# Qualité des données

Source Helena Galhardas,Carlo Batini

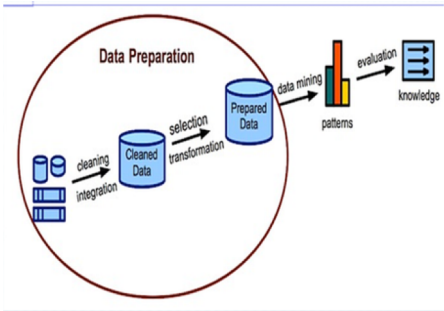
## Plan

- Introduction
- Problèmes de la qualité
- Dimensions de la Qualité
- Activités pertinentes dans la qualité des données

# Introduction

Les données sources sont « nettoyées » avant traitement

Les applications (traitements) sont effectuées sur les données préparées




# Problèmes de qualité

La donnée, dans le monde réel, est **erronnée (sale)**

- **Incomplète** : Manque valeurs d'attribut, manque certains attributs d'intérêt, ou contenant seulement des données agrégées
  - par exemple, la profession = « »
- **Bruitée** : Contenant des erreurs ou des valeurs aberrantes (orthographe, erreurs de phonétique et de dactylographie, transpositions, valeurs multiples dans un seul champ de format libre)
  - par exemple, Salaire = « - 10 »
- **Inconsistante** : Contenant des anomalies dans des codes ou des noms (synonymes et surnoms, variations de préfixe et suffixe, abréviations, troncature et initiales)
  - Exemple : âge = « 24 » et dateDeNaissance = « 03/07/1998 »
  - Exemple : la note « 1,2,3 », est remplacée par « A, B, C »
  - Exemple,- : écart de valeur entre les enregistrements en double

## Origine des erreurs

- **Données incomplètes** :
  - valeur de données non disponibles au moment de la collecte
  - critères différents entre le moment où les données ont été recueillies et lors de leur analyse.
  - Problèmes humain / matériel / logiciels
- **Données bruitées** :
  - collecte de données : instruments défectueux
  - saisie des données: erreurs humaines ou informatiques
  - transmission de données
- **Données incompatibles (et redondantes) :**
  - Les sources de données sont différentes
    - conventions de nommage non uniforme
  - violation de dépendance fonctionnelle / d'intégrité référentielle

## Objectif qualité

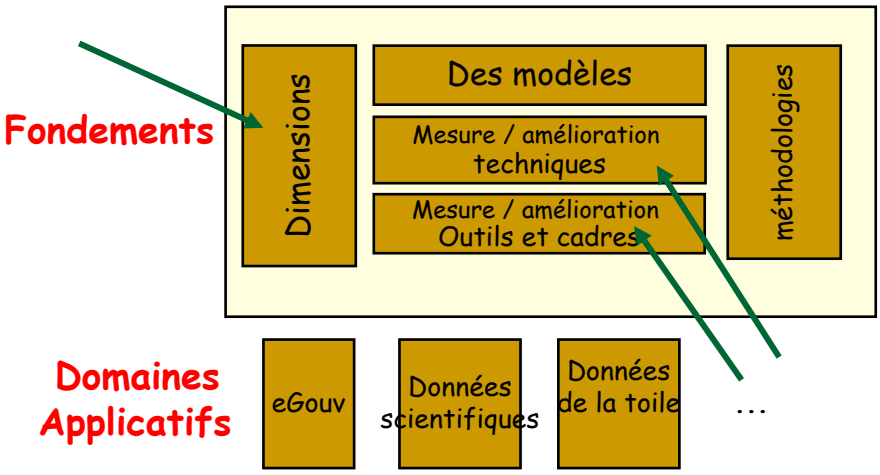
Conversion de données source en des données cible sans erreurs, doublons et les incohérences, à savoir,

Nettoyage et transformation pour obtenir ...

