doit perdurer, même si l'équipe entière part, se disperse ou est écrasée par un bus en traversant la rue. Vous pouvez réussir à le maintenir avec une bonne documentation de l'infrastructure.
- Consultants externes – Le management doit savoir qu'il y a du travail supplémentaire à faire et que des conflits peuvent émerger après une fusion. Des consultants peuvent fournir la continuité nécessaire pour réussir la transition.
- Gestion de données agile – Les solutions open source de data management Talend permettent de garder votre entreprise flexible, vous donnant ainsi la possibilité de diviser et conquérir la charge de travail sans licence coûteuse d'applications commerciales.

Le changement est bon… Sauf pour la qualité de données

Les entreprises subissent des changements de leurs processus métier afin de s'améliorer. C'est une bonne chose, n'est-ce pas ? Les principaux exemples comprennent :

- Expansion de l'entreprise sur de nouveaux marchés
- Nouveaux accords de partenariats
- Nouvelles lois de régulation sur le reporting
- Reporting financier à une entreprise mère
- Réduction d'effectifs

Si la qualité de données est définie comme l'adaptation aux besoins (« fitness for purpose »), que se passe-t-il lorsque l'objectif change ?

C'est cette nouvelle utilisation de données qui apporte des changements dans le niveau perçu de qualité de données, même si les données sous-jacentes sont les mêmes. Il est naturel que les données changent. Lorsqu'elles le font, les règles de qualité de données, les règles métier et les couches d'intégration de données doivent également changer.

Stratégie de l'entreprise

- Gouvernance des données – En mettant en place une équipe de gouvernance de données transverse aux métiers, vous aurez toujours une équipe qui contrôlera les changements que subit votre entreprise et examinera leur impact sur les informations. Cela devrait être dans la charte d'une équipe de gouvernance de données.
- Communication – Une communication régulière et un modèle de métadonnées bien documenté simplifient le changement.
- Flexibilité des outils – L'un des défis lorsque vous achetez des outils de qualité de données embarqués dans des applications d'entreprise est qu'ils peuvent ne pas fonctionner dans toutes les applications d'entreprise. Lorsque vous choisissez des outils, soyez sûr qu'ils soient suffisamment flexibles pour fonctionner avec les données de n'importe quelle application et que l'entreprise soit attachée à la flexibilité et à l'ouverture.

Code caché

Les bases de données commencent rarement leur vie en étant vides. Le point de départ est généralement une conversion de données à partir d'une source de données déjà existante. Le problème étant que, si les données peuvent fonctionner parfaitement dans l'application source, elles peuvent échouer dans la cible. Il est difficile de voir tout le code applicatif maison et les processus spéciaux qui se déroulent derrière les données, sauf si vous les profilez.

Stratégie de l'entreprise

- Profilez tôt et souvent – Ne partez pas du principe que vos données sont adaptées à l'objectif visé parce qu'elles fonctionnent dans l'application source. Le profiling vous donnera une évaluation exacte de la forme et de la syntaxe des données dans la source. Il vous permettra également de savoir la quantité de travail que vous devrez fournir pour les faire fonctionner dans la cible.

- Appliquez des outils de qualité de données lorsque c'est possible – Plutôt que du code personnalisé dans l'application, une meilleure stratégie est de laisser les outils de qualité de données appliquer les standards. Les outils de qualité de données appliquent des standards d'entreprise de manière uniforme, ce qui simplifie le partage des données.

Transition des transactions

De plus en plus de données sont échangées entre les systèmes via des interfaces en temps réel (ou quasiment en temps réel). Dès que les données entrent dans une base de données, cela déclenche des procédures nécessaires à l'envoi de transactions à d'autres bases de données en aval. L'avantage est la propagation immédiate des données à toutes les bases de données correspondantes.

Mais que se passe-t-il lorsque les transactions tournent mal ? Un système qui fonctionne mal peut causer des problèmes aux applications en aval. En fait, même un petit changement dans un modèle de données peut poser des problèmes.

Stratégie de l'entreprise

- Vérifications des schémas – Effectuez des vérifications de schémas dans les flux de vos Jobs afin de vous assurer que vos applications en temps réel produisent des données cohérentes. Les vérifications de schémas vont effectuer des tests simples afin de vérifier que vos données sont complètes et correctement formatées, avant de les charger.

- Monitoring de données en temps réel – Le niveau supérieur aux vérifications de schémas est le monitoring de données proactif avec des outils de profiling et de monitoring de données. Les outils comme Talend Data Quality Portal vous assurent que vos données contiennent le bon type d'informations. Par exemple, si vos références ont toujours une forme et une longueur particulières et qu'elles contiennent un ensemble fini de valeurs, toute variation de cet attribut peut être monitorée. Lorsque des variations se produisent, le logiciel de monitoring vous envoie une notification.

Métamorphose de métadonnées

Le référentiel des métadonnées doit pouvoir être partagé entre de nombreux projets, avec un audit régulier sur l'utilisation et l'accès au référentiel. Par exemple, votre entreprise peut avoir des références et des descriptions qui sont universelles pour le CRM, pour la facturation, pour les systèmes ERP, etc. Lorsqu'une référence devient obsolète dans le système ERP, le système CRM doit le savoir. Les métadonnées changent et doivent être partagées.

En théorie, documenter en totalité ce qu'il se passe dans la base de données et combien sont interdépendants les nombreux processus vous permettrait d'atténuer complètement le problème. Les descriptions et les références doivent être partagées entre toutes les applications concernées. Pour commencer, vous pouvez analyser les implications relatives à la qualité de données de tout changement dans le code, les processus, la structure des données, ou les procédures de collection des données, ce qui vous permet d'éliminer les erreurs de données inattendues. En pratique, cette tâche représente un énorme travail.

Stratégie de l'entreprise

- Modèles de données prédéfinis – De nombreuses entreprises possèdent des définitions très simples du contenu de chaque ensemble de données. Par exemple, l'industrie automobile suit certains standards ISO 8000. L'industrie de l'énergie suit les standards Petroleum Industry Data Exchange (PIDX). Cherchez un modèle de données dans votre domaine pour vous aider.
- Gestion de données agile – La réussite de la gouvernance de données se fait en commençant petit et en construisant un processus qui règle d'abord les problèmes les plus importants du point de vue métier. Vous pouvez tirer parti des solutions agiles

de Talend afin de partager des métadonnées et de mettre en place des processus facultatifs à travers l'entreprise.

Définition de la qualité de données

De plus en plus d'entreprises reconnaissent le besoin de qualité de données, mais il y a différents moyens de nettoyer des données et d'améliorer leur qualité. Vous pouvez :

- Ecrire du code et nettoyer manuellement
- Manipuler la qualité de données dans l'application source
- Acheter des outils pour nettoyer les données

Considérez ce qu'il se passe lorsque vous avez au moins deux de ces types de processus de qualité de données ajustant les données. L'équipe commerciale a une définition du client, la comptabilité en a une autre. A cause des différents processus, les deux équipes ne s'accordent pas sur le fait que les deux enregistrements sont des doublons.

Stratégie de l'entreprise

- Outils de standardisation – Choisissez des outils qui ne sont pas liés à une solution particulière. De la qualité de données uniquement dans SAP, par exemple, ne va pas améliorer vos ensembles de données Oracle, Salesforce et MySQL. Lorsque vous sélectionnez une solution, prenez-en une comme Talend Data Quality, capable d'accéder à toutes les données, n'importe où et n'importe quand. Comme la solution de Talend est flexible, bénéficier d'une solution commune comme celle-ci, à travers différentes plateformes et solutions ne sera pas très coûteux.

- Gouvernance de données – En mettant en place une équipe de gouvernance de données transverse, vous avez les bonnes personnes au bon endroit afin de définir un modèle de données commun.

Perte d'Expertise

Dans presque tous les projets ayant de très grands volumes de données, il y a une personne dont l'expertise concernant les données héritées est exceptionnelle. Ce sont ces gens qui comprennent pourquoi la date d'embauche de certains employés est stockée dans le champ de la date de naissance et pourquoi certains attributs de noms contiennent également des numéros de TVA.

Les données peuvent être une sorte d'historique pour une entreprise. Elles peuvent provenir de systèmes hérités. Dans certains cas, la même valeur dans le même champ signifie quelque chose de totalement différent dans un autre enregistrement. Avoir conscience de ces anomalies permet aux experts d'utiliser les données correctement.

Si vous vous trouvez dans cette situation, voici quelques processus métier que vous pouvez suivre.

Stratégie de l'entreprise

- Profiler et monitorer – Profiler les données vous permet d'identifier la plupart de ces types de problèmes. Par exemple, si un numéro de TVA se trouve dans un champ de nom, l'analyse vous permet de le remarquer rapidement. Le monitoring prévient les récurrences.
- Documenter – Même s'ils peuvent être réticents par peur de perdre la sécurité de leur emploi, assurez-vous que les experts

documentent toutes les anomalies et les transformations devant advenir à chaque déplacement des données.

- Utiliser les consultants – Les employés experts peuvent être si précieux et si occupés qu'ils n'ont pas le temps de documenter les anomalies héritées. Les entreprises de consulting externes sont généralement très bonnes pour documenter les problèmes et fournir une continuité entre l'héritage et les nouveaux employés.

Conclusion

Aujourd'hui, la plupart des entreprises ont compris que leur succès est étroitement lié à la qualité de leurs informations. Les entreprises comptent sur les données pour prendre des décisions importantes pouvant affecter le service clients, la conformité aux réglementations, la chaîne d'approvisionnement et bien d'autres domaines encore. Puisque vous collectez de plus en plus d'informations concernant les clients, les produits, les fournisseurs, les transactions et la facturation, vous devez vous attaquer à la racine des problèmes de la qualité de données. Les outils de qualité de données sont simplement cela, des outils pour attaquer les causes à la racine et résoudre les problèmes n'ayant pas été détectés à la racine. Les outils vont de pair avec les gens et les processus pour conduire à une information de haute qualité et de haute valeur ajoutée.

Vous souhaitez en savoir plus à propos de la suite open source Talend Data Quality ? Visualisez les Webinars en ligne ou téléchargez la dernière version du logiciel open source de data management sur Talend.com.

A propos de Talend

Talend est le leader reconnu du marché en data management open source. Un studio de développement unique fournit une cohérence entre les projets d'intégration, de qualité et de gestion de données, afin que les ressources puissent être partagées et réutilisées, tout en restant ouvertes, intuitives and économiques. Relever les défis de la gestion de données ne signifie pas être excessivement cher ou restreint à une application. Talend bouleverse le modèle propriétaire traditionnel en fournissant une technologie open source établie sur la base des performances, de la simplicité d'utilisation, de l'extensibilité et de la robustesse.

Les produits de qualité de données de Talend fournissent aux entreprises une vue et un monitoring détaillés des données métier et incluent un ensemble complet de fonctionnalités permettant d'améliorer la qualité et l'efficacité des actifs de données critiques.

Talend propose deux options de qualité de données : Talend Open Profiler, un produit de profiling de données open source, disponible sur le site Web de Talend en téléchargement gratuit et Talend Data Quality, qui comprend des fonctionnalités supplémentaires de niveau Entreprise, notamment le nettoyage et la standardisation, le matching et le dédoublonnage, un outil embarqué d'intégration de données pour effectuer rapidement et facilement des transformations de données, et des dashboards de qualité de données basés Web.

