Master d'Informatique

BDR - 4I803 – COURS 5

Conception de BD réparties

Fragmentation et réplication

2016

1

Bases de Données Réparties

- Définition
- Conception
- Décomposition
- Fragmentation horizontale et verticale
- Outils d'interface SGBD
  - extracteurs, passerelles
- Réplication
- SGBD répartis hétérogènes

2

BD réparties (1)

- Principe :
  - Un site héberge une BD : accès local rapide, interne au site.
  - Accès global possible à des BD situées sur des sites externes
- Plusieurs niveaux d'intégration (ordre croissant d'intégration) :
  1. Client/serveur : BD centralisée, seuls certains traitements (interface, p.ex.) sont locaux.
  2. Accès distant (Remote Data Access)
  3. Vues réparties : extension du mécanisme de vues pour définir des vues sur plusieurs sites.
  4. médiateurs
  5. BD réparties/fédérées

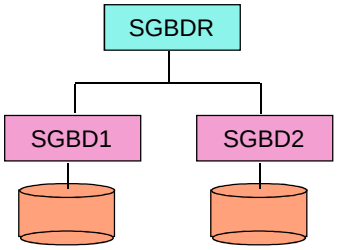
3

BD Réparties (2)

- BD réparties :
  - Plusieurs bases sur plusieurs sites, mais une seule BD « logique ».
  - Fédérée : intègre des bases et des schémas existants
  - Répartie « pur » : conçue répartie. Pas d'accès locaux
- Les ordinateurs (appelés **sites**) communiquent via le réseau et sont faiblement couplés
  - pas de partage de MC, disque, au contraire de *BD parallèles*
- Chaque site
  - contient des données de la base,
  - peut **exécuter** des transactions/requêtes locales et
  - participer à l'exécution de transactions/requêtes globales

4

SGBD réparti



Rend la répartition (ou distribution) *transparente*

– dictionnaire des données (catalogue, métabase) réparties
– traitement des requêtes réparties
– gestion de transactions réparties
– gestion de la cohérence et de la sécurité

5

Paramètres à considérer

- Coût et temps de communication entre deux sites
  - Accès réseau (longue distance, WAN, MAN) beaucoup plus coûteux que accès disque
- Fiabilité
  - fréquence des pannes des sites, du réseau (cf. P2P)
- Accessibilité aux données
  - accès aux données en cas de panne des sites, du réseau.
  - accès aux sites les moins encombrés, les plus puissants

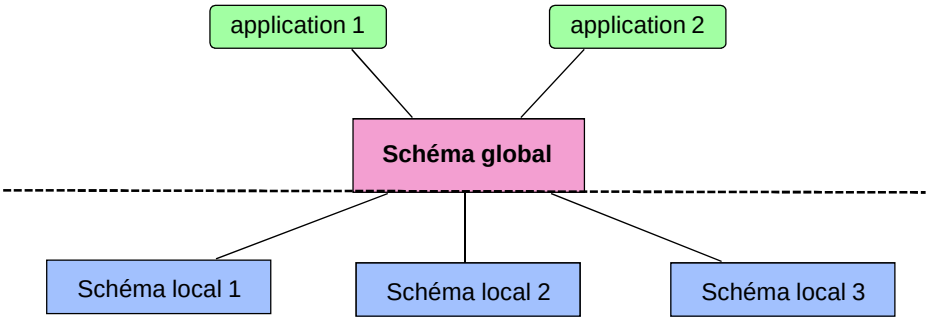
6

Evaluation de l'approche BDR

- avantages
  - extensibilité
  - partage des données hétérogènes et réparties
  - performances avec le parallélisme
  - Disponibilité et localité avec la réplication
- inconvénients
  - administration complexe
  - complexité de mise en œuvre
  - distribution du contrôle
  - surcharge (l'échange de messages augmente le temps de calcul)

7

Architecture de schémas



- indépendance applications / bases locales
- schéma global lourd à gérer

8

Schéma global

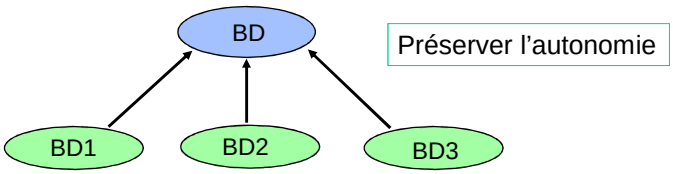
- Schéma conceptuel global
  - description globale et unifiée de toutes les données de la BDR
    - Nom des relations avec leurs attributs
  - Fournir l'indépendance à la répartition
- Schéma de placement
  - règles de correspondance avec les données locales
    - Vues globales définies sur les relations locales (*cf.* global as view)
  - Fournit l'indépendance à la localisation, la fragmentation et la duplication

