La qualité des SI

Master SID: Qualité des SID

Samira SI-SAID CHERFI (sisaid@cnam.fr)
Maître de conférences
Conservatoire National des Arts et Métiers

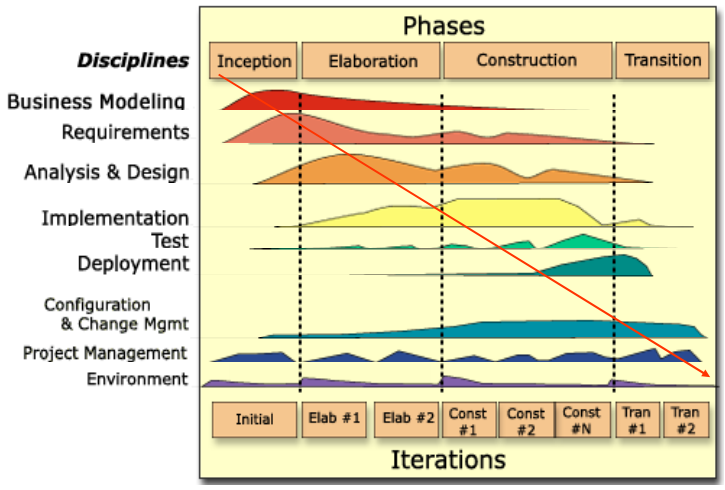
Objectifs du cours

- Présenter un tour d'horizon sur la qualité des données
- Présenter un tour d'horizon sur la qualité des modèles
- Identifier les enjeux et les opportunités
- Définir les directions de recherche.

