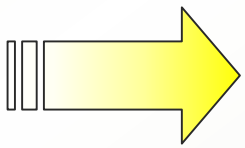
Sélection automatique d'index dans les entrepôts de données

Atelier de recherche
Fouille de données complexes
EGC'04

Kamel Aouiche, Jérôme Darmont, Omar Boussaid
Laboratoire ERIC, Lyon 2
{kaouiche, jdarmont,boussaid}@eric.univ-lyon2.fr

Introduction

- Administration des bases de données
  - **La conception logique et physique des bases de données**
  - **Définition des fichiers et des disques de stockage**
  - **Réglage des performances : sélection de vues, d'index, etc.**
- Minimiser et automatiser les fonctions d'administration



Sélection automatique d'index

Introduction

Définitions

- Index
  - **Structure physique permettant d'accélérer l'accès aux données**
  - **Les index accélèrent les requêtes d'interrogation mais ralentissent les requêtes de mises à jour**
- Technique d'indexation
  - **Structure d'un index**
  - **Bases de données (b-arbre, index de hachage)**
  - **Entrepôts de données (index bitmap, index de jointure…)**
- Charge
  - **Ensemble de requêtes résolues par le système**
- Coût d'utilisation d'un index
  - **Temps d'exécution d'une requête donnée en présence d'un index**
- Gain
  - **Différence entre le temps d'exécution d'une requête sans index et avec index**

Introduction

- Sélection d'index dans les bases de données
  - **Problème NP - complet**
  - **Nombre d'index exponentiel en nombre total d'attributs dans la base**

- Sélection d'index dans les entrepôts de données
  - **Données volumineuses → Index volumineux**
  - **Sélection sous la contrainte de l'espace de stockage disponible partagé avec les vues matérialisées**
  - **Mises à jour prises en compte différemment (rafraîchissements des entrepôts)**

Plan

- **État de l'art**
  - **Sélection d'index dans les bases de données**
  - **Sélection d'index dans les entrepôts de données**
- **Fouille de données pour la sélection d'index**
  - **Améliorations de nos travaux sur la sélection d'index**
  - **Adaptations à l'environnement des entrepôts de données**
- **Conclusion et perspectives**

État de l'art

- Sélection d'index dans les bases de données
  - **Utilisation d'une fonction mathématique pour estimer le coût des requêtes exploitant un ensemble d'index (Kratica *et al.*, 2003 ; Feldman *et al.*, 2003 …)**
  - **Appel à l'optimiseur de requêtes pour évaluer ce coût (Frank *et al.*, 1992 ; Chaudhuri *et al.*, 1997-2001 ; Valentin *et al.*, 2000 …)**

