Einsatz und Realisierung von Datenbanksystemen

ERDB Übungsleitung

Alice Rey, Maximilian Reif, Tobias Goetz

i3erdb@in.tum.de

Folien erstellt von Maximilian Bandle & Alexander Beischl

Organisatorisches
Disclaimer

Die Folien werden von der Übungsleitung allen Tutoren zur Verfügung gestellt.

Sollte es Unstimmigkeiten zu den Vorlesungsfolien von Prof. Kemper geben, so sind die Folien aus der Vorlesung ausschlaggebend.

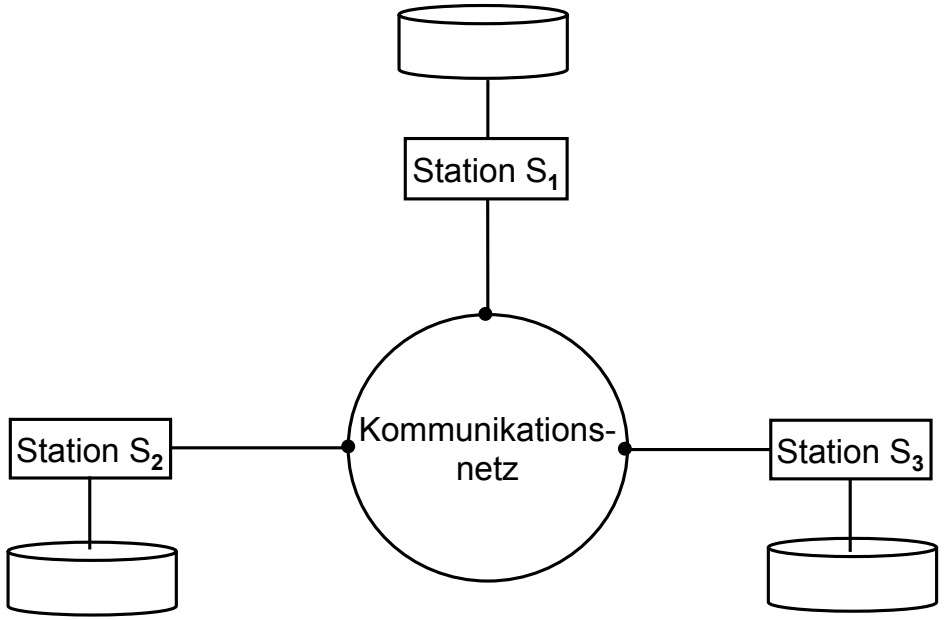
Falls Ihr einen Fehler oder eine Unstimmigkeit findet, schreibt an i3erdb@in.tum.de mit Angabe der Foliennummer.

