EVALUATION DE LA QUALITE DANS LES SYSTEMES D'INTEGRATION DE DONNEES

Mokrane Bouzeghoub

http://www.prism.uvsq.fr

UNIVERSITE DE VERSAILLES SAINT-QUENTIN-EN-YVELINES

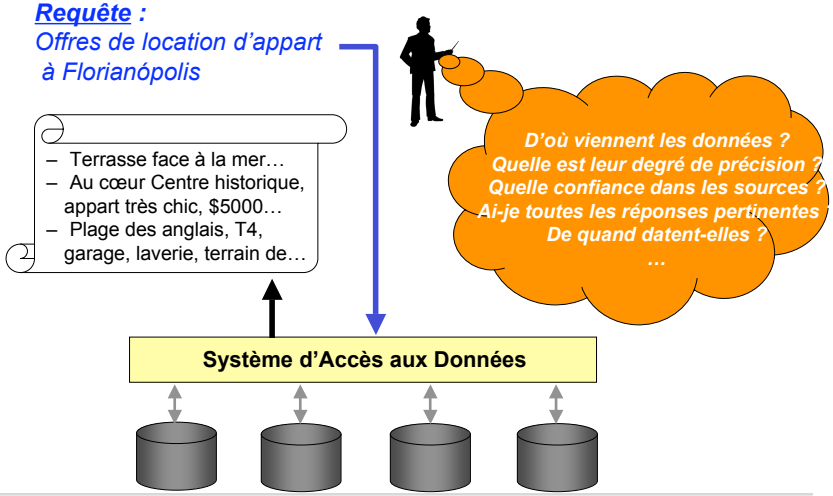
Laboratoire PRISM

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

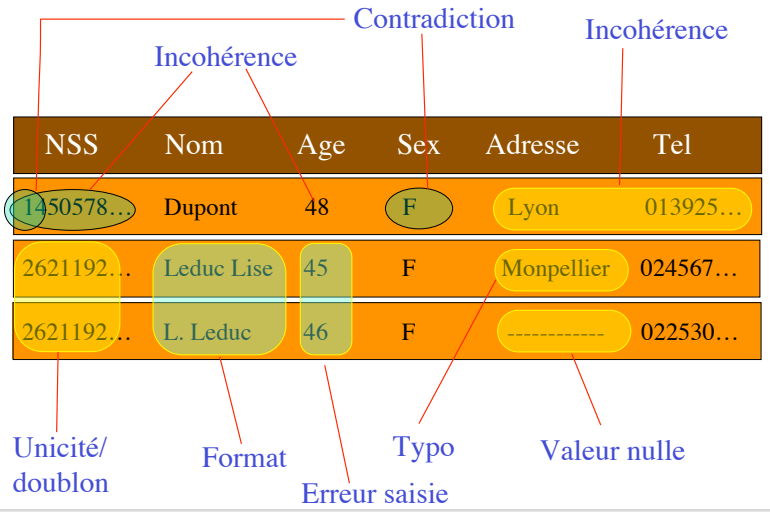
Sommaire

1. Introduction et motivation
2. Terminologie
3. Evaluation de la fraîcheur des données
4. Evaluation de l'exactitude des données

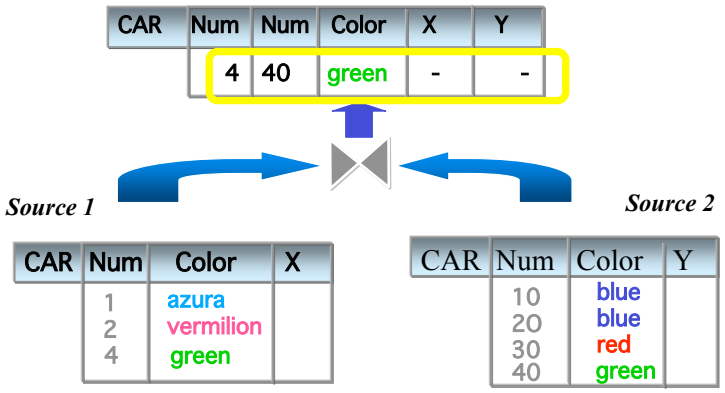
Introduction - motivation



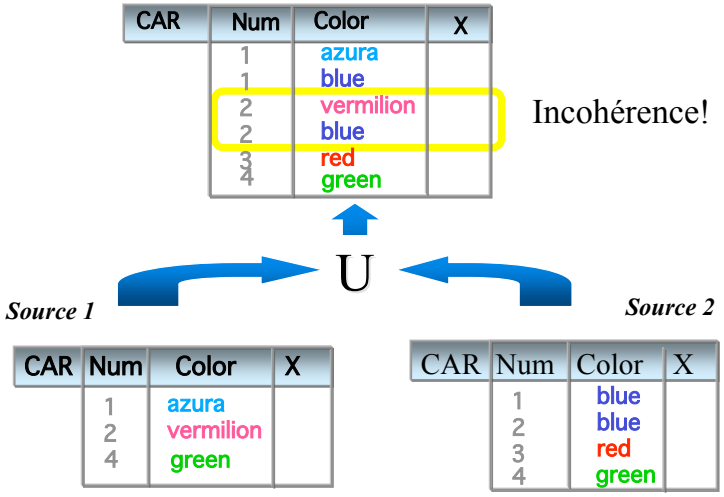
Anomalies sur les données



Exemple 2: sémantique des opérations (jointure)



