Systèmes d'intégration de ressources hétérogènes et distribuées

0. Un exemple d'un système d'intégration de données (Voir fichier Complément SI)

1. Introduction & Contexte : Intégration virtuelle de données

2. Architecture Médiateur-Adaptateur

3. Problèmes posés par la conception d'un système d'intégration

4. Principe d'évaluation et d'optimisation d'une requête

1. Introduction : Contexte, Motivation et Objectif

Ressources informatiques sont connectées à l'Internet
==> Accessibles à travers des protocoles de comm. TCP/IP, HTTP
Ressources : {Données, Programmes applicatifs } que le propriétaire souhaite partager avec d'autres utilisateurs.

• Caractéristiques des ressources :

- **Autonomes :**
  Développées et maintenues en isolation, et indépendamment des applications pouvant les utiliser
  ==> Capacités de traitement d'une ressource sont décidées uniquement par leurs propriétaires
- **Distribuées :**
  Ressources pertinentes d'un même sujet se trouvent souvent sur des sites distants
- **Hétérogènes :**
  Structures de données (relationnel, objet, XML, HTML, ...)
  Capacités de traitement de requêtes.

• Contexte d'intégration virtuelle à grande échelle

- **Approche alternative : Entrepôt de Données ED**
  consiste à répliquer toutes les sources dans un seul site, sur lequel sont évaluées toutes les requêtes.
- **Avantages de l'approche ED : Meilleure(s)**
  - **Performances : absence de comm. Inter-sites**
  - **Fiabilité : sources répliquées sont sous le contrôle d'un site**
- **Inconvénients de l'approche ED :**
  - **Impossible de dupliquer toutes les ressources d'intérêt sur un site, pour des raisons de confidentialité, de droits intellectuels ou sécurité**
  - **Passage à l'échelle est difficile lorsqu'on utilise un grand nombre de ressources**
  - **Programmes dépendent souvent de l'env. particulier du site de leur propriétaire et donc ne peuvent pas être utilisées ailleurs.**

➡Les ressources seront exploitées à partir de leurs sites d'origines

2. Architecture Médiateur-Adaptateur M-A (Rappel)

- **Rôle d'un Médiateur :**
  - **interagir avec les ressources distribuées et avec les appl. qui demandent d'accéder à l'info intégrée**
  - **fournit un point d'accès unique et uniforme aux ressources**
  - **fournit une interface d'interrogation que les appli. peuvent utiliser**
  - **la présence de couche intermédiaire fournit par le médiateur permet aux appli. utilisant l'info. intégrée d'être indépendantes par rapport à l'autonomie et à l'évolution des ressources**
- **Rôle d'un Adaptateur**
  - **agit comme un intermédiaire entre des ressources et des médiateurs**
  - **cache aux médiateurs le format interne et la mise en œuvre de leurs ressources**
  - **fournit aux médiateurs une description de ressources dans le format du médiateur**
  - **accepte de traiter des sous-requêtes sur ces ressources (à la demande du médiateur)**
  - **le traitement de requêtes dans un système M-A est divisé entre les adaptateurs et les médiateurs**
  - **connecté directement aux diverses ressources via une interface adaptée à chaque ressource**
  - **exporte :**
    - **des méta-données concernant les ressources, par ex. . description du contenu de chaque source de données**
    - **une description des capacités de traitement de requêtes, et**
    - **des statistiques sur des données ou des paramètres de coût associés à leurs capacités de traitement**
- **Schéma Local et Schéma Global**
  - **l'interface uniforme entre les adaptateurs et les médiateurs est basée sur un modèle de données commun dans lequel toutes les ressources sont décrites**
  - **chaque adaptateur présente au médiateur un schéma local**

- **le médiateur fournit aux applications utilisant ses services un schéma global**

3. Problèmes posés lors de la conception d'un système d'intégration basé sur une architecture M-Adaptateurs

- **Un modèle commun doit être établi pour les ressources hétérogènes :**
  - ***riche*** **pour modéliser les aspects intéressantes des ressources**
  - ***évolué*** **pour permettre l'optimisation de requêtes**
- **Un adaptateur est nécessaire pour chaque type de ressource, afin de la présenter sous le modèle commun choisi**
  - **la mise en œuvre des adaptateurs est une tâche difficile**
  - **sa difficulté est un obstacle majeure pour le développement d'application d'intégration de ressources à grande échelle**
- **Conception d'un schéma global :**
  - **par un processus d'intégration de schéma**
    - **entraîne la résolution des différences de domaines**
    - **l'identification des entités communes dans des sources de données différentes (utilisation des clés communes)**
- **Méthodologies de traitement de requêtes**
  1. **la requête, exprimée/schéma global, doit être re-formulée en termes de schémas locaux (schémas des adaptateurs)**
  2. **le médiateur décompose la requête ainsi re-formulée dans**
     a. **plusieurs sous-requêtes à envoyer aux adaptateurs et**
     b. **quelques opérations que le médiateur va exécuter sur les résultats des sous-requêtes des adaptateurs**
  3. **Optimisation de req. distribuées (décomposition n'est pas unique)**
  4. **les sous-requêtes sont envoyées aux adaptateurs pour les exécuter**
  5. **les résultats des sous-requêtes des adaptateurs sont fournis au médiateur, qui exécute certaines opérations nécessaires et retourne le résultat complet.**
- **Remarques : Dans le cas d'un SI à plusieurs médiateurs**
  - **les étapes ci-dessus sont les mêmes**
  - **si plusieurs médiateurs participent à l'exécution de la requête, ils pourraient se voir attribuer des sous–requêtes à traiter (Comme les adaptateurs)**
  - **un seul médiateur coordonne l'exécution d'une requête.**