Verteilte Datenbanken

Verteilte Datenbanksysteme

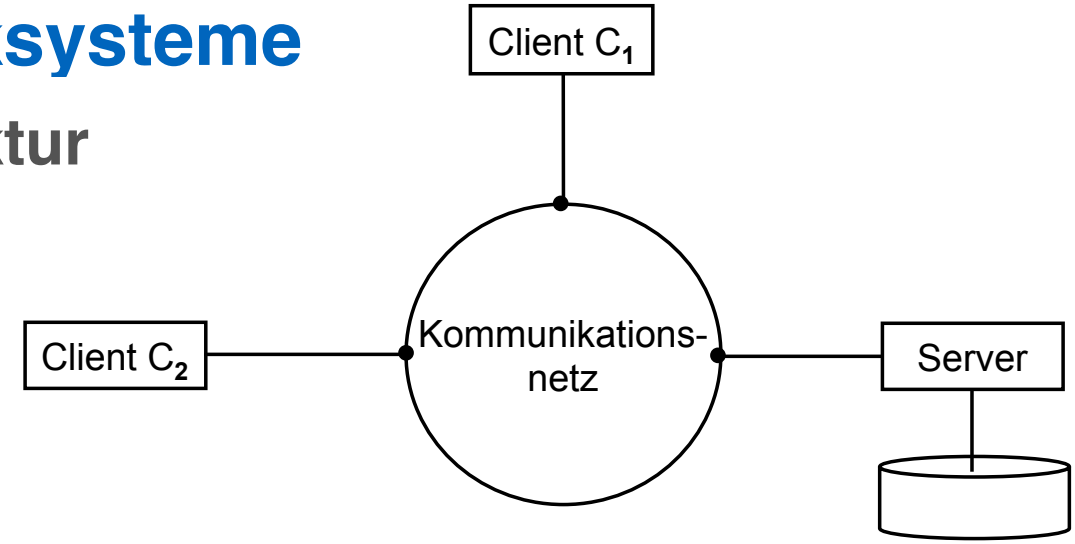
Motivation

- Globale Gesamtinformation auf Stationen (Sites) verteilt
- Daten werden von verteilten Datenbankverwaltungssystemen (VDBMS) verwaltet
- Stationen dürfen lokale Daten bearbeiten
- Kommunikationsverbindung (LAN, WAN, Telefonverbindungen)



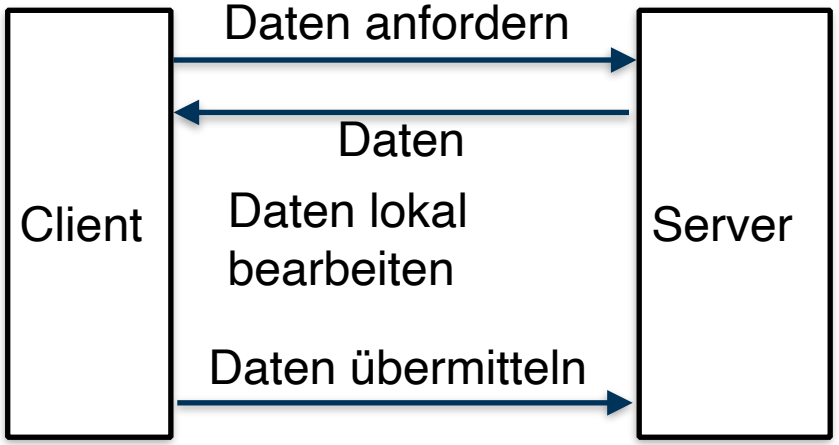
Verteilte Datenbanksysteme

Client-/Server-Architektur



- Degradiertes verteiltes Datenbanksystem

- Nur Server darf Daten abspeichern



Verteilte Datenbanksysteme
Fragmentierung

Horizontale Fragmentierung:

- Relation wird in disjunkte Tupelmengen geteilt

Vertikale Fragmentierung:

- Relation wird nach Attributen geteilt (durch Projektionen)

- Kombinierte Fragmentierung möglich

Aufgabe 1

Professoren						
PersNr	Name	Rang	Raum	Fakultät	Gehalt	Steuerklasse
2125	Sokrates	C4	226	Philosophie	85000	1
2126	Russel	C4	232	Philosophie	80000	3
2127	Kopernikus	C3	310	Physik	65000	5
2133	Popper	C3	52	Philosophie	68000	1
2134	Augustinus	C3	309	Theologie	55000	5
2136	Curie	C4	36	Physik	95000	3
2137	Kant	C4	7	Philosophie	98000	1

Gehen Sie von folgender kombinierter Fragmentierung der in Abbildung 1 dargestellten Relation *Professoren* aus:

1. Zuerst erfolgt eine vertikale Fragmentierung in

$$\begin{aligned} \text{ProfVerw} &:= \Pi_{\text{PersNr, Name, Gehalt, Steuerklasse}}(\text{Professoren}) \\ \text{Profs} &:= \Pi_{\text{PersNr, Name, Rang, Raum, Fakultät}}(\text{Professoren}) \end{aligned}$$

2. Das Fragment Profs wird weiter horizontal fragmentiert in

$$\begin{aligned} \text{TheolProfs} &:= \sigma_{\text{Fakultät = 'Theologie'}}(\text{Profs}) \\ \text{PhysikProfs} &:= \sigma_{\text{Fakultät = 'Physik'}}(\text{Profs}) \\ \text{PhiloProfs} &:= \sigma_{\text{Fakultät = 'Philosophie'}}(\text{Profs}) \end{aligned}$$

Übersetzen Sie aufbauend auf dieser Fragmentierung die folgende SQL-Anfrage in die kanonische Form.

```
select Name, Gehalt Rang
from Professoren
where Gehalt > 80000;
```

Optimieren Sie diesen kanonischen Auswertungsplan durch Anwendung algebraischer Transformationsregeln (Äquivalenzen).