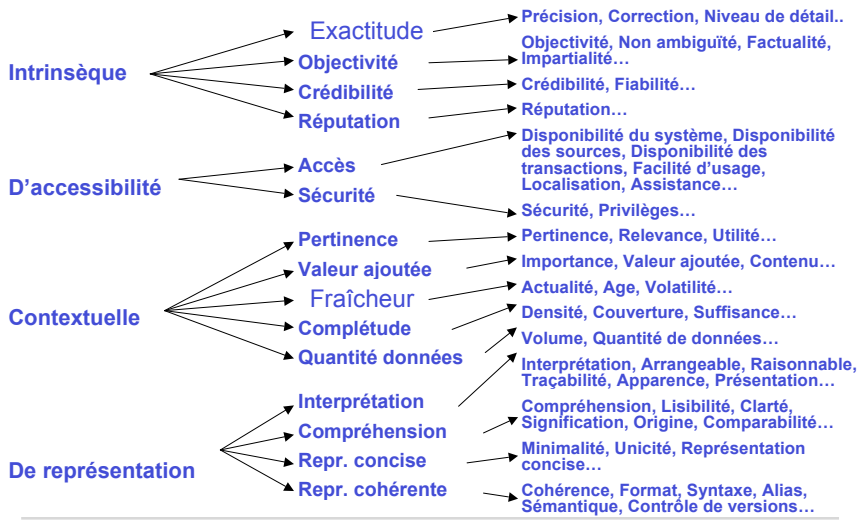
Exemple 3 : sémantique des opérations (union)



Notion de qualité des données

- **Qualité des données : qualifie tout ce qui concerne les types de données et leurs instances quels que soient les ressources décrites et les niveaux de leur description**
- **La qualité des données est basée sur deux concepts**
  - le facteur de qualité
  - le but de qualité
- **La qualité des données doit être évaluable de façon objective**
  - mais l'interprétation des valeurs est dépendante de l'observateur (stakeholder)

Multitude des facteurs



Difficulté de l'évaluation

- Domaine très ouvert vers une multitude de critères et une multitude de perceptions
- Etat de l'art très disparate, allant de méthodes d'estimation empiriques à des modèles d'évaluation très formels et très complexes
- La qualité des données peut rarement s'évaluer *'de visu',*
  - Soit on caractérise le processus de production de l'information
  - Soit on analyse les corrélations entre données
- Les outils d'évaluation de la qualité sont
  - Soit enfouis dans les systèmes (compilation, exécution, correction)
  - Soit externes aux systèmes (outils d'observation, inspection, diagnostique)

Deux grandes approches

- Approche orientée données
  - Inspection des données (détection d'anomalies)
  - Nettoyage des données (actions correctives)

ETL

- Approche orientée processus
  - Analyse des activités et détection des chemins critiques
  - Amélioration du système (évolution, maintenance)

Les deux approches peuvent être combinées mais en pratique rarement

- Complexité des systèmes
- Coût de l'évaluation v.s. Urgence des demandes

Terminologie

- **Facteur de qualité:**
  attribut significatif qui caractérise la qualité d'un objet ou d'un ensemble d'objets dans le contexte d'un but précisé. Il a une valeur d'un certain domaine, mesurée avec une certaine métrique

  Métrique:
  c'est un instrument qui définit l'unité de mesure. Une métrique a un domaine où elle prend ses valeurs et un agent/process qui évalue cette valeur. Le même facteur de qualité peut avoir plusieurs métriques

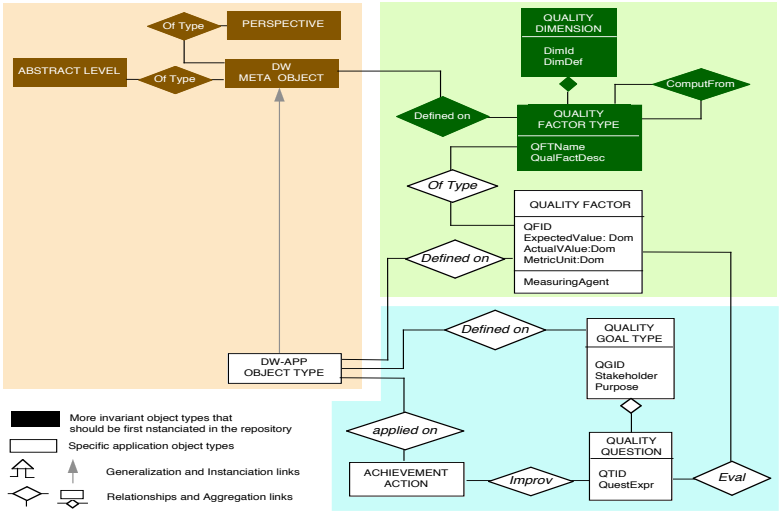
  Dimension:
  C'est une classe de facteurs de même nature. Peut varier d'une classification à l'autre.