- Le schéma global fait partie du dictionnaire de la BDR et peut être conçu comme une BDR (dupliqué ou fragmenté)

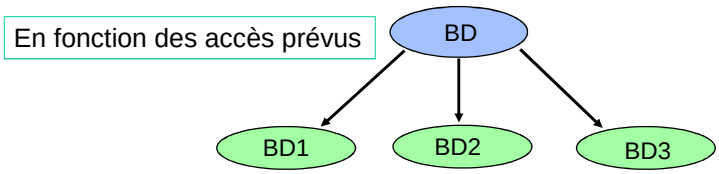
9

Migration vers une BDR : 2 approches

Intégration logique des BD locales existantes (fédérée, médiateur)

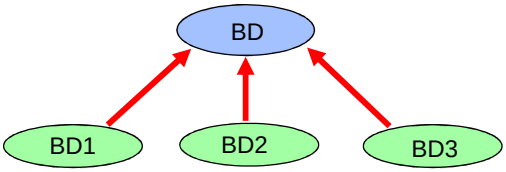


Décomposition en BD locales : répartie « pur »



10

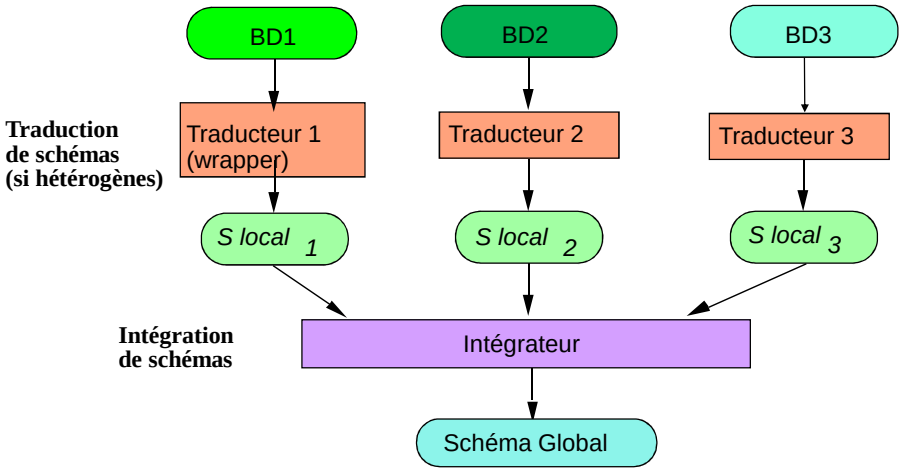
Intégration de BD existantes



11

Conception d'une BDR par intégration

Approche médiateur



12

Intégration de schémas

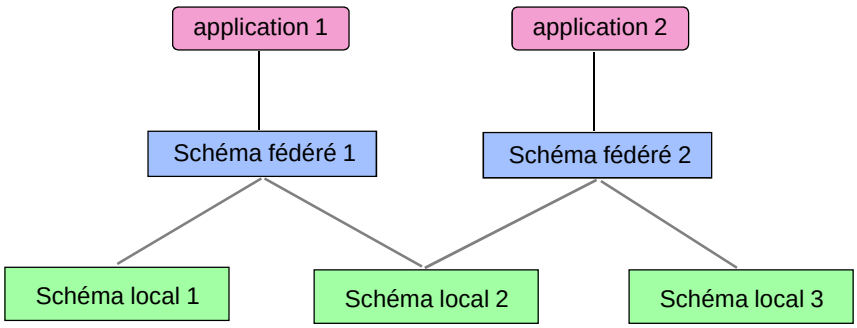
1. pré-intégration
   - Les schémas sont transformés pour les rendre plus homogènes
   - *identification* des éléments reliés (e.g. domaines équivalents) et établissement des règles de conversion (e.g. 1 inch = 2,54 cm)
   - Pbs : hétérogénéité des modèles de données, des puissances d'expression, des modélisations

2. comparaison
   - *identification* des conflits de noms (synonymes et homonymes) et des conflits structurels (types, clés, dépendances)

3. conformance
   - *résolution des* conflits de noms (renommage) et des conflits structurels (changements de clés, tables d'équivalence)
   - Définition de règles de traduction entre le schéma intégré et les schémas initiaux.