Verteilte Datenbanksvsteme

Fragmentierung

Professoren						
PersNr	Name	Rang	Raum	Fakultät	Gehalt	Steuerklass
2125	Sokrates	C4	226	Philosophie	85000	1
2126	Russel	C4	232	Philosophie	80000	3
2127	Kopernikus	C3	310	Physik	65000	5
2133	Popper	C3	52	Philosophie	68000	1
2134	Augustinus	C3	309	Theologie	55000	5
2136	Curie	C4	36	Physik	95000	3
2137	Kant	C4	7	Philosophie	98000	1

1. Zuerst erfolgt eine vertikale Fragmentierung in

$$\begin{aligned} \text{ProfVerw} &:= \Pi_{\text{PersNr, Name, Gehalt, Steuerklasse}}(\text{Professoren}) \\ \text{Profs} &:= \Pi_{\text{PersNr, Name, Rang, Raum, Fakultät}}(\text{Professoren}) \end{aligned}$$

Verteilte Datenbanksvsteme

Fragmentierung Vertikal

ProfVerw			
PersNr	Name	Gehalt	Steuerklasse
2125	Sokrates	85000	1
2126	Russel	80000	3
2127	Kopernikus	65000	5
2133	Popper	68000	1
2134	Augustinus	55000	5
2136	Curie	95000	3
2137	Kant	98000	1

Profs				
PersNr	Name	Rang	Raum	Fakultät
2125	Sokrates	C4	226	Philosophie
2126	Russel	C4	232	Philosophie
2127	Kopernikus	C3	310	Physik
2133	Popper	C3	52	Philosophie
2134	Augustinus	C3	309	Theologie
2136	Curie	C4	36	Physik
2137	Kant	C4	7	Philosophie

2. Das Fragment Profs wird weiter horizontal fragmentiert in

$$\begin{aligned} \text{TheolProfs} &:= \sigma_{\text{Fakultät} = \text{'Theologie'}}(\text{Profs}) \\ \text{PhysikProfs} &:= \sigma_{\text{Fakultät} = \text{'Physik'}}(\text{Profs}) \\ \text{PhiloProfs} &:= \sigma_{\text{Fakultät} = \text{'Philosophie'}}(\text{Profs}) \end{aligned}$$

Verteilte Datenbanksvsteme

Fragmentierung Horizontal

Profs				
PersNr	Name	Rang	Raum	Fakultät
2125	Sokrates	C4	226	Philosophie
2126	Russel	C4	232	Philosophie
2127	Kopernikus	C3	310	Physik
2133	Popper	C3	52	Philosophie
2134	Augustinus	C3	309	Theologie
2136	Curie	C4	36	Physik
2137	Kant	C4	7	Philosophie

2. Das Fragment Profs wird weiter horizontal fragmentiert in

$$\begin{aligned} \text{TheolProfs} &:= \sigma_{\text{Fakultät} = \text{'Theologie'}}(\text{Profs}) \\ \text{PhysikProfs} &:= \sigma_{\text{Fakultät} = \text{'Physik'}}(\text{Profs}) \\ \text{PhiloProfs} &:= \sigma_{\text{Fakultät} = \text{'Philosophie'}}(\text{Profs}) \end{aligned}$$

Verteilte Datenbanksvsteme

Fragmentierung Horizontal

TheolProfs				
PersNr	Name	Rang	Raum	Fakultät
2134	Augustinus	C3	309	Theologie

PhysikProfs				
PersNr	Name	Rang	Raum	Fakultät
2127	Kopernikus	C3	310	Physik
2136	Curie	C4	36	Physik

PhiloProfs				
PersNr	Name	Rang	Raum	Fakultät
2125	Sokrates	C4	226	Philosophie
2126	Russel	C4	232	Philosophie
2133	Popper	C3	52	Philosophie
2137	Kant	C4	7	Philosophie