Part III

Gestion de la qualité des données

Chapter 5

Les approches pour traiter la qualité des données

Nous avons vu que la qualité des données pouvait être décrite de différentes façons avec des critères et des indicateurs assez différents. De la même façon il existe différentes approches pour traiter de la qualité des données.

La gestion de la qualité des données est à la convergence de différentes disciplines, telles que les bases de données, les statistiques, la gestion des processus, ...
Une telle gestion suppose d'identifier, de mesurer et enfin de résoudre les problèmes de qualité des données. Pour cela les entreprises ont souvent développé de manière empirique des techniques d'amélioration de la qualité des données pour répondre à un problème spécifique à un instant donné. Les spécialistes de domaine (sociétés IT ou universitaires) se sont aussi intéressé à ce problème et ont proposé des approches de gestion de la qualité des données.
Ces différentes approches supposent la réalisation de 4 étapes :

1. Définition de la mesure de la qualité des données en fonction des besoins des utilisateurs, et choix des axes prioritaires de travail.
2. Mesure de la qualité des données.
3. Évaluation de l'impact de la non-qualité et proposition d'un plan d'amélioration.
4. Réalisation du projet d'amélioration et vérification des résultats.

5.1 La mesure de la qualité des données

Dans une démarche qualité des données il faut commencer par définir clairement les caractéristiques attendues ainsi que les critères d'évaluation de la qualité des données (choix des dimensions). Pour chaque entreprise il convient de définir les indicateurs spécifiques et de contrôler leur évolution dans le temps par des mesures.
Chaque mesure peut être subjective, lorsqu'elle mesure la perception des utilisateurs par exemple, ou objective lorsqu'elle correspond au résultat de suivis automatiques de certains indicateurs.

5.1.1 Les mesures subjectives

En 2002, suite au constat que les mesures, analyses et amélioration de la qualité des données étaient le résultat de techniques systématiques, *Lee & al.* proposent un instrument de mesure subjectif de la qualité, appelé *"Information Quality Assessment"* (IQA) ([10]). Ils proposent de mesurer la perception d'un ensemble de 69 caractéristiques, regroupées en 16 dimensions (modèle de Wang et Strong) par des acteurs.
Pour réaliser ces mesures, des **questionnaires spécifiques** sont mis en place avec les utilisateurs des données et des spécialistes de la qualité de l'information. Les mesures obtenues sont quantitatives mais dépendent bien de la perception de la personne qui a répondu.

5.1.2 Les mesures objectives

Ces mesures supposent de proposer des définitions rigoureuses des dimensions et caractéristiques mesurées. Pour commencer on considère que *"le système d'information est là pour fournir une représentation d'un domaine d'application"* [15]. Les problèmes de qualité des données correspondent alors à une représentation inachevée, ambiguë, sans signification ou incorrecte.

5.1.3 Les mesures combinées

Pour obtenir une mesure globale de la qualité des données, *Pipino & al.* ([11]) ont proposé l'utilisation d'une grille permettant de combiner les évaluations subjectives et objectives.

5.1.4 La qualité des données à différents niveaux

La qualité des données peut être abordée en étudiant différents niveaux :

1. la qualité de la représentation des données dans le système d'information, niveau MCD ;
2. la qualité de la gestion des données, niveau processus ;
3. la qualité des données, niveau instance/valeur.

Les principales dimensions mesurables objectives sont décrites dans le tableau 5.1 . Nous présentons dans ce tableau certaines des dimensions déjà présentées au chapitre 2, mais ici nous les associons à l'un des trois niveaux permettant d'aborder la qualité des données.

Niveau	Dimension	Descriptif
MCD	Lisibilité	Donne au MCD une facilité de lecture par sa clarté et sa minimalité
MCD	Complétude	Donne au MCD une couverture de l'ensemble des besoins
MCD	Expressivité	Donne au MCD une richesse descriptive pour représenter naturellement les besoins et la réalité
MCD	Correction	Donne au MCD une conformité par rapport aux spécifications
MCD	Traçabilité	Documentation détaillée et historique de la conception et de l'évolution du MCD
MCD	Simplicité	Restreint le MCD à un ensemble d'éléments nécessaires
Processus	Sécurité	Ensemble des facteurs portant sur l'aptitude du système à préserver les données de toute manipulation hasardeuse ou malveillante
Processus	Fiabilité	Ensemble des facteurs portant sur l'aptitude du système à maintenir les données dans des conditions précises et pendant une période déterminée
Processus	Accessibilité	Ensemble des facteurs portant sur l'aptitude du système à rendre les données consultables et manipulables dans des temps adéquats
Processus	Disponibilité	Ensemble des facteurs portant sur l'effort nécessaire pour l'utilisation des données et sur l'évaluation individuelle de cette utilisation par un ensemble d'utilisateurs
Processus	Maintenabilité	Ensemble des facteurs portant sur l'effort nécessaire pour faire des modifications sur les données et leur schéma
Processus	Interopérabilité	Ensemble des facteurs portant sur l'aptitude du système à permettre et faciliter l'échange de données
Processus	Confidentialité	Ensemble des facteurs portant sur l'aptitude du système à assurer que les données ne sont accessibles que par les utilisateurs dont l'accès est autorisé
Instances	Complétude	Quantité de valeurs renseignées
Instances	Cohérence	Quantité de valeurs satisfaisant l'ensemble des règles de gestion définies
Instances	Exactitude	Quantité de valeurs correctes et sans erreur
Instances	Fraicheur	Ensemble des facteurs qui capturent le caractère récent et d'actualité d'une donnée entre l'instant où elle a été extraite ou créée dans la BD et l'instant où elle est présentée à l'utilisateur

Table 5.1 : Principales dimensions de la qualité

5.2 Les différents types d'approches

Les approches concernant l'évaluation et le contrôle de la qualité des données peuvent être classées selon 4 grandes catégories.

5.2.1 Les approches préventives

Ces approches s'appuient sur l'ingénierie des systèmes d'information et le contrôle des processus. Elles utilisent des techniques permettant d'évaluer la qualité des modèles conceptuels de données, des développements logiciels et des processus traitant les données. Elles effectuent en amont du stockage des données, des évaluations au niveau des modèles et des processus mis en uvre.

5.2.2 Les approches diagnostiques

Ces approches s'appuient principalement sur des méthodes statistiques, d'analyse et de fouille de données exploratoire pour détecter les anomalies dans les données et notamment les erreurs présentes dans de grandes quantités de données.

5.2.3 Les approches correctives

Elles essaient de détecter les erreurs en les comparant à des valeurs issues de la réalité (dites "données de terrain") et proposent des corrections. Ces approches sont basées sur des techniques de nettoyage et de consolidation de données, elles utilisent en particulier des langages de manipulation des données étendus et des outils de type ETL.

5.2.4 Les approches adaptatives

Ces approches, dites aussi actives, proposent des traitements dynamiques de vérification en temps réel de contraintes garantissant la qualité des données. Elles sont généralement appliquées lors de la médiation ou l'intégration de données.

5.3 La gouvernance

Une fois les premières étapes de la mise en place d'une approche de la gestion de la qualité des données réalisées :

- définition de la mesure de la qualité des données,
- choix des indicateurs pour mesurer la qualité,
- mesures et analyse des résultats,

il faut définir des plans d'action à mettre en oeuvre pour corriger la situation. Ici intervient un paramètre dont nous n'avons pas parlé jusqu'à présent : **la gouvernance**. Il faut ainsi formaliser un modèle de pilotage des acteurs, processus et techniques pour assurer la maitrise des données de l'entreprise.
Une telle démarche doit être poussée par la direction et voir l'implication de tous les acteurs de l'entreprise. Il est important de créer un **comité Qualité des données**, sous la responsabilité d'un membre de la direction, qui porte la responsabilité de la qualité des données.
Ce comité doit être composé de correspondants représentant chaque partie de l'entreprise et qui sont responsables des données relevant de leur domaine, ainsi que de la définition des mesures et indicateurs de qualité de leur service (règles métier). Un analyste s'occupe alors de la mise en oeuvre des règles métiers dans des outils de profilage et de nettoyage.

5.4 Les outils de gestion de la qualité des données

La mise en place d'une gestion de la qualité des données passe par la mise en place de solutions technologiques ([9]) qui doivent permettre :

- de faire des diagnostics et d'évaluer les problèmes de qualité,
- de supporter l'intégration de données,
- d'automatiser le traitement des erreurs dans les processus d'extraction et de rechargement de données,
- définir un framework pour capturer et gérer les erreurs liées à la mauvaise qualité des données,
- de fournir un cadre pour mesurer l'évolution des indicateurs dans le temps,
- de fournir des indicateurs de qualité sur les données utilisées.

Les solutions technologiques de gestion de qualité des données se basent sur des outils qui permettent de réaliser du profilage, de la standardisation, du nettoyage, du rapprochement, de l'enrichissement, de la décomposition et de la surveillance.

5.4.1 Le profilage ou *Profiling*

Le profilage est le processus qui consiste à récolter les données dans les différentes sources de données existantes (bases de données, fichiers,...) et à collecter des statistiques et des informations sur ces données.
Plus précisément, les résultats d'un profilage permettent de définir des indicateurs clés associés aux valeurs de chaque donnée, tel que le nombre et le pourcentage de champs nuls ou remplis, le nombre de valeurs uniques, le nombre de fréquences pour chaque valeur et les patterns, les valeurs maximales ou minimales, l'information sur le type de la donnée et la longueur des chaînes de caractères. On peut même obtenir un niveau de détails supplémentaires sur les dépendances entre les colonnes et les relations entre tables, notamment.
L'objectif est d'identifier les anomalies ou partager les spécificités des éléments des données telles que :

- des valeurs manquantes,
- des valeurs présentes mais qui auraient dû être absentes,
- des valeurs qui apparaissent à une fréquence imprévues, qu'elle soit basse ou haute,
- des valeurs qui ne respectent pas un pattern ou un format donné,
- des données aberrantes qui sont bien trop basses ou trop élevées pour la plage définie.

Le profilage est assez proche de l'analyse des données. Il permet d'analyser la qualité des données afin de définir les domaines d'amélioration en étudiant la structure des tables et les relations entre les tables, la pertinence des données (colonnes utilisées, poids des colonnes vides...) et la validité de formats (adresses, informations d'identification...).
Il a pour objectif :

- d'identifier les données réutilisables pour d'autres fins ;
- d'avoir des mesures sur la qualité des données et sur la conformité par rapport aux standards de l'entreprise ;
- d'évaluer les risques engendrés par l'intégration de ces données dans de nouvelles applications ;

- d'évaluer si les métadonnées décrivent correctement les données sources ;
- d'avoir une bonne compréhension de l'enjeu des données sources sur les projets à venir afin d'anticiper de mauvaises surprises ;
- d'avoir une vue globale des données pour permettre la gestion des données de référence ou la gouvernance des données afin de renforcer la qualité des données.

5.4.2 La standardisation

La standardisation des données a pour objectif d'assurer une interopérabilité optimale des données, en vue de leur réutilisation. En utilisant les règles métiers on automatise le processus de vérification et de correction des données afin notamment que les données soient correctement orthographiées, que les abréviations soient standardisées et que les modèles de formatage soient correctement utilisés.
Il existe tout un éventail de normes ISO, prévues pour un usage général dans les divers domaines scientifiques et techniques :

- Références bibliographiques (ISO 690:2010) : donne des principes directeurs pour la rédaction des références bibliographiques, en organisant un ordre dans les mentions. Exemple pour un livre : Nom, Prénom. Titre. Édition, collection, année.
- Représentation des pays (ISO 3166-3:2013) : énonce les principes pour une représentation des pays, *BE* pour Belgique, *FR* pour France ...
- Représentation des monnaies (ISO 4217:2015) : définit le code de trois lettres attribué aux devises dans le monde, *EUR* pour l'euro, *USD* pour le dollar américain, ...
- Représentation normalisée de la localisation des points géographiques par coordonnées (ISO 6709:2008): spécifie notamment la représentation des coordonnées, dont la latitude et la longitude, utilisées pour l'échange de données.
- ...