Terminologie (suite)

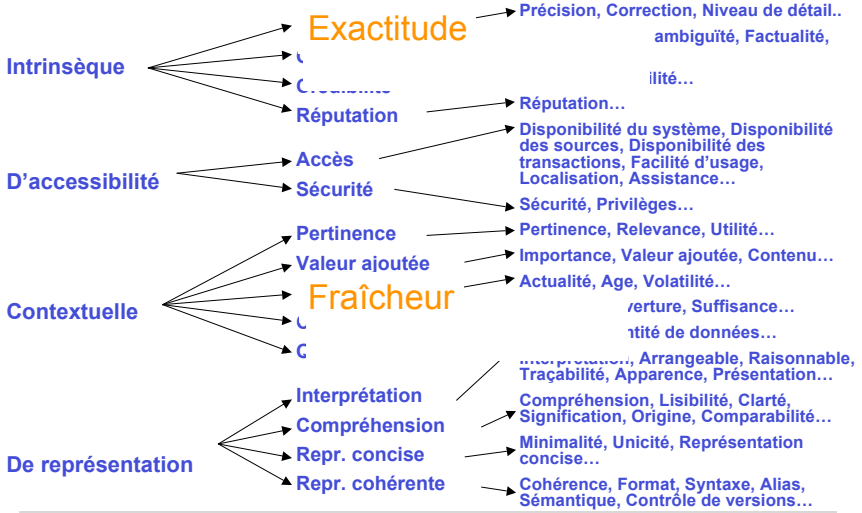
- **Valeur attendue** :
  c'est généralement une valeur ou un intervalle de valeurs qui délimite les valeurs minimale et maximale d'un facteur de qualité.
- **Valeur effective:**
  c'est la valeur calculée pour un facteur de qualité, en utilisant une métrique et un programme particulier. Elle peut être aussi donnée arbitrairement.
- **But de qualité:**
  c'est un projet déterminé par un observateur (stakeholder) qui veut atteindre un niveau de qualité sur une partie ou la totalité d'un système. Si la qualité n'est pas satisfaisante, il doit exister des actions d'amélioration de cette qualité.
- **Question**:

  c'est une subdivision d'un but de qualité en sous-buts élémentaires dont l'évaluation coincide avec un sel facteur de qualité.

Structuration des concepts



Etude de deux facteurs



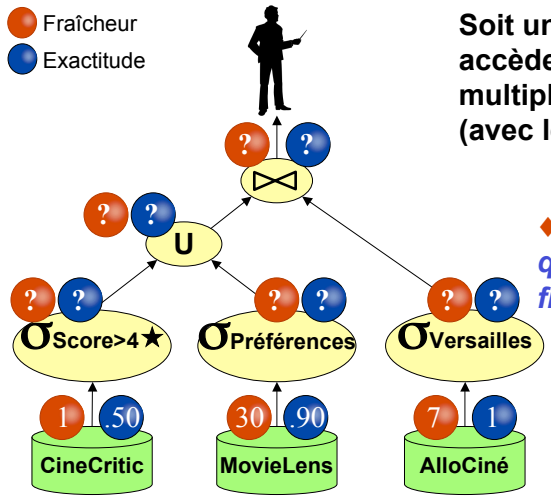
Intérêt de la fraîcheur et de l'exactitude

La fraîcheur et l'exactitude sont des dimensions de qualité importantes :

- ***Plusieurs études montrent leur importance pour les utilisateurs***
- ***Plusieurs types de systèmes visent à les améliorer***

- **Fraîcheur des données :**
  - **Entrepôts de données**
  - **Réplication de données**
  - **Caching**
- **Exactitude des données :**
  - **Une multitude d'applications: CRM, systèmes décisionnels, B2B…**
  - **Nettoyage de données**

Problématique (interrogation)

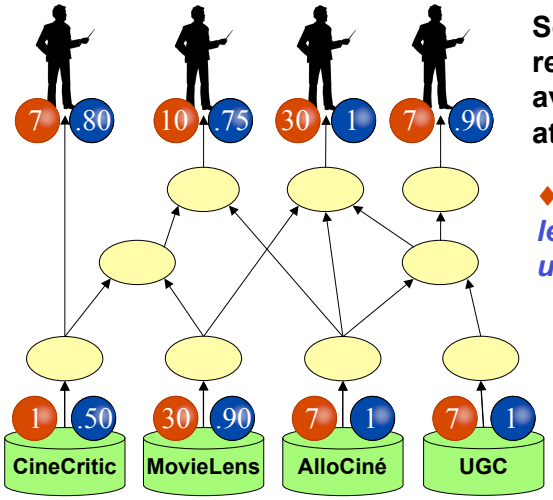


Soit une requête qui accède à des sources multiples (avec leur propre qualité)

♦ ***Comment calculer la qualité des résultats finaux ?***

♦ ***Comment calculer la qualité des résultats intermédiaires ?***

Problématique (conception)



Soient plusieurs requêtes avec différentes attentes de qualité

- ***Comment satisfaire les besoins des utilistaeurs ?***
- ***Comment déterminer les contraintes sur les sources ?***

Plusieurs questions

- **Quelles définitions précises donne-t-on aux concepts de fraîcheur et d'exactitude ?**
- **De quels paramètres dépend la qualité des résultats ?**
- **Comment évaluer la qualité des résultats d'un opérateur/activité ?**
- **Comment évaluer la qualité globale d'une requête/processus ?**
- **A quel moment a-t-on besoin d'évaluer la qualité ?**

Démarche

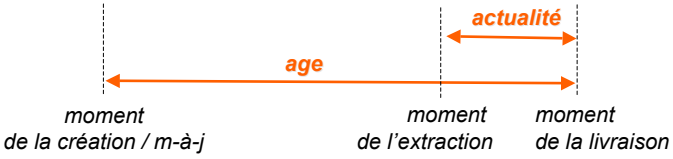
- **Développer un canevas pour :**
  - **Fournir un support formel pour l'évaluation de la qualité**
  - **Analyser des facteurs et des métriques de qualité**
  - **Identifier les propriétés du SID qui influencent ces facteurs**
  - **Développer des algorithmes d'évaluation**
- **Pour chacun des facteurs**
  - **Analyser les définitions et métriques**
  - **Analyser les dimensions qui les influencent**
  - **Définir les algorithmes d'évaluation**
  - **Définir les actions d'amélioration**

