

Mobile und Verteilte Datenbanksysteme

Roland Hediger, Egemen Kaba

5. Juli 2014

Inhaltsverzeichnis

I. Theorie	4
1. Einführung	5
1.1. Fundamental Definitionen	5
1.2. Verteilte Datenbank Definitionen	5
1.2.1. Klassifizierung	5
1.3. Datas 12 Regeln	5
1.4. Aspekte Verteilte Datenbanksysteme	6
1.5. Parallele Datenbanksysteme	6
1.5.1. Architekturen	6
1.5.2. Aspekte	6
1.6. Nosql Systeme	7
1.6.1. Konzepte	7
1.7. Mobile Datenbank Systeme	7
1.8. Semantic Web	7
2. Trigger	8
2.1. Triggertypen	8
2.2. DML Trigger auf Tabellen	8
2.3. PLSQL allgemein	9
2.4. Trigger Beispiele	9
3. Entwurf Verteilte Datenbanken	12
3.1. Einführung	12
3.2. Entwürfe	13
3.2.1. Top Down	13
3.2.2. Bottom Up	14
3.3. Korrektheit der Fragmentierung	14
3.4. Fragmentierungsabkürzungen	14
3.5. PHF	14
3.6. PHF Schritt für Schritt	15
3.7. PHF Beispiel Bikes	15
3.8. DHF	16
4. Vertikale Fragmentierung	18
4.1. Zugriffseigenschaften	18
4.2. Cluster Methode	19
4.2.1. Affinität	19
4.2.2. Vertauschen bei Affinität	20
4.3. Bond Energie Algorithmus	20
4.4. VF - Vertical Fragmentation Example	20
4.5. VF : Cluster möglichketen	21
4.6. Schluss : VF	21

5. Verteilte Anfrage Verarbeitung	24
5.1. Query Processor	24
5.2. Zentralisiert vs Dezentralisiert	24
5.3. Überblick	24
5.3.1. Ziel der Optimierung	24
5.3.2. Komplexität	24
5.4. Methode	25
5.4.1. Verteilte Verarbeitung	25
5.4.2. Zerlegung	25
5.5. Lokalisierung	25
5.5.1. PHF	25
5.5.2. Schritten für Lokalisierung PHF	26
5.5.3. Reduktion mit Join für PHF	26
5.6. Reduktion für VF	26
6. Concurrency	30
6.1. Notation + Definitionen für Concurrency	30
6.2. Einführung	30
6.2.1. Aspekte der Transaktionsverarbeitung	31
6.3. Architektur	31
6.4. Nebenläufigkeit	31
6.4.1. Beispiel : Global nicht serialisierbar	32
6.5. Realisierung	32
6.5.1. Zentrales 2PL	33
6.5.2. Primary Copy 2pl	33
6.5.3. Verteiltes 2PL	34
6.6. Deadlocks	34
6.6.1. Deadlock Erkennung	34
6.6.2. Verteilte Deadlocks	34
7. Verteilte Transaktionen 2	36
7.1. Lokale oder Verteilte Zuverlässigkeit	36
7.2. Lokale Wiederherstellung	36
7.3. Zuverlässigkeit	37
7.3.1. Komponenten	37
7.4. 2 Phase Commit	37
7.4.1. Zustandsübergänge	38
7.4.2. Recovery Protocols	39
7.4.3. Ausfälle	40
7.4.4. Probleme mit 2pc	40
7.5. 3 Phase Commit	40
8. Replikation I	41
8.1. Einleitung	41
8.1.1. Gründe für Replikation	41
8.1.2. Ausführungsmodell	41
8.2. Konsistenzmodelle	41
8.3. Update Propagation Strategies	42
8.4. Replikationsstrategien	43
9. Replikation II (Mat Views)	45
9.1. Komponenten	45
9.2. Komponenten	45
9.3. Benutzer und Rollen	45
9.4. Master Site Strukturen	46
9.5. Materialized View Site Strukturen	46
9.6. Konfliktauflösung	46
10. Nosql	47
10.1. Einführung	47
10.1.1. Motivation	47

10.2. Cassandra	49
10.3. Neo4J	50
10.4. Bigtable	52
10.5. MongoDB	54
11. Semantic Web	58
11.1. World Wide Web	58
11.2. Idee des Semantic Webs	58
11.3. Ziele des Semantic Webs	58
11.4. Datenmodell WWW und semantic web	59
11.5. Technologien	59
11.5.1. RDF - Resource Description Framework	59
11.6. RDF und OWL	60
II. Labs	61
12. Lab 1 - Trigger	62
12.1. Event Logging	62
12.1.1. Aufgabenstellung	62
12.1.2. Lösung	62
12.1.3. Lösungsbeschrieb	64
12.2. Referential Integrity	64
12.2.1. Aufgabenstellung	64
12.2.2. Lösung	64
12.2.3. Lösungsbeschrieb	65
13. Lab 3 - Verteilter Datenbankentwurf	66
13.1. Aufgabenstellung	66
13.2. Lösung	66
13.3. Lösungsbeschrieb	66
13.3.1. Weg zur globalen Affinität	66
14. Lab 8 - Replikation und Konflikte	70
14.1. Setup Master Site	70
14.2. Setup Materialized View Site	70
14.3. Erstellen von Master Group	71
14.4. Erstellen von Master View Group	72
14.5. Lösungen der Konflikte	73
14.5.1. Update Konflikt Lösung	73
14.5.2. Uniqueness Konflikt Lösung	74
14.5.3. Delete Konflikt	74
15. Lab 11 - Relationale Operationen mit Map Reduce	76
15.1. Aufgabenstellung	76
15.2. Group by mit Count	76
15.3. Join	77
15.3.1. Lösungsbeschrieb	80
15.3.2. Join	80

Teil I.

Theorie

1. Einführung

Was wird Verteilt

- Aufbau-logik, Verarbeitungselemente
- Funktion
- Daten
- Steuerung

1.1. Fundamental Definitionen

Fundamental Principle To the user, a distributed system should look exactly like a nondistributed system.

Data Processing and Mobility It has become a common approach to turn any location and situation a job of ce.

Warum Parallel Performance Performance And more Performance.

Big Data Volume, Variety, Velocity.

1.2. Verteilte Datenbank Definitionen

Verteilte Datenbank - DDB ine verteilte Datenbank ist eine Sammlung mehrerer, untereinander logisch zusammengehöriger Datenbanken, die über ein Computernetzwerk verteilt sind.