13

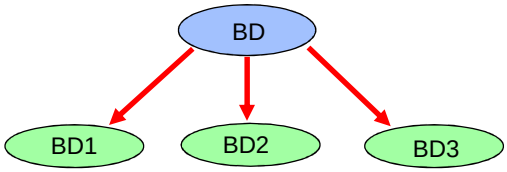
Architecture fédérée



moyen contrôlé de migration

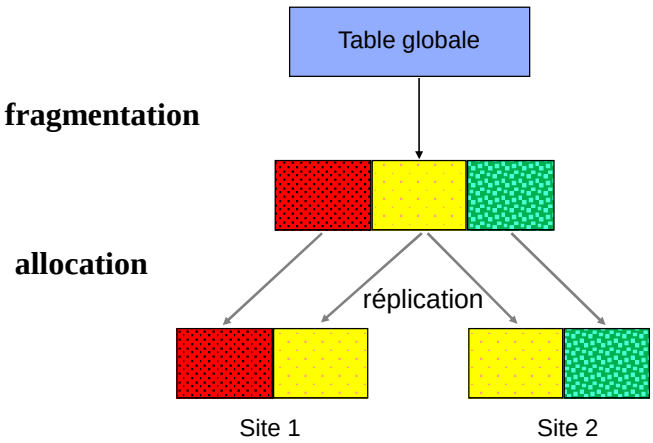
14

Décomposition



15

Conception par décomposition



16

Objectifs de la décomposition

Fragmentation

- Trois types : horizontale, horizontale dérivée, verticale
  - Possibilité de composer plusieurs fragmentations: mixte
- Performances en favorisant les accès (et traitements) locaux
- Equilibrer la charge de travail entre les sites (parallélisme)
- Contrôle de concurrence plus simple pour les accès à un seul fragment

Trop fragmenter : BD éclatée, nombreuses jointures réparties

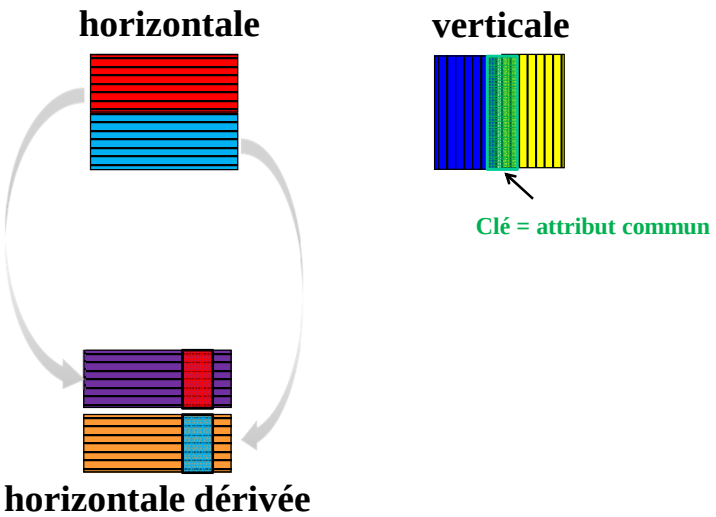
Duplication (ou réplication)

- favoriser les accès locaux
- augmenter la disponibilité des données

Trop répliquer : surcoût de maintenir cohérence des répliques

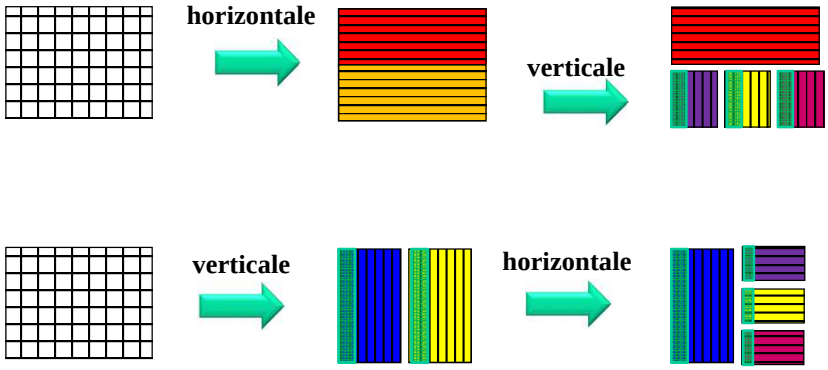
17

Types de Fragmentation



18

Fragmentation Mixte



19

Fragmentation correcte

Complète

- chaque élément de R doit se trouver dans un fragment

Reconstructible

- on doit pouvoir recomposer R à partir de ses fragments (ressemble à décomposition de schéma vue en Li341 pour fragmentation verticale)