Fraîcheur des données (currency, freshness)

- **Plusieurs définitions de fraîcheur :**
  - Actualité (currency): **distance extraction – livraison**
    - **Exemple: solde bancaire**
    - **Métrique: le temps passé après l'extraction**

    Fréquence d'extraction
  - Age (timeliness): **distance création/m-à-j – livraison**
    - **Exemple: Top 10 CDs**
    - **Métrique: le temps passé après la création/m-à-j**

    Fréquence de mise à jour



Dimensions influant sur la fraîcheur

- **Plusieurs paramètres influent sur l'évaluation de la fraîcheur**
- **Ils sont classés en 3 dimensions :**
  - Nature des données
  - Types d'architectures
  - Politiques de synchronisation

Dimension 1: Nature des données

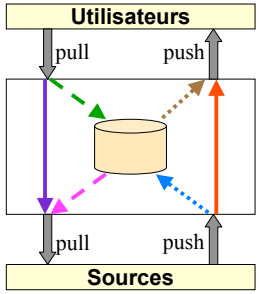
- **Toutes les données ne changent pas selon le même rythme; la fréquence du changement est un élément déterminant dans l'évaluation de la fraîcheur**
  - Données stables : ex. noms villes, codes postales
  - Changements peu fréquents : ex. adresses employés
  - Changements fréquents : ex. stock, info en temps réel
- **Il est quelquefois possible d'anticiper la fraîcheur des données en corrélant l'état actuel de la BD et le cycle de changement des données :**
  - Événements du changement
    - Ex. état civil (mariage, divorce, …)
  - Fréquence de changement
    - Ex. films à l'affiche (tous les mercredis)

Dimension 2: Types d'architectures

- **Les processus qui extraient, intègrent et délivrent des données peuvent introduire des délais**
  - Importants: influant sur la fraîcheur des données
  - Négligeables: par rapport au cycle de vie des données
- **La fraîcheur dépend aussi des modalités de réplication des données :**
  - Virtuelles (coût d'exécution et communication)
  - Caching (temps de vie)
  - Matérialisées (période de rafraîchissement)

Dimension 3: Politiques de synchronisation

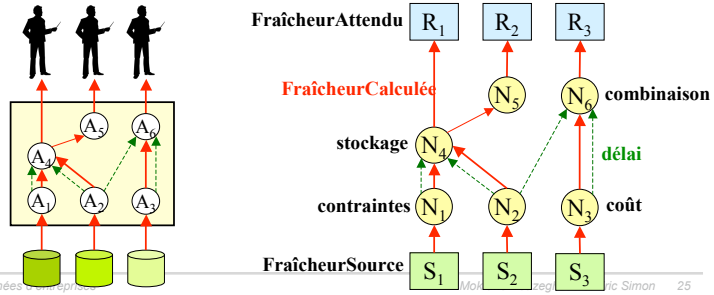
- **2 niveaux de synchronisation:**
  - Sources ↔ DIS ↔ utilisateurs (pull / push)
    ❶ ❷
- **Catégories :**
  - Politiques synchrones :
    - Pull-pull, push-push
  - Politiques asynchrones :
    - Pull/pull, pull/push, push/pull, push/push



- **Les politiques asynchrones introduisent des délais additionnels (fréquence de rafraîchissement)**

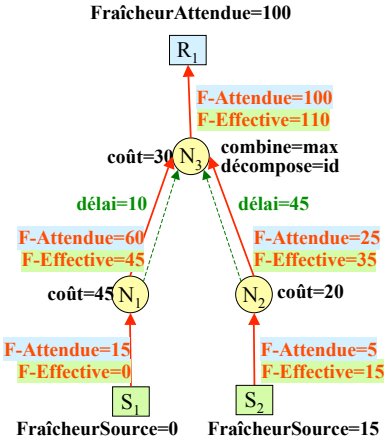
Support de raisonnement : Processus applicatif

- **L'évaluation de la fraîcheur se fait sur une représentation abstraite du processus d'intégration de données**
  - **Graphe de processus**
    - **Nœuds: sources, activités, requêtes**
    - **Arcs: synchronisation, flux de données**
  - **Graphe de qualité (même topologie que celui du processus)**
    - **Nœuds: paramètres sur les sources, activités, requêtes**
    - **Arcs: délais de synchronisation, fraîcheur des flux de données**



Étiquettes associées à la fraîcheur

- **Dérivées des 3 dimensions ou calculées**
  - **Nœuds sources :**
    - Fraîcheur **des données** sources
  - **Nœuds requêtes :**
    - Fraîcheur attendue **par les utilisateurs**
  - **Nœuds activités :**
    - Coût **d'exécution d'une activité**
    - **Fonction de** combinaison **de valeurs de fraîcheur**
    - **Fonction de** décomposition **de contraintes de fraîcheur**
  - **Arcs de contrôle :**
    - Délai **entre l'exécution de 2 activités**
  - **Arcs de données :**
    - Fraîcheur effective **produite par une activité**
    - Fraîcheur attendue **pour une activité**