**des données de qualité!**

- Pas de donnée de qualité, pas de décision de qualité!
  - Les décisions de qualité doivent être fondées sur des données de bonne qualité (les données en double ou manquantes peuvent provoquer des statistiques incorrectes ou même trompeuses)

## Composants «génériques» d'un système de gestion de la qualité




## Contextes applicatifs

- **Intégrer** les données provenant de différentes sources
  - Alimenter un entrepôt de différents magasins de données opérationnelles
- **Éliminer les erreurs et les doublons**
  - Les doublons dans un fichier de clients
- **Migrer** les données d'un schéma source vers un schéma cible fixe
  - Les packages d'applications héritées (« legacy »)
- **Convertir des données mal structurées** en des données structurées
  - Le traitement des données recueillies à partir du Web

# Dimensions de la qualité

- Précision (Accuracy)
  - Des erreurs dans les données

  Exemple: » JHN » par rapport à « John »
- Pertinence (Currency)
  - Le manque de données mises à jour

  Exemple: Mise à jour d'une adresse (permanente)
- Cohérence (Consistency)
  - Anomalies dans les données

  Exemple: Code postal et ville incohérents
- Complétude (Completeness)
  - Manque de données
  - Connaissance partielle des enregistrements dans une table ou des attributs d'un enregistrement

# Complétude : exemples

| Prefix | StreetName | Number | ZipCode | City |
|---|---|---|---|---|
| Via | Salaria | 113 | 00198 | Roma |

| Prefix | StreetName | Number | ZipCode | City |
|---|---|---|---|---|
| | Salaria | | | Roma |

Attribute Completeness

| Prefix | StreetName | Number | ZipCode | City |
|---|---|---|---|---|
| Via | Salaria | 113 | 00198 | Roma |
| Via | Gracchi | 74 | 00193 | Roma |

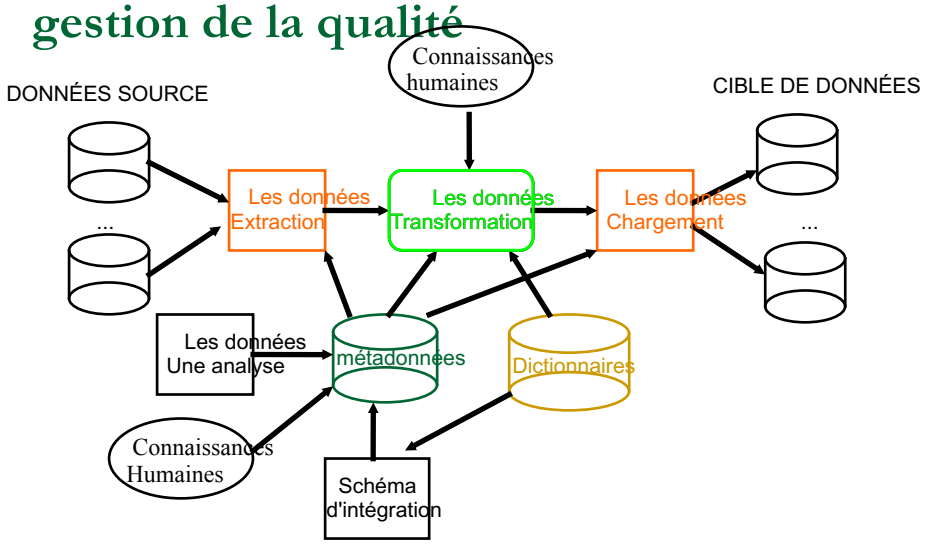
| Prefix | StreetName | Number | ZipCode | City |
|---|---|---|---|---|
| Via | Salaria | 113 | 00198 | Roma |

Entity Completeness

# Technologies pour assurer la qualité

- **programmes ad hoc** écrits dans un langage de programmation comme C ou Java ou en utilisant un langage propriétaire SGBDR
  - Programmes difficiles à optimiser et mainteni
- **mécanismes de SGBDR** pour garantir les contraintes d'intégrité
  - Ne tarite pas  les problèmes d'instance
- Les scripts à base d'**ETL (Extract-Transform-Load)**
- **Nouvelles technologies à base de Machine Learning**