5.4.3 Le nettoyage ou *Cleansing*

Il permet de détecter et corriger les données corrompues ou inexactes. L'objectif du nettoyage est de rendre la source de données cohérente avec les autres sources de données de l'entreprise. Ce type d'opération est effectué a postériori sur les données contrairement à la standardisation qui a lieu lors de la saisie des données.

5.4.4 Le rapprochement ou *Matching*

Il s'agit de comparer et rapprocher des données de sources différentes pour détecter d'éventuels doublons dans de grands ensembles de données que ce soit dans une ou plusieurs bases de données.
Après avoir identifié les doublons, ou les doublons possibles, le rapprochement de données permet d'appliquer des mesures telles que la fusion des deux entrées identiques ou similaires en une seule. Il permet aussi d'identifier les non-duplicatas, ce qui peut être important pour savoir que deux données similaires ne sont pas les mêmes.

5.4.5 L'enrichissement

Il utilise des sources externes pour compléter les données. L'objectif de l'enrichissement de données est de rechercher le plus grand nombre de données, sans aucun a priori, afin de permettre aux algorithmes la découverte de corrélations qui nont pas pu être observées jusque-là. Des données qui semblent pour certaines sans relation avec d'autres peuvent se révéler fondamentales dans les résultats obtenus. Il est donc important d'enrichir les données avec tous types de données, sans aucune limite et a priori.

5.4.6 La décomposition ou *Parsing*

Permet l'identification, la vérification et la décomposition des éléments des zones de saisie libre un par un. Le parsing consiste à analyser un flux de caractères en entrée et à le segmenter en éléments caractéristiques plus petits (adresse, nom, . . .).

5.4.7 La surveillance ou *Monitoring*

Elle permet d'identifier les problèmes et de réagir avant qu'ils ne génèrent des non-qualités. Cette solution permet de suivre l'évolution des données dans le temps, d'identifier les tendances et d'alerter sur la violation de règles de qualité.

5.5 Quelques bonnes pratiques

Pour obtenir et conserver des données de qualité, il est important d'adopter quelques bonnes pratiques pour la définition, la création et la mise à jour des données ([4]). Ci-dessous sont présentés quelques exemples classiques et relativement simples de bonnes pratiques pour assurer la qualité des données.

5.5.1 La compréhension des besoins

La base de toute activité de conception ou de maintenance d'un système d'information est une bonne compréhension des besoins des utilisateurs. L'analyse et la formalisation des besoins est une phase essentielle ; l'utilisation de diagrammes et de modèles permet de formaliser et valider les besoins des utilisateurs.
L'une des difficultés de cette étape est le dialogue entre les utilisateurs et les informaticiens. Les premiers possèdent une expertise dans leur domaine, mais ne savent pas toujours exprimer formellement leurs besoins. Les informaticiens doivent poser des questions et reformuler les réponses.
Lors de l'étape d'expression des besoins, il est important d'essayer d'avoir une vision globale d'un processus et se focaliser sur ce qui est important, d'autre part il faut chercher à simplifier et éviter la construction de solutions trop complexes de type *"usines à gaz"*. Il faut de plus éliminer les étapes sans valeur ajoutée, identifier les cas d'erreurs possibles et les exprimer clairement, penser les données dès la conception de la solution et consacrer du temps à la documentation.

5.5.2 La codification des données

La codification permet de définir le format et les valeurs possibles des données (attributs).
Elle doit être effectuée avec soin et en tenant compte des évolutions possibles du système d'information.
Il est important de s'appuyer sur des normes lorsqu'il en existe (ISO 80000-3:2006 pour les mesures, ISO 3166 pour les codes de pays, ...). L'utilisation de normes permet l'usage d'un langage reconnu par la communauté et facilite les échanges de données entre systèmes.
Il faut veiller, entre autres, à :

- bien structurer des éléments tels que les adresses et tenir compte des normes dans les différents pays ;
- éviter les codes mnémoniques qui constituent des codifications peu évolutives et ne peuvent se substituer à un descriptif de l'entité qu'ils identifient ;
- définir des règles claires et simples pour les descriptions et libellés. Il faut notamment éviter les abréviations ;
- prendre en compte les aspects multi-langues. En particulier la codification des entités doit pouvoir être reconnue par tous les utilisateurs quelque soit leur pays.

5.5.3 La documentation des données

Il est important de documenter l'ensemble des données, de décrire leur utilisation et leurs principales règles de gestion.
Pour chaque donnée on peut décrire : sa définition, l'utilisation qui en est faite, ses grandes règles de gestion, sa codification et ses contraintes particulières. L'intérêt d'une telle documentation est de centraliser les informations sur les données du système et de pouvoir les mettre à disposition des informaticiens et des utilisateurs. Elle permettra notamment, en termes de qualité de données, de comparer la représentation et la réalité.

5.5.4 L'administration des données

Lors de la mise en place d'une application il est important de définir les procédures de gestion des données et de préciser qui fait quoi.
Il faut ainsi prendre en compte : l'architecture logique de l'application et l'organisation des données. L'architecture logique définit les systèmes logiques nécessaires au fonctionnement des différentes applications. Ainsi, si une application doit être implantée dans différents systèmes, il peut être intéressant de mettre en place un système de données de référence qui centralise les données communes à ces systèmes.

5.5.5 L'organisation de la gestion des données

Il faut ici définir les acteurs qui maintiennent les données et décrire les procédures de gestion des données. En général, on essaie de déléguer la gestion des données à leurs utilisateurs naturels. Par contre dans le cas des données de référence qui sont centralisées, il faut définir proprement quels acteurs peuvent créer, modifier, supprimer une donnée, sous quelles conditions et comment on peut mettre en place des contrôles pour vérifier la correction de ces données.
Toutes les données au sein d'une entreprise ne peuvent pas être centralisées il faut donc étudier chaque type de donnée pour décider ou non de sa centralisation. Le tableau 5.2 présente les

avantages et inconvénients d'une gestion centralisée ou décentralisée des données au sein d'un système.

Gestion	Avantages	Inconvénients
Centralisée	- Respect des normes - Evite les doublons - Allège le travail	- Réactivité - Coût (besoin d'une équipe dédiée)
Décentralisée	- Réactivité - Autonomie - Meilleure appropriation de la donnée par l'utilisateur	- Risque de dériver (langage commun) - Risque de doublon accru - Pollution de la BD (doublons, ...)

Table 5.2 : Gestion centralisée versus gestion décentralisée de la donnée

5.6 Exemples d'approches

Divers organismes, entreprises, ... ont défini des cadres de référence qui permettent de mettre en place des principes de gestion de la qualité des données selon une méthode clairement définie. Chacun de ces cadres contient des solutions qui lui sont propres, mais plusieurs contiennent également des méthodes communes qui peuvent, de ce fait, être considérées comme des incontournables en qualité de données. Le but recherché par ces cadres de référence est habituellement de contrôler la qualité des données comprises dans les systèmes d'information. Pour en arriver à ce contrôle, plusieurs activités sont nécessaires.
Les dimensions qui ont été présentés dans une deuxième partie de ce cours ne sont pas utilisées dans toutes les approches présentées, ou en tout cas pas nommées dimension.

5.6.1 TDQM

Le modèle TDQM ou *Total Data Quality Management* est une adaptation de la gestion totale de la qualité (TQM) de Deming adaptée aux données. Elle a été développée dans les années 90 au MIT (Massachusetts Institute of Technology). Les données ont leur propre cycle pour la qualité : définies, mesurées, analysées et améliorées. Ce modèle propose un parallèle entre la fabrication d'un produit physique et la fabrication d'une donnée qui est traitée comme un produit.
La première étape de cette méthode comprend des phases telles que :

1. identifier les dimensions clés ;
2. donner des définitions précises et significatives pour chacune des dimensions ;
3. définir les mesures de ces dimensions ;
4. développer un algorithme pour calculer la qualité des données.

Quinze dimensions sont utilisées et réparties en quatre catégories (intrinsèque, accessibilité, contextuelle et représentationnelle).

5.6.2 TIQM

Cette méthode est inspirée de la proposition de Deming concernant la qualité qui présente une liste de 14 recommandations en gestion. Elle créé une correspondance de ses processus sur les méthodes : 1) définir; 2) mesurer; 3) analyser; 4) améliorer; 5) contrôler. Elle se caractérise par cinq étapes distinctes de mesures et d'amélioration ainsi que par un procédé général qui permet de suivre et de gérer l'ensemble du projet. On utilise des catégories pour identifier les problèmes de qualité. Dans chacune de ces catégories, on utilise des attributs similaires aux dimensions. Les attributs tels que la précision, la cohérence, la validité, la complétude et l'unicité sont fréquemment utilisés.

5.6.3 ICIS

L'Institut canadien d'information sur la santé est considéré comme un pionnier dans les techniques de gestion de la qualité des données dans le milieu de la santé [12]. Il a mis en place, dans les années 2000, un cadre sur la qualité des données. Ce cadre comprend un outil d'évaluation qui permet de mesurer et de documenter les limites et les forces comprises dans les banques de données de l'ICIS. Cet outil d'évaluation utilise cinq dimensions pour lesquelles 19 caractéristiques et 61 critères ont été définis. Les dimensions utilisées sont l'exactitude, l'actualité, la comparabilité, la facilité d'utilisation et la pertinence.

5.6.4 MDM

L'approche MDM, *Master Data Management* ou Gestion des données de référence, permet de mettre en place un référentiel de données transversales ainsi qu'une organisation adaptée qui permet de gérer ces données. L'objectif est de mutualiser les efforts et d'assurer la synchronisation, le partage et le contrôle des données de référence.

5.7 En résumé

La gestion de la qualité des données n'est pas en soi une tâche insurmontable mais elle nécessite des qualités particulières que ce soit chez les informaticiens ou dans toute l'organisation.
Outre la compétence en informatique, l'une des qualités les plus recherchées chez un informaticien est l'aptitude à communiquer et éventuellement l'association avec une compétence métier (production, environnement, ...).
"Il ne faut pas sous-estimer l'élégance toute simple de l'amélioration de la qualité. Outre le travail en équipe, la formation et la rigueur, elle n'exige pas de talents particuliers. Quiconque le veut peut en être un contributeur efficace." [12].

Chapter 6

Master Data Management ou MDM

Nous présentons dans ce chapitre le **Master Data Management** ou **MDM**, qui est une approche qui s'intéresse à la qualité de certaines données. Cette partie est intégralement issue de l'excellent document "Master Data Management - Mise en place d'un référentiel de données" [18]

L'enjeu du Master Data Management (MDM) est de faciliter la gestion des données de référence, de façon transversale à différentes applications en mettant en place une organisation de circonstance supportée par un référentiel de données. La mise en place d'un tel référentiel permet de se réapproprier ses propres données métier, de les enrichir et d'assurer leur pérennité, indépendamment des processus qui les manipulent. D'un point de vue opérationnel, l'intérêt de l'approche MDM est de mutualiser les efforts et d'assurer la synchronisation, le partage et la qualité des données à travers plusieurs silos d'informations en quasi temps réel.