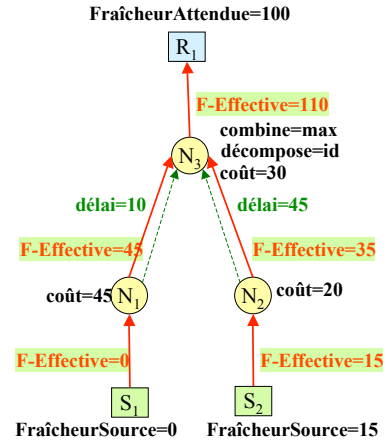


Approche de construction du graphe

- **Entrée: graphe du processus**
  - **Identification des activités (processus)**
  - **Identification des sources**
- **Définition du graphe de qualité**
  - **Définition des exigences utilisateurs (fraîcheur attendue)**
  - **Instanciation des propriétés du graphe (bornes)**
  - **Calcul de la fraîcheur**
    - Calcul par propagation directe **(forward propagation)**
    - Calcul par propagation inverse **(backward propagation)**

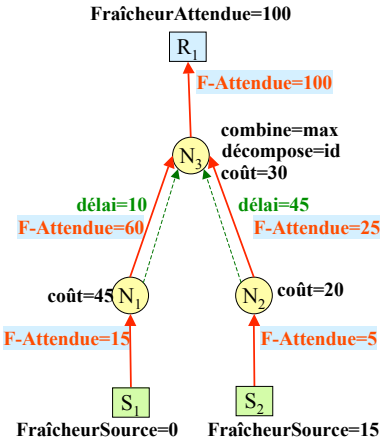
Propagation directe (forward)

- **Permet :**
  - **Fixer les bornes de fraîcheur**
  - **Vérifier la conformité du graphe par rapport aux besoins des utilisateurs**
- **Fonctionnement :**
  - **Propagation de valeurs de fraîcheur le long du graphe**
  - **Calcul de la fraîcheur produite par chaque nœud**



Propagation inverse (backward)

- **Permet :**
  - Fixer des contraintes de fraîcheur sur les sources
  - Vérifier la conformité du graphe par rapport à la fraîcheur des sources
- **Fonctionnement :**
  - Propagation des valeurs de fraîcheur le long du graphe
  - Calcul des contraintes de fraîcheur pour chaque nœud



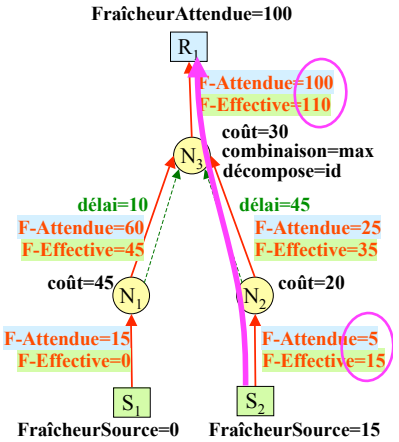
Analyse du graphe de qualité

- **Un graphe est satisfaisant si**

  pour chacun de ses utilisateurs, il produit une fraîcheur conforme à leurs exigences

- **Si une exigence n'est pas satisfaite**

  il faut détecter les chemins critiques déterminant le sous graphe à restructurer

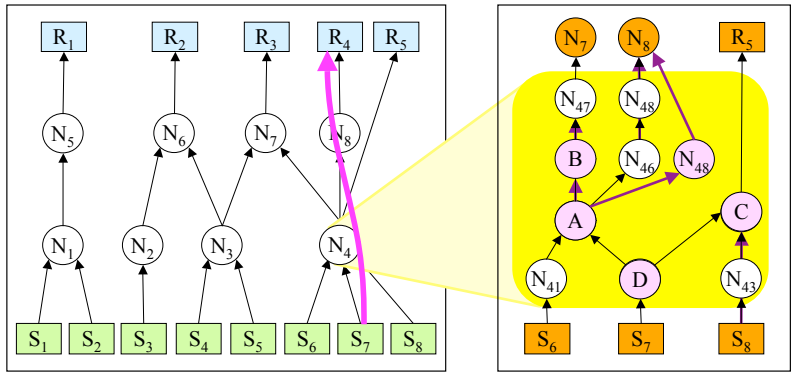


Actions d'amélioration

- **Améliorer la conception des activités pour réduire leurs coûts**
- **Réduire les délais de synchronisation**
- **Décomposer les activités pour augmenter le //**
- **Négocier avec les utilisateurs pour réduire leurs exigences**
- **Négocier avec les fournisseurs pour relaxer les contraintes sur les sources**

Approche de raffinage et restructuration

Hiérarchie d'activités + opérations de navigation + opérations de restructuration



Opérations de navigation et restructuration

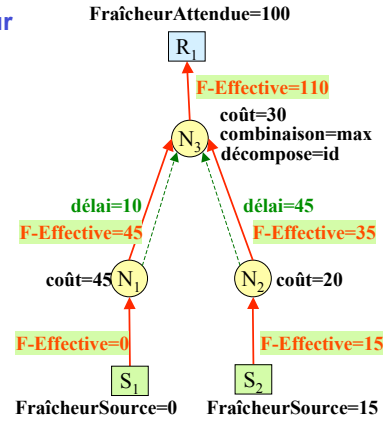
- **Primitives de navigation :**
  - Focus+, focus–
  - Zoom+, zoom–
- **Primitives de restructuration :**
  - Ajouter nœud / arc / étiquette
  - Enlever nœud / arc / étiquette
- **Macro opérations de restructuration :**
  - Décomposer nœud
  - Paralléliser nœuds
  - Fusionner nœuds
  - Remplacer nœud / sous graphe
  - …