Verteiltes Datenbankverwaltungssystem (D-DBMS) Ein verteiltes Datenbankverwaltungssystem ist die Software, die die verteilte Datenbank verwaltet und gegenüber den Nutzern einen transparenten Zugang erbringt.

Verteiltes Datenbank System $DDBS = DBS + D-DBMS$

1.2.1. Klassifizierung

Heterogenität Hardware, Netzwerkprotokolle, Datenverwaltung, Datenmodell, Abfragesprache, Transaktionsverwaltung 2 Ausprägungen: homogen, heterogen.

Verteilung Verteilung: betrifft die Verteilung der Daten 2 Ausprägungen: verteilt, zentral

Autonomie Autonomie: betrifft die Verteilung der Steuerung 3 Ausprägungen: stark integriert, halbautonom, isoliert

1.3. Dates 12 Regeln

1. Lokale Autonomie
2. Unabhängigkeit von zentralen Systemfunktionen
3. Hohe Verfügbarkeit
4. Ortstransparenz
5. Fragmentierungstransparenz
6. Replikationstransparenz
7. Verteilte Anfragebearbeitung
8. Verteilte Transaktionsverarbeitung
9. Hardware Unabhängigkeit

10. Betriebssystem Unabhängigkeit
11. Netzwerkunabhängigkeit
12. Datenbanksystem Unabhängigkeit

1.4. Aspekte Verteilte Datenbanksysteme

Verteilter Datenbankentwurf

- wie die Datenbank verteilen
- Verteilung der DB mit Replikaten oder ohne
- Verzeichnisverwaltung

Anfragebearbeitung

- Zerlegung der Anfragen in ausführbare Instruktionen
- Anfrageoptimierung
- berücksichtigen von Verarbeitungs- und Datentransferkosten

Nebenläufigkeit

- Synchronisation konkurrierender Transaktionen
- Konsistenz und Isolation
- Deadlock Erkennung

Zuverlässigkeit

- Robustheit gegenüber Fehler
- Atomarität und Dauerhaftigkeit

1.5. Parallele Datenbanksysteme

Parallele DB Systeme kombinieren Datenbankverwaltung und Parallel Verarbeitung zur Verbesserung der Performance und der Verfügbarkeit.

1.5.1. Architekturen

- Shared Memory Architecture
- Shared Disk Architecture
- Shared Nothing Architecture

1.5.2. Aspekte

- Parallele Anfrageverarbeitung
- Daten Partitionierung
- Parallelisierung von Operationen
- Lastausgleich
- Verfügbarkeit

1.6. Nosql Systeme

- Big Data
- Performance vs Scalability
- Latency vs Throughput
- Availability vs Consistency

1.6.1. Konzepte

- CAP Theorem ¹
- ACID vs BASE²
- Speicherstrukturen : Keyvalue, Document, Wide Column, Graph DB.
- Map / Reduce : Verarbeitung
- Consistent Hashing : Verteilung
- Multiversion Concurrency Control
- Paxos³

1.7. Mobile Datenbank Systeme

Mobile Datenbank Systeme sind Verteilte Datenbank Systeme mit zusätzlichen Eigenschaften und Einschränkungen:

- beschränkte Ressourcen
- häufig nicht verbunden
- verlangt andere Transaktions Modelle
- verlangt andere Replikationsstrategien
- Ortsabhängigkeit

1.8. Semantic Web

Die aktuelle Web Infrastruktur unterstützt ein verteiltes Ge echt von Webseiten, die gegenseitig mittels den sog. Uniform Resource Locators (URL) verknüpft sein können:

- Das Semantic Web unterstützt das Web auf der Ebene der Daten statt nur auf der Ebene der Darstellung
- Datenelemente können gegenseitig verknüpft sein, nicht nur Webseiten
- Datenelemente können gegenseitig verknüpft sein, nicht nur Webseiten
- Information über einzelne Entitäten können verteilt sein
- Daten Modell: Resource Description Framework (RDF)

$$3^2 + 4x + 3x_{bla}$$

¹Consistency, Availability , Partition Tolerance, will never achieve all three

²Basically Available, Soft State, Eventual Consistency

³Protokolle : Verfahren definiert - Griechisches Parlament lausig - wie kann man zu Entscheidung kommen

2. Trigger

Was sind Trigger?

Prozeduren in der Datenbank hinterlegt. Ereignisse feuern die Trigger, keine explizite Ausführung.

Wozu dienen Trigger?

- Realisieren aktive Datenbanksysteme
- Berechnung abgeleiteter Attribute
- Überprüfen komplexer Integritätsbedingungen
- Implementierung von Geschäftsregeln
- Protokollierung, Statistiken
- Überprüfen von Integritätsbedingungen in verteilten
- Datenbanken Synchronisation von Replikaten

Trigger Konzept: ECA Prinzip - Event, Condition Action

2.1. Triggertypen

TriggerTypen (ereignis) DML-Trigger können auf Tabellen oder Views definiert werden. SQL-1999 beschreibt nur DML-Trigger auf Tabellen.

- DML: UPDATE, DELETE, INSERT
- DDL: CREATE, ALTER, DROP, ...
- Datenbank: SERVERERROR, LOGON, LOGOFF, STARTUP, SHUTDOWN, ...

TriggerTypen (Timing) • BEFORE Trigger

- AFTER Trigger
- INSTEAD OF Trigger

Der INSTEAD OF Trigger ist ein Spezialfall. Die Triggeraktion ersetzt den triggernden Befehl und wird nur bei Views eingesetzt. SQL-1999 beschreibt nur BEFORE und AFTER Trigger.