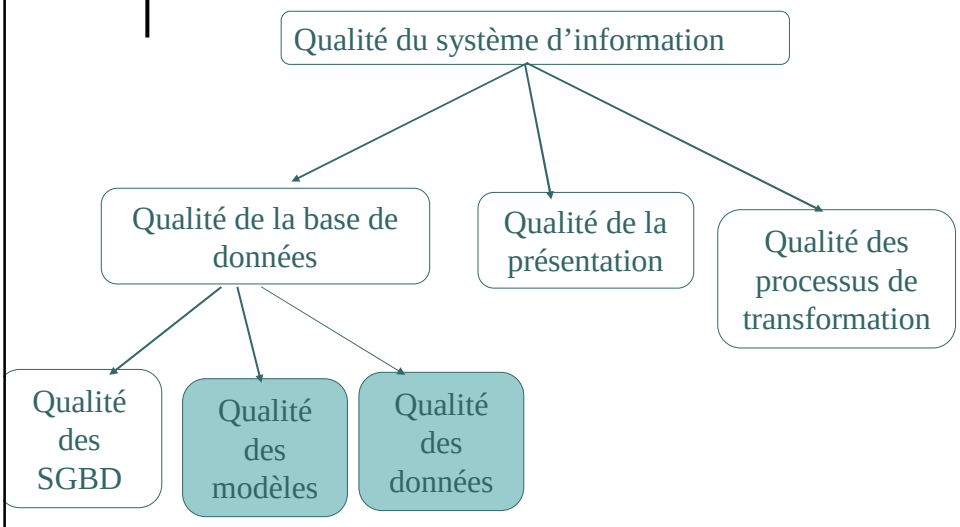
2

Notre vision

Gestion de la qualité

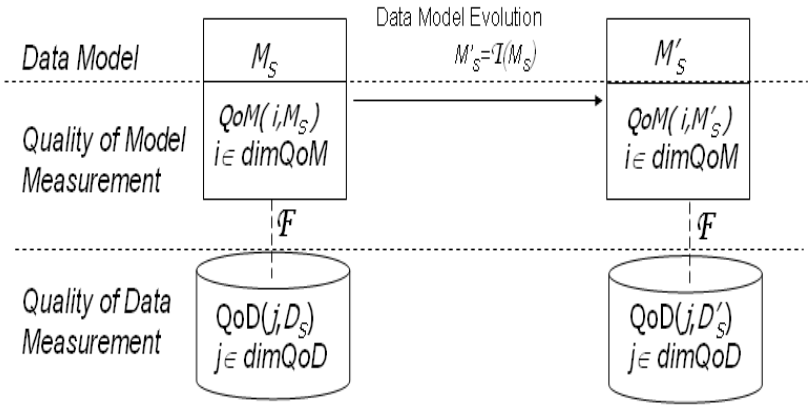


Introduction



4

Hypothèses



5

La qualité des données

La qualité des données

- Qualité des données
  - Motivation
  - Définition
  - Une nouvelles vision du problème de la qualité des données
  - Métriques pour la qualité des données
- Le processus de qualité des données
  - Un cadre pour les problèmes et les solutions
- Les approches pour la qualité des données
  - Approches par les méta données
  - Approches par la gestion du processus
  - Approches correctives par les techniques de bases de données
  - Les approches statistiques

Motivation

- La qualité des données dans les bases, les entrepôts de données ou plus généralement dans les systèmes d'informations est un enjeu majeur. 50 à 80 % du temps dans les projets de fouille de données.
- La donnée est indispensable à la prise de décision
- De plus en plus d'applications utilisent les données en ligne, or ces données souffrent aujourd'hui d'un manque de fiabilité : erreurs, données isolées, doublons, incohérences, valeurs manquantes, incomplètes, incertaines, obsolètes, ou peu fiables.
- Les problèmes liés à la qualité des données sont coûteux et universels

Définition

- Une données n'est pas de qualité lorsqu'elle ne correspond plus aux attentes:
  - Elle ne répond pas aux spécifications
    - Obsolète ou incohérente
    - Ne respecte pas les types ou les schémas
  - Incompréhensible: complexe, absence de méta données
- Plusieurs raisons possibles
  - Erreurs humaines, traitement incomplet de la donnée etc.

9

Exemple 1 (source these V. Peralta)

stid	name	address	telephone	interview	test
21	María Roca	Carrasco	6001104	low	1.0
22	Juan Pérez	Colonia 1280/403	9023365	medium	.5
43	Emilio Gutiérrez	Irigoitía 384	3364244	high	.8
56	Gabriel García	Propios 2145/101		low	.5
57	Laura Torres	Maldonado & Yaro	099628734	medium	.7
58	Raúl González	Rbla Rca Chile 1280/1102	4112533	high	9
101	Carlos Schnider	Copacabana 1210	094432528	high	.9701
102	Miriam Revoir		9001029	medium	.7945
103	A. Benedetti	Charrúa 1284/1	7091232	low	.9146
104	Luis López	Sixtina s/n		high	.8220

Extrait de la base de données

stid	name	address	telephone	interview	Test
43	Emilio Gutiérrez	Potosí 934	6019945	high	.8000
56	Gabriel García	Batlle y Ordoñez 2145/101	5143029	low	.5130
57	Laura Torres	Maldonado 864	099628734	medium	6965
58	Horacio Acher	Soca 2315	7079428	medium	.7600
59	Renzo Quinteros	Juan Paullier 635/001	4037690	low	.5505
101	Carlos Schnider	Copacabana 1210	094432528	high	.9701
102	Miriam Revoir	Canelones 1524	9001029	medium	.7945
103	Ana Benedetti	Charrúa 1284/1	7091232	low	.5146
104	Luis López	Sixtina s/n		high	.8218

Information du monde réel

Exemple 1

- Analyse de l'exemple
  - Du point de vue sémantique
    - Les étudiants 21 et 22 n'existe pas dans le réel
    - L'étudiant 58 fait référence au mauvais étudiant
    - Certains autres étudiants ont une valeur fausse sur un attribut ou des valeurs absentes.
  - Du point de vue syntaxique
    - Certains noms et adresses des étudiants ne respectent pas la description standard (nom du 103, adresse du 57, erreurs dans adresse de 22 (colinia au lieu de coloniaa, valeur de la note de 58 hors intervalle)
  - Du point de vue de la précision
    - Certaines valeurs de notes sont arrondies. Pour 21, carasso est le nom d'un quartier
- Données erronées, imprécises ou absentes
  - types, formats, absence de valeur ou valeur par défaut
- Meta donnée ou expertise du domaine
    - Comment interpréter l'absence de donnée ?