6.1 L'approche MDM

Nous présentons ici cette approche et en quoi elle est utile pour assurer la qualité des données.

6.1.1 Qu'est ce que l'approche MDM ?

Quelle que soit sa complexité, un système d'information ne fournit une aide efficace que si il peut proposer et traiter des données pertinentes et de qualité. Il est donc primordial de pouvoir mesurer cette qualité, de l'améliorer de manière continue et de la piloter suivant les besoins et usages (*fitness for use*). Cependant, on ne peut mesurer ce qu'on ne contrôle pas et on ne peut contrôler ce qu'on ne connaît pas ou plus.
L'échange contrôlé de données entre applications est un problème complexe. En effet, les applications constituant les systèmes d'information des grandes entreprises ou des administrations publiques sont fortement hétérogènes. Cette hétérogénéité engendre une perte de la maîtrise des données qui, dans la majorité des cas, restent cloisonnées au sein des différentes applications existantes qui sont difficilement interopérables. Au final, les données sont souvent dupliquées dans plusieurs silos fonctionnels, chacun exploitant sa propre base de données avec ses propres structures de données, sa propre interprétation de leur contenu et ses propres règles métier.

L'enjeu du MDM est de pouvoir mettre en place un référentiel de données ainsi qu'une organisation adaptée qui permettront de gérer les données transversalement aux différents projets et applications. Ainsi l'objectif de l'approche MDM est de mutualiser les efforts et d'assurer la synchronisation, le partage et le contrôle des données à travers les différents silos en quasi temps réel.

Un **référentiel de données** consiste essentiellement en une application qui supervise la gestion d'une banque de données alimentée par plusieurs fournisseurs et est consultable par des différents utilisateurs ou consommateurs. Ce référentiel se focalise sur les données à haute valeur ajoutée dont la qualité et l'accessibilité sont cruciales pour les partenaires métier. Ces données sont aussi appelées **données de référence** ou **master data**. L'objectif du référentiel est d'intégrer et d'uniformiser les différentes données reçues et/ou collectées pour ensuite les rendre facilement accessibles. Cette intégration peut-être réalisée de manière logique ou physique :

- *Intégration Logique* : le référentiel de données joue un rôle d'**annuaire de données** permettant aux consommateurs d'identifier à quel(s) fournisseur(s) de données ils doivent s'adresser. Cet annuaire suit le même principe qu'un annuaire téléphonique ; il facilite les échanges de données sans s'occuper de leur contenu.
- *Intégration Physique* : le référentiel de données joue un rôle de **consolidateur de données** et une base de données spécifiquement dédiée à la gestion des données de référence est créée afin de faciliter leur consolidation.

Depuis l'ouverture des systèmes d'information vers l'extérieur, la valorisation et le partage des données sont devenus des éléments primordiaux. Le cloisonnement des données constitue cependant un frein majeur à cette valorisation :

- Dans le meilleur des cas, les applications tentent de garantir la synchronisation des différentes données qu'elles partagent (produits, code pays, clients, employés, patients, etc.). Au fil du temps, cette synchronisation est malheureusement rarement conservée.
- Dans le pire des cas, les différents acteurs n'envisagent même pas la synchronisation de certaines de leurs données simplement en raison du fait qu'ils ne parlent pas le même langage. Les données qu'ils s'échangent ne peuvent être confrontées/comparées car ils ne les interprètent pas de la même manière. Dès lors, sans un référentiel commun pour partager leurs données, toute échange devient infructueux.

Les problèmes liés à la synchronisation et la qualité des données ne sont souvent identifiés que tardivement par les utilisateurs finaux, ce qui est évidemment la pire des situations. La qualité des données ne peut pas être garantie que par le biais d'une opération curative et limitée dans le temps. L'enjeu est de gérer les données transversalement aux applications réparties dans différents services et sur différents sites géographiques. Gérer transversalement des données signifie pouvoir les partager, contrôler leur synchronisation, identifier, en quasi temps réel, leurs problèmes de qualité et les corriger conformément à un processus bien défini. Cette gestion transversale des données amène à investir dans la gestion des données de référence : le **Master Data Management** ou **MDM**.
Le MDM n'est ni une technologie, ni un logiciel mais une méthode qui se focalise sur la rationalisation de la gestion des données partagées au sein d'une organisation ou entre plusieurs organisations. L'objectif du Master Data Management est de gérer de manière unifiée et transversale les données partagées. Malheureusement, elles sont souvent hétérogènes et

dispersées dans plusieurs bases de données non synchronisées. Le MDM veut pallier cette problématique en :

- définissant un **référentiel** commun pour les données partagées,
- automatisant le **partage** et la **synchronisation** des données,
- déterminant les règles de **gouvernance** associées aux données : Qui peut y accéder ?, Qui peut les modifier ?, Qui peut faire du reporting d'anomalies ?, ...
- garantissant la **qualité** des données dans le temps,
- favorisant **l'intégration** des données : soit physiquement dans une base de données commune, soit logiquement dans un annuaire de données déterminant les redirections vers les fournisseurs de données, soit en combinant une intégration physique et logique.

L'approche MDM est avant tout une méthode, qui doit être supportée par des solutions MDM spécifiquement adaptées à cet effet. Au départ, les solutions MDM étaient spécifiques à certains domaines et à certains types de données. Deux catégories de solutions MDM dites **"verticales"** se sont distinguées :

- *la gestion des catalogues produits* ou *Product Information Management* (PIM), notamment dans les domaines de la grande distribution et du manufacturing,
- *l'intégration des données clients* ou *Customer Data Integration* (CDI), particulièrement pour l'administration de grosses bases de données transactionnelles (gestion des doublons, vérification et homogénéisation des adresses, ...) dans des domaines tels que les banques ou les assurances.

À l'heure actuelle, les solutions MDM tendent de plus en plus vers des solutions **"horizontales"** et génériques qui prennent en compte tous types de données et qui couvrent l'ensemble de leur cycle de vie. Généralement, on distingue deux catégories de solutions MDM :

- Le **MDM analytique** où les solutions MDM se limitent principalement à faciliter les prises de décision sur des ensembles de données spécialement adaptées à ce type d'analyse.
- Le **MDM opérationnel** où les solutions MDM permettent de définir, créer et synchroniser les données de référence de qualité nécessaires au bon fonctionnement d'un système transactionnel et délivrées en quasi temps réel.

6.1.2 Pourquoi utiliser l'approche MDM ?

Au fil du temps, les organisations ont constaté qu'elles perdaient le contrôle de leurs données. L'origine de cette perte de contrôle s'explique au travers de différents facteurs.

6.1.2.1 Volume et complexité

Le **volume** et la **complexité** des données ne cessent de croître. La quantité d'information à stocker est de plus en plus importante, le niveau de détails exigé pour décrire une donnée est de plus en plus fin et les structures de données deviennent de plus en plus complexes.

6.1.2.2 Interdépendance des données

Les données sont de plus en plus **interdépendantes**. En effet, les contraintes métier imposent souvent des dépendances fortes entre les données. À tout moment il faut pouvoir être capable de vérifier que ces contraintes sont respectées. L'augmentation du volume de données implique l'augmentation du nombre de ces contraintes, ce qui entraîne une explosion de la complexité.

Dès lors, lorsqu'une donnée est modifiée, il devient de plus en plus difficile d'identifier quels seront les impacts éventuels sur d'autres données et de vérifier que les contraintes métier sont toujours satisfaites.

6.1.2.3 Partage des données

Dans un monde de plus en plus ouvert, où les données sont une ressource à part entière, elles doivent être **partagées**. Le problème est similaire à celui du jeu du téléphone arabe. Si on énonce une phrase quelconque à la première personne en tête d'une file et que je lui demande de la répéter à son voisin et ainsi de suite, j'ai la quasi certitude, que, lorsque cette phrase arrivera à la fin, sa signification aura probablement changé du tout au tout. En effet, plus il existe d'intermédiaires, plus le contenu de la donnée risque d'être altéré ; soit parce que les intermédiaires n'utilisent pas le même langage, soit parce qu'ils estiment que certaines corrections peuvent être apportées ou que certaines données sont négligeables.

6.1.2.4 Qualité des données

Si aucune mesure n'est prise, la **qualité des données** manipulées a naturellement toujours tendance à se détériorer. Soit les données sont mal introduites, soit elles sont incomplètes, soit elles ne sont plus à jour, soit elles sont corrompues lors de mauvaises manipulations, ...

6.1.2.5 Dispersion et duplication

Certaines données sont **dispersées** et **dupliquées**. Les données peuvent être dupliquées pour des raisons d'efficacité ou de facilité d'accès. Cependant, il faut toujours s'assurer que ces données dupliquées restent synchronisées et qu'elles n'évoluent pas de manière anarchique. Malheureusement, cette synchronisation n'est souvent pas garantie, ce qui entraîne l'apparition de données **hétérogènes**. Si l'adresse d'un même citoyen est différente d'une application à l'autre, il faut en étudier la raison avec précaution. On doit être capable, d'une part, de retrouver et de rendre accessibles ses données où qu'elles se trouvent et, d'autre part, de détecter et de lever les divergences entre données hétérogènes. Les origines de ces divergences se retrouvent soit au niveau de leur contenu (valeur), soit dans la manière dont elles sont interprétées (définition) :

- D'une part, une donnée peut avoir une définition univoque mais des valeurs hétérogènes ; soit parce qu'une faute de frappe a été malencontreusement introduite, soit parce que le rythme de mise à jour des données est différent. Par exemple, l'adresse du domicile légal d'un citoyen a une définition juridique univoque mais un même citoyen peut avoir trois adresses différentes selon les applications considérées. Dans la première application, il est domicilié " 176 avenue des alouettes ", dans la seconde, " 176 avenue des allouettes " et dans la troisième, " 13 rue des mésanges ". Si la première adresse est considérée comme son véritable domicile légal, on peut présupposer qu'une erreur de frappe a été introduite dans la deuxième application, tandis que la troisième application n'a pas encore été avertie du changement d'adresse.
- D'autre part, une même donnée peut avoir des définitions différentes suivant l'application ou le domaine métier considéré. Par exemple, la notion d'adresse peut être interprétée soit comme l'adresse effective d'un citoyen soit comme l'adresse de son domicile légal. Suivant la définition utilisée, la valeur de la donnée est correcte mais les applications ne

l'interprètent pas de la même manière. Si ces différences d'interprétation ne sont pas rendues explicites, les données deviendront difficilement exploitables pour les consommateurs de données. Dans ce cas de figure, même si les données transitent d'une application à l'autre, aucune des deux applications ne peut en tirer profit, que du contraire. En effet, ces données constituent du " bruit " qui risque de perturber le bon fonctionnement des applications.

L'apparition de données hétérogènes ou de mauvaise qualité engendre notamment des dysfonctionnements opérationnels dans des processus métier critiques, des choix stratégiques se basant sur des données potentiellement erronées et la mobilisation d'importantes ressources afin de résoudre les problèmes liés aux données. Les difficultés à synchroniser et à homogénéiser ces données ont naturellement mis en lumière la nécessité de reprendre leur contrôle. Ces difficultés s'accentuent en fonction de l'éparpillement des données dans les applications et du nombre d'applications impliquées dans les échanges.