Zerlegung in disjunkte Tupelmengen

Verteilte Datenbanksvsteme

Fragmentierung Vertikal & Horizontal

ProfVerw			
PersNr	Name	Gehalt	Steuerklasse
2125	Sokrates	85000	1
2126	Russel	80000	3
2127	Kopernikus	65000	5
2133	Popper	68000	1
2134	Augustinus	55000	5
2136	Curie	95000	3
2137	Kant	98000	1

TheolProfs				
PersNr	Name	Rang	Raum	Fakultät
2134	Augustinus	C3	309	Theologie

PhysikProfs				
PersNr	Name	Rang	Raum	Fakultät
2127	Kopernikus	C3	310	Physik
2136	Curie	C4	36	Physik

PhiloProfs				
PersNr	Name	Rang	Raum	Fakultät
2125	Sokrates	C4	226	Philosophie
2126	Russel	C4	232	Philosophie
2133	Popper	C3	52	Philosophie
2137	Kant	C4	7	Philosophie

Übersetzen Sie aufbauend auf dieser Fragmentierung die folgende SQL-Anfrage in die kanonische Form.

```
select Name, Gehalt Rang
from Professoren
where Gehalt > 80000;
```

Aufgabe 2

MatrNr	Name	Note	Standort
10101	Philipp	1,0	München
10102	Magdalena	1,0	Garching
10103	Erik	1,0	Garching
10104	Josef	1,0	Garching
10105	Alex	1,0	Garching
10106	Maxmilian	1,0	München

Für eine verteilte Datenbank soll die Tabelle geeignet fragmentiert werden. Ziel ist, Namen mit Standort der Studenten lokal und die Noten getrennt abzupeichern.

1) Fragmentieren Sie die Relation geeignet *vertikal*.
   a) Geben Sie das Schema für die zwei resultierenden Relationen $KlausurV_1$ und $KlausurV_2$ an. Unterstreichen Sie jeweils den Primärschlüssel.
   b) Geben Sie in SQL-92 die zwei resultierenden Relationen `KlausurV1` und `KlausurV2` als Hilfstabellen (mittels `with`) an.
2) Die geeignetere der beiden resultierenden Relationen soll *horizontal* fragmentiert werden.
   a) Geben Sie das Prädikat der Selektion an, mit dem fragmentiert wird.
   b) Geben Sie in SQL-92 die zwei resultierenden Relationen `KlausurH1` und `KlausurH2` als Hilfstabellen (mittels `with`) an.
3) Schreiben Sie eine SQL-Abfrage, die die Ursprungsrelation aus den Teilrelationen zusammensetzt.

Aufgabe 3

Für die Rekonstruierbarkeit der Originalrelation R aus vertikalen Fragmenten $R_1, \ldots, R_n$ reicht es eigentlich, wenn Fragmente paarweise einen Schlüsselkandidaten enthalten. Illustrieren Sie, warum es also nicht notwendig ist, dass der Durchschnitt aller Fragmentschemata einen Schlüsselkandidaten enthält. Es muss also nicht unbedingt gelten

$$R_1 \cap \cdots \cap R_n \supseteq \kappa,$$

wobei κ ein Schlüsselkandidat aus R ist.

Geben Sie ein anschauliches Beispiel hierfür – am besten bezogen auf unsere Beispiel-Relation *Professoren*.

Verteilte Datenbanksysteme

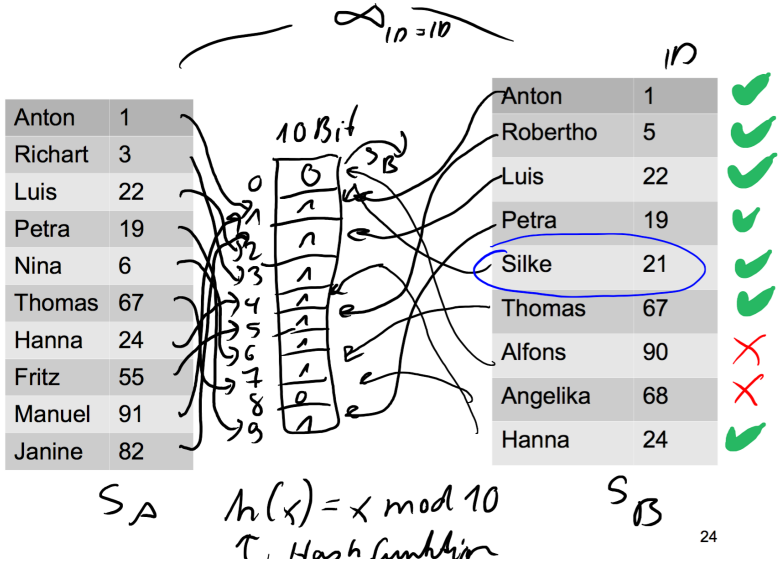
Bloom-Filter

- Einsatz bei sehr voluminösen Join-Attributen (z.B. lange Strings)