11

Exemple 2 (source: centre d'études statistiques – Canada)

- Voici un exemple de questionnaire ayant servi à une enquête:

Question	**Response**
How did you get to work on Census Day 2001? *If you used more than one method of transportation, mark **all** relevant circles.*	( ) Car, truck or van—as a driver ( ) Car, truck or van—as a passenger ( ) Public transit (*e.g., bus, street car, subway, light rail transit, commuter train, ferry*) ( ) Walked to work ( ) Bicycle (**x**) Motorcycle ( ) Taxicab ( ) Other method (**x**) Worked from home (**x**) Did not go to work

12

Erreurs communes

- Changements automatiques
  - Changer le type des données
    - Entier en chaine, modifier la position d'un champs.
  - Changement de format ou d'échelle
    - Dollars vs. euros
  - Remplacement temporaire par les valeurs par défaut
    - Echec d'une tâche
  - Absence de valeur ou valeur par défaut
    - 0 représente à la fois l'absence de valeur ou la valeur par défaut
  - Trous dans les séries temporelles
    - Surtout si les enregistrements correspondent à une évolution dans le temps.

13

Quelques définitions de la qualité des données

- Précision et exactitude (accuracy)
  - Données correctement enregistrées, valides et précises
- Complétude (completeness)
  - Toutes les données pertinentes sont stockées
- Fraicheur (Freshness)
  - Les données sont mises à jour et suffisamment récentes pour l'usage auquel elles sont destinées.
- Cohérence (consistency)
  - Données répondant aux contraintes

14

Limite de ces définitions

- Difficile à mesurer
  - L'exactitude et la complétude sont difficiles à mesurer.
- Indépendante du contexte
  - Lorsque ce qui nous intéresse est le calcul de données agrégées ou des tendances, la précision de l'information est moins importante
- Incomplète
  - Facilité d'interprétation, l'existence de méta-données, l'analyse etc.
- Vague
  - Les définition conventionnelles ne fournissent pas d'aide quant à la manière d'améliorer la qualité des données.

15

Objectifs

- Proposer une mesure objective de la qualité des données
- Offrir le moyen de mesurer le degré de confiance à accorder aux données
- Evaluer la validité et l'intérêt des connaissances induites par les données
- Aider à mettre en place des moyens pour améliorer cette qualité

16

Causes de la non qualité des données

- Erreurs lors de la conception des données (non respect des règles de normalisation, non mise en place de procédures de contrôle et de maintien de l'intégrité)
- Mauvaise interprétation des données
- Erreurs lors de la collecte des besoins ou lors du passage de l'analyse à la conception
- Erreurs lors de la saisie des données
- Erreurs liées à la sécurité des données

17

Causes de la non qualité des données

- Inadéquation ou absence de procédures de mise à jour des données
- Conséquence d'un processus d'intégration (hétérogénéité, niveaux de qualité inégales, techniques d'intégration inadaptées etc.)
- Perte de données ou introduction d'erreurs suite à une migration
- Incohérences liées à des réplications mal gérées
- Erreurs ou inadéquation des processus de traitement de l'information

18

Vers une nouvelle vision

- Nous avons besoin de définitions qui
  - Reflètent l'usage des données
  - Aident à améliorer la qualité des données
  - Soient mesurables (via la définition de métriques)
- Il est d'abord indispensable de cerner comment et quand est ce que des problèmes de qualité interviennent

19

Transformation des données; un processus continu

La continuité des données

- La données et l'information ne sont pas statiques. Elle résultent d'un processus à travers lequel elles évoluent:
  - Création
  - Collecte et livraison
  - Stockage
  - Intégration
  - Extraction
  - Recherche et analyse