6.1.2.6 Avantages de la méthode MDM

La donnée est en quelque sorte la matière première de tout système d'information et doit clairement être au centre des préoccupations des entreprises et des institutions. Celles-ci vont devoir se doter d'architectures efficaces pour valoriser les données accumulées. À l'heure actuelle, les systèmes d'information sont de plus en plus ouverts et nécessitent de s'échanger des données valides mais aussi consolidées et cohérentes entre elles. L'objectif est d'offrir à l'ensemble des acteurs impliqués une vision unique et authentique des données offrant la possibilité :

- de garantir une meilleure réactivité des agents, notamment grâce à la facilité de mener une investigation,
- de répondre de manière plus rapide à des changements de réglementation,
- de restituer des données complètes et de qualité,
- d'échanger plus facilement les données entre applications hétérogènes sous un format standardisé,
- de fournir les données pertinentes rapidement et sous différentes formes afin d'améliorer la capacité décisionnelle du métier et d'augmenter le crédit accordé à ces données.

Un des principaux avantages de l'approche MDM, en particulier la mise en uvre de référentiels centraux de données, est de réduire les coûts des services informatiques :

- Réduire les coûts d'interfaces applicatives en rationalisant les flux de données partagés par différents processus métier et en réduisant le nombre d'interactions entre applications.
- Réduire les coûts des redondances de données en limitant les acquisitions dupliquées de données, réalisées par les différents départements d'une même organisation et en constituant, grâce au référentiel central, un point unique d'acquisition, de stockage et de distribution de ces données pour l'ensemble des fournisseurs-consommateurs dans une ou plusieurs organisations.
- Réduire les coûts de nettoyage de données en centralisant les initiatives d'amélioration de qualité des données, en évitant la multiplication d'initiatives spécifiques à chaque application et en permettant l'identification de doublons inter-applications.
- Réduire les coûts de traitement et de nettoyage de données externalisées en mutualisant les efforts de nettoyage (en utilisant un outil de gestion de qualité des données par exemple)et en les mettant à disposition de tous au moyen d'un répertoire central partagé.

- Réduire les coûts de licence, de support et de matériel des systèmes redondants en réduisant le nombre d'entrepôts de données et en rendant obsolètes ceux contenant des données dupliquées.
- Réduire les coûts de développement et de maintenance en utilisant une plateforme MDM configurable et évolutive offrant des services standardisés d'accès et de modification des données de référence.
- Réduire les coûts de livraison d'information en mettant en uvre un référentiel délivrant des données de qualité, à jour et traçable, et évitant ainsi les aller-retour entre le métier et les services informatiques pour discuter de l'origine, de la pertinence ou de la fraîcheur des données.

6.2 Les concepts

Nous présentons dans cette partie, les concepts de base du MDM, ainsi les différents types d'architecture MDM.

6.2.1 Concepts fondamentaux

Nous décrivons ici le concept de données de référence, comment le gérer et les 3 approches fondamentales de l'approche MDM : la gouvernance, l'intégration et la qualité des données.

6.2.1.1 Données de référence

Assurer la synchronisation et l'intégration des données a un coût non négligeable et il est illusoire de vouloir appliquer ces principes à l'ensemble de ses données. Il faut se concentrer sur un sous-ensemble de celles-ci appelées **données de référence** ou **master data**.
Une donnée de référence est une information de base, fondamentale pour l'activité de l'entreprise, et partagée ou dupliquée dans plusieurs systèmes. Cette donnée métier doit être identifiable et reconnue comme telle partout dans l'organisation, quel que soit le service qui en est responsable, le système d'information, le serveur ou le logiciel qui l'héberge, la traite ou l'enregistre, la division ou la filiale qui la produit.
Les données de référence s'opposent aux données dites transactionnelles qui se réfèrent aux événements relatifs à ces objets métier. Elles possèdent un long cycle de vie et sont sujettes aux changements.
Généralement, ces données concernent des personnes (employés, consommateurs, patients, ...), des objets (produits, immeubles, ...), des lieux (succursales, bureaux, pays, ...) ou des concepts (ventes, contrats, licences, ...).
Gérer des données de référence génère des coûts supplémentaires non négligeables, en conséquence, tous les objets métier ne doivent pas être pris en compte et doivent donc être sélectionnés avec précaution. Les principales caractéristiques d'une donnée de référence sont qu'elle est partagée et/ou échangée avec des tiers, qu'elle possède une haute valeur ajoutée et que sa (ses) source(s) authentique(s) est (sont) reconnue(s).

6.2.1.2 Gestion des données de référence

La gestion des données de référence ou MDM n'est ni une technologie ni un logiciel, mais une démarche, la gouvernance des données, qui met en oeuvre des procédures durables. Cette gouvernance est assurée par une organisation de circonstance, composée d'individus aux tâches précises, et assistée par des outils dédiés en vue d'améliorer la qualité et le partage des données transversalement à l'organisation.
Il n'existe pas de définition de l'approche MDM qui soit communément acceptée par l'ensemble de la communauté. Différentes définitions existent, en voici deux habituellement citées :

1. *MDM is the authoritative, reliable foundation for data used across many applications constituencies with the goal to provide a single view of the truth no matter where it lies.* (MDM est le fondement fiable et faisant autorité pour les données utilisées dans de nombreuses applications et groupes dans le but de fournir une vue unique de la vérité, sans tenir compte de où elle se trouve).

2. *Master Data Management (MDM) is a discipline in which the business and the IT organization work together to ensure the uniformity, accuracy, semantic persistence, stewardship and accountability of the enterprise's official, shared master data. Organizations apply MDM to eliminate endless, time-consuming debates about whose data is right, which can lead to poor decision making and business performance*. (Le Master Data Management (MDM) est une discipline dans laquelle l'entreprise et l'organisation informatique travaillent ensemble pour assurer l'uniformité, l'exactitude, la persistance sémantique, la gérance et la responsabilité des données de base officielles et partagées de l'entreprise. Les organisations appliquent le MDM pour éliminer les débats interminables et fastidieux sur "quelles données sont exactes", ce qui peut conduire à de mauvaises prises de décisions et à de mauvaises performances commerciales.)

Évidemment, ces définitions peuvent prêter à controverse. Ainsi, il est difficilement imaginable de fournir une vue unique de la vérité. La volonté est plutôt d'expliciter une vue commune entre les différents acteurs qui pourra ensuite être contextualisée par ceux-ci suivant leur besoins spécifiques. De plus, les débats concernant la validité et/ou l'authenticité des données ne vont pas disparaître comme par enchantement mais seront rationalisés au sein de l'organisation gérant le référentiel de données.

À ce stade, il est important de distinguer l'approche MDM d'autres approches qui partagent les mêmes idées et principes. À première vue, toutes ces approches semblent répondre à la même problématique. Il faut donc déterminer ce qui est couvert par chacune de ces approches et s'intéresser à la véritable valeur ajoutée du MDM.
Le MDM est une approche définissant un ensemble de bonnes pratiques et de moyens facilitant la gestion à la fois opérationnelle et analytique des données de référence de manière générique et transversale aux applications. La solution préconisée est de centraliser ces données dans un référentiel maître et indépendant des systèmes applicatifs afin d'en garantir sa pérennité et de mutualiser les efforts de contrôle et d'amélioration de la qualité. Pour cela l'approche MDM regroupe un ensemble de démarches et d'outils préexistants afin de centraliser et de rationaliser la gestion et le partage des données critiques.
L'approche MDM se base principalement sur trois approches fondamentales que sont la "*Data Governance*", la "*Data Quality*" et la "*Data Integration*".

6.2.1.3 Gouvernance des données ou *Data Governance*

La *Data Governance* définit un ensemble de bonnes pratiques et de moyens facilitant la gestion des données. Plus les données sont partagées entre différents intervenants plus cette problématique devient cruciale. L'important est de gouverner ses données et d'éviter l'anarchie. Mais qu'entend-on par *"gouverner ses données"* ? Gouverner c'est essentiellement prévoir, négocier, soulever et résoudre des problèmes :

- **Prévoir**, c'est établir une stratégie de gestion des données déterminant les procédures afin de mettre à jour, partager, assurer la sécurité, contrôler l'accessibilité des données, préserver les droits de leur(s) propriétaire(s), etc.
- **Négocier**, c'est mettre les gens autour de la table afin de trouver des compromis permettant de gérer ces données tout en respectant les exigences des fournisseurs et consommateurs de données ainsi que les réglementations en vigueur.
- **Soulever les problèmes**, c'est être capable de détecter des anomalies relatives aux données et à leur utilisation.
- **Résoudre les problèmes**, c'est être capable de prendre les mesures nécessaires afin de corriger ces anomalies ou d'en diminuer au maximum les effets négatifs.

La gouvernance des données consiste principalement à superviser l'exploitation des données et englobe les éléments qui permettent de gérer de manière optimale les dimensions de qualité telles que l'exactitude, la disponibilité, la sécurité et la conformité réglementaire des données. Les principaux moyens préconisés sont :

- la création d'un comité de pilotage et de surveillance,
- l'identification des "propriétaires", des "fournisseurs" et des "consommateurs" des données,
- la définition des rôles et responsabilités,
- la description des données (glossaire métier, dictionnaires, métadonnées,...),
- la définition des politiques et des processus de gestion des données,
- l'établissement des normes et des procédures pour l'utilisation des données,
- la mise en oeuvre des vérifications et des contrôles des données.

6.2.1.4 L'intégration de données ou *Data Integration*

L'intégration de données définit un ensemble de processus permettant de migrer, combiner et consolider des données provenant de différentes parties du système d'information. L'intégration de données consiste habituellement à extraire des données de différentes sources (bases de données, fichiers, applications, services web, emails, ...), à leur appliquer des transformations (jointures, déduplication, calculs, ...), et à envoyer les données résultantes vers les systèmes cibles. Suivant le niveau d'intégration demandé et le niveau de disponibilité des données, différentes technologies/outils peuvent être utilisés :

- **Extraction Transformation Loading ou ETL** : cet outil permet d'effectuer des synchronisations massives de données d'une base de données vers une autre. Différentes techniques d'extraction, de réplication, de transformation et de conversion de données sont utilisées afin de (re)peupler différentes bases de données.
  Les avantages de cette technique sont la forte intégration des données, la grande disponibilité des indicateurs et agrégats et son adéquation avec des outils d'analyse.

Les désavantages sont les coûts souvent très élevés en matériel, logiciel, maintenance et service ainsi que le délai de rafraîchissement trop long.

- **Enterprise Information Integration ou EII** : cette technologie permet d'interroger plusieurs sources de données afin d'obtenir une vue unifiée et intégrée des données de l'entreprise. Le rôle du serveur EII est de faire la médiation entre les différentes sources de données en décomposant correctement les requêtes qu'il reçoit et en les redirigeant vers les différentes bases de données concernées pour ensuite rassembler les différents résultats avant de les renvoyer.
  Les solutions EII permettent l'accès en quasi temps réel aux données, contrairement aux solutions ETL qui accèdent aux données de manière périodique.
- **Enterprise Application Integration ou EAI** : cette technologie propose une architecture intergicielle (middleware) permettant à des applications hétérogènes de gérer l'échange et la conversion des données en quasi temps réel. Ces applications peuvent être développées indépendamment et peuvent utiliser des technologies différentes.

Une solution MDM combine généralement ces trois techniques. Les techniques EII sont utilisées comme requêteur pour retrouver les différentes données dispersées dans plusieurs bases de données et les intégrer de manière logique. Les techniques ETL sont utilisées comme extracteur pour récupérer les données dispersées et les intégrer de manière physique dans une nouvelle base de données qui constituera la base de données centrale de référence. Les techniques EAI/ESB sont utilisées comme transporteur pour échanger les données entre applications consommatrices et/ou fournisseuses de données.