Résumé du processus global

- **Construire le graphe de qualité :**
  - Identifier la topologie (activités, sources)
  - Fixer les étiquettes sur chaque nœud / arc
  - Définir les fonctions de calcul de la fraîcheur sur chaque nœud
- **Exécuter les algorithmes de propagation**
- **Vérifier la conformité des résultats obtenus par rapport aux préférences utilisateurs / contraintes sources**
  - Rapport de conformité
- **Pour chaque résultat non conforme :**
  - Déterminer les chemins critiques
  - Analyser les chemins critiques
  - Proposer des actions de restructuration

Exploitation durant l'exécution

- **Acquisition de valeurs réelles**
  - Routines de mesure de la fraîcheur des sources
  - Coût et délais mesurés pendant l'exécution de la requête / processus
- **Apprentissage pour affiner les bornes de la construction initiale**
  - Exécution de requêtes de test
  - Statistiques de coûts et délais
  - Statistiques de mesures de la fraîcheur des sources



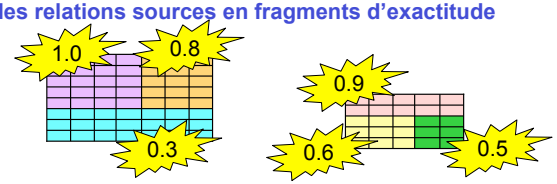
Résumé des contributions sur la fraîcheur

- **Clarification des définitions et métriques de fraîcheur**
- **Étude des dimensions qui influent sur la fraîcheur**
  - Étiquetage du graphe de qualité
- **Deux algorithmes d'évaluation**
  - Selon deux objectifs de qualité :
    - Qualifier les données : propagation de la fraîcheur des sources
    - Pousser les contraintes : propagation de la fraîcheur attendue
  - Paramétrables
- **Outil interactif de diagnostique :**
  - Analyse des chemins critiques
  - Opérations de navigation et restructuration du graphe de qualité

Exactitude des données (accuracy)

- **Problème à résoudre :**
  - Informer l'utilisateur de l'exactitude des résultats obtenus à partir d'une ou plusieurs sources de données
- **But: Définir un cadre d'évaluation**
  - Paramétrable (dimensions qui influent sur l'exactitude)
  - Evolutif en fonction des besoins et des ources
- **Spécificité:**
  - Données sources avec des exactitudes très différents
  - → Partitionnement des relations sources en fragments d'exactitude homogène

Comment combiner toutes ces valeurs ?



Exactitude

- **Plusieurs définitions de l'exactitude :**
  - Correction Sémantique :
    - Les données représentent des états du monde réel
    - Ex: L'adresse de Marcel est "45, av. des États-unis" ?
  - Correction Syntaxique :
    - Les données n'ont pas d'erreurs syntaxiques
    - Ex: fautes d'orthographe/frappe, discordances de format
  - Précision :
    - Degré de proximité des valeurs
    - Ex: "2000 euros" vs "2018 euros"

Comparaison avec le monde réel

Règles de vérification

Métriques d'exactitude

- **Trois familles de métriques:**
  - Booléens :
    - Indique si une donnée est exacte ou pas
    - Ex: numéros de téléphone
  - Déviations de valeurs :
    - La distance entre une donnée et une valeur exacte
    - Ex: distance ('rue St Lazar', 'rue St Lazare')
  - Degrés :
    - La confiance sur l'exactitude d'une donnée
    - Ex: reconnaissance de caractère avec exactitude 0.8, 1.0 et 0.6.

Dimensions influant sur l'exactitude

- **Plusieurs paramètres influent sur l'évaluation de l'exactitude**
- **Ils sont classés en 4 dimensions :**
  - Granularité de la mesure
  - Types d'erreurs
  - Types de données
  - Techniques architecturales

Dimension 1: Granularité de la Mesure

- **Catégories :**
  - Cellule
    - Une mesure pour chaque cellule
  - Objet
    - Une mesure pour chaque tuple
  - Table
    - Une mesure pour chaque relation (ou vue)
- **Le granule le plus fin fournit une meilleure observation de la distribution des erreurs dans les sources de données**

Dimension 2: Types d'erreurs

- **Catégories :**
  - Erreurs de valeur :
    - Domaines des attributs, erreurs de frappe, valeurs hors champ (ville='France')…
  - Erreurs de standardisation :
    - Abréviations ('DB prog'), synonymes, transpositions de mots ('V. Peralta'; 'Peralta V.'), formats/unités pas standard…
  - Valeurs imbriquées :
    - Cellules qui correspondent à des valeurs multiples (nom='J. Smith 12.02.70 Paris')
  - Valeurs fausses :
    - Valeurs qui ne correspondent pas au monde réel
- **Les types d'erreurs vont influer sur le type d'outil d'évaluation à utiliser**