21

Création des données

- Comment est ce que les données entrent dans le système.
- Sources de problèmes
  - Saisie manuelle
  - Absence de standards pour les contenus et les formats
  - Entrée de doublons
  - Approximations, contraintes matérielles ou logicielles
  - Erreurs de mesure

22

Solutions

- Solutions potentielles:
  - Préventives:
    - Architecture du processus (intégrer les contrôles de qualité)
    - Gestion du processus (valoriser l'entrée de données correctes)
  - Rétrospective (correctives):
    - Nettoyage de données (éliminer les doublons, résolution du problème de rapprochement des noms, standardisation des valeurs des champs)
    - Diagnostiques (détection automatique des erreurs).

23

Collecte et livraison

- Destruction ou mutilation des données par des prétraitements inappropriés
  - Agrégation inappropriées
  - Conversions inappropriés (interprétation des valeurs nulles)
- Perte de données:
  - Dépassement de capacité
  - Problème de transmission
  - Absence de vérifications

24

Solutions

- Etablir des protocoles de transmission fiables
  - Utiliser des serveurs relais
- Vérification
  - calculs, analyseurs
  - Utilisation de patterns de vérification
- Interdépendances
  - Rechercher les relations entre les flots de données et les étapes de traitements
- Interface d'accord
  - Validation des données par le demandeur.

25

Stockage des données

- Problèmes du stockage physique
  - Pouvait être un problème par le passé (prix des supports)
- Problèmes du stockage logique(ER → relations)
  - Pauvreté des méta données.
    - Les données proviennent souvent des programmes ou de vieilles bases.
    - Modèles et structures de données inappropriés.
    - Absence de datation des données, problèmes de normalisation etc.
  - modifications Ad-hoc.
    - Restructurer la données pour répondre aux besoins de l'interface
  - Contraintes logicielles et matérielles.
    - Problème de l'an 2000

26

Solutions

- Meta données
  - Documenter et publier les spécifications des données.
- Planification
  - Prévoir le pire.
  - Souvent très difficile.
- Exploration des données
  - Utiliser la fouille et l'analyse des données pour examiner les données
    - Est-ce qu'elles répondent aux spécifications pré-établies?
    - Y a-il eu des changements?

27

Intégration des données

- Données multi sources
- Problèmes communs
  - Données hétérogènes : clés différentes, formats différents
    - Correspondance approximative
  - Différents sens
    - Qu'est ce qu'un client,: un compte, un individu, une famille, …
  - Synchronisation temporelle
    - Est ce que la données est liée à une période donnée? Y a-il une compatibilité entre les périodes?
  - Données ancienne
    - IMS, tableurs, structures ad-hoc
  - Facteurs sociologiques
    - Réticence à partager les données – perte de pouvoir.

28

Solutions

- Outils du commerce
  - Il existe des résultats significatifs dans le domaine de l'intégration des données
  - Il existe des outils pour la migration, le profilage et le nettoyage de données.
- Exploration et analyse des données
  - Nécessité d'extraire les méta données (sens des données).
  - Visualiser avant et après l'intégration : est ce que l'intégration suit le déroulement attendu?

29

Extraction des données

- Les données exportées sont souvent l'image des données actuelles. Les problèmes surviennent parce que:
  - La source de données est mal comprise.
  - Le besoin d'extraction est mal compris.
  - Des erreurs
    - Interprétation des valeurs nulles (pour inapplicabilité ou absence temporaire)
  - Contraintes de calcul
    - Coût élevé des calculs conduit à se contenter d'un résultat partiel

30

Solutions

- Outils – utiliser des outils d'extraction adéquats (ETL, outils XML)
- Tester – établir des requêtes de test dont le but est de vérifier que ce que l'on obtient correspond à ce qui est attendu

31

Recherche et analyse des données

- Problèmes durant l'analyse
  - Erreur humaine
  - Passage à l'échelle et performance
  - Seuils de confiance
  - L'utilisation des techniques de boite noires
  - L'utilisation de modèles statistiques
  - Expertise de domaine insuffisante
  - Manque de familiarité avec les données

32

Solutions

- Exploration des données
  - Déterminer les techniques et modèles appropriés, trouver les erreurs, développer l'expertise du domaine.
- Analyse continue
  - Est ce que les résultats sont stables? Comment est ce qu'ils changent?
- Systématiser l'analyse
  - Faire de l'analyse une partie intégrante de la boucle de rétro-action.

