

Bases de Données Avancées

S. Cerrito

Année 2007-2008, Evry

1 Médiation de données ayant des sources différentes

Architecture générale

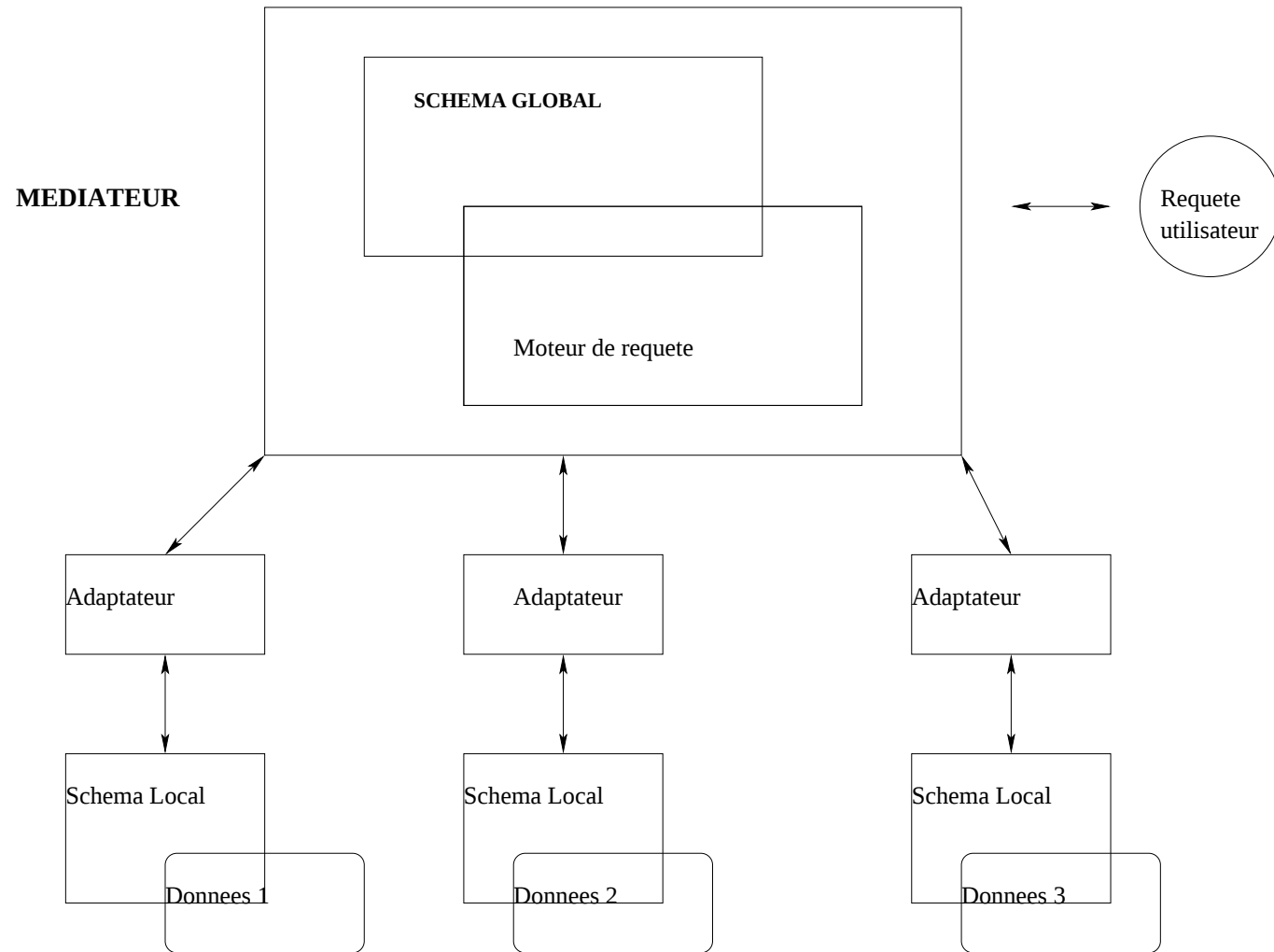


Schéma global : on décrit dans un formalisme spécifique le vocabulaire que l'utilisateur pourra utiliser pour exprimer sa requête. (Et il faut aussi un formalisme de description du contenu des sources). L'utilisateur exprime sa requête en termes du schéma global.