4. Evaluation et optimisation de requêtes dans un système d'intégration

4.1. Principe d'évaluation (voir pages séparées)

4.2. Optimisation : = <Espace de recherche, Stratégies de recherche, Modèle de coûts>

• Espace de recherche

- **Nature de l'espace de recherche : arbres linéaires et ramifiés**
- **Type de requête**
- **Localisation des ressources utilisées par la requête**
- **Capacités limitées de traitement de requêtes des adaptateurs**
  - **Ex. Adaptateur du site Web de type Pages Jaunes PJ**
    Exécute une classe de requêtes assez <u>restreintes</u>
- **Classification des capacités de traitement de requêtes.**
  - **Capacités de traitement de *<u>requêtes positives</u>***
    Ex . SGBDR classique
  - **Capacités de traitement de *requêtes. négatives***
    - **l'adaptateur peut imposer des restrictions lors de l'accès aux ressources qu'il gère**
    - **Ex : le service Web PJ, on peut exiger un nom et un code postal afin d'obtenir le Numéro de Tél.**
    - **Certaines requêtes pourraient être *<u>rejetées</u>* par le système : car l'adaptateur ne l'accepte pas et le médiateur n'a pas accès aux sources de données**

• Stratégies de recherche

La présence des capacités limitées des sources, implique une modification de la stratégie de recherche de l'optimiseur afin d'explorer uniquement l'espace des plans faisables.

➡Différentes sources de difficultés dans l'exécution de requêtes : < Distribution, Hétérogénéité, Contrôle>

- **Distribution**

  ==> Pb. de comm. : Min (Coûts de Comm.)
  - **Application des techniques d'exécution validées dans les SGBDR : Jointure distribuée, Jointure à base de semi-jointure, ...**
- **Hétérogénéité des capacités de traitement de requêtes**
  - **Adaptation aux capacités de l'adaptateur (traduction de la sous requête dans le langage de l'adaptateur)**
  - **Intégration des opérateurs adaptatifs en raison de l'autonomie des sites d'exécution (ré-optimisation, construction d'un nouveau plan, ...)**
- **Manque de contrôle sur les sites distants :**
  - **Difficulté de transférer des prog. d'un site Si vers un site Sj ➡Obligé(s) d'envoyer les arguments du prog sur le site où le prog. peut être exécuté.**
  -

- **Modèle de coûts**
  - **Difficulté d'obtenir des statistiques sur les données :**
    - **Nature de ressources utilisées : < CPU, RAM, BP (R & D), SD>**
    - **Restrictions sur l'accès à ces ressources**
    - **Changements fréquents dans les contenus des SD**
    - **Instabilité des ressources de calcul RC ➔ impact sur les valeurs des paramètres modélisant les RC**
  - **Formules de Coûts FC :**
    - **Chaque écrivain d'adaptateur peut fournir des FC pour son propre adaptateur ou**
    - **Une FC générique peut être utilisée pour tous les adaptateurs et calibrée par l'exécution d'un ensemble de requêtes de calibrage.**

4.3. Opérateurs relationnels à accès restreint

Soit une source de données modélisée sous forme d'une BD relationnelle composée de 2 relations : R (nom^{f}, adresse^{f}, profession^{f}) et S (nom^{b}, num_tel^{f}) avec :
<attribut>^{f} signifié que la valeur de l'attribut est libre (free) pouvant être donnée en entrée (e.g. dans la clause Where du langage SQL) ou affichée /sélectionnée (en sortie) et
<attribut>^{b} signifié que les valeur de l'attribut doivent être instanciées (bound) et l'attribut (instancié) ne peut être utilisé qu'en entrée comme filtre (e.g. dans la clause Where du langage SQL).

Remarque : Dans SGBD relationnel classique (requêtes positives) tous les attributs sont tous implicitement de type « libre », comme la relation R ci-dessus.

1. Opérateur de Sélection

La requête "Selection num-tel From S,where nom= »Durand »" est valide (ou faisable). Cependant, la requête "Selection nom from S,where num-tel= 11111 » n'est pas valide (ou rejetée), car, le type de "nom" est « bound » (instancié).

2. Opérateur de jointure, appelée Jointure dépendante

Donner/afficher le num-tel et l'adresse des abonnés ?

Dans ce cas, comme les types d'attributs (valeurs) de jointure entre R et S sont respectivement « libre » et « instancié » on doit alors faire R Jointure S : on part de R pour explorer les tuples de S. Mais, on ne peut pas faire S Jointure R : on part de S pour explorer les tuples de R, car les valeurs de nom^{b} doivent être, d'abord, instanciées. En conséquence, la jointure de deux relations dans un système d'intégration pourrait ne pas être commutative. Cela dépendra du type d'attribut « libre » ou « instancié ».