TriggerTypen (granularität)

- STATEMENT Trigger
- ROW Trigger

2.2. DML Trigger auf Tabellen

SQL-Skript

```
1 CREATE [OR REPLACE ] TRIGGER{ BEFORE|AFTER } events
2 [WHEN ( condition ) ]
3 pl / sql_block
```

SQL-Skript

```
1 { DELETE|INSERT|UPDATE
2 [OF column [ , column ] . . . ] }
3 [OR { DELETE|INSERT|UPDATE
4 [OF column [ , column ] . . . ] } ] . . .
5 ON table [FOR EACH ROW]
```



```

DECLARE
    Deklarationsteil

BEGIN
    Programmteil

EXCEPTION
    Ausnahmebehandlung

END;
/

```

Abbildung 2.1.: figure

2.3. PLSQL allgemein

Ausgabe Prozeduren

```

dbms_output.put_line (item IN VARCHAR2);
dbms_output.put_line (item IN NUMBER);
dbms_output.put_line (item IN DATE);
dbms_output.put (item IN VARCHAR2);
dbms_output.put (item IN NUMBER);
dbms_output.put (item IN DATE);
dbms_output.new_line;

```

Abbildung 2.2.: figure

mit EXECUTE

```

EXECUTE dbms_output.put_line ('Hello world');

```

als Block

```

BEGIN
    dbms_output.put_line ('Hello world');
END;

```

Abbildung 2.3.: figure

Datentypen:

- SQL Datentypen wie VARCHAR2, DATE, NUMBER
- reine PL/SQL Datentypen wie BOOLEAN, PLS_INTEGER
- strukturierte Datentypen wie TABLE, VARRAY, RECORD
- Datentypen für Spalten und Zeilen aus Tabellen %ROWTYPE, %TYPE

2.4. Trigger Beispiele

SQL-Skript

```

1  CREATE OR REPLACE TRIGGER BEFORE UPDATE ON filme
2  filme_stat_upd
3  DECLARE
4  msg VARCHAR2 ( 30 ) := ' filme_stat_upd ' ;
5  BEGIN
6  dbms_output . put _ line ( msg ) ;
7  END;

```

Zuweisung

```
variable := expression;
```

Die Zuweisung kann auch im Deklarationsteil verwendet werden.

im Deklarationsteil

```
DECLARE
  last_name  VARCHAR2(30) := 'Wyss';
  first_name VARCHAR2(25) := 'Bernhard';
```

Abbildung 2.4.: figure

If - then	If - then - elsif
<pre>IF condition THEN ... END IF;</pre>	<pre>IF condition THEN ... ELSIF condition THEN ... ELSE ... END IF;</pre>
<pre>IF condition THEN ... ELSE ... END IF;</pre>	

Abbildung 2.5.: figure

SQL-Skript

```
1  UPDATE v i d e o f i l m e
2  SET m i n d e s t a l t e r = 10
3  WHERE k a t a l o g n r = 2028
```

SQL-Skript

```
1  CREATE OR REPLACE TRIGGER filme_rows_upd
2  BEFORE UPDATE ON f i l m e
3  FOR EACH ROW
4  DECLARE
5  msg VARCHAR2 ( 3 0 ) : = ' filme_rows_upd ' ;
6  BEGIN
7  dbms_output . p u t _ l i n e ( msg ) ;
8  END;
```

SQL-Skript

```
1  CREATE OR REPLACE TRIGGER regdatum_test
2  BEFORE INSERT ON r e g i s t r i e r u n g e n
3  FOR EACH ROW
4  DECLARE
5  msg VARCHAR2 ( 3 0 ) : = 'Datum f a l s c h ' ;
6  BEGIN
7  I F : new . datum > SYSDATE THEN
8  RAISE_APPLICATION_ERROR(-20005,msg);
9  END I F ;
10 END;
```

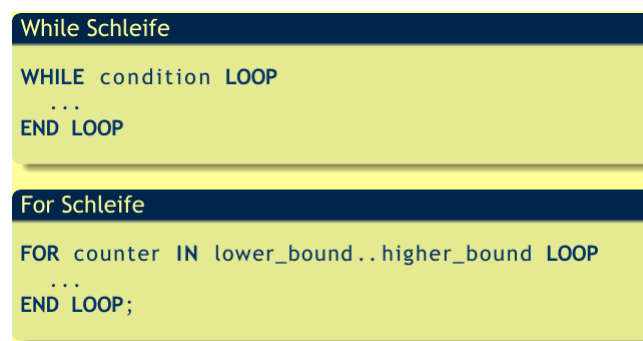


Abbildung 2.6.: figure

3. Entwurf Verteilte Datenbanken

BIKES					KUNDEN		
BNr	BName	Preis	Typ	Bestand	KNr	KName	Ort
B5	MCD03	4490.00	Road	2	K1	Bike Outlet	Basel
B4	Siena	2390.00	Mountain	4	K2	Swiss Bike	Olten
B2	City Cross	2190.00	Trekking	3	K3	Borer Velos	Basel
B3	Valiant	1090.00	Trekking	7	K4	City Bikes	Zürich
B1	Luxor	980.00	City	10	K5	MyBike	Aarau
B6	Atlanta	890.00	Trekking	8			
B7	Striker	890.00	Mountain	7			

AUFTRAEGE			APOSTEN		
ANr	Datum	KNr	ANr	BNr	Menge
A1	12.07.2009	K1	A1	B1	2
A2	01.09.2009	K2	A2	B1	4
A3	23.10.2009	K3	A2	B2	1
A4	05.11.2009	K4	A3	B3	3
A5	10.11.2009	K2	A3	B4	1
A6	19.11.2009	K1	A4	B5	1
A7	14.12.2009	K1	A5	B6	2
A8	17.12.2009	K2	A6	B6	4
			A7	B7	3
			A8	B3	4

Abbildung 3.1.: Beispiel DB

Anwendungen auf verschiedenen Knoten benötigen verschiedene Daten, unterschiedlich häufig:

A1: Liste der Aufträge mit günstigen Bikes (auf ORION in Basel)

A2: Änderungen an Bikes (auf CALYPSO in Aarau)

A3: Liste der Aufträge mit teuren Bikes vom Typ Mountain (auf TELESTO in Olten)

A4: Liste aller Kunden (auf ANANKE in Zürich)

Die Daten sollten dort abgelegt sein, wo sie am häufigsten gebraucht werden.

3.1. Einführung

Anweisungen ■ Entscheid über die Platzierung von Daten auf den Knoten eines Computernetzwerks

- beeinflusst die Performance der DDB und der Anwendungen
- lokaler Zugriff ist günstiger als Zugriff auf entfernte Knoten
- Analyse:
 - welche Anwendungen (Queries)
 - auf welchen Knoten
 - benötigen welche Daten
 - mit welcher Häufigkeit

- Resultat:
Menge von Fragmenten (Ausschnitte der Daten)
zugeteilt auf verschiedene Knoten

3.2. Entwürfe

3.2.1. Top Down

beim Entwurf from scratch in homogenen Systemen nachgelagert an den konzeptionellen Entwurf.