Il faut aussi une correspondance entre relations globales et relations locales. Deux approches :

1. Approche *Global As View* (GAV) : toute relation locale est une vue sur les locales.
2. Approche *Local As View* (LAV) : vice-versa, i.e. ce sont les relations locales qui sont des vues sur les globales.

Adaptateurs (Wrappers) : interfaces entre les sources et le médiateur.

Programmes qui transforment les requêtes (élaborées par le médiateur) en des requêtes exécutables sur les sources.

Exemple de Schéma Global/Relation Locales

Pour simplifier, prenons un formalisme logique pour exprimer les relations globales et les locales aussi.

– **Schéma Global, relations globales :**

$\text{emp}(E)$: E est un employé, $\text{tél}(E,N)$: N est le num. de tél. de E, $\text{bureau}(E,B)$: B est le num. de bureau de E, $\text{dir}(E,D)$: D est le directeur de E, $\text{dept}(E,De)$: De est le département de E.

$\text{même-bur}(E,E') \equiv \text{Emp}(E) \ \& \ \text{Emp}(E') \ \& \ \text{bureau}(E,B) \ \& \ \text{bureau}(E',B)$

$\text{emp-tél-bur}(E,N,B) \equiv \text{Emp}(E) \ \& \ \text{Emp}(E') \ \& \ \text{tél}(E,N) \ \& \ \text{bureau}(E,B)$

$\text{emp-dept-dir}(E,De,D) \equiv \text{Emp}(E) \ \& \ \text{Emp}(E') \ \& \ \text{dept}(E,De) \ \& \ \text{dir}(E,D)$

– **Sources, relations locales :**

Source 1 : $s1(E,N,D)$ fournit infos sur les employés, leur tél et leur directeur.

Source 2 : $s2(E,B,De)$ fournit infos sur les employés, leur bureau et leur département.

Source 3 : $s3(E,N)$ fournit infos sur les num de tél des employés du dépt. “jouets”.

Exemple de Schéma Global/Relation Locales : suite

Correspondance relations globales et locales : Approche GAV

On exprime les relations globales en fonction des relations locales, par ex. :

$\text{emp}(E) =_{DEF} s1(E,N,D) \text{ OR } s2(E,B,De) \text{ OR } s3(E,N)$

$\text{emp-tél-bur}(E,N,B) =_{DEF} (s1(E,N,D) \ \& \ s2(E,B,De)) \text{ OR } (s3(E,N) \ \& \ s2(E,B,De)) \text{ OR } (s3(E,N) \ \& \ s2(E,B,\text{”jouets”}))$

R : donner les employés du bureau N. 10 et leur tél. : $\text{emp-tél-bu}(E, \text{tél}, \text{”10”})$.

Pour calculer R, dans un formalisme logique, on peut faire du chaînage en arrière, à la Prolog, par ex.

Exemple de Schéma Global/Relation Locales : suite

Correspondance relations globales et locales : *Approche LAV*

On exprime les relations locales en fonction des relations globales :

$s1(E,N,D) =_{DEF} emp(E) \ \& \ tél(E,N) \ \& \ dir(E,D)$

$s2(E,B,De) =_{DEF} emp(E) \ \& \ bureau(E,B) \ \& \ dept(E,De)$

$s3(E,N) =_{DEF} emp(E) \ \& \ tél(E,B) \ \& \ dept(E,"jouets")$

Requête : donner les employés du bureau N. 10 et leur tél., *i.e.* emp-tél-bur(E,N,10)
(langage du schéma global).

On doit calculer les valeur E, N des relations locales :

a) $s2(E,10,De) \ \& \ s1(E,N,D)$ sur les sources 2 et 1

b) $s2(E,10,De) \ \& \ s3(E,N)$ sur les sources 2 et 3

(a) et (b) se *réécrivant* en :

a') $\text{emp}(E) \ \& \ \text{bureau}(E,B) \ \& \ \text{dept}(E,De) \ \& \ \text{emp}(E) \ \& \ \text{tél}(E,N) \ \& \ \text{dir}(E,D)$

b') $\text{emp}(E) \ \& \ \text{bureau}(E,B) \ \& \ \text{dept}(E,De) \ \& \ \text{emp}(E) \ \& \ \text{tél}(E,B) \ \& \ \text{dept}(E, \text{''jouets''})$.

