

Physische Datenorganisation

Themenbereiche bisher:

- Speicherhierarchie
- Hintergrundspeicher / RAID
- Speicherstrukturen
- ISAM
- B-Bäume

Nun:

- Hashing
- Clustering

Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015

15.12.2014

Partitionierung

- Bäume brauchen im Schnitt $\log_k(n)$ Seitenzugriffe, um ein Datenelement zu lesen (k=Verzweigungsgrad, n=Anzahl indexierter Datensätze)
 - Hashtabellen (partitionierende Verfahren) brauchen im Schnitt zwei Seitenzugriffe
- warum B-Bäume und nicht Hashing?

Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015

15.12.2014

Was ist Hashing?

- to hash = zerhacken
- Speicherung der Tupel in einem festgelegten Speicherbereich
- Hashfunktion: Abbildung von Tupeln (Schlüsselwerte) in eine festgelegte Menge von Funktionswerten
- optimale Hashfunktion:
 - injektiv (keine gleichen Funktionswerte für unterschiedliche Argumente)
 - surjektiv (kein Speicherverschnitt)
- typische Hashfunktion $h: h(x) = x \bmod N$
Menge von Funktionswerten somit $\{0, \dots, N-1\}$

Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015

15.12.2014

Beispiel Hashing

- Beispiel-Hashfunktion $h(x) = x \bmod 3$

0	
1	(27550, 'Schopenhauer', 6)
2	(24002, 'Xenokrates', 18) (25403, 'Jonas', 12)

Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015

15.12.2014

Kollisionen

- Kollisionsbehandlung

0		
1	(27550, 'Schopenhauer', 6)	•
2	(24002, 'Xenokrates', 18) (25403, 'Jonas', 12)	• →
	(26120, 'Fichte', 10) (28106, 'Carnap', 3)	• → ...

- Ineffizient bei nicht vorhersehbarer Datenmenge
- Ausweg: erweiterbares (dynamisches) Hashing
→ zusätzliche Indirektion über Directory

Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015

15.12.2014

Vor- / Nachteile Hashing

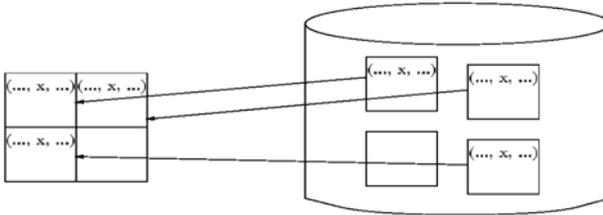
- + wenig Hintergrundspeicherzugriffe
 - konstanter Aufwand: $O(1)$, i.d.R. 1-2
 - + einfache Implementierung
- Kollisionsbehandlung notwendig
 - Vorreservierung des Speicherbereichs
 - nicht dynamisch bzw. nur mit Nachsteuerung
 - **keine Bereichsanfragen, nur Punktanfragen**

Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015

15.12.2014

Objektballung Clustering logisch verwandter Daten

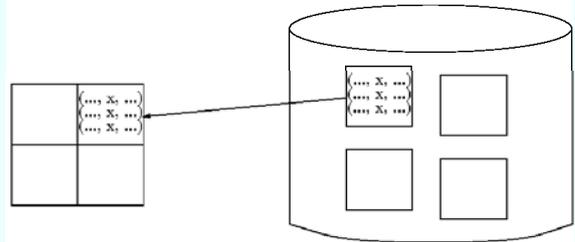
```
select *
from R
where A = x;
```



Hauptspeicher ← Zugriffslücke → Hintergrundspeicher
Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015 15.12.2014

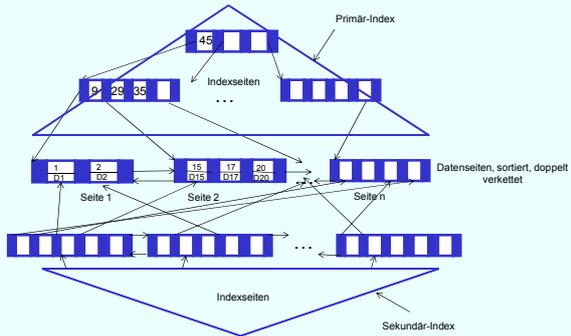
Objektballung Clustering logisch verwandter Daten

Besser:



Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015 15.12.2014

Erinnerung



Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015 15.12.2014

Verschränkte Speicherung

Seite P_i	
2125 ◦ Sokrates	◦ C4 ◦ 226 •
5041 ◦ Ethik	◦ 4 ◦ 2125 •
5049 ◦ Mäeutik	◦ 2 ◦ 2125 •
4052 ◦ Logik	◦ 4 ◦ 2125 •
2126 ◦ Russel	◦ C4 ◦ 232 •
5043 ◦ Erkenntnistheorie	◦ 3 ◦ 2126 •
5052 ◦ Wissenschaftstheorie	◦ 3 ◦ 2126 •
5216 ◦ Bioethik	◦ 2 ◦ 2126 •

Seite P_{i+1}	
2133 ◦ Popper	◦ C3 ◦ 52 •
5259 ◦ Der Wiener Kreis	◦ 2 ◦ 2133 •
2134 ◦ Augustinus	◦ C3 ◦ 309 •
5022 ◦ Glaube und Wissen	◦ 2 ◦ 2134 •
2137 ◦ Kant	◦ C4 ◦ 7 •
5001 ◦ Grundzüge	◦ 4 ◦ 2137 •
4630 ◦ Die 3 Kritiken	◦ 4 ◦ 2137 •
⋮	

Datenbanksysteme für Hörer anderer Fachrichtungen WS 2014/2015 15.12.2014