# Architecture typique d'un système de gestion de la qualité



# Problèmes de Qualité

## Problèmes de qualité (1/3)

- **niveau du schéma**
  - Problèmes peuvent être résolus avec une meilleure conception/traduction/intégration de schémas.
- **niveau de l'instance**
  - erreurs et incohérences des données (valeurs) qui ne sont pas résolues au niveau du schéma

# Problèmes de qualité (2/3)

- **Au niveau du schéma**
  - Evité par un SGBDR
    - Données manquantes - *prix du produit non rempli*
    - Mauvais type de données - *« Abc » dans le prix du produit*
    - valeur de données erronées - *0,5 dans les taxes de produit (IVa)*
    - copie exacte des données - *différentes personnes avec le même N° Sécu.*
    - contraintes de domaine génériques - *prix facture incorrect*
  - Non évité par un SGBDR
    - Données erronées catégoriques - *pays et états correspondants*
    - données temporelles invalides – contrainte du *just-in-time*
    - données spatiales incompatibles - *coordonnées et formes géo.*
    - Conflits de noms - *personne versus personne ou personne versus client*
    - Conflits structurels - *adresses*

# Problèmes de qualité(3/3)

- **niveau de l'instance problèmes de qualité des données**
  - enregistrement unique
    - Les données manquantes dans un champ *non nul* N°SS *: -9999999*
    - données erronées - *prix: 5, mais le prix réel: 50*
    - fautes d'orthographe *José Maria Silva vs José Maria Sliva*
    - Données enfouies : *Prof. José Maria Silva*
    - valeurs de champs incorrectes : *ville: France*
    - données ambigües: *J. Maria Silva; Miami en Floride, Ohio*
  - Enregistrements multiples
    - Les doublons : *Nom: Jose Maria Silva, Naissance: 01/01/1950 et Nom: José Maria Sliva, Naissance: 01/01/1950*
    - Valeurs contradictoires : *Nom: José Maria Silva, Naissance: 01/01/1950 et Nom: José Maria Silva, Naissance: 01/01/1956*
    - Données non normalisées: *José Maria Silva vs Silva, José Maria*

# Dimensions de la qualité

## Dimensions “classiques”

- **Précision**
- **Complétude**
- dimensions temporelles: **Pertinence, Ponctualité**, et **Volatilité**
- **Cohérence**
- Leurs définitions ne fournissent pas de mesures quantitatives lorsque une ou plusieurs **métriques** doivent être associés
  - Pour chaque mesure, une ou plusieurs **méthodes de mesure** doivent être fournies : (i) où la mesure est effectuée; (Ii) quelles données sont incluses; (Iii) quel est le dispositif de mesure; et (iv) l'échelle à laquelle les résultats sont rapportés.
- Les dimensions de la qualité du schéma sont également définies

# Précision

- **La proximité entre une valeur v et une valeur v'**, considérée comme la représentation correcte du phénomène du monde réel que v vise à représenter.
  - Ex: pour un nom de personne « John », v '= John est correct, v = Jhn est incorrect
- **Précision syntaxique**: Proximité d'une valeur v avec les éléments du domaine de définition
  - Ex: si v = Jack, même si v '= John, v est considéré comme syntaxiquement correct
  - Mesurée au moyen de fonctions de comparaison (Par exemple, la distance d'édition) qui retourne un score
- **Précision sémantique**: La proximité de la valeur v à la valeur réelle v'
  - Mesurée avec un <oui, non> ou <correct, pas correct>
  - Coincide avec exactitude
  - La valeur réelle correspondante doit être connue

# Ganularité de la précision

- La précision peut se référer à:
  - une seule valeur d'un attribut de relation
  - un attribut ou d'une colonne
  - une relation
  - la base de données entière