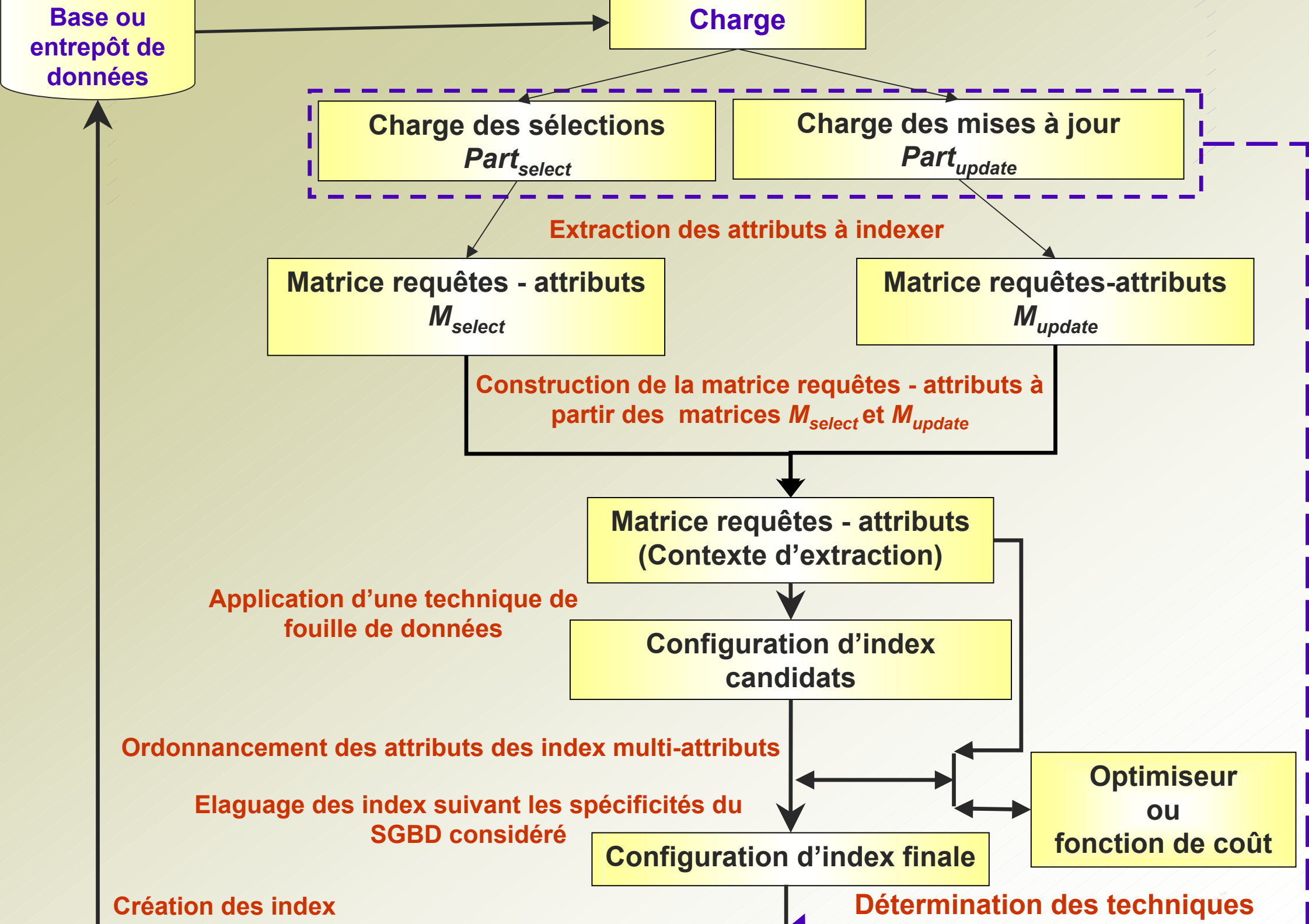
- Sélection d'index dans les entrepôts de données
  - **Optimisation sous la contrainte de l'espace de stockage**
  - **Optimiser le temps de maintenance**
  - **Optimiser le temps d'exécution des requêtes (Gupta *et al.*, 1997 ; Agrawal *et al.*, 2001 ; Golfarelli *et al.*, 2002 …)**

Fouille de données pour la sélection d'index

- Aouiche *et al.* 2003
  - Utilisation de la fouille de données pour la sélection d'index dans les bases de données
  - Analyse de la charge
  - Utilisation de techniques rudimentaires pour élaguer les index les moins avantageux

Fouille de données pour la sélection d'index dans les entrepôt de données

- Améliorations apportées
  - Analyse plus fine de la charge
  - Détermination des techniques d'indexation
  - Construction du contexte d'extraction
  - Ordonnancement des attributs des index multi-attributs
  - Elagage d'index de la configuration en faisant appel à l'optimiseur ou à une fonction de coût
- Adaptations aux environnement des entrepôts de données
  - Analyse de la charge
  - Détermination des techniques d'indexation
  - Contrainte de l'espace de stockage


Base ou entrepôt de données
Charge
Charge des sélections
Part_select
Charge des mises à jour
Part_update
Extraction des attributs à indexer
Matrice requêtes - attributs
M_select
Matrice requêtes-attributs
M_update
Construction de la matrice requêtes - attributs à partir des matrices M_select et M_update
Matrice requêtes - attributs (Contexte d'extraction)
Application d'une technique de fouille de données
Configuration d'index candidats
Ordonnancement des attributs des index multi-attributs
Elaguage des index suivant les spécificités du SGBD considéré
Optimiseur ou fonction de coût
Configuration d'index finale
Création des index
Détermination des techniques

Conclusion

- Technique de fouille de données pour la sélection d'index
- Compléter nos travaux :
  - Analyse plus fine de la charge
  - Détermination des techniques d'indexation
  - Etablissement d'un ordre dans les attributs des index multi-attributs
  - Prise en compte des spécifiés du SGBD considéré
- Adaptation de ces travaux pour la sélection d'index dans les entrepôts de données :
  - Choix des attributs indexables
  - Détermination des techniques d'indexation spécifiques aux entrepôts
  - Prise en compte de la contrainte de l'espace de stockage disponible

Perspectives

- Mettre en œuvre notre approche au sein d'un SGBD
- Coupler la sélection d'index avec la sélection des vues matérialisés (classification des requêtes)
- Exploiter l'ensemble des motifs fréquents pour cibler les index de jointure en étoile
- Valider notre approche
  - Comparaison avec nos travaux antérieurs
  - Comparaison avec les index proposés par un expert
  - Comparaison avec les autres méthodes de sélection d'index

(1) Select name, price from Item
where catalog-number like `999%'
and price > 1000 order by name