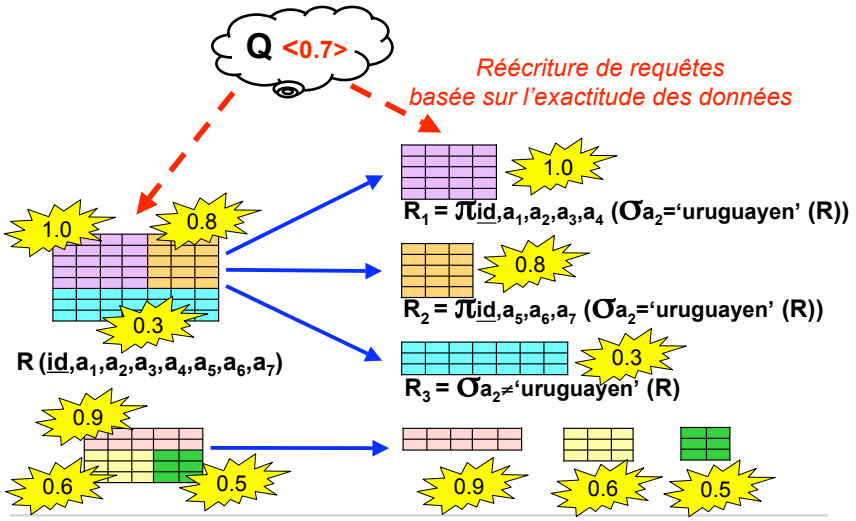
Dimension 3: Types de données

- **Catégories :**
  - Numérique
  - Date
  - Énumération
  - Chaînes de caractères
- **Fournissent des moyens pour contrôler les erreurs de syntaxe et de précision**
  - Différents types d'erreurs (ex. intervalles, formats invalides, expressions régulières)
  - Distances (ex. différence arithmétique)

Dimension 4: Techniques architecturales

- **Les processus du SID peuvent introduire ou corriger des erreurs (ex: nettoyage)**
  - Virtuelles (transformations arithmétiques et formatage)
  - Caching (pas de transformations)
  - Matérialisées (transformations complexes, nettoyage de données, normalisation de format)
- **Certaines techniques de mesure très coûteuses ont du sens seulement pour les données matérialisées**
  - ex. comparaison avec le monde réel

Estimation de l'exactitude



Support du raisonnement

- **Graphe de qualité similaire à celui utilisé pour la fraîcheur**
  - **Nœuds sources : fragments**
    - Exactitude **du fragment**
  - **Nœuds requêtes :**
    - Exactitude attendue **par les utilisateurs**
  - **Nœuds activités : opérateurs de la requête**
    - **Fonction de** combinaison **de plusieurs valeurs d'exactitude**
  - **Arcs :**
    - Exactitude **de chaque fragment résultat d'un opérateur**
    - Exactitude effective **produite par l'opérateur**

Estimation de l'exactitude d'une relation

à partir de l'exactitude des fragments

fragment$_1$: exactitude = 0.9
fragment$_2$: exactitude = 0.7 } ***exactitude global ?***
fragment$_3$: exactitude = 0.6

- **Fonction de combinaison qui dépend de l'application :**
  - **Ex: produit, moyenne…**
  - **Ex: moyenne pondérée par le nombre d'attributs / tuples / cellules**
- **Les paramètres peuvent s'obtenir de :**
  - **Intension du fragment : attributs, …**
  - **Statistiques: estimations de cardinalités, …**
  - **Extension du fragment : nombre de tuples, nombre de cellules**
  - **…**

Exemple (conception)

Requête :
Titre, ciné, horaire, score
des films à l'affiche à Versailles
avec score > 4★

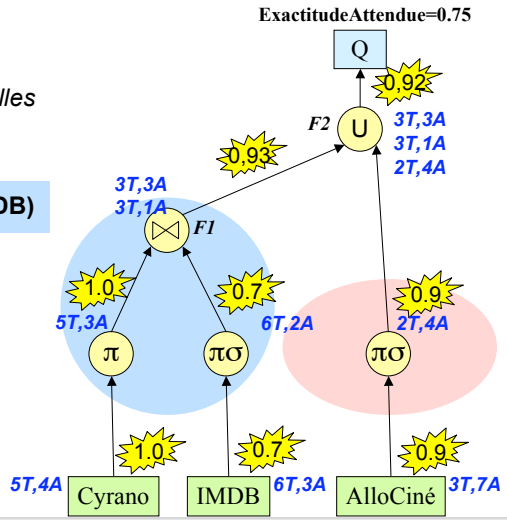
Réécriture 1
$R_1 = \pi(Cyrano) \bowtie \pi(\sigma(IMDB))$

Réécriture 2
$R_2 = \pi(\sigma(CinéCité))$

Réécriture 3
$Q = R_1 \cup R_2$

Graphe des réécritures

5T,4A
→ estimation de la sélectivité
→ Nombre attributs



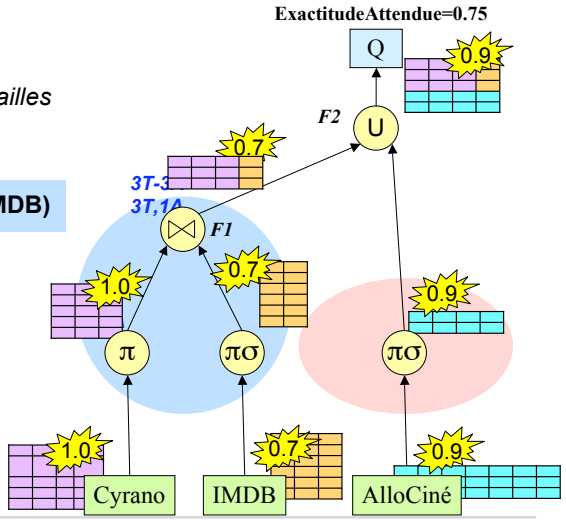
Exemple (interrogation)

Requête :
Titre, ciné, horaire, score des films à l'affiche à Versailles avec score > 4★