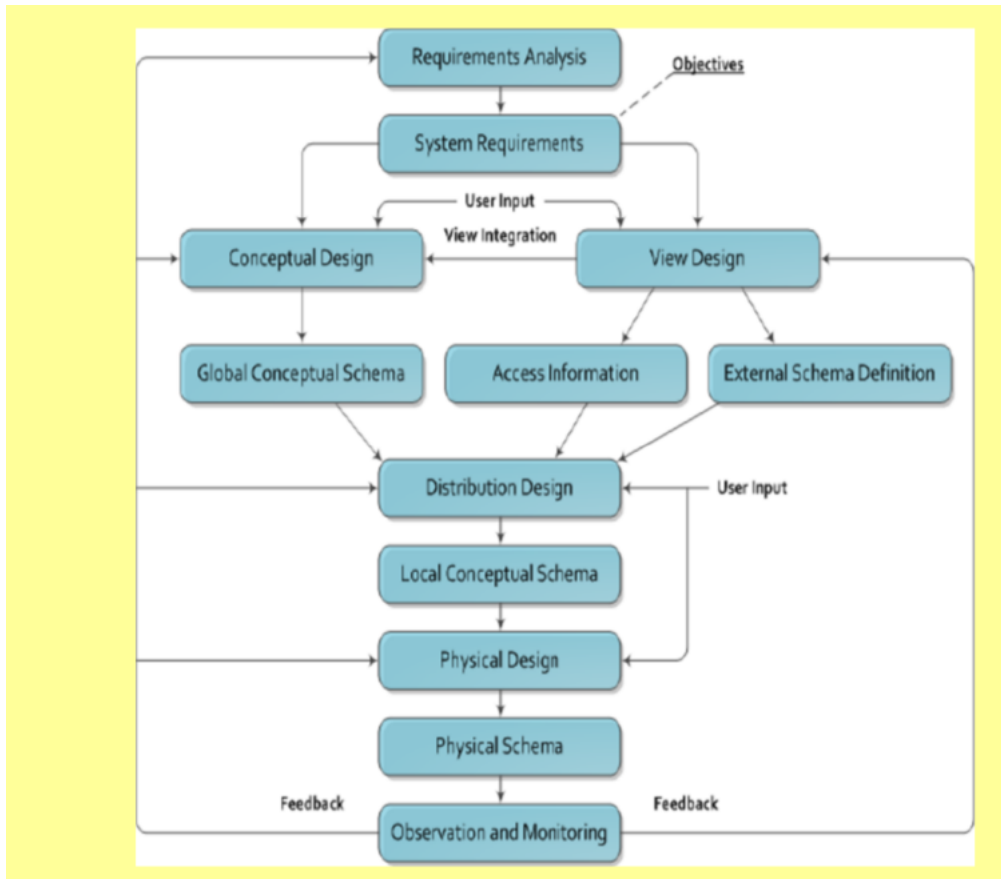


Abbildung 3.2.: figure

1. wozu überhaupt fragmentieren?
2. wie fragmentieren?
3. wie viel fragmentieren?
4. wie prüfen der Korrektheit?
5. wie Fragmente zuteilen?
6. welche Informationen werden benötigt?

Grad der Fragmentierung

- Vollständige Relationen sind zu grob, einzelne Attributswerte sind zu fein.
geeignete Teile (Fragmente) der Relationen bestimmen
Nebenläufigkeit steigern bei mehreren Transaktionen mit Zugriff auf verschiedene Fragmente
Zusatzkosten für Transaktionen mit Zugriff auf Fragmente an verschiedenen Knoten

Horizontal

Zerlegt aufgrund Riehen, mit Where klausel.

Projektion - Vertikal

Spaltenliste. In allen Fragmente PK muss dabei sein damit es rekonstruierbar ist- Kann kombiniert werden mit Horizontal für optimale Verteilung.

3.2.2. Bottom Up

- Multidatenbank Anwendungen
- Datenbanken schon auf verschiedenen Knoten vorhanden
- Problem der Datenintegration, Schemaintegration
- Web Services

3.3. Korrektheit der Fragmentierung

Vollständig Wenn R zerlegt wird $R_1, R_2 \dots R_n$ dann muss jedes Datenelement aus R in einem R_i enthalten sein. (Join mit gemeinsamen Pk - wiederzusammenstellung). (Horizontal Umkehroperation - Union)

Rekonstruierbar Wenn R zerlegt wird in , dann muss es $R_1, R_2 \dots R_n$ relationale Operatoren geben, so dass R wiederhergestellt werden kann.

Disjunkt wenn R horizontal zerlegt wird in $R_1, R_2 \dots R_n$, dann müssen die Fragmente paarweise disjunkt sein. wenn R vertikal zerlegt wird in $R_1, R_2 \dots R_n$, dann müssen die Fragmente bezogen auf die nichtprimen Attribute paarweise disjunkt sein.

3.4. Fragmentierungsabkürzungen

- Horizontale Fragmentierung (HF)
 - Primäre horizontale Fragmentierung (PHF)
 - Abgeleitete horizontale Fragmentierung (DHF) Fragmente gehören zu andere Fragmente wegen PK oder FK. Ähnliche Fragmentierung sinnvoll.
- Vertikale Fragmentierung (VF)
- Gemischte Fragmentierung (MF)

3.5. PHF

Quantitative Informationen:

Bedingungen in Queries bestehen aus:

Simple Predicate p Vergleich eines Attributs mit einem Wert : Name = "bla" Preis < 2000. Allgemein bestehende Bedingungen aus booleschen Kombination von simple Predicates : $P = \{p_1, p_2 \dots, p_m\}$ vom simple predicates bilden wir

Mintem predicate $M(P)$ Verknüpfungen aller simple predicates aus P mit AND und NOT. bname= bla and not preis < 2000 z.B