(2) Select * from Item
where delivery-time > 100 or price > 100

(3) Select Item.name from Item, Item-Warehouse
where Item-Warehouse.catalog-number = Item.catalog-number
and Item-Warehouse.quantity > 10000
and Item.supplier-code = 4 and Item.price > 1000

(4) Select Item.name, Warehouse.name, Item-Warehouse.quantity,
Item.prace from Item, Warehouse, Item-Warehouse
where Item.catalog-number = Item-Warehouse.catalog-number
and Item-Warehouse.warehouse-code =Warehouse.warehouse-code
and Item.price>1000 and Warehouse.Warehouse-code like `%12'
and Item-Warehouse.quantity<1000

(5) Update Item set delivery-time = delivery-time - 1

(6) Update Item-Warehouse set quantity = 50
where catalog-number = 10

(7) Delete from Item where delivery-time = 0

Charge des selections

(5) Update Item set delivery-time = delivery-time - 1

(6) Update Item-Warehouse set quantity = 50
where catalog-number = 10

(7) Delete from Item where delivery-time = 0

Exemple d'une charge

Charge des mises à jour

La matrice M_{Select}

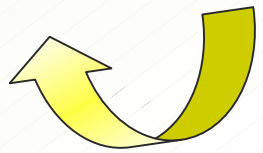
Tables	Item				
Requêtes	price	name	delivery-time	catalog-number	supplier-name
(1)	1	1	0	0	0
(2)	1	0	1	0	0
(3)	1	0	0	1	1
(4)	1	0	0	1	0

Tables	Item-Warehouse			Warehouse
Requêtes	Catalog-number	quantity	warehouse-code	warehouse-code
(1)	0	0	0	0
(2)	0	0	0	0
(3)	1	1	0	0
(4)	1	1	1	1

La matrice M_{update}

Attributs mis à jour					
Table	Item				
Requêtes	catalog-number	name	price	dupplier-code	delivery-time
(5)	0	0	0	0	1
(6)	0	0	0	0	0
(7)	1	1	1	1	1

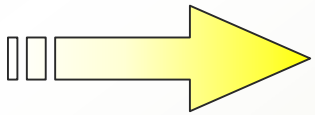
	Attributs mis à jour	Attributs de recherche	
Table	Item-Warehouse	Item	Item-Warehouse
Requêtes	quantity	delivery-time	catalog-number
(5)	0	0	0
(6)	1	0	0
(7)	0	1	1



Détermination des techniques d'indexation

Cas des bases de données

Attributs	Techniques d'indexation
Des clauses <attribut> = <constante> | <sous - requête> <attribut> in <liste_de_valeurs>|<sous-requête>	Index de hachage si les tables sont statiques (la fréquence d'insertions et de suppressions faible)
Des autres clauses	Index en b-arbre

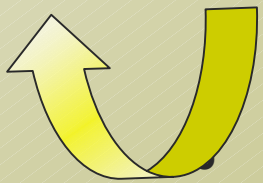


Consultation de la matrice M_{Update}

Détermination des techniques d'indexation

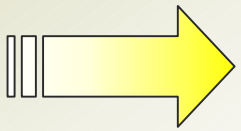
Cas des entrepôts de données

- Facteurs déterminant les techniques d'indexation :
  - **Caractéristiques des attributs indexés**
    - **la cardinalité : le nombre de valeurs distinctes d'un attribut**
    - **la distribution : la fréquence des occurrences des valeurs distincts d'un attribut**
    - **la portée : la différence entre les valeurs maximum et minimum d'un attribut**
  - **Caractéristiques liées à l'usage des attributs indexés**
    - **Conditions de jointure**
    - **Prédicats de restriction**
    - **…**



Ordre des attributs dans les index multi-attributs

- Index mono-attributs
- Index multi-attributs
  - Etablissement de l'ordre des attributs



1. **Garantir qu'un index est utilisé par un grand nombre de requêtes**
2. **Gérer plus efficacement l'espace de stockage des index**

Select * from t where a=4 and b=5

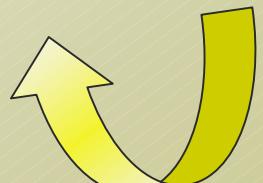
Select * from t where b=3

Select * from t where a=3

Index ba

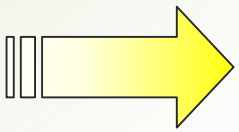
Index a

Index b



Elagage d'index de la configuration

- Charge volumineuse ➔ Nombre d'index candidats est important
- Nombre d'index autorisé par table est limité (Oracle autorise 16 index par table)



Elaguer les index les moins avantageux

Appel à l'optimiseur

Fonction de coût

Privilégier les index occupant moins d'espace