33

La qualité des données: pourquoi ?

- Il existe de nombreux types de données ayant différents usages et des problèmes de qualité spécifiques
  - Données fédérées
  - Les données multidimensionnelles
  - Données descriptives
  - Données géographiques
  - Flots de données
  - Données du web
  - Données textuelles, numériques etc.

34

La qualité des données: pourquoi ?

- Il existe de nombreux usages des données
  - Calculs
  - Analyse des agrégats
- Interprétation des données: La donnée n'est utile que si l'on connaît le sens qu'elle véhicule.
- Pertinence des données : Est ce que les données dont on dispose suffisent pour répondre aux besoins
  - Utilisation de données mandatée (concept de confiance en la source de données)

35

Métriques pour la qualité des données

Contraintes et qualité des données

- Nombreux problèmes de qualité des données peuvent être résolus par des contraintes sur les schémas
  - Non autorisation des valeurs nulles, valeurs par défaut, contraintes de domaines, clés étrangères
- Beaucoup d'autres problèmes sont dûs aux flots de données (workflow) et peuvent donc être résolus par des contraintes dynamiques
- Les contraintes suivent la règle des 80-20
  - Un petit nombre de contraintes peut résoudre la majeure partie des cas, mais il faut des milliers de contraintes pour résoudre les quelques cas restants.
- Les contraintes sont mesurables: des métriques pour mesurer la qualité?

37

Métriques pour la qualité des données

- On souhaite des quantités mesurables
  - Qui montrent ce qui est erroné et comment y remédier
- Types de métriques
  - Statiques vs. dynamiques
  - Opérationnelles vs. diagnostiques
- Les métriques doivent être initialement correctes avec une possibilité de les améliorer .
- Un large éventail de métriques est possible
  - Savoir choisir le plus pertinent

38

Exemple de métriques

- Utilité et fiabilité des données
  - Conformité avec le schéma (statique)
    - Évaluer les contraintes sur un échantillon.
  - Conformité avec les règles métier (dynamique)
    - Évaluer les contraintes en opérant des modifications sur la base de données.
  - Précision
    - Effectuer un inventaire (coûteux), ou utiliser des données mandatées (tracer les insatisfactions). Audit ?
  - Accessibilité
  - Facilité d'interprétation
  - Nombre d'erreurs dans l'analyse
  - Nombre de succès des processus de bout en bout, etc.

39

Quelques approches de la qualité des données

Approches préemptives centrées méta-données

- Utilisation des méta données
  - Données sur les données
  - Types de données, domaines, ou contraintes sur les données peuvent aider mais ne sont pas toujours suffisantes
  - Interprétation des valeurs
    - Échelle de valeurs, unités de mesure
  - Interprétation des tables
    - Fréquence des mises à jour, définition de vues

41

Exemple : XML

- Format d'échange de données fondé sur le langage SGML
- A une structure arborescente
  - Champs à plusieurs valeurs, structures complexes, etc.
- "auto-descriptif" : le schéma fait partie de l'enregistrement
  - Attributs d'un champs
- DTD (*Document Type Definition*) : grammaire permettant de vérifier la conformité du document XML.

```
<tutorial>
 <title> Data Quality and Data Cleaning: An Overview <\title>
 <Conference area="database"> SIGMOD <\Conference>
 <author> T. Dasu
    <bio>  Statistician <\bio> <\author>
 <author> T. Johnson
    <institution> AT&T Labs <\institution> <\author>
 <\tutorial>
```