+ Verringerung der Transferkosten/ Netzwerkauslastung durch Tupelvorauswahl mittels Hashfunktion

+ Filter wird kompakter (Bitvektor V)

- Filterpräzision geht verloren



Verteilte Datenbanksysteme

Bloom-Filter

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

V
0
1
2
3
4
5
6
7
8
9

S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

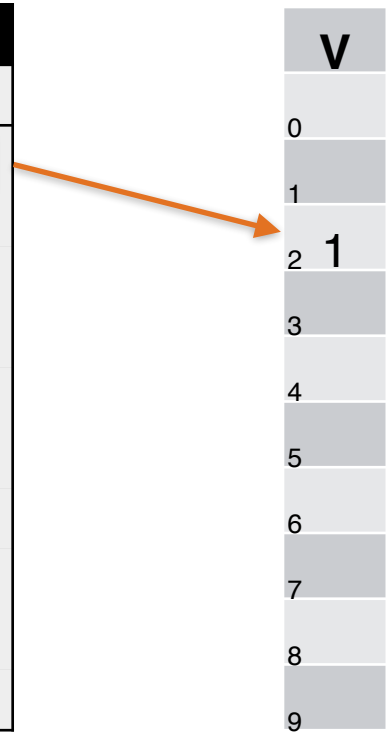
$$h(x) = x \bmod 10$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

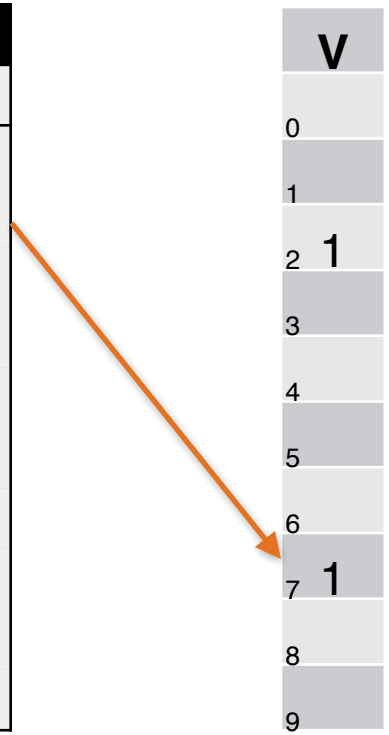
$$h(2) = 2 \bmod 10 = 2$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

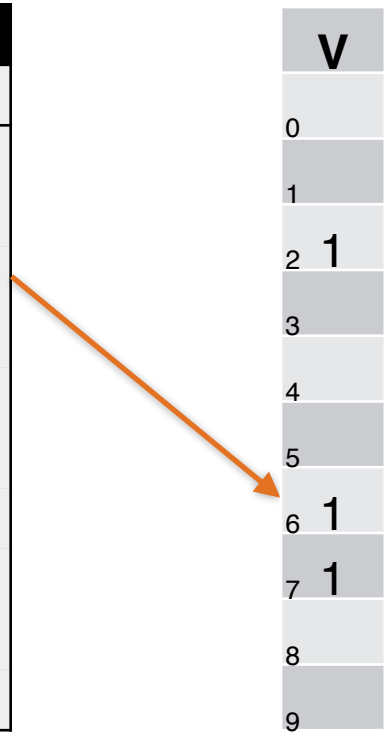
$$h(7) = 7 \bmod 10 = 7$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