[Disjointe] /*si on veut éviter réplication pour cohérence */

- chaque élément de R ne doit pas être dupliqué (sauf clé en cas de fragmentation verticale)

20

Fragmentation Horizontale

Fragments définis par **sélection**

$Client_1 = \sigma_{ville = 'Paris'} Client$
$Client_2 = \sigma_{ville \neq 'Paris'} Client$

Inférence : correcte

Reconstruction par union
$Client = Client_1 \cup Client_2$

En SQL :
```
create view Client as
   select * from Client1
   union
   select * from Client2
```

Client

nclient	nom	ville
C 1	Dupont	Paris
C 2	Martin	Lyon
C 3	Martin	Paris
C 4	Smith	Lille

Client1

nclient	nom	ville
C 1	Dupont	Paris
C 3	Martin	Paris

Client2

nclient	nom	ville
C 2	Martin	Lyon
C 4	Smith	Lille

21

Fragmentation Horizontale par sélection

- Fragmentation de R selon *n* prédicats
  - les prédicats $\{p_1, \ldots, p_n\}$ ex: {a<10, a > 5, b='x', b='y'}
- L'ensemble *M* des prédicats de fragmentation est :
  - $M = \{ m \mid m = \bigwedge_{1 \leq k \leq n} p_k^* \}$ avec $p_k^* \in \{p_k, \neg p_k\}$
  - Eliminer les *m* de sélectivité nulle ex: $a>10 \wedge a < 5$
  - Simplifier: $a<10 \wedge a \leq 5 \wedge b='x' \wedge b \neq 'y'$ devient $a \leq 5 \wedge b='x'$
- Construire les fragments $\{R_1, \ldots, R_k\}$
  - Pour chaque m_i , $R_i = \sigma m_i (R)$
- Minimalité
  - Ne pas avoir 2 fragments toujours lus ensemble
- Choisir les p_i des requêtes les plus fréquentes

- Ref biblio récente (2015) sur la fragmentation et le choix des fragments optimaux. Regroupement hiérarchique d'ensembles de nuplets issus de requêtes fréquentes.
  - Waterloo Univ: G. Aluç, M. T. Özsu, K. Daudjee and O. Hartig. "Executing Queries over Schemaless RDF Databases", ICDE 2015 (Int'l Conf. on Data Engineering)

Fragmentation Horizontale **Dérivée**

Fragments définis par semi jointure
Cde1 = select Cde.*
from Cde, Client 1
where Cde.nclient = Client1.nclient

$Cde_i = Cde \ltimes Client_i$ pour ***i*** dans{1; 2}

Reconstruction par **union**

$Cde = Cde_1 \cup Cde_2 = \cup_i Cde_i$

Cde

ncde	nclient	produit	qté
D 1 D 2 D 3 D 4	C 1 C 1 C 2 C 4	P 1 P 2 P 3 P 4	10 20 5 10

Cde1

ncde	nclient	produit	qté
D 1 D 2	C 1 C 1	P 1 P 2	10 20

Cde2

ncde	nclient	produit	qté
D 3 D 4	C 2 C 4	P 3 P 4	5 10

23

Propriétés de la fragmentation horizontale dérivée

R: fragmentation horizontale → fragments R_i
S: fragmentation horizontale dérivée → fragments $S_i = S \ltimes_A R_i$

- Complète
  - Chaque tuple de S doit joindre avec au moins un tuple de R
    - $\forall s \in S, \exists t \in S_i, s = t$
- Disjointe
  - $\forall i,j$ tq $i \neq j$, $S_i \cap S_j = (S \ltimes R_i) \cap (S \ltimes R_j) = S \ltimes (R_i \cap R_j) = \varnothing$
    - Rappel: $R1 \cap R2 \Leftrightarrow R1 \ltimes R2$
- Reconstructible
  - $\cup_i S_i = (S \ltimes R_1) \cup (S \ltimes R_2) \cup \ldots \cup (S \ltimes R_n) = S \ltimes (\cup_i R_i) = S$
- $\Rightarrow$ contrainte d'intégrité référentielle
  - A = clé de R
  - S.A référence R.A
    - $\forall s \in S, \exists r \in R, s.A = r.A$