6.2.1.5 Qualité des données ou *Data Quality*

La *Data Quality* définit un ensemble de bonnes pratiques et de moyens en adéquation avec les usages améliorant la qualité des données stockées dans une base de données ou dispersées dans plusieurs bases de données. Différentes techniques sont utilisées tel que le profilage, la surveillance, la standardisation et le rapprochement des données.
Au-delà de la technologie, le contrôle et l'amélioration de la qualité des données sont des éléments primordiaux à toute approche MDM. Le taux d'anomalies augmente fortement en fonction du nombre de sources de données et du degré d'hétérogénéité de celles-ci. D'un côté, l'efficacité de l'échange des données dépend principalement de la confiance accordée par les consommateurs à ces données. D'un autre côté, la qualité de l'intégration des données dépend largement de la qualité des données de départ. L'intégration de données de mauvaise qualité ne peut générer que des données de mauvaise qualité pour lesquelles il sera plus complexe d'identifier l'origine du problème.

6.2.1.6 L'originalité de l'approche MDM

Produire un annuaire de données de référence ou constituer une base de données rassemblant les données de référence ne suffit pas. Il faut pouvoir assurer la maintenance et garantir la qualité des données de référence sur le long terme. Les efforts investis dans la création de données de référence cohérentes et de qualité ne doivent pas être réduits à néant par un manque de bonnes pratiques favorisant la préservation de cette cohérence et de cette qualité.
Dans la majorité des cas, des changements importants doivent avoir lieu au niveau des processus métier et des outils adéquats doivent être mis en place. Néanmoins, les défis les plus importants

sont bien souvent plus organisationnels que techniques.
L'exploitation des données de référence nécessite de les partager et de les tenir à jour de manière collaborative. Les données de référence passent du statut de données stockées dans de simples fichiers plats difficilement exploitables au statut de données réellement valorisables pour l'organisation. En conséquence, la gestion de ces données doit être rationalisée. Il est primordial de faciliter leur échange, de contrôler et d'améliorer leur qualité de manière continue.
Pour ce faire, l'approche MDM met en avant quatre principes fondamentaux :

1. Les données étant généralement dispersées et donc difficilement valorisables, il est indispensable de pouvoir **créer une vision unique de ces données**. Unique ne veut pas dire qu'elle est imposée à tous, mais que l'ensemble des acteurs a trouvé un consensus sur le contenu, le format et la sémantique des données échangées. C'est une vision commune et neutre qui a l'avantage d'être explicite et de permettre l'intégration des données. Chacun est libre de la contextualiser à sa guise suivant ses besoins métier.
2. Une fois que la vision consolidée est jugée exploitable il faut pouvoir la **partager**. La création de cette vue consolidée est une étape préalable qui peut se révéler très complexe, surtout dans un contexte multi-organisationnel. Elle requiert un travail d'harmonisation qui peut prendre de nombreuses années.
3. Dans le cadre du partage des données, l'approche MDM met en avant les notions de référentiel de données, de **qualité des données** et de **gouvernance des données**.
4. Créer et partager la vue unique ne suffit pas, il faut aussi pouvoir **la gérer et la faire évoluer dans le temps** en fonction des besoins métier. Un référentiel n'est pas qu'une simple banque de données logique ou physique, c'est aussi une application qui permet de gérer les échanges de données entre les différentes applications qui y sont connectées.
   Par exemple, le référentiel doit être capable de garder pour chaque donnée de référence les liens entre les identifiants (primary key) utilisés chez les fournisseurs, chez les consommateurs et dans le référentiel.

6.2.2 Les architectures MDM

Les architectures MDM sont des architectures d'échange qui déterminent où les données de référence vont être stockées et comment elles vont être partagées entre les différents fournisseurs et consommateurs de données. Chaque architecture possède ses avantages et ses inconvénients. Le choix d'une architecture est souvent délicat, car ce choix peut évoluer au cours du temps et dans certaines situations une combinaison de différentes architectures peut se révéler nécessaire.

6.2.2.1 Répertoire virtuel

La première architecture consiste à créer un **annuaire de données** dont le rôle principal est de rediriger les requêtes des consommateurs de données vers les fournisseurs de données adéquats. Son activité principale consiste à gérer un index reprenant les clefs d'accès aux données sources. Cette architecture offre une plateforme d'échange qui s'occupe d'aiguiller les messages vers les fournisseurs de données sans tenir compte de leur contenu (cf. figure 6.1).
L'objectif est de créer un point d'entrée unique auquel se référer pour consulter les données de référence et ainsi permettre leur **intégration logique** alors qu'elles demeurent dans les applications sources d'origine.

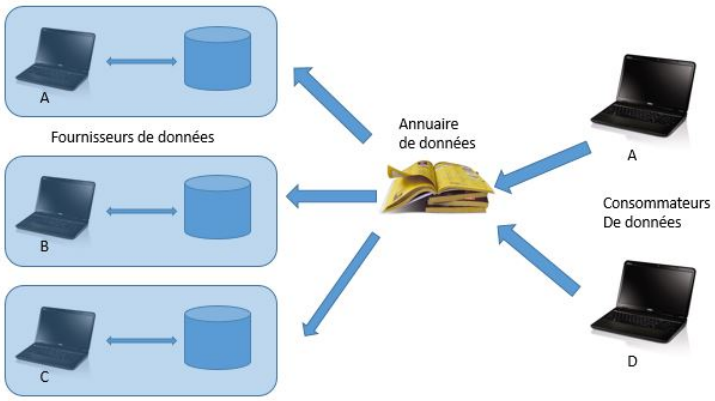

Figure 6.1 : Répertoire virtuel

6.2.2.2 Centralisation

À l'autre extrême, se trouve l'architecture dite de "centralisation". Cette architecture n'intègre plus les données logiquement mais physiquement (cf. figure 6.2). Les données de référence et les attributs nécessaires au bon fonctionnement des applications sont centralisés dans une base de données unique.
Cette base de données centrale devient la seule source de vérité. Une seule et même application peut être utilisée pour l'acquisition, la validation et la consultation des données. Dès lors, fournisseurs et consommateurs de données utilisent la même application centrale pour gérer ou interroger les données de référence.

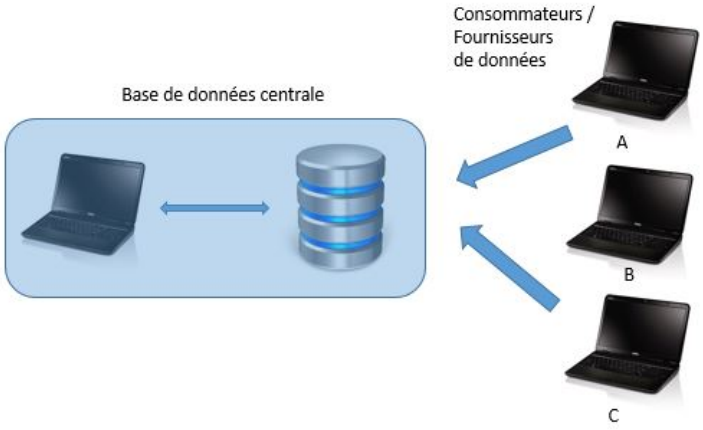

Figure 6.2 : Architecture centralisée

6.2.2.3 Coopération

Au croisement des deux architectures précédentes est née une architecture intermédiaire dite de "coopération" (Figure 6.3). Les objectifs de cette architecture sont :

- de diminuer la charge de travail liée à la mise en place d'une architecture de centralisation,
- de diminuer la complexité liée à la gestion et à l'évolution d'un répertoire virtuel,

- de minimiser les impacts sur les applications fournisseuses,
- de créer une base de données commune et neutre contenant une version intégrée des données de référence.

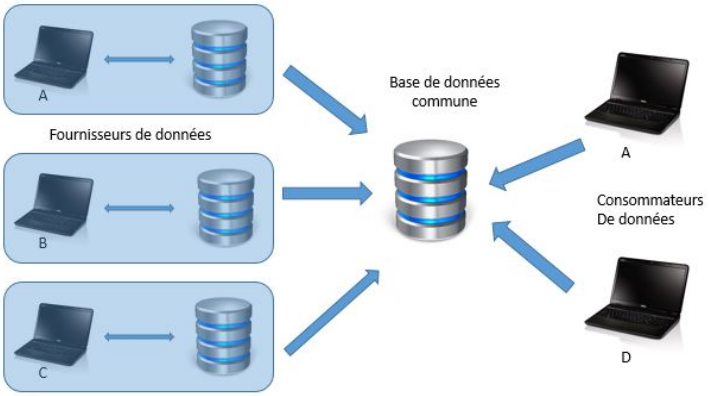

Figure 6.3 : Architecture de coopération

6.2.2.4 Choix d'une architecture MDM

Chacune des trois architectures MDM présentée possède des avantages et des inconvénients. L'architecture de centralisation semble la plus efficace opérationnellement car les applications consommatrices ont toujours accès à une seule source de données de référence consistante et à jour. Cependant, le coût de la mise en place d'une telle architecture est extrêmement élevé et les impacts sur les applications fournisseuses de données sont souvent trop importants voir même radicaux. Les deux autres architectures autorisent la duplication des données ce qui implique, d'une part, une latence entre la mise à jour des données sources et la mise à jour des données de référence et, d'autre part, l'obligation de mettre en place des techniques complexes assurant la consistance et la synchronisation des données.
Le choix d'une architecture de type répertoire virtuel semble le plus indiqué lorsque :

- la gouvernance des données est faible,
- les données de référence évoluent peu,
- le nombre de sources authentiques différentes pour la même donnée est limité et la cohérence entre ces différentes sources est importante,
- le niveau de disponibilité requis pour les données de référence n'est pas primordial,
- la complexité des requêtes de consultation des données de référence est faible,
- un retour sur investissement à court terme est primordial.

Le choix d'une architecture de type centralisation semble le plus indiqué lorsque:

- la gouvernance des données est forte,
- le nombre d'applications impliquées dans le projet est limité,
- le contrôle sur les applications fournisseuses est important,
- la contextualisation des données de référence n'est pas un besoin capital pour les applications consommatrices,

- la valeur ajoutée apportée par un répertoire central opérationnel justifie les coûts et le temps nécessaires à sa mise en place.

Le choix d'une architecture de type coopération semble le plus indiqué lorsque :

- le niveau de gouvernance sur les données doit être flexible,
- le contrôle sur les applications sources n'est pas garanti,
- la création d'une source de données neutre est nécessaire,
- le niveau de disponibilité des données est important,
- la cohérence des données apporte une grande valeur ajoutée,
- l'architecture de type répertoire virtuel doit évoluer de manière incrémentale vers une architecture plus centralisée.

