# Base for Data Quality Systems
# BDQS

***Une gestion opérationnelle de la qualité de données***

***Gilles Amat***

***www.bdqs.com***

***gamat@bdqs.com***

18 Janvier 2005

# BDQS
# Pourquoi ? (1)

- BDQS a été conçu sur le constat suivant :
  - Lorsqu'on manipule une donnée, on a souvent besoin de lui associer 'une note qualité' :
    - Par exemple, un Téléphone n'aura pas la même valeur, la même utilisation fonctionnelle selon sa probabilité à être valide.
  - Cette note qualité, ou attribut, est dépendante de la ou les variables auxquelles elle se rapporte
    - Si le téléphone change, la note qualité afférente doit être réévaluée

# BDQS
# Pourquoi ? (2)

- Les dictionnaires des Bases de Données actuelles, à notre connaissance, ne comportent pas cette notion d'attribut qualité dépendant d'une variable
- *'Current commercial relational database management systems and their underlying relational model are based on the assumption that data stored in the databases are correct'.*
  - *Richard Wang, Data Quality, 2001*

  - Le moyen de pallier cela consiste à créer des triggers qui mettent à jour les attributs qualité : l'intégrité demeure fonctionnelle, et non pas structurelle.

  - Il n'existe pas de langage de manipulation de l'information dans le contexte de l'attribut qualité

# BDQS
# Pourquoi ? (3)

- ➔ Naissance de BDQS
  - Intégrer la notion d'attribut qualité sur une base 'maîtrisée', dont on possède les sources.
  - Base retenue : un moteur de comptage optimisé pour les requêtes rapides en Marketing Direct
  - BDQS est donc plutôt un prototype, utilisé dans un monde réel, opérationnel afin d'en tirer des règles

# BDQS
## Les différents modules

- BDQS comporte une série de modules dédiés aux différentes phases de prise en compte , mesure, manipulation et diffusion de la qualité de données :
  - BDQS Load Module
  - BDQS Publication Module
  - BDQS Query Module
  - BDQS Update
  - BDQS Evolutions

# BDQS

## Exemples de règles qu'on pourrait construire ….

- Un attribut qualité est lié à une variable ou un bloc de variables
- Toutes les variables d'un bloc sont mises à jour dans une même transaction
- Le changement d'un élément de calcul d'un attribut déclenche l'annulation ou le recalcul de l'attribut selon une propriété de l'attribut

# BDQS
## Load Module

- Module qui permet la prise en compte de données et le calcul d'attributs qualité.

- *Qu'est-ce qu'un attribut qualité : un indicateur affecté à une variable ou un ensemble de variables qui restitue leur niveau qualité.*

- Par exemple :
  - La présence d'un SIRET dans le référentiel SIRENE et le degré de correspondance en nom/adresse
  - La présence d'un couple (CP,Ville) dans un référentiel postal.
  - La cohérence de l'information entre le client et ses actes d'achats

# BDQS
## Load Module

- BDQS Load Module comporte 3 catégories de fonctions :
  - Niveau enregistrement : les valeurs de l'attribut ne s'appuient que sur des informations de l'enregistrement traité.
  - Niveau table : les valeurs de l'attribut s'appuient sur l'ensemble de la table (calcul de doublons par exemple)
  - Niveau base : les valeurs de l'attribut s'appuient sur plusieurs tables (calculs de cohérence par exemple)

# BDQS
## Load Module

- Exemples de fonctions :
  - Contrôle de syntaxe
  - Contrôle dans un référentiel et calcul de proximité sur des éléments d'identité du référentiel (par exemple nom,adresse)
  - Calcul de doublons approximatifs à travers une fonction de Dédoublonnage intégrant elle-même les attributs qualité.
  - *L'interface est ouverte à toute fonction spécifique*

# BDQS
## Load Module

- Liste des principales fonctions :
  - Groupe 1 : Valider la syntaxe d'un champ, la cohérence,..
  - Groupe 2 : Contrôle dans une liste de valeurs
    - Il existe des fonctions différentes dans ce groupe, optimisées pour la recherche :
      - Depuis une recherche dans une simple liste de valeurs
      - Jusqu'au contrôle dans un référentiel tel que le Téléphone : recherche dans le référentiel Téléphone, et comparaison des noms/adresses

Exemple :

| | | | | |
|---|---|---|---|---|
| 0139239300 | AID | 2 rue Henri Le Sidaner | 78000 | Versailles **(EXTERNE)** |
| 0139239300 | Analyse Informatique de Données | 4 rue Henri le Sidaner | 78000 | Versailles **(REFERENTIEL)** |

# BDQS
## Load Module

- Liste des principales fonctions :
  - Groupe 3 : Doublons égaux. Par exemple, lier les records qui ont le même téléphone.
  - Groupe 4 : Doublons similaires. Par exemple, lier les records avec un nom, un prénom, une adresse similaires. Pour cette fonction, c'est le logiciel de dédoublonnage Proxim de la société A.I.D. qui est utilisé.

Exemple :

| | | | |
|---|---|---|---|
| 0139239300 Dupont Raoul | 12 Allée des Acacias | 78000 | Versailles |
| 0139239300 Dupon**d R**. | **10** Allée des Acacias | 78000 | Versailles |

# BDQS
## Load Module

- Ce module, actuellement disponible, est utilisé en opérationnel chez A.I.D.
  - Les temps de chargement, contrôle, calculs d'agrégats sont optimisés
  - Il est enrichi régulièrement par des fonctions métier
  - Un de ses points forts est la traçabilité des erreurs : les variables, les enregistrements erronés sont disponibles dans la Base de Données chargée, avec des attributs indiquant leur niveau de problème.

# BDQS

## *Opérationnel*

- L'information peut ainsi être gérée par des opérateurs classiques dans sa dimension 'Qualité' :
  - Compter les enregistrements erronés
  - Croiser , analyser selon les sources
  - Créer des agrégats par grand segment de client / prospect
  - Publier par BDQS Publication les niveaux qualité

  - Exemple de Publication : ***www.bdqs.com***

***Login : bdqs***

***Passwd : xy7895***