Peu importe si l'approche est GAV ou LAV,
l'hétérogénéité des sources donne lieu à des difficultés :

- 1 Modèles de données $\neq$: Relationnel, Objet, Semi-structuré, même HTML ou des fichiers plats.
- 2 Structuration : Même dans le cadre du même type de modèle, possibilité de 2 schémas différents. Par ex. une source relationnelle stocke les infos sur les clients dans 1 seule table, une autre en plusieurs tables.
Et puis, une source peut ne pas avoir de schéma (HTML, par ex.)
- 3 Problèmes de sémantique.
 - i) Par ex. d'un attribut : pour la source A altitude est mesuré en mètres, pour B en *miles*. Pire : pour A le salaire est net, pour B est brut.
 - ii) Possibilité d'informations contradictoires.

1.1 Que peut-il apporter XML s'il est choisi comme langage du schéma global ?

Tout d'abord : Expressivité Riche.

Par ex., dans un approche LAV, si on veut passer du relationnel au semi-structuré, les schémas relationnels :

Cours(NumCours,NomProf,Salle)

Inscriptions(NomEtudiant,NumCours)

ont une écriture simple en un DTD, par ex. :

⟨ !ELEMENT Cours (NomProf,Salle) ⟩

⟨ !ATTLIST Cours NoCours ID #REQUIRED ⟩ ⟨ !ELEMENT Inscriptions
(NomEtudiant) ⟩

⟨ !ATTLIST Inscriptions NoCours IDREF #REQUIRED ⟩

Reprise des difficultés, une par une

1 Modèles de données $\neq$

XML fournit une syntaxe neutre pour décrire des données structurés, conceptuellement, par des graphes, permettant l'imbrication et les références.

Plusieurs structures de données peuvent être ainsi représentées.

Ce point est lié à l'expressivité de XML. *Ici, la contribution de XML est importante.*

Reprise des difficultés, une par une

2 Structuration : Même dans le cadre du même type de modèle, possibilité de 2 schémas différents. Puis, une source peut ne pas avoir de schéma du tout.

A nouveau, on peut exprimer plusieurs structurations de données sous forme de graphes, et les DTD ou les XML-schémas sont très expressifs.

Ceci dit, la nécessité de créer des mappings entre les plusieurs formes et cas de schémas, éventuellement à l'aide de Xquery, reste.

Reprise des difficultés, une par une

3 Problèmes de sémantique. Attribut ayant domaines différents, informations contradictoires.

XML fournit un mécanisme permettant d'attacher des *metadonnées*, décrivant les données, ce qui aide à déterminer quelle transformations doivent être faites pour intégrer plusieurs sources. Par ex., ici on indique ce que l'altitude est

< Altitude > 500

< UniteMes > miles </UniteMes >

</Altitude >

Mais le problème reste complexe.

Réalités Industrielles (liste qui ne se veut pas exhaustive)

Enosys Software.

An extrait de <http://xml.coverpages.org/Enosys-XQuery.html> :

an XML-based information integration software provider, ..., the latest evidence of the technology industry's adoption of XQuery as the new standard for enterprise information integration (EII).

"By providing a standard method for query and retrieval of XML data, XQuery is playing the role that SQL did for relational databases in the late eighties," said Dave Banks, CEO of Enosys.

Enosys technology has deep roots in research from Stanford and University of California, where query-able real-time XML views of multiple data sources were first proposed and researched by the founders and key members of the Enosys team. Enosys was the first to harness the power of XQuery for enterprise information integration. With an XQuery-based EII offering, users spend significantly less money and less time to achieve real-time integration of data from disparate sources.

Actuate (achat de Nimble, qui avait développé de la technologie pour l'intégration des données :

[http ://www.actuate.com/products/data-integration/index.asp](http://www.actuate.com/products/data-integration/index.asp)

Technologie utilisant, entre outre, Xpath.

Liquid Data

Derivé de Enosys

Fournit une vue fondée sur XML de toutes les données d'une entreprise. Les sources apparaissent soit comme des documents virtuels XML soit comme une collection de fonctions qui prennent des paramètres XML et produisent des résultats XML (*i.e.* des requêtes XQUERY).