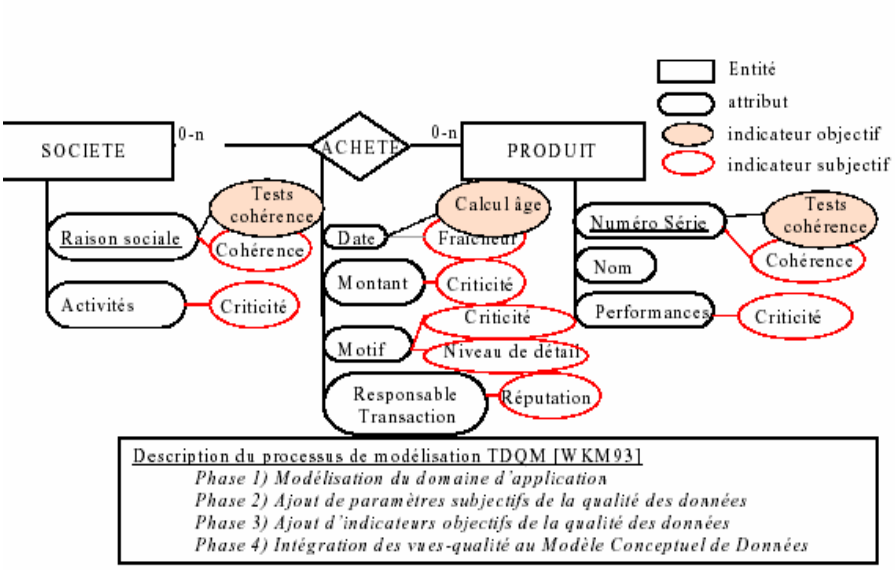
42

Conception des données:
TDQM Wang et al

- Propose une méthodologie de modélisation de la qualité qui:
  - Guide l'ajout de la dimension qualité sur les éléments du modèle
    - Modélisation du domaine
    - Rendre explicite les paramètres de qualité subjectifs
    - Rajouter des paramètres de qualité objectifs (critères subjectifs mesurables)
    - Intégration des différentes vues de la qualité

43

Conception des données:



Remarques sur la méthode

- Cette approche semble difficile à réaliser vu les coûts liés à la création et à la maintenance de tels modèles.

45

Autres approches par les méta données

- Documents et données semi-structurées
  - Annotation qualitatives des données (ex. celles issues du web) par la définition de critères (politique de diffusion, contenu, forme, qualité de la documentation, etc.)
- Entrepôts de données
  - Garder la trace des transformations subies par les données. Ces métadonnées sont ensuite utilisées pour le déboguage des erreurs.

46

Approches centrées sur le processus

- Des processus métier qui promeuvent la qualité des données.
  - Récompenser (financièrement) les efforts pour la qualité
  - Standardisation des contenus et des formats
  - Mettre en place des contrôles pour la saisie unique et correcte des données
  - Automatisation
  - Assignation des responsabilités : garants de la qualité
  - Audit de bout en bout
    - Surtout lorsqu'il y a échange de données entre des organisations.
  - Écoute et surveillance de données pour comparer l'état actuel des données par rapport à ce qui est attendu
  - Boucle de retro-action pour corriger les erreurs