Fragmentation **Verticale**

Fragments définis par **projection**

Cde1 = $\pi_{ncde,\ nclient}$Cde

Cde2 = $\pi_{ncde,\ produit,\ qté}$ Cde

Cde

ncde	nclient	produit	qté
D 1	C 1	P 1	10
D 2	C 1	P 2	20
D 3	C 2	P 3	5
D 4	C 4	P 4	10

Reconstruction par jointure

Cde = Cde1 ⋈ Cde2

En SQL :

create view Cde as

select * from Cde_1 ,Cde_2

where Cde_1.ncde = Cde_2.ncde

Cde1

ncde	nclient
D 1	C 1
D 2	C 1
D 3	C 2
D 4	C 4

Cde2

ncde	produit	qté
D 1	P 1	10
D 2	P 2	20
D 3	P 3	5
D 4	P 4	10

25

Fragmentation Verticale

Comment définir une fragmentation verticale ?

- Affinité des attributs : mesure la proximité sémantique des attributs (combien « ils vont ensembles »)
  - Soit par connaissance de l'application,
  - soit par analyse des requêtes (on mesure combien de fois deux attributs donnés ont été interrogé ensembles)
  - Résultat sous forme de matrice d'affinité
- 2 approches : regroupement, partitionnement (grouping – splitting)
  - Idem que pour optimisation de schéma relationnel SPI, SPD
- Algorithme de regroupement des attributs bien adapté
  - BEA : bond energy algorithm (Mc Cormick et al. 72) : $O(n^2)$
  - insensible à l'ordre de départ des attributs
  - Part des attributs individuels et effectue des regroupements de groupes
- Algorithme de partitionnement :
  - Part d'une relation et observe le bénéfice qu'on peut tirer à partitionner

26

Matrice d'affinité des attributs

Matrice A

aij = affinité de Ai avec Aj

ex: nb de requêtes qui accèdent Ai et Aj

	A1	A2	A3	A4
A1	45	0	45	0
A2		80	5	75
A3			53	3
A4				78

Matrice d'affinité regroupement des attributs

Matrice A

	A1	A3	A2	A4
A1	45	45	0	0
A3		53	5	3
A2			80	75
A4				78

Allocation des Fragments aux Sites

Non-répliquée

- partitionnée : chaque fragment réside sur un seul site

Dupliquée

- chaque fragment sur un ou plusieurs sites
- maintien de la cohérence des copies multiples : coûteux
- **(le fameux) Compromis Lecture/écriture**:

  + le ratio Lectures/màj est > 1, + la duplication est avantageuse

29

Allocation de Fragments

Problème: Soit

F un ensemble de fragments

S un ensemble de sites

Q un ensemble d'applications et leurs caractéristiques

trouver la distribution "optimale" de *F* sur *S*

Optimum

- coût minimal de communication, stockage et traitement
- Performance = temps de réponse ou débit

Solution

- allouer une copie de fragment là où le bénéfice est supérieur au coût

30

Exemple d'Allocation de Fragments

Client1

nclient	nom	ville
C 1 C 3	Dupont Martin	Paris Paris

Client2

nclient	nom	ville
C 2 C 4	Martin Smith	Lyon Lille

Cde1 = Cde ⋉ Client1

ncde	client	produit	qté
D 1 D 2	C 1 C 1	P 1 P 2	10 20

Site 1

Cde2

ncde	client	produit	qté
D 3 D 4	C 2 C 4	P 3 P 4	5 10

Site 2

31

Exemple

Trois universités parisiennes (Jussieu, Sorbonne, Dauphine) ont décidé de mutualiser leurs équipements sportifs (locaux) et les entraîneurs. La gestion commune est effectuée par une base de données répartie, dont le schéma **global** est le suivant :

PROF (Idprof, nom, adresse, tél, affectation, salaire)
ETUDIANT (Idetu, nom, adresse, assurance, police, université, équipe)
LOCAUX (Idlocal, adresse, université)
EQUIPE (équipe, sport, niveau)
HORAIRE (Idlocal, équipe, jour, heure_début, heure_fin, prof)

- Chaque université rémunère ses profs en envoyant un chèque à leur adresse, mais aussi elle doit pouvoir contacter tout prof qui utilise ses locaux.
- Chaque équipe correspond à un sport. La plupart des équipes ont droit à un (seul) créneau (jour, heure) dans un des locaux communs pour leur entrainement. Cependant, pour le sport «cyclisme », il n'y pas besoin de locaux.
- Chaque université gère évidemment ses propres étudiants, ainsi que ses locaux et les créneaux correspondants.
- Les équipes ne sont associées à aucune université en particulier. Cependant, pour des questions d'assurance, chaque université doit aussi gérer les étudiants qui utilisent ses locaux. Pour le cyclisme , c'est Dauphine qui en a la charge.
- Les relations globales sont fragmentées et réparties sur les différents sites.