## Quantifier la précision

- Faible erreur de précision
  - Caractérise les erreurs de précision qui ne touchent pas l'identification des tuples
- Forte erreur de précision
  - Caractérise les erreurs de précision qui affectent l'identification des tuples
- Pourcentage de tuples précis
  - Caractérise la fraction de tuples appariés

## Complétude

- «Les données sont d'une largeur, profondeur et portée suffisantes pour réaliser une tâche".
  - **De schéma**: taux d'absence des concepts et leurs propriétés dans un schéma
  - **De colonne** : pourcentage de valeurs manquantes pour une colonne dans une table.
  - **D'une population**: taux de valeurs manquantes par rapport à une population de référence

## Complétude relationnelle

- **La complétude d'une table** caractérise le fait qu'une table représente le monde réel.
  - **présence / absence et signification des valeurs nulles**

  Exemple:

  Personne (nom, prénom, date de naissance, e-mail),

  - e-mail vaut NULL pour indiquer que la personne n'a pas de e-mail (pas d'incomplétude)
  - e-mail existe, mais pas renseigné (incomplétude)
  - On ne sait pas si une personne a un e-mail (pas d'incomplétude)

## Quantification de la complétude (1)

- **Modèle sans valeurs nulles avec Hypothèse du Monde Clos (HMC)**
  - Besoin d'une référence de relation r » pour une relation r, qui contient tous les tuples qui satisfont le schéma de r

  Exemple:

  Si Marseille compte 1 M d'habitants, et si une entreprise possède une base de données avec un nombre égal à 800.000, alors le taux de complétude est de 0,8.

## Quantification de la complétude (2)

- **Modèle avec des valeurs nulles avec HMC**: Définitions spécifiques pour des granularités différentes
  - **valeurs**: Pour capturer la présence de valeurs nulles pour certains champs d'un tuple
  - **tuple**: Pour caractériser l'intégralité d'un tuple par rapport aux valeurs de tous ses domaines:
    - Évalue le% des valeurs indiquées dans le tuple par rapport au nombre total d'attributs du tuple lui-même

    Exemple: Étudiant (stID, nom, prénom, vote, examdate)

    Vaut 1 pour (6754, Mike, Collins, 29, 17/07/2004)

    Vaut 0,8 pour (6578, Julliane, Merrals, NULL, 17/07/2004)

## Quantification de la complétude (3)

- **Attribut :** mesure le nombre de valeurs nulles d'un attribut spécifique dans une relation
  - Pourcentage des valeurs indiquées dans une colonne par rapport au nombre total de valeurs qui auraient dû être indiquées.

  Exemple: Pour le calcul de la moyenne des voix des étudiants, la complétude de l'attribut Vote est nécessaire
- **Relations**: Capture le taux des valeurs (non) nulles dans l'ensemble de relation
  - Mesure le taux d'information présente dans une relation par rapport à un contenu "maximal" possible, c'est à dire sans valeurs nulles.

## Dimensions temporelles

- **Pertinence** : Concerne la façon dont les données sont mises à jour sans délai

  Exemple: Si l'adresse résidentielle d'une personne est mise à jour (l'adresse où la personne vit), la pertinence est élevée
- **Volatilité** : fréquence de variation des données dans le temps

  Exemple: Les dates de naissance (zéro volatilité) vs les cours boursiers (degré élevé de volatilité)
- **Ponctualité** : Degré “d’actualité” des données pour réaliser une tâche donnée, à un instant donné

  Exemple: Le calendrier des cours universitaires peut être pertinent car il contient les données les plus récentes, mais il ne sera pas ponctuel s’il est publié après le début des cours.