Le choix d'une architecture MDM peut aussi être déterminé par les objectifs métier liés au projet. Nous considérons ici 7 objectifs :

- Améliorer le **contrôle des données**. Le contrôle des données est directement lié au niveau de gouvernance souhaité.
- Améliorer la **disponibilité des données**. La disponibilité dépend de nombreux paramètres. Le premier est le nombre d'intervenants par lesquels les données vont transiter. Pour améliorer la disponibilité, il faut éviter de faire des appels en cascade à différents services d'accès aux données provenant de différentes applications.
- Améliorer la **cohérence/qualité des données**. La cohérence des données n'est pas assurée avec l'utilisation d'un répertoire virtuel. De plus, la qualité des données échangées n'est pas connue. Ainsi des données de mauvaise qualité risquent d'être diffusées vers les applications consommatrices. L'architecture de coopération permet de créer une vue unique cohérente et intégrée des données de référence. Cependant, la synchronisation entre les fournisseurs de données et la vue unique n'est pas toujours garantie. L'architecture de centralisation élimine ce problème de synchronisation et assure l'accès à des données de référence uniques, cohérentes et à jour.
- Améliorer **l'enrichissement des données**. L'enrichissement et l'intégration des données sont fortement dépendants des données de départ. Si les données à enrichir sont incohérentes ou de mauvaise qualité le résultat de l'intégration sera forcément discutable. C'est pourquoi les architectures de type coopération et centralisation devraient être privilégiées si l'objectif principal est de favoriser un enrichissement optimal des données de référence.
- Préserver **l'indépendance** des applications fournisseuses de données. Plus la gestion des données de référence se centralise, moins l'indépendance des applications fournisseuses pourra être préservée.
- Minimiser les changements nécessaires à la **mise en place de la solution MDM**. La mise en place d'un répertoire virtuel est nettement moins onéreuse que la mise en place d'une architecture de centralisation. Dans le premier cas, la difficulté est de créer un annuaire de données performant. Dans le deuxième cas, la difficulté est d'adapter radicalement les applications existantes en revoyant complètement la gestion et l'implémentation de leurs accès aux bases de données.
- Diminuer les **coûts de gestion des données**. Une fois que la solution MDM a été mise en place, l'architecture de type centralisation favorise les économies d'échelle. Les

données ne sont plus dupliquées et aucun mécanisme de synchronisation n'est plus nécessaire. Lorsque les données de référence évoluent, les changements sont apportés une et une seule fois. Dans le cas du répertoire virtuel, chaque application continue à devoir assurer la cohérence et la qualité de ses données de manière locale. On échange des données mais on ne mutualise pas les efforts pour leur gestion. Dans le cas de la coopération, on mutualise les efforts pour gérer les données, mais on les duplique dans une base de données commune, ce qui augmente inévitablement les coûts de gestion.

Ces 7 objectifs sont repris dans le tableau 6.1 .

	Répertoire Virtuel	Coopération	Centralisation
Contrôle	–	+	++
Disponibilité	–	+	++
Cohérence	–	+	+++
Enrichissement	–	++	++
Indépendance	++	–	– – –
Mise en place	++	–	– – –
Coûts de gestion	– –	–	+++

Table 6.1 : Choisir son architecture MDM

6.3 La mise en place d'une solution MDM

Le taux d'échec des projets MDM reste important. La mise en place et l'utilisation de bonnes pratiques est essentielle à la conduite de tels projets.
Un projet MDM passe par différentes étapes similaires à celles d'un projet de développement de système d'information. Cependant les projets MDM ont des particularités du fait de leurs caractères transversal et multi-organisationnel.
Les grandes phases du développement d'un tel projet sont bien évidemment l'analyse, la conception et l'implémentation (6.4).

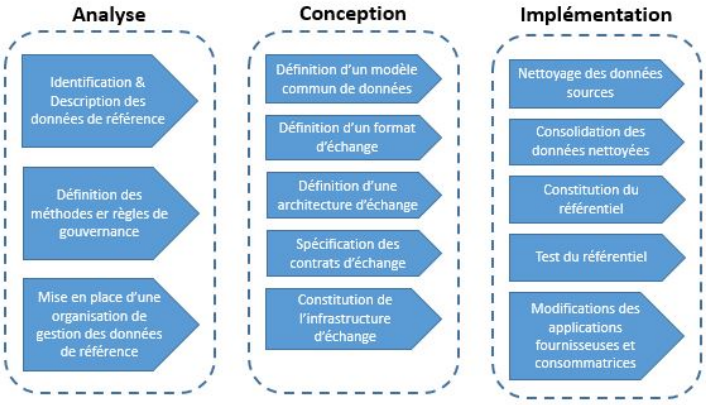

Figure 6.4 : Les étapes de la mise en place d'une solution MDM

6.3.1 Phase d'analyse

Aborder l'approche MDM comme un problème avant tout technologique est une erreur qui peut être lourde de conséquence. L'approche MDM consiste essentiellement à comprendre des processus et des systèmes métier complexes et à mettre en place des processus afin de gérer leurs données.
Au départ, il faut identifier :

- quelles données vont être partagées,
- pourquoi elles doivent être partagées,
- qui va en être la source authentique,
- où vont-elle être stockées,
- qui en est le propriétaire,
- qui peut y accéder,
- et quels sont les obstacles (techniques, conceptuels ou politiques) qui pourraient empêcher le partage de ces données.

Ensuite, il faut s'interroger sur les règles à appliquer afin d'assurer la gestion et la maintenance de ses données de référence. Enfin, il faut déterminer les parts de responsabilité dans la gestion journalière des données de référence et favoriser la coordination des différents intervenants en créant une organisation spécifiquement dédiée à cette tâche.

6.3.1.1 Identifier et décrire les données de référence

Le périmètre des données de référence doit être délimité avec précaution. Cette première étape d'analyse est généralement révélatrice de la complexité d'un projet MDM. L'objectif est de définir un périmètre optimal pour les données de référence et d'obtenir un aperçu détaillé de leur statut actuel.
L'identification des données de référence dépend principalement du domaine d'application, des besoins métier et du type de données échangées.
Généralement, une donnée de référence possède les caractéristiques suivantes :

- Elle a une **valeur métier** importante. Il faut distinguer les données critiques pour le métier, des données moins sensibles tant du point de vue de leur qualité que de leur disponibilité.
- Une donnée de référence est réutilisée, **partagée** entre différentes applications et/ou échangée avec des tiers.
- La **durée de vie** d'une donnée de référence est considérée comme longue. En opposition aux données dites transactionnelles, une donnée de référence a généralement un long cycle de vie. Cette durée dépend principalement du métier considéré. De plus, un même type de données peut avoir des cycles de vie et des rythmes de mise à jour différents suivant les contextes d'utilisation de cette donnée.
- Le **volume** des données de référence est suffisamment élevé. L'utilité de la création d'un référentiel dépend de la quantité de données qui vont y être stockées. Outre le volume des données, le volume des transactions est également à prendre en considération.

Une fois ces données identifiées, il faut les expliciter et les faire valider par les acteurs impliqués.

6.3.1.2 Déterminer les méthodes et les règles de gouvernance

Sans une gouvernance efficace et appropriée, les initiatives MDM rencontreront des difficultés principalement liées à des conflits politiques et à la résistance des fournisseurs de données. Cette résistance se justifie par le manque d'une explicitation claire des règles régissant l'utilisation de leurs données et sur la responsabilité qu'ils devront endosser. L'objectif est de définir une stratégie de gouvernance dès les premières phases du projet de manière à augmenter ses chances de réussite. Définir une stratégie de gouvernance consiste à :

- Définir les règles de qualité concernant la consistance, la précision et la complétude des données.
- Définir des indicateurs de mesure de la qualité des données.
- Définir les moyens mis en oeuvre pour détecter les anomalies.
- Définir les règles d'arbitrage pour traiter ces anomalies.
- Définir les règles d'intégration et d'élimination des doublons.
- Définir les règles de sécurité en termes de protection et d'accessibilité des données.
- Définir un processus de maintenance et de gestion du changement afin de préserver la qualité des données et d'assurer leur disponibilité.

6.3.1.3 Mettre en place une organisation pour gérer les données de référence

Au terme de la phase d'analyse, il est nécessaire de mettre en place une organisation afin de mener la suite du projet et d'assurer la gestion journalière des données de référence ainsi que le respect et l'évolution des règles de gouvernance. L'objectif est de constituer une équipe de personnes en charge de la gestion du référentiel aussi bien pour sa constitution que sa maintenance.
Cette organisation est dotée de différentes prérogatives :

- la gestion et la centralisation des données de référence,
- l'élaboration et l'évolution des règles de gouvernance,
- la supervision des données de référence (utilisation, accès, préservation de la cohérence et du caractère privé, ...),
- l'arbitrage des problèmes relatifs au référentiel et aux données gérées par celui-ci,
- faciliter l'utilisation du référentiel et le valoriser.

6.3.2 Phase de conception

Une fois la phase d'analyse terminée, différentes décisions doivent être prises afin de concevoir le référentiel. Dans un premier temps, il est nécessaire de définir un modèle commun pour les données de référence. Ce modèle permettra à tous les intervenants de comprendre la structure et la sémantique des données qui vont être échangées. Dans un deuxième temps, un format standard d'échange doit être défini afin de servir de langage commun entre les différents interlocuteurs. Ensuite, il faut déterminer l'architecture d'échange et spécifier les contrats d'échanges établis entre les différents intervenants. Enfin, une infrastructure d'échange doit être proposée afin de pouvoir satisfaire au mieux les règles de gouvernance et vérifier leur mise en oeuvre correcte.

6.3.2.1 Définir un modèle commun des données de référence

L'objectif de cette étape est de définir une modèle logique standard pour les données de référence explicitant leurs attributs, le type des attributs, les valeurs par défaut, les valeurs autorisées, les contraintes, les relations entre les données, la signification des données, ...
L'interprétation des données doit être claire pour tous les utilisateurs. Des techniques telles que la modélisation des données de référence, la constitution d'une ontologie et/ou la constitution d'un glossaire métier sont fortement recommandées afin de minimiser les ambiguïtés et de s'assurer de la convergence des points de vue sur les données de référence.

6.3.2.2 Standardiser le format d'échange des données de référence

L'étape préalable à l'échange effectif de données est la définition d'un format standard (généralement basé sur les technologies XML/XSD) qui sert de langage commun pour toutes les applications fournissant ou consommant ces données. L'objectif est de définir un langage commun pour toutes les applications leur permettant de pouvoir comprendre, publier et retraiter les données de référence. Ce langage commun est généralement défini à partir du modèle commun de données et détermine la structure des données et leur format. Deux alternatives sont envisageables pour définir ce type de langage :

- le standard est un **langage pivot** permettant la traduction d'un format de données vers un autre. L'avantage de cette solution est que ce standard reste transparent pour les applications. La solution MDM traduit les données dans le format spécifique à l'application consommatrice.
- le standard est un **langage commun** et la solution MDM fournit simplement les données de référence sous ce format qui devra pouvoir être compris et donc exploité par toutes les applications.

6.3.2.3 Définir une architecture pour échanger les données de référence

Des considérations à la fois techniques et business entrent en ligne de compte dans la sélection d'une architecture. Le critère le plus discriminant est souvent le niveau de contrôle sur les données. L'architecture de centralisation permet un contrôle fin et poussé des données de référence, de leur qualité et de leur intégration. À l'opposé, l'architecture de type répertoire virtuel limite le contrôle aux échanges de données et à la redirection des demandes d'accès vers les données. Entre les deux se situe l'architecture de coopération.

6.3.2.4 Spécifier les contrats d'échange

L'élaboration des contrats d'échange est généralement une étape obligatoire pour spécifier la qualité des échanges effectués entre les différents partenaires. Ces contrats d'échange déterminent le niveau de service sur lequel s'accordent les parties concernées afin de partager et administrer les données de référence.
Durant cette étape, on explicite de manière contractuelle un *Service Level Agreement* ou SLA qui définit essentiellement :

- le propriétaire du service,
- les modalités de l'échange : output, input, type de données, fréquences, volumétrie, dates de mise à disposition, ...