32

Données réparties avec Oracle : Database link

Lien à une table dans une BD distante spécifié par :

- nom de lien
- nom de l'utilisateur et password
- Infos de connexion (protocole client-serveur d'oracle)

Exemple de syntaxe :

```
create database link Site2
connect to E1234  identified by "E1234" using 'ora10';
```

Create synonym Emp2 for Emp@Site2;

Ou

Create view Emp2 as select * from Emp@Site2;

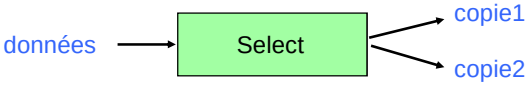
Vue répartie

33

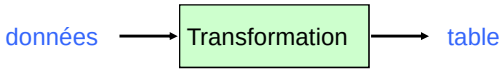
Outils d'interface SGBD

Passerelle

Réplicateur



Extracteur



34

Passerelles

- Fonctions
  - définition des procédures de transformation (dictionnaire) et exécution dans l'environnement cible
  - conversion de formats et de valeurs
  - filtrage et fusion de fichiers ou de tables
  - données calculées et résumés

35

La réplication

Plan:

Objectifs
Fonctions
Modèles d'appartenance
  - fixe, dynamique ou partagé

Détection des modifications

36

Objectifs de la réplication

+ Accès simplifié, plus performant pour les lectures
+ Résistance aux pannes
+ Parallélisme accru
+ Evite des transferts

- Overhead en mise à jour
- Cohérence des données
- Toujours bien si on privilégie les lectures et/ou si peu de conflit entre màj

37

Objectifs de la réplication

Problème : comment partager des données entre p sites ?

- Solution 1 : sans duplication
  - stockage sur un site et accès réseau depuis les autres sites
  - problèmes de performances et de disponibilité
- Solution 2 : duplication synchrone
  - propagation des mises à jour d'un site vers les autres par une transaction multi-site avec validation 2PC ou communication de groupe
  - problèmes liés au 2PC : bloquant et cher. Communication de groupe seulement pour LAN
- Solution 3 : réplication asynchrone
  - Non bloquant mais uniquement cohérence à terme
  - Contrôle de la fraîcheur, traitement spécifique des requêtes read-only

Aussi mono-maître,multi-maîtres

38

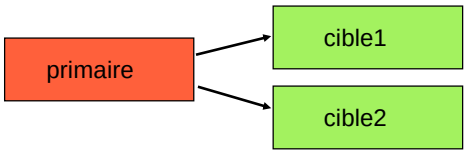
Fonctions d'un réplicateur

- Définition des objets répliqués
  - table cible = sous-ensemble horizontal et/ou vertical d'une ou plusieurs tables
- Définition de la fréquence de rafraîchissement
  - immédiat (après mise à jour des tables primaires)
  - à intervalles réguliers (heure, jour, etc.)
  - à partir d'un événement produit par l'application
- Rafraîchissement
  - complet ou partiel (propagation des modifications)
  - push (primaire -> cibles) ou pull (cible -> primaire)

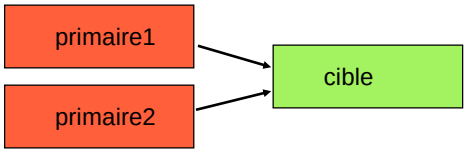
Réplication Monomaître

- Seul le site **primaire** peut recevoir les transactions des applications
  - insert, update, delete
- les sites **cibles** ne reçoivent que des requêtes en lecture seule
  - select

Diffusion



Consolidation

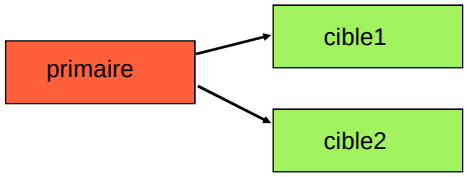


40

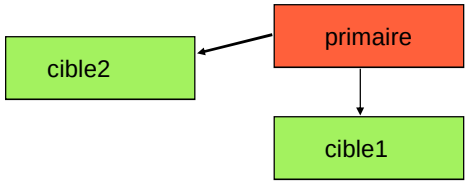
Monomaitre
avec appartenance dynamique

Le site **primaire** peut être différent au cours du temps, en fonction d'événements: panne d'un site, état de la données, etc.