$R_1 = \pi(\text{Cyrano}) \bowtie \pi(\sigma(\text{IMDB}))$

$R_2 = \pi(\sigma(\text{CinéCité}))$

$Q = R_1 \cup R_2$



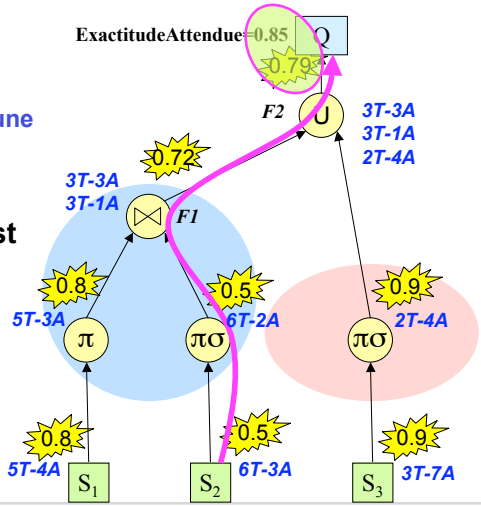
5T-4A
→ estimation de 5 tuples
→ 4 attributs

Algorithme de propagation (forward)

- **Permet :**
  - Fixer les bornes d'exactitude
  - Vérifier la conformité du graphe par rapport aux besoins des utilisateurs
- **Fonctionnement :**
  - Propagation de valeurs de exactitude le long du graphe
  - Calcul de l'exactitude des fragments résultants et de l'exactitude globale de chaque opération.
- **Utilisable en conception et exécution**

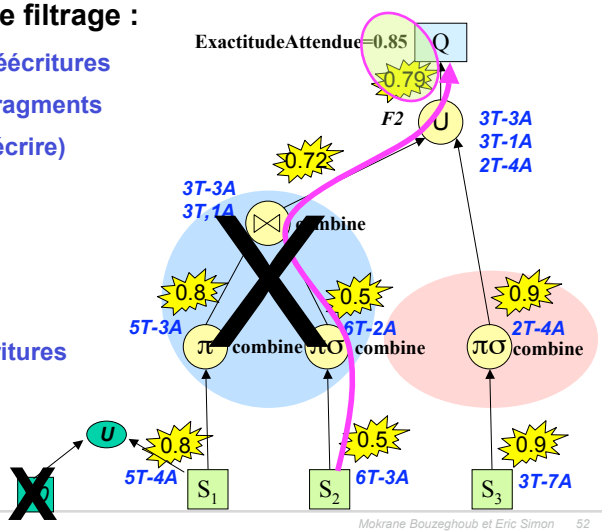
Analyse de l'exactitude

- **Un graphe est satisfaisant si**
  - pour chacun de ses utilisateurs, il produit une exactitude conforme à leurs exigences
- **Si une exigence n'est pas satisfaite**
  - il faut détecter les chemins critiques déterminant le sous graphe à restructurer

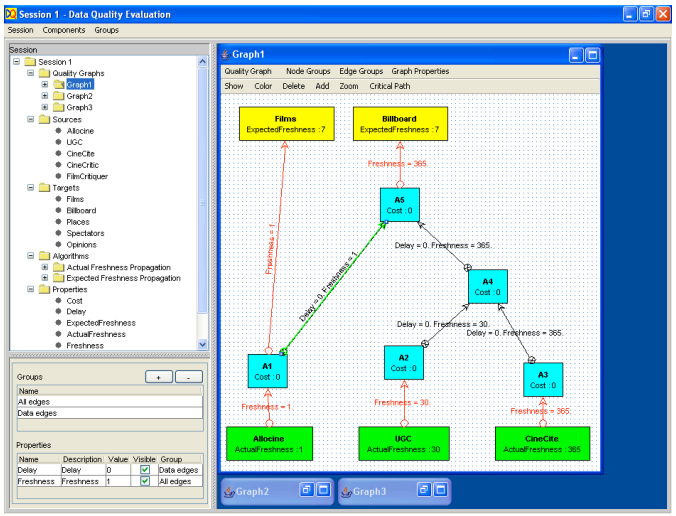


Opérations de restructuration

- **Opérations de filtrage :**
  - Filtrage de réécritures
  - Filtrage de fragments (avant de réécrire)
- **Alternative :**
  - Tri des réécritures



Outil réalisé



Expérimentations

- **Prototype DQE (Data Quality Evaluation)**
  - Graphes de qualité + exécution d'algorithmes d'évaluation
    - Pas limité à la fraîcheur et l'exactitude
  - Outil interactif de diagnostique
    - Fournit des facilités de visualisation (ex. chemin critique)
- **Performances et limitations :**
  - Supporte des configurations de grande taille avec des temps d'évaluation raisonnables, par exemple :
    - 300 graphes de 200 nœuds : 800 millisecondes
    - 700 graphes de 100 nœuds : 1 seconde
    - 100 graphes de 800 nœuds : 2 secondes
- **Utilisation illustrée par 3 applications**

Bilan sur la qualité

- **Une étude de la fraîcheur et de l'exactitude**
  - Définitions, métriques, dimensions qui les influencent
- **Un canevas pour l'évaluation de la qualité**
  - Graphes de qualité
  - Algorithmes d'évaluation de la fraîcheur et de l'exactitude
    - Paramétrables selon le scénario d'application
    - Propagation de valeurs de qualité le long du graphe de qualité
  - Une approche pour l'amélioration de la qualité (des actions d'amélioration)
  - Adaptable à d'autres facteurs de qualité
- **Un prototype de plateforme d'aide à l'évaluation et d'amélioration de la qualité**