- la qualité du service fourni : disponibilité, temps de réponse, engagement sur la qualité des données, ...
- les procédures relatives au traitement des non-conformités : délais de dépannage, instances d'arbitrage, mesures de reprise, ...

6.3.2.5 Constituer l'infrastructure d'échange des données de référence

Une fois que la phase d'analyse permet d'avoir une vue globale de la situation et qu'une architecture d'échange a été choisie, il est nécessaire de mettre en place une infrastructure adaptée et efficace. L'objectif est de construire une infrastructure qui supporte l'architecture choisie et satisfait aux exigences notamment en terme de gouvernance, de scalabilité et de reliability.

6.3.3 Phase d'implémentation

Une fois que l'infrastructure et les outils la supportant sont opérationnels, il est temps de les utiliser pour produire les données de référence et les partager effectivement entre les différents fournisseurs et consommateurs de données. Ce processus d'implémentation est généralement itératif et a pour objectif final de fournir en sortie un référentiel de données respectant l'architecture choisie et satisfaisant les SLA spécifiés lors de la phase de conception.

6.3.3.1 Nettoyer et transformer les données sources

L'objectif de cette étape est d'améliorer la qualité des données sources et de se conformer au format standard d'échange défini durant la phase de conception.
Le nettoyage et les transformations appliquées aux données sources sont similaires aux opérations d'*Extract Transform and Load* utilisées afin d'alimenter un entrepôt de données (normalisation des formats de date, insertion de valeur par défaut, correction des codes postaux, ...).
Le nettoyage comprend également la détection de doublons pouvant apparaître au sein même d'une base de données. La plupart des outils et technologies sont capables d'identifier un grand nombre d'erreurs potentielles. Néanmoins, en ce qui concerne leur correction, seules les plus triviales peuvent être prises en charge de manière automatisée. Certaines erreurs nécessiteront toujours des investigations humaines déterminant les actions à prendre, en accord avec le métier.
L'utilisation d'outils adaptés à l'analyse de grandes bases de données est fortement recommandée, notamment, pour les fonctionnalités de profilage, standardisation, rapprochement et nettoyage.

6.3.3.2 Consolider les données sources

L'objectif de la consolidation est d'identifier les doublons existant entre différents fournisseurs de données ou au sein du même fournisseur de données pour ensuite les fusionner. La consolidation des doublons est une étape délicate qui doit être menée par le métier et supportée par des outils facilitant l'identification et la résolution des doublons. Des outils sont nécessaires car la phase d'identification peut nécessiter l'analyse et la comparaison de milliers d'enregistrements.
Néanmoins, même avec des outils, la tâche demeure complexe car si on fusionne trop de données on perd de l'information et, d'un autre côté, si on oublie certains doublons on

risque des désynchronisations et l'apparition d'incohérences entre les données utilisées par les différentes applications. Dans tous les cas, les bonnes pratiques préconisent de toujours conserver l'historique des versions afin d'éviter que des opérations de fusionnement ou de défusionnement ne deviennent irréversibles.
De plus, l'élimination des doublons ne suffit pas, il est nécessaire d'étudier comment les relations existant entre les données vont évoluées lors de la consolidation.

6.3.3.3 Constituer le référentiel

Suivant l'architecture et l'infrastructure d'échange choisies, différentes politiques sont envisageables pour constituer un référentiel. Une fois que le nettoyage, la transformation et potentiellement la consolidation des données ont été effectués, il faut déterminer comment les données de référence vont être stockées et rendues disponibles. Elles peuvent être stockées à leur emplacement initial, ou être dupliquées et/ou migrées vers une base de données spécifiquement dédiée au stockage des données de référence. La migration des données de référence n'est pas improbable et, dans ce cas de figure, il est primordial de minimiser les impacts sur les applications sources. Enfin, il faut offrir différents services (*Data Services*) permettant de consulter les données du référentiel, de les filtrer, de les créer, de les supprimer, de les mettre à jour, de vérifier certaines anomalies (automatiquement ou manuellement), ...

6.3.3.4 Tester et évaluer le référentiel

Une fois le référentiel constitué et avant de le mettre en production, il est nécessaire de tester la disponibilité des données, les mécanismes de synchronisation, la charge sur les bus de données, les temps de réponse, la fiabilité. Les utilisateurs de données doivent vérifier que les données qu'ils réceptionnent sont correctement interprétables et contextualisables suivant les besoins qu'ils ont spécifiés.
Dans cette étape on ne vérifie pas que l'infrastructure, on vérifie également :

- la qualité des données et la manière dont elles ont été consolidées,
- le résultat du nettoyage et de la consolidation et leur conformité par rapport aux attentes des utilisateurs du référentiel,
- l'utilisabilité du référentiel par rapport aux scénarii d'utilisation et services spécifiés au préalable,
- la disponibilité des données auprès des fournisseurs de données.

6.3.3.5 Modifier les applications fournisseuses et consommatrices

L'objectif est de faire évoluer les applications fournisseuses et consommatrices en fonction des choix architecturaux et d'infrastructure pour tenir compte du référentiel et satisfaire les contrats d'échanges. Les principales améliorations à apporter sont :

- la capacité à envoyer et recevoir des messages conformes au standard d'échange communément accepté,
- la capacité des applications à mettre en place des mécanismes de synchronisation avec les données de référence,
- la capacité des applications fournisseuses de données à améliorer la qualité de leurs données en interne avant de les exporter vers les données de référence,

– la capacité des applications à d'abord consulter les données de référence avant de créer un nouvel enregistrement.

6.4 En résumé

L'approche MDM n'est pas simplement une approche de gestion de la qualité des données telle que nous l'avons présenté au chapitre 5. Elle se focalise sur la gestion des données de référence.

Part IV

Annexes

Bibliography

Les références données dans cette page correspondent à des supports qui ont été utilisés lors de la conception de ce cours.
Il n'y a aucun besoin de consulter ces documents pour atteindre les objectifs du module de base de données avancées. Par contre ces supports peuvent être intéressants pour les étudiants qui souhaitent aller plus loin.

Qualité des données

- Laure Berti-Equille *Qualité de données multi-sources et recommandation multi-critères.* 1999- Actes du XVIIème Congrès INFORSID, La Garde, France, 1-4 juin, 1999

- Laure Berti-Equille *La qualité des données comme condition à la qualité des connaissances : un état de l'art* - 2004 MQPFD, pp.95-118

- Laure Berti-Equille *Qualité des données* - 2004 Ingénierie des Systèmes d'Information. 9. 117-143. doi:10.3166/isi.9.5-6.117-143

- Christophe Brasseur *Data Management : qualité des données et compétitivité* 2005 - Hermès Sciences - Collection Management et Informatique - ISBN2-7462-1210-2

- M.L. Brodie *Data quality in information systems* 1980 - Information and management vol. 3, pp. 245-258

- Artur D. Chapman *Les principes de qualité des données* 2005 - version 1.0. Trad. Chenin, N. Copenhague: Global Biodiversity Information Facility https://www.gbif.org/document/80626/les-principes-de-qualite-des-donnees

- G.P.A. Delen and B.B. Rijsenbrij *The specification, engineering, and measurement of information systems quality* 1992 - J. Systems Soffware, vol. 17, no. 3, pp. 205-217

- Institut canadien dinformation sur la santé *Le cadre de la qualité des données de l'ICIS* 2009 - Ottawa

- Informatica France *Des données de qualité - Exploitez le capital de votre organisation* 2008 - Livre Blanc JEMM research

- Yang W. Lee, Diane M. Strong, Beverly K. Kahn, Richard Y. Wang *AIMQ: a methodology for information quality assessment* 2002 - Information Management, Vol 40, Issue 2, doi.org/10.1016/S0378-7206(02)00043-5

- Pipino, Leo L. and Lee, Yang W. and Wang, Richard Y. *Data Quality Assessment*

2002 - Commun. ACM, doi.acm.org/10.1145/505248.50601

- Thomas C. Redman, Michael Daugherty, Mike Daugherty *Data Quality: The field Guide* 2001 - Sagebrush Education Resources - ISBN 978-0-61-391717-9

- Talend Open studio *Les 10 causes principales des problèmes de qualité de données* 2013 - Livre blanc Talend Open Studio.

- Richard Y. Wang, Veda C. Storey, Christopher P. Firth *A Framework for Analysis of Data Quality Research* 1995 - IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 7, NO, 4, AUGUST 1995

- Y. Wand and R.Y. Wang *Anchoring data quality dimensions in ontological foundations* 1996 - Commun. ACM 39, 11 pages 86-95. DOI=http://dx.doi.org/10.1145/240455.240479

- Sabrina Zaïdi-Chtourou *La qualité de l'information dans les systèmes d'information marketing* 2009 - Thèse de l'université Jean Moulin Lyon 3

MDM

- Franck Régnier-Pécastaing, Michel Gabassi, Jacques Finet *MDM : enjeux et méthodes de la gestion des données* 2008 - DUNOD - ISBN 978-2-10-053555-2

- Jean-Christophe Trigaux *Master Data Management - Mise en place d'un référentiel de données* 2009 - http://documentation.smals.be

Autres

- Gabriel Siméon *Données le vertige* 2012 - Article Libération (http://www.liberation.fr/futurs/2012/12/03/donnees-le-vertige_864585)

- Smile *E-commerce Open Source* 2009 - Livre blanc Smile

- Smile *Décisionnel : Le meilleur des solutions open source* 2012 - Livre blanc Smile

- J.M. Juran *Managerial Breakthrough* 1964 - New York: McGraw-Hill

- Rolande Marciniak, Frantz Rowe *Systèmes dinformation, dynamique et organisation* 2009 - 3e éd., Economica, Paris.

Quelques sites

- Bilan *Les ventes sur Internet ont grimpé de 10% en 2017* 20 Février 2018 - Bilan La référence Suisse en économie - http://www.bilan.ch/economie/ventes-internet-ont-grimpe-de-10-2017

- FEVAD *Bilan 2017 du e-commerce en France : les ventes sur internet en hausse de 14% sur un an* 6 février 2018 - https://www.fevad.com/bilan-2017-e-commerce-france-ventes-internet-hausse-de-14-an/

- Forrester *Site Forrester : définition du e-commerce*

https://www.forrester.com/eCommerce

- *Les meilleures solutions de Business Intelligence selon Forrester* 2015 - Journal du Net - https://www.journaldunet.com/solutions/saas-logiciel/meilleures-outils-de-bi-business-intelligence.shtml

- Wikipedia *Mars Climate Orbiter* https://fr.wikipedia.org/wiki/Mars_Climate_Orbiter

- *Problèmes de câblage pour l'A380 ?* 2009 - https://www.ladepeche.fr/article/2009/03/15/573829-problemes-de-cablage-pour-l-a380.html

- Mary Shacklett *La mauvaise qualité des données est un problème coûteux : 3 conseils pour progresser* 2015 - http://www.zdnet.fr/actualites/la-mauvaise-qualite-des-donnees-est-un-probleme-couteux-3-conseils-pour-progresser-39830220.htm

- Jean-Michel Franco *La qualité des données, produites par les collaborateurs des entreprises, est essentielle pour maximiser les effets positifs de l'IA* 2017 - https://www.journaldunet.com/solutions/expert/68190/a-l-heure-de-l-automatisation-le-role-de-l-humain–gardien-du-temple–est-renforce.shtml