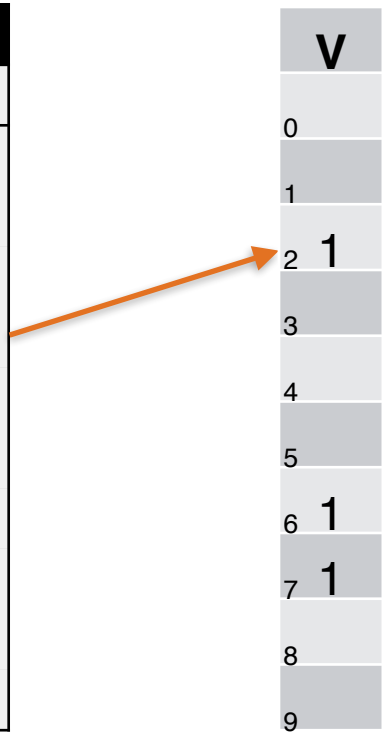
$$h(6) = 6 \bmod 10 = 6$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

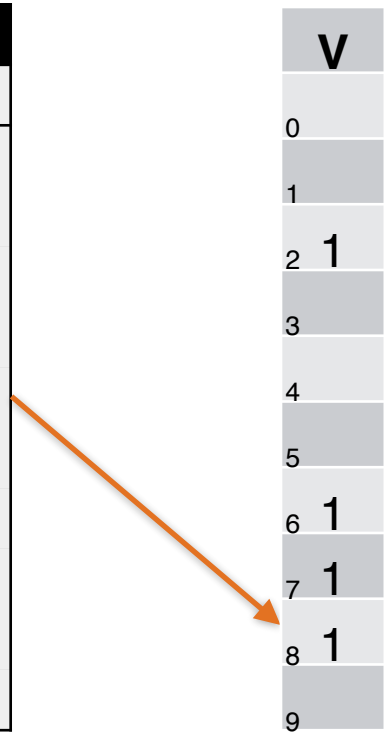
$$h(2) = 2 \bmod 10 = 2$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

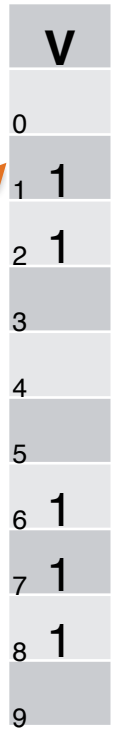
$$h(8) = 8 \bmod 10 = 8$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

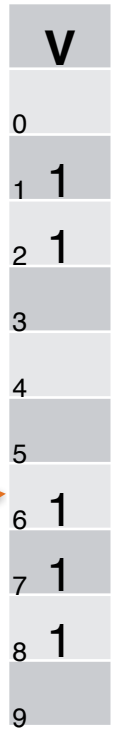
$$h(1) = 1 \bmod 10 = 1$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

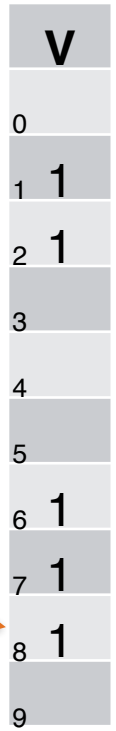
$$h(16) = 16 \bmod 10 = 16$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

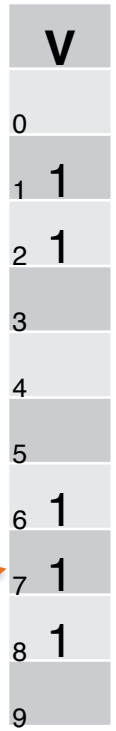
$$h(8) = 8 \bmod 10 = 8$$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5



S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

$h(7) = 7 \bmod 10 = 7$

Verteilte Datenbanksysteme

Bloom-Filter

1. Tabelle R mit h(x) auf V mappen:

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

$$h(5) = 5 \bmod 10 = 5$$

Verteilte Datenbanksysteme

Bloom-Filter

2. Felder in V ohne hash-Treffer mit 0 füllen

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

	V
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

$h(x) = x \bmod 10$

Verteilte Datenbanksysteme

Bloom-Filter

3. Bitvektor V an S schicken

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

	v
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

$h(x) = x \bmod 10$

Verteilte Datenbanksysteme

Bloom-Filter

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

	V
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S	
Raum	Gebäude
1	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