## Métriques des dimensions temporelles

- Pertinence : dernière mise à jour des métadonnées
- Volatilité : durée de validité des données
- Ponctualité : pertinence + vérification que les données sont disponibles avant l'heure d'utilisation prévue

## Cohérence

- capture la violation des règles sémantiques définies sur un ensemble d'éléments de données, où les éléments de données peuvent être des tuples, des tables ou des enregistrements relationnels dans un fichier
  - Les contraintes d'intégrité
    - Les contraintes de domaine, clé, inclusion et dépendances fonctionnelles
  - La vérification des données: règles sémantiques dans les statistiques

## Evolution des dimensions

- Les dimensions traditionnelles sont la précision, la complétude, la ponctualité, et la cohérence

1. Avec l'avènement des réseaux, les sources augmentent considérablement, et les données deviennent souvent des « données trouvées ».
2. la collecte et l'analyse des données sont souvent déconnectés.
   - Nécessité de considerer de Nouvelles dimensions de qualité

## Autres dimensions

- Interprétabilité : Concerne la documentation et les métadonnées disponibles pour interpréter correctement la signification et les propriétés des sources de données
- Synchronisation entre différentes séries chronologiques :
  - bonne intégration des données horodatées.
- Accessibilité : Mesure la capacité de l'utilisateur à accéder aux données selon son environnement (culturel, technique, ...)

# Activités orientées qualité

## Activités orientées Qualité

- **Standardization / normalisation**
- **Appariement d'entités / Identification d'objet ou entité / Matching d'enregistrements**
- **Intégration de données**
  - Correspondances entre schémas
  - Résolution de conflits (entre instances)
  - Sélection de sources (à intégrer)
  - Fusion de résultats
  - Composition de la qualité
- **localisation des erreurs / données d'audit**
  - Edition-Imputation de données / Détection de déviation
- **Le profilage des données**
  - induction de la structure
- **Correction / Nettoyage des données**
- **Nettoyage de schéma**

## Normalisation

- Modification des données avec de nouvelles données selon les normes définies ou des formats de référence

Exemple:

- Changer « Pl. de Gaulle» en « Place de Gaulle»
- Changer « Bob» en « Robert»

## Appariement d'entités / Identification d'objet ou entité / Matching d'enregistrements

- Activité nécessaire pour déterminer si les données de la même source ou dans des sources différentes représentent le même objet du monde réel

## Intégration de données

- Présenter une vue unifiée des données appartenant à des sources de données hétérogènes et distribuées
- Deux sous-activités:
  - **traitement des requêtes dirigé par la qualité** : fournir des résultats de requête sur la base d'une caractérisation de la qualité des données au niveau des sources
  - **résolution des conflits d'instances** : identifier et résoudre les conflits de valeurs se rapportant aux mêmes objets du monde réel.

## Résolution des conflits d'instances

- Trois types de conflits d'instance :
  - **conflits de représentation**
    - Dollar par rapport à l'Euro
  - **conflits d'équivalence de clés**
    - même monde réel mais les objets ont des identifiants différents
  - **conflits de valeurs d'attibuts**
    - les instances correspondant aux mêmes objets, ayant la même clé mais ont des valeurs diferentes pour les autres attributs

## Localisation des erreurs / données d'audit

- Étant donnés 1, 2 ou n tables ou groupes de tables, et un groupe de contraintes d'intégrité / qualités (complétude, exactitude), trouver des documents qui ne respectent pas les contraintes / qualités.
  - édition-imputation des données
    - Mettre l'accent sur les contraintes d'intégrité
  - détection de déviation
    - vérification des données qui marquent les déviations comme des erreurs de données possibles

## Le profilage des données

- L'évaluation **propriétés statistiques et propriétés intensionnelles des tables et des dossiers**
- Induction d'une structure
  - **description structurelle**, c'est à dire « toute forme de régularité qui peut être trouvée »

## Correction /Nettoyage des données

- Étant donnés 1, 2 ou n tables ou groupes de tables, et une série d'erreurs de qualité identifiées dans les tuples, générer des corrections probables et corriger les tuples, de telle sorte que les nouveaux tuples respectent les qualités.

## Nettoyage de schéma

- Transformer le schéma conceptuel pour atteindre ou optimiser un ensemble de qualités donné (par exemple, la lisibilité, la normalisation), tout en conservant d'autres propriétés (par exemple, l'équivalence de contenu)