47

Surveillance des données

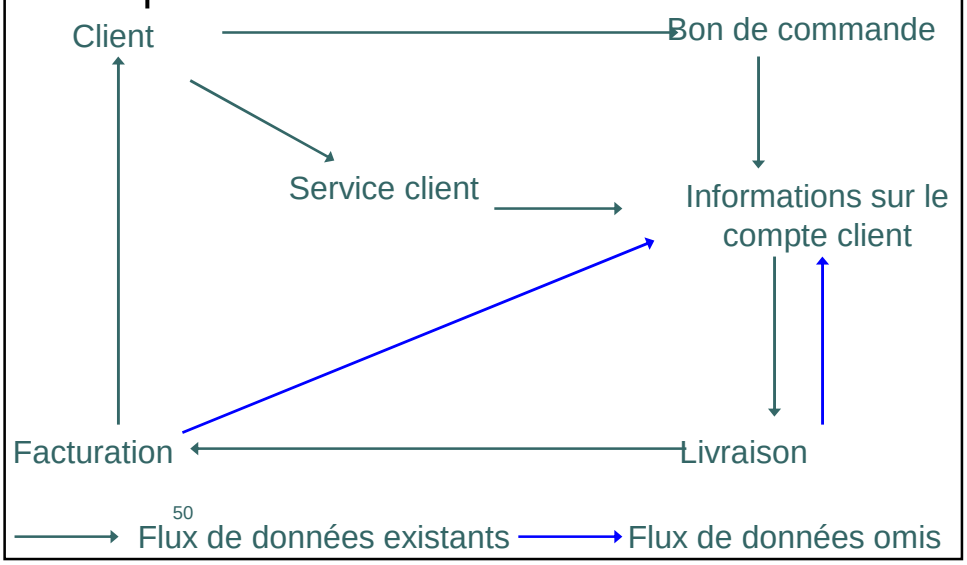
- Traquer les données pour compléter les boucles de retro-action
- Méthodes:
  - Audit et surveillance des données
    - Suivre un échantillon de transactions dans un workflow .
  - Réconcilier les BD mises à jour de manière incrémentale avec les sources originelles
  - Publication des données

48

Exemple (1/2)

- Vente, approvisionnement, et facturation de services de télécommunications
  - Diverses étapes faisant intervenir à la fois des organisations et des bases de données.
- La transition entre diverses unités organisationnelles est souvent source d'erreurs.
- Il existe des flux de retour de données naturels
  - Le client qui se plaint si la facture est trop élevée
- Il existe aussi des flux manquants
  - Le client ne se plaindra généralement pas si la facture est en dessous du prix initialement prévu.

Exemple (2/2)



La technologie des BD et la qualité

- La plupart des données sont dans des bases de données
  - Des outils puissants pour l'analyse et le requêtage des données
  - Des outils et utilitaires qui facilitent import/export et l'accès aux données
  - Moyens pour valider les données.
  - Intégration de données de diverses sources

51

La technologie des BD et la qualité

- Les SGBD fournissent des moyens de garantir une certaine qualité des données
  - Types de données
    - String, date, float, integer
  - Domaines (restreindre les valeurs prises par un champs)
  - Contraintes
    - Contraintes sur les colonnes
      - Not Null, Unique
    - Contraintes sur les tables
      - Contraintes de clé primaire et de clé étrangères
  - Un langage de requêtes puissant et performant (SQL)
  - Les triggers

52

Pourquoi existe-t-il des bases non propres

- Contraintes souvent omises
  - Parfois les DBA lèvent les contraintes pour faire des manipulations rapides.
- Données complexes, hétérogènes et mal comprises
  - Issues de fusions, d'intégration ou du web.
- Problèmes non détectés
  - Valeurs incorrectes ou omises
- Non maintien des méta données
  - Modification des données sans répercussion sur les schémas.

53

Métriques pour évaluer la qualité d'une BD relationnelle

Exemple Naumann, leser 1999

Spécifique à la source de données	Facilité de compréhension: jugement utilisateur de 1 à 10 Fiabilité: sore de 1 à 10 fondé sur des préférences des méthodes expérimentales Actualité: fréquence de màj mesurée en nb jours
Spécifique aux requêtes	Disponibilité: % du temps pendat lequel la donnée est accessible Coût: coût d'une requête en $ Temps de réponse: temps d'attente moyen par requête Précision: % de données sans erreurs Pertinence: % du monde réel représenté dans la source
Spécifique aux attributs	Complétude: pourcentage de valeurs non nulles

54

Approches correctives par les techniques de BD

- Nettoyage des données par extraction et transformation
  - Exploiter les métriques afin d'améliorer la qualité des données
  - Le nettoyage consiste à:
    - Auditer les données pour détecter les incohérences
    - Choisir les transformations pour les corrections
    - Appliquer les transformations choisies
  - On utilise des outils d'audit (ACR) et les ETL (extraction-transformation-loading) pour la mise en œuvre du nettoyage