Appartenance à l'instant t1



Appartenance à l'instant t2



41

Propagation des mises à jour

- Propagation
  - gérée par le SGBD réparti
  - Synchronisée, ou non, avec la transaction

Propagation synchrone

- Avant la validation de la transaction
- L'application obtient une réponse **APRES** la propagation
  1) transaction locale → 2) propagation → 3) validation → 4) réponse
- Propagation : nb de messages échangés entre sites
  - Immédiate après chaque opération (L,E)
    - 1 message par opération (L/E): interaction linéaire
  - Différée juste avant la fin de la transaction
    - 1 message par transaction : interaction constante
      - Contenu du message: SQL ou Log
- Validation: décision prise
  - par plusieurs sites (vote, ex 2PC)
  - par chaque site séparément (sans vote)

Propagation asynchrone

- L'application obtient une réponse **AVANT** la propagation
  - 1) transaction locale → 2) validation locale → 3) réponse
  - 4) propagation → 5) validation

  Etapes (1,2,3) indépendantes de (4,5)

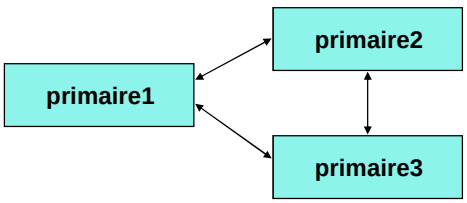
Réplication multi-maîtres

- Plusieurs maîtres pour une donnée (R1, R2)
- Mises à jour sur R1 et R2
- Pb: propagation des mise à jour:
  - Les mises à jour de R sont réparties sur **plusieurs** maîtres R1, … Rn
  - Une réplique (r) doit recevoir toutes les mises à jour reçues par **les maîtres**

Réplication Multimaîtres

Une donnée appartient à plusieurs sites, qui peuvent chacun mettre à jour et diffuser aux autres sites

- **augmente la disponibilité**
- **Optimiste : peut produire des conflits, qui doivent être détectés et résolus**
- **Préventif : éviter les conflits**



46

Classification des solutions

- Monomaître / Multimaîtres
- Asynchrone / Synchrone
  - Synchrone : avec intéraction constante / linéaire
  - Asynchrone : optimiste / pessimiste
- Validation: avec vote / sans vote
- Rmq
  - Primary copy = Monomaître
  - Update Everywhere = Multi-maître
  - Eager Replication = Repl. avec propagation synchrone
  - Lazy Replication = Repl. avec propagation asynchrone

Détection des modifications

- Solution 1 : utilisation du journal
  - les transactions qui modifient écrivent une marque spéciale dans le journal
  - détection périodique en lisant le journal, indépendamment de la transaction qui a modifié
  - modification de la gestion du journal
- Solution 2 : utilisation de triggers
  - la modification d'une donnée répliquée déclenche un trigger
  - mécanisme général et extensible
  - la détection fait partie de la transaction et la ralentit
- Solution 3: détecter les conflit potentiels (a priori)
  - Parser le code (pas possible pour transactions interactives)
  - Graphe d'ordonnancement global

48

Réplication totale vs. partielle

- Evaluation de requêtes, gestion du répertoire (catalogue)
  - Le plus facile est réplication totale
  - Réplication partielle ou fragmentation: difficulté. équiv.
- Contrôle de concurrence
  - Fragmentation > répli. totale > répli. Partielle (+difficile)
- Fiabilité
  - Répli. totale > répli. partielle >> fragmentation
- Réalisme : le plus réaliste est réplication partielle

49

Conclusions et perspectives

Applications classiques
- décisionnel (data warehouse)
- transactionnel

Applications nouvelles
- intégration de données du Web
  - grand nombre de sources
  - hétérogénéité très forte
- intégration des données semistructurées (HTML, XML)
- intégration de la recherche documentaire
- Intégration de services Web (ex. agence de voyage)

50