$h(x) = x \bmod 10$

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

V	
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S	
Raum	Gebäude
1 ✔	IMETUM
2	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

$$h(1) = 1 \bmod 10 = 1$$

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

v	
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S	
Raum	Gebäude
1 ✔	IMETUM
2 ✔	MI Büro
4	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

$h(2) = 2 \bmod 10 = 2$

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

V	
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S	
Raum	Gebäude
1 ✔	IMETUM
2 ✔	MI Büro
4 ✘	Physik
6	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

$h(4) = 4 \bmod 10 = 4$

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

	V
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S	
Raum	Gebäude
1 ✔	IMETUM
2 ✔	MI Büro
4 ✘	Physik
6 ✔	MW
7	MI Raum
8	ERI
9	MI Bib
10	Physik
11	Chemie

h(6) = 6 mod 10 = 6

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

	V
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S		
Raum		Gebäude
1	✔	IMETUM
2	✔	MI Büro
4	✘	Physik
6	✔	MW
7	✔	MI Raum
8		ERI
9		MI Bib
10		Physik
11		Chemie

h(7) = 7 mod 10 = 7

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

v	
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S		
Raum		Gebäude
1	✔	IMETUM
2	✔	MI Büro
4	✘	Physik
6	✔	MW
7	✔	MI Raum
8	✔	ERI
9		MI Bib
10		Physik
11		Chemie

h(8) = 8 mod 10 = 8

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

	V
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S		
Raum		Gebäude
1	✔	IMETUM
2	✔	MI Büro
4	✘	Physik
6	✔	MW
7	✔	MI Raum
8	✔	ERI
9	✘	MI Bib
10		Physik
11		Chemie

h(9) = 9 mod 10 = 9

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

	V
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S		
Raum		Gebäude
1	✔	IMETUM
2	✔	MI Büro
4	✘	Physik
6	✔	MW
7	✔	MI Raum
8	✔	ERI
9	✘	MI Bib
10	✘	Physik
11		Chemie

$h(10) = 10 \bmod 10 = 0$

Verteilte Datenbanksysteme

Bloom-Filter

False positive

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

4. S überprüft mit h(x) den Bitvektor V

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

	V
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S		
Raum		Gebäude
1	✔	IMETUM
2	✔	MI Büro
4	✘	Physik
6	✔	MW
7	✔	MI Raum
8	✔	ERI
9	✘	MI Bib
10	✘	Physik
11	✔	Chemie

$h(11) = 11 \bmod 10 = 1$

Verteilte Datenbanksysteme

Bloom-Filter

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

5. Übermitteln der Treffer zur Station R

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

v	
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S		
Raum		Gebäude
1	✔	IMETUM
2	✔	MI Büro
4	✘	Physik
6	✔	MW
7	✔	MI Raum
8	✔	ERI
9	✘	MI Bib
10	✘	Physik
11	✔	Chemie

$$h(x) = x \bmod 10$$

Verteilte Datenbanksysteme

Bloom-Filter

False positive Rate 1/6

✔ Tupel wird zur Station mit R geschickt
✘ Tupel wird nicht übermittelt

False positives werden übermittelt und von R beim Join verworfen.

R	
Pers	Raum
Max	2
Magda	7
Tom	6
Alex	2
Julius	8
Kathi	1
Anna	16
Gregor	8
Thuy	7
Domi	5

v	
0	0
1	1
2	1
3	0
4	0
5	1
6	1
7	1
8	1
9	0

S		
Raum		Gebäude
1	✔	IMETUM
2	✔	MI Büro
4	✘	Physik
6	✔	MW
7	✔	MI Raum
8	✔	ERI
9	✘	MI Bib
10	✘	Physik
11	✔	Chemie

$h(x) = x \bmod 10$

Aufgabe 4

Gegeben seien die Tabellen `Studenten` und `Punkte` mit Schlüssel `MatrNr`, wobei `Punkte` auf einem separaten Rechner gespeichert ist. Es soll folgende Anfrage ausgeführt werden:

```
SELECT Name, Bonus FROM Student s, Punkte p WHERE s.MatrNr = p.MatrNr;
```

Der Datenbankadministrator entscheidet sich für einen Bloom-Filter zur Vorauswahl der Tupel. Auf `MatrNr` wird die Hash-Funktion $h(x) = x \mod 5$ angewendet.

`Studenten`

MatrNr	Name	Hashwert
27	Magda	
4	Josef	
19	Erik	
95	Philipp	

`Punkte`

MatrNr	Bonus	Hashwert
27	ja	
16	nein	
25	nein	
95	ja	

a) Berechnen Sie die Hash-Werte und tragen Sie diese in die obige Tabelle ein.

b) Füllen Sie den von `Studenten` zu übertragenden Bitvektor aus. Verwenden Sie 0 oder 1.

c) Geben Sie basierend auf dem Bitvektor an, welche Tupel aus `Punkte` übertragen werden (nur `MatrNr` angeben).

d) Geben Sie die Falsch-Positiv-Rate (false positive rate) an.

e) Nehmen Sie an, dass jedes Tupel 8 Byte und der Bloomfilter selbst 1 Byte groß ist. Berechnen Sie zunächst die übertragenen Bytes ohne und mit Einsatz des Bloom-Filters.

Aufgabe 5

Überlegen Sie sich, welche Tupel bei der Anwendung des bloomfilterbasierten Joins in Abbildung 2 übertragen werden. Markieren Sie insbesondere, welche Tupel übertragen werden, obwohl sie keinen Joinpartner finden (sog. *false drops*). Wie kann die Anzahl dieser *false drops* verringert werden? Welche Eigenschaften sollte die Hashfunktion h(c) die bei dieser Joinbearbeitung verwendet wird erfüllen?

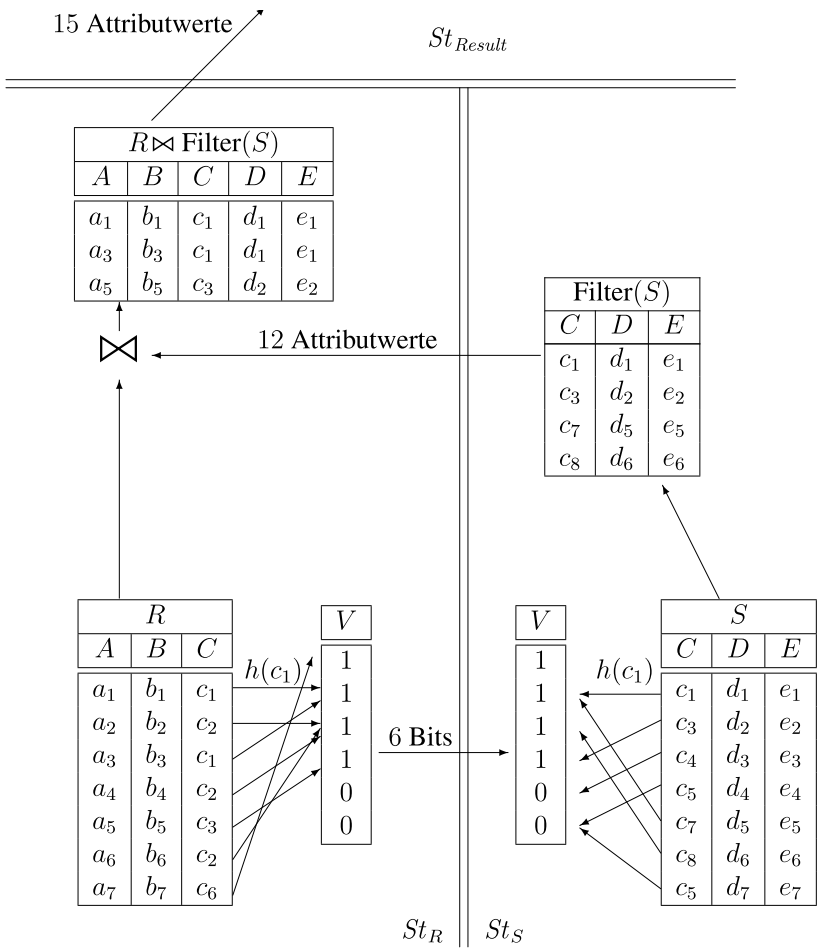


Abbildung 2: Beispiel einer verteilten Joinbearbeitung mit Bloomfilter.

Fragen?