55

Approches correctives par les techniques de BD: Appariement approximatif et élimination des doublons

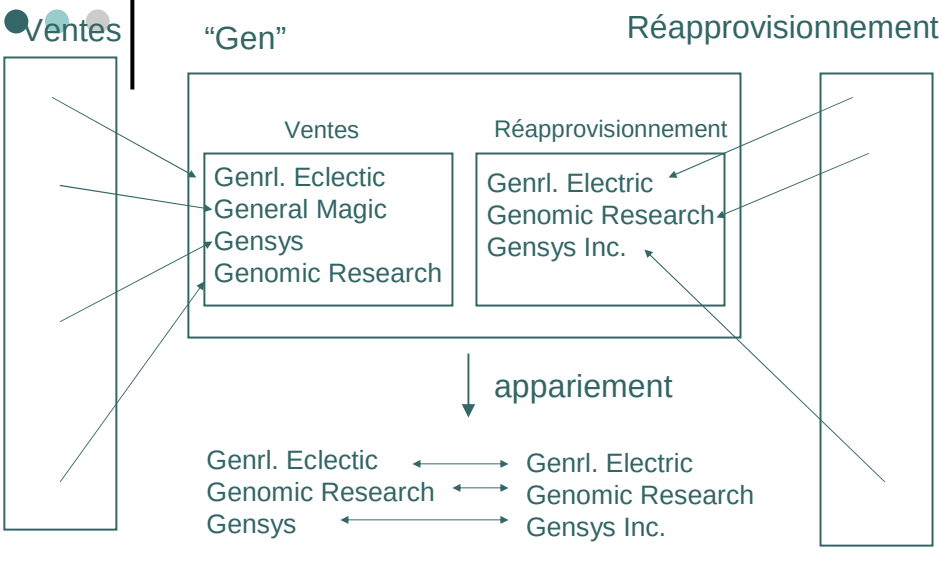
- Relier les tuples dont les champs sont proches
  - Appariement des champs texte
    - Utilisation des distance fondées sur les opération d'édition, ex. "telecom" et "telefon" sont a une distance d'édition de 2 (2 caractères différents)
  - Appariement d'arbres
    - Pour XML
    - Plus coûteuse que celle des chaînes
  - Appariement ad-hoc
    - Chercher la solution la plus probante

56

Approches correctives par les techniques de BD: Appariement approximatif et élimination des doublons

- Jointure approximative et élimination des doublons
  - Jointure approximative : entre deux tables (rapprochement puis fonction de hachage)
  - Eliminer les doublons : dans une même table
- Plus général que l'appariement approximatif
  - **Corrélation d'informations** : utiliser d'autres sources
  - **Données absentes** : croisement de plusieurs recherches

Exemple de jointure approximative



58

Approches statistiques et fouille de données exploratoire

- Absence de méthode de qualité des données explicite
  - Des données statistiques sont recueillies par des expérimentation
- Quatre catégories peuvent être adaptées à la problématique de la qualité des données
  - Les techniques dirigées par les modèles
  - Données absente, incomplète, ambigües, altérée ex. Donnée tronquée ou censurée.
  - Données anormale ou suspecte (données isolées ou outliers)
  - Test de validité (Goodness-of-fit) pour vérifier l'indépendance des attributs

59

Conclusion

- Rien n'est complètement garanti
  - Les méta données sont souvent absentes ou incomplètes
  - Le transfert de données réussi rarement du premier coup
  - Les saisies manuelles sont source d'erreurs
- La définition de métriques pour la qualité des donnée est la solution
  - Définir et mesurer le problèmes
  - Créer les méta données.
- Organiser la qualité des données volumineuse
  - Il est important de monter le gain
  - Désigner des chefs d'orchestre pour les processus de bout en bout
  - Publier les donner pour avoir un retour et faciliter la correction des données

60