Bemerkung :

p_1 : bname = 'Sienna'

p_2 : preis < 2000

$p_1 \wedge p_2$

$p_1 \wedge \neg p_2$

$\neg p_1 \wedge p_2$

$\neg p_1 \wedge \neg p_2$

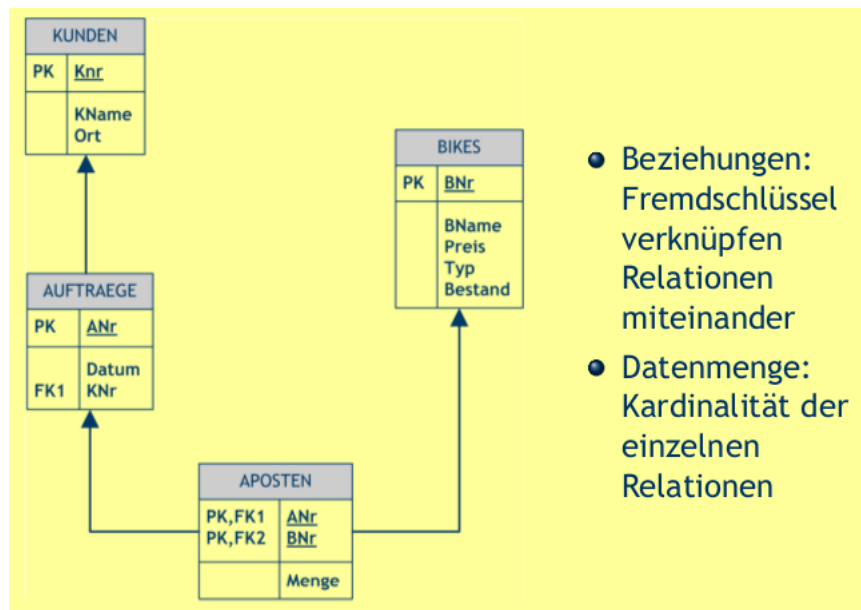


Abbildung 3.3.: figure

Quantitative Informationen : Brauche ich das minterm oder nicht?

Minterm selectivity $sel(m_i)$ Anzahl Tupel die mit dem Minterm m_i ausgewählt werden.

Access frequency $acc(m_i)$ Häufigkeit der Anwendungen auf Daten mit dem minterm m_i zugreifen

3.6. PHF Schritt für Schritt

Vollständigkeit Eine Menge von simple predicates ist vollständig genau dann, wenn auf beliebige 2 Tupel im gleichen Fragment von allen Anwendungen mit der gleichen Häufigkeit zugegriffen wird

Minimalität Wird durch ein simple predicate ein Fragment weiter aufgeteilt, dann muss es mindestens eine Anwendung geben, die auf diese Fragmente verschieden zugreift Ein simple predicate soll also relevant sein für die Bestimmung einer Fragmentierung Sind alle simple predicate eine Menge P relevant, dann ist P minimal

- Menge von Simple predicates mit der Eigenschaft der Vollständigkeit und Minimalität.

3.7. PHF Beispiel Bikes

Zugriffshäufigkeit gegeben je nach art des Fahrrads.

Anwendung 2 : Bikes aufgrund preise verwaltet.

Predicates aus beobachtungen

Anwendung 1: Zugriffshäufigkeit der Anwendung in excel, schauen. Wenn grosse verschieden - schlussfolgerung kann unterscheiden durch eine WHERE bedingung Anwendung 2 : Preis spielt hier eine rolle - ist relevante Bedingungen. Simple predicates aufnehmen die RELEVANT sind.

Verfahren erklären : Minimale Menge von simple predicates ausarbeiten aufgrund der Beobachtung Anwendungen. wähle eines aus von allen predicates das relevant ist - typ = mounten. ergänze meine menge p' um die nächsten nur relevanten

sehe phf beispiel.

p1..p6 = P p1 ist relevant wege häufigkeit.

p2, p3 hinzufügen relevant

p4 nicht relevant weil : typ =city kann ich ausdrücken im minterm von anderen (Restmenge) p4 ist aber einfacher zu benutzen.

entweder p5, oder p6 aufnehmen sehe oben. MINIMALE MENGE. 2 HOCH 6 Kombinationen. durch impl und Widersprüche nur 8 sinnvoll.

Beispiel :

Typ = Road und Typ = Mountain ist ein Widerspruch. Bike für beides geht nicht.



Abbildung 3.4.: figure

Impl beispiel : Typ = road folgt nicht alle anderen Typen. alle Extrabedingungen mit nots reduzieren.
(Schritt 3) Fragmente bilden siehe Folie.

homogen zugriff auf Fragment ?? Nutzung alle Tupel gleich. Zugriffshäufigkeit konform pro Fragment.
wichtigsten Anwendung betrachte 20 Prozent 80 Prozent last. kleinste sinnvolle Einheit.

3.8. DHF

- ist eine horizontale Fragmentierung, die auf einer horizontalen Fragmentierung einer übergeordneten Relation basiert
- stellt sicher, dass Fragmente verschiedener Relationen, auf die häufig im Verbund zugegriffen wird, dem gleichen Knoten zugeteilt werden
- definieren mit einem Semijoin (Dreieckzeichen) ¹: Kann als normalen Join geschrieben werden : Projektion $\pi_{knr, datum, auftrage.knr} (\text{Aufträge} \bowtie \text{Kunden}_1)$

¹Mit subqueries definiert: `SELECT * FROM AUFTRÄGE WHERE EXISTS (SELECT 1 FROM KUNDEN1 WHERE KUNDEN1.KNR = AUFTRÄGE.KNR)`

- wird Relation KUNDEN fragmentiert durch
KUNDEN₁: $\sigma_{\text{Ort} = \text{'Basel'}}$ (KUNDEN)
KUNDEN₂: $\sigma_{\text{Ort} = \text{'Olten'}}$ (KUNDEN)
KUNDEN₃: $\sigma_{\text{Ort} = \text{'Zürich'}}$ (KUNDEN)
KUNDEN₄: $\sigma_{\text{Ort} = \text{'Aarau'}}$ (KUNDEN)
- dann wird AUFTRAEGE fragmentiert mit
AUFTRAEGE₁: (AUFTRAEGE) \triangleright (KUNDEN₁)
AUFTRAEGE₂: (AUFTRAEGE) \triangleright (KUNDEN₂)
AUFTRAEGE₃: (AUFTRAEGE) \triangleright (KUNDEN₃)
AUFTRAEGE₄: (AUFTRAEGE) \triangleright (KUNDEN₄)

Abbildung 3.5.: figure

4. Vertikale Fragmentierung

Geht um "ähnliche Nutzung".

- Vertikale Fragmentierung einer einzelnen Relation
- Modellieren des Zugriffs auf die Relationen
 - Welche Anwendungen (Queries) verwenden welche Attribute
- und werden auf welchen Knoten
- wie häufig ausgeführt
- mit dem Ziel zu entscheiden
 - welche Attribute werden häufig zusammen verwendet und sollen deshalb zum selben Fragment gehöre

BIKES					KUNDEN		
BNr	BName	Preis	Typ	Bestand	KNr	KName	Ort
B5	MCD03	4490.00	Road	2	K1	Bike Outlet	Basel
B4	Siena	2390.00	Mountain	4	K2	Swiss Bike	Olten
B2	City Cross	2190.00	Trekking	3	K3	Borer Velos	Basel
B3	Valiant	1090.00	Trekking	7	K4	City Bikes	Zürich
B1	Luxor	980.00	City	10	K5	MyBike	Aarau
B6	Atlanta	890.00	Trekking	8			
B7	Striker	890.00	Mountain	7			

AUFTRÄGE			APOSTEN		
ANr	Datum	KNr	ANr	BNr	Menge
A1	12.07.2009	K1	A1	B1	2
A2	01.09.2009	K2	A2	B1	4
A3	23.10.2009	K3	A2	B2	1
A4	05.11.2009	K4	A3	B3	3
A5	10.11.2009	K2	A3	B4	1
A6	19.11.2009	K1	A4	B5	1
A7	14.12.2009	K1	A5	B6	2
A8	17.12.2009	K2	A6	B6	4
			A7	B7	3
			A8	B3	4

Abbildung 4.1.: figure

4.1. Zugriffseigenschaften

Die Verwendung (Usage) der Attribute durch diese Anwendungen (Queries) wird in einer Matrix dargestellt: Matrix U.

	BName	Preis	Typ	Bestand
q_1	1	0	0	1
q_2	0	1	0	1
q_3	0	1	1	0
q_4	0	0	1	1

eine 1 bedeutet: Query verwendet Attribut eine 0 bedeutet: Query verwendet Attribut nicht

Anwendung wird ebenfalls in einer Matrix festgehalten: Matrix Acc.

Für die Relation BIKES sind die Anwendungen als Queries gegeben:

q₁ SELECT bestand FROM bikes WHERE bname = ?	q₂ SELECT bestand, preis FROM bikes
q₃ SELECT preis FROM bikes WHERE typ = ?	q₄ SELECT AVG(bestand) FROM bikes WHERE typ = ?

Abbildung 4.2.: figure

	S1	S2	S3
q ₁	15	20	10
q ₂	5	0	0
q ₃	25	25	25
q ₄	3	0	0

die Zahl 15 oben links besagt z.B.: Query q₁ wird auf Knoten S1 15 Mal pro Tag ausgeführt

4.2. Cluster Methode

- Der Zugriff auf jede Attributsmenge durch die Anwendungen ist sehr unterschiedlich
Dadurch würde jedes Attribut sein eigenes Fragment bilden, was wenig sinnvoll ist
Besser wird nach Attributsmengen gesucht, auf die ähnlich zugegriffen wird
- Die Clustering Methode erkennt ähnliche Zugriffsmuster auf Attribute
Für jedes Paar von Attributen bestimmen, wie oft auf sie gemeinsam zugegriffen wird
Cluster bilden von Attributen mit hoher Affinität

4.2.1. Affinität

$$aff(A_i, A_j) = \sum_{k: U_{k_i}=1, U_{k_j}=1} \sum_{l=1}^s acc_{kl} \quad (4.1)$$

	BName	Preis	Typ	Bestand
BName	45	0	0	45
Preis	0	80	75	5
Typ	0	75	78	3
Bestand	45	5	3	53

Matrix AA Zeigt Affinität zweier Attribute :

Skalarprodukt benachbarter Spalten misst die Ähnlichkeit des Zugriffsmusters: Affinität der Nachbarschaft (bond):

$$\sum_{z=1}^n aff(A_z, A_x) * aff(A_z, A_y) \quad (4.2)$$

Die Summe über alle Skalarprodukte ist die globale Affinität der Nachbarschaft

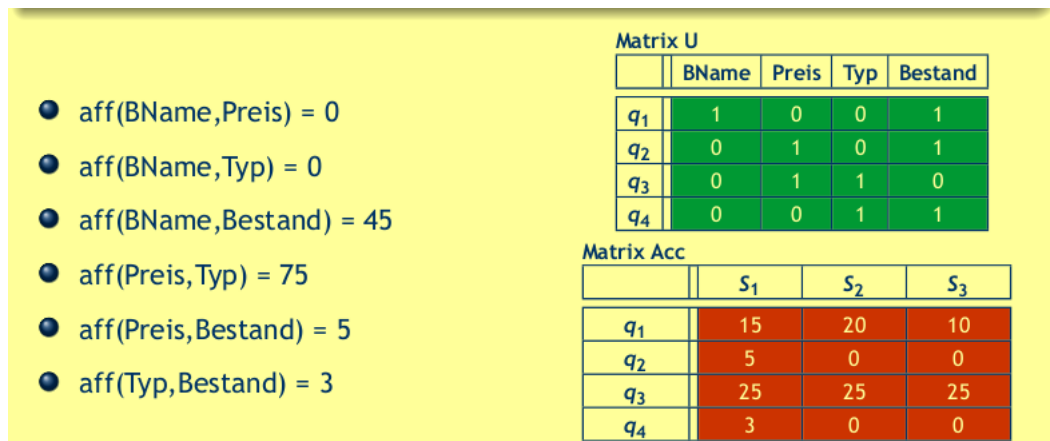


Abbildung 4.3.: figure

BName, Preis: $45 \times 5 = 225$

Preis, Typ:

$80 \times 75 + 75 \times 78 + 5 \times 3 = 11865$

Typ, Bestand: $75 \times 5 + 78 \times 3 + 3 \times 53 = 768$

globale Affinität: 12858

Vertauschen ist auch möglich:

4.2.2. Vertauschen bei Affinität

Durch Austausch von Spalten (und den entsprechenden Zeilen) in der Matrix AA verändert sich die globale Affinität der Nachbarschaft:

	BName	Preis	Typ	Bestand
BName	45	0	0	45
Preis	0	80	75	5
Typ	0	75	78	3
Bestand	45	5	3	53

Preis, BName: $5 \times 45 = 225$

BName, Typ: $45 \times 3 = 135$

Typ, Bestand:

$75 \times 5 + 78 \times 3 + 3 \times 53 = 768$

Globale Affinität: 1128

4.3. Bond Energie Algorithmus

- Gegeben $n \times n$ Matrix AA der Affinitäten
Beliebige 2 Spalte aus AA wählen und in Resultats Matrix CA stellen
- Iteration:
eine der übrigen $n - i$ Spalten so in Resultats Matrix positionieren ($i + 1$ mögliche Positionen), dass sich der grösste Beitrag an die globale Affinität der Nachbarschaft ergibt
Die Zeilen entsprechend den Spalten anordnen
- Beitrag einer Spalte A_k wenn zwischen A_i und A_j platziert:
 $\text{cont}(A_i, A_k, A_j) = \text{bond}(A_i, A_k) + \text{bond}(A_k, A_j) - \text{bond}(A_i, A_j)$

4.4. VF - Vertical Fragmentation Example

Spalte Preis an 1. Position platzieren und Beitrag an globaler Affinität berechnen:

$\text{cont}(_, \text{Preis}, \text{BName}) = 0 + 225 - 0 = 225$

Matrix AA				
	BName	Preis	Typ	Bestand
BName	45	0	0	45
Preis	0	80	75	5
Typ	0	75	78	3
Bestand	45	5	3	53

Abbildung 4.4.: Gegebene Matrix der AA Affinitäten

Matrix CA				
	BName	Typ		
BName	45	0		
Preis	0	75		
Typ	0	78		
Bestand	45	3		

Abbildung 4.5.: Aus der Affinitäts Matrix AA werden die Spalten BName und Typ beliebig gewählt

Spalte Preis an 2. Position platzieren und Beitrag an globaler Affinität berechnen:

$\text{cont}(\text{BName}, \text{Preis}, \text{Typ}) = 225 + 11865 - 135 = 11955$

Spalte Preis an 3. Position platzieren und Beitrag an globaler Affinität berechnen:

$\text{cont}(\text{Typ}, \text{Preis}, _) = 11865 + 0 - 0 = 11865$

Preis an der 2. Position bietet den grössten Beitrag.

4.5. VF : Cluster möglichkeiten

- Trennpunkt kann entlang der Diagonalen verschoben werden
- Für jede der Möglichkeiten berechnen, wie oft:
 - ausschliesslich auf Attribute eines Fragments zugegriffen wird (gut)
 - auf Attribute beider Fragmente zugegriffen wird (schlecht)
- Trennqualität $sq = acc(VF_1) \times acc(VF_2) - acc(VF_1, VF_2)^2$ maximieren
 - $acc(VF_1)$: Summe Zugriffshäufigkeiten der Queries ausschliesslich auf VF 1
 - $acc(VF_2)$: Summe Zugriffshäufigkeiten der Queries ausschliesslich auf VF 2
 - $acc(VF_1, VF_2)$: Summe Zugriffshäufigkeiten der Queries auf VF 1 und VF 2

4.6. Schluss : VF

Aus der erhaltenen Clustering der Attribute, ergänzt mit dem Primärschlüssel ergeben sich die Fragmente:

BIKES 1 : $\pi_{B\#r, B\#ame, Bestand}(BIKES)$

BIKES 2 : $\pi_{B\#r, Preis, Typ}(BIKES)$

- Mit der Spalte *Bestand* ebenso verfahren
- Die 2. Position für *Bestand* gibt den grössten Beitrag
- Anordnen der Zeilen entsprechend den Spalten

In der Resultats Matrix CA sind die Clusters erkennbar für eine optimale vertikale Fragmentierung
Es sind aber noch 2 weiter Clusterungen möglich
Wie kann die optimale berechnet werden?

Matrix CA

	BName	Bestand	Preis	Typ
BName	45	45	0	0
Bestand	45	53	5	3
Preis	0	5	80	75
Typ	0	3	75	78

Matrix CA

	BName	Bestand	Preis	Typ
BName	45	45	0	0
Bestand	45	53	5	3
Preis	0	5	80	75
Typ	0	3	75	78

Abbildung 4.6.: figure

1. Variante:

VF_1 : BName

VF_2 : Bestand, Preis, Typ

	BName	Bestand	Preis	Typ
BName	45	45	0	0
Bestand	45	53	5	3
Preis	0	5	80	75
Typ	0	3	75	78

Matrix CA

	BName	Bestand	Preis	Typ
BName	45	45	0	0
Bestand	45	53	5	3
Preis	0	5	80	75
Typ	0	3	75	78

Matrix CA

	BName	Bestand	Preis	Typ
Name	45	45	0	0
Bestand	45	53	5	3
Preis	0	5	80	75
Typ	0	3	75	78

2. Variante:

VF_1 : BName, Bestand

VF_2 : Preis, Typ

3. Variante:

VF_1 : BName, Bestand, Preis

VF_2 : Typ

Abbildung 4.7.: figure

Verwenden der Matrix U und Acc

Matrix U

	BName	Preis	Typ	Bestand
q_1	1	0	0	1
q_2	0	1	0	1
q_3	0	1	1	0
q_4	0	0	1	1

Matrix Acc

	S_1	S_2	S_3
q_1	15	20	10
q_2	5	0	0
q_3	25	25	25
q_4	3	0	0

1. Variante:

$$acc(VF_1) = 0$$

$$acc(VF_2) = 83$$

$$acc(VF_1, VF_2) = 45$$

$$sq = -45^2$$

2. Variante:

$$acc(VF_1) = 45$$

$$acc(VF_2) = 75$$

$$acc(VF_1, VF_2) = 8$$

$$sq = 45 * 75 - 8^2$$

3. Variante:

$$acc(VF_1) = 50$$

$$acc(VF_2) = 0$$

$$acc(VF_1, VF_2) = 78$$

$$sq = -78^2$$

Variante 2 hat demnach die höchste Trennqualität.

Abbildung 4.8.: figure

5. Verteilte Anfrage Verarbeitung

Beispiel Query: $KUN_1 = \pi_{KNR, KNAME}(KUN)$
 $KUN_2 = \pi_{KNR, ORT}(KUN)$
 $KUN = KUN_1 \bowtie KUN_2$

5.1. Query Processor

Umformen der SQL Queries in ausführbare Operationen für den Datenzugriff:

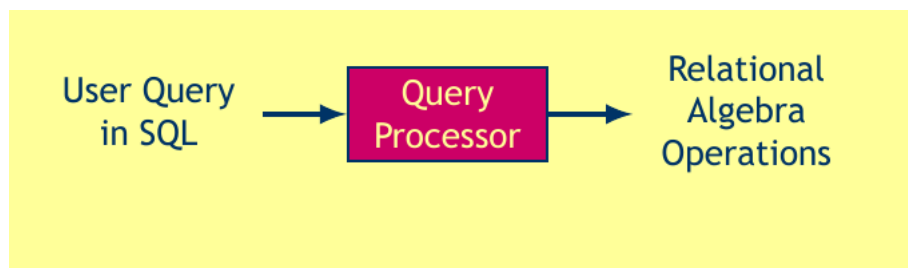


Abbildung 5.1.: figure

5.2. Zentralisiert vs Dezentralisiert

Zentralisiertes System

- Umformen der SQL Anfrage in ausführbare Einheiten - Operationen der Rationalen Algebra.

- Auswählen des besten Ausführungsplans

verteiltes System

- berücksichtigen der Kommunikationskosten

- auswählen des besten Knotens.

Relationen werden mit den ersten drei Buchstaben von der DB Tabellennamen zukünftig benannt.

5.3. Überblick

5.3.1. Ziel der Optimierung

- durch Umformen und Optimieren soll eine Kostenfunktion minimiert werden
 - CPU Zeit
 - IO Zeit
 - Comm Zeit
- verschiedene Gewichte in unterschiedlich verteilten Umgebungen. WAN - Kommzeit LAN - ähnlich gewichtet.

5.3.2. Komplexität

$\theta\pi$ mit Duplikate : $O(n)$

π Ohne Duplikate - GROUP $O(n \log n)$

$\bowtie \cup \cap \div$: $O(n \log n)$

\times : $O(n^2)$

Query auf Knoten 5

- SQL:
 SELECT typ
 FROM bikes JOIN aposten USING (bnr)
 WHERE menge > 2
- Relationale Algebra:
 $\pi_{\text{Typ}}(\text{BIK} \bowtie (\sigma_{\text{Menge} > 2}(\text{APO})))$

Fragmente, Zuteilung

- BIK_1 : $\text{BNr} \leq 'B3'$: Knoten 3
- BIK_2 : $\text{BNr} > 'B3'$: Knoten 4
- APO_1 : $\text{BNr} \leq 'B3'$: Knoten 1
- APO_2 : $\text{BNr} > 'B3'$: Knoten 2

Abbildung 5.2.: figure

5.4. Methode

5.4.1. Verteilte Verarbeitung

5.4.2. Zerlegung

Sql Query zerlegen und umformen in relationale Algebra unter Verwendung des globalen Schemas :

- Normalisierung (Bedingung in Where Klausel)
- Analyse um inkorrekte Queries zurückzuwechseln.
 Analyse nach Typ
 Analyse nach Semantik
- Vereinfachung, Redundanz beseitigen.
- Umformen in optimalen Ausdruck der relationalen Algebra.

5.5. Lokalisierung

- verwenden des Fragmentierungsschema.
- verteilte Anfrage mit globalen Relationen abbilden in Anfragen mit Fragmenten:
 Ersetzen der globalen Relation mit den Fragmenten :
 \cup für Horizontale Fragmentierung
 \bowtie für Vertikale Fragmentierung.
- Optimierung der lokalisierten Anfrage durch Reduktion mit Selektion oder Join.

5.5.1. PHF

Bemerkung: Eine Selektion auf einem Fragment mit einer Bedingung, die der Bedingung für die Fragmentierung widerspricht , ergibt leere Resultsrelationen.

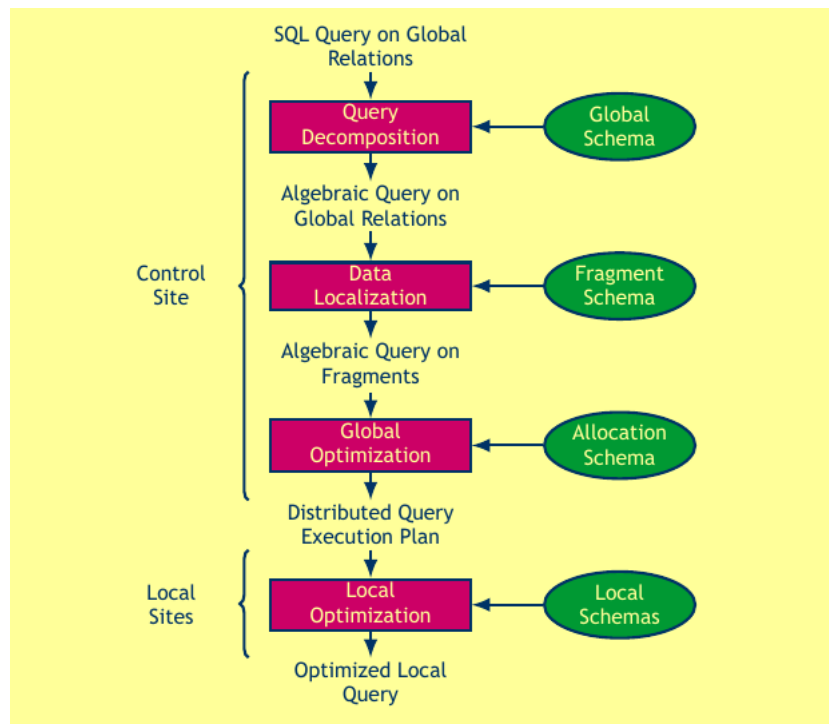


Abbildung 5.3.: figure

Beispiel

AUF(ANr, Datum, KNr) ist fragmentiert:

$$AUF_1 = \sigma_{ANr \leq A3}(AUF)$$

$$AUF_2 = \sigma_{A3 < ANr \leq A6}(AUF)$$

$$AUF_3 = \sigma_{ANr > A6}(AUF)$$

Rekonstruktion mit:

$$AUF = AUF_1 \cup AUF_2 \cup AUF_3$$

Abbildung 5.4.: figure

5.5.2. Schritten für Lokalisierung PHF

Schritt 1 globaler Query Baum erstellen

Schritt 2 globale Relation mit Fragmenten ersetzen.

Schritt 3 Reduktion:

Eine Reduktion ist möglich wenn die Fragmentierung mit dem Join Attribute erfolgte. Wenn sich die Bedingungen der Fragmente widersprechen, resultiert eine leere Relation. Vor einer möglichen Reduktion wird das Distributivgesetz angewendet :

$$(R_1 \cup R_2) \bowtie S = (R_1 \bowtie S) \cup (R_2 \bowtie S)$$

5.5.3. Reduktion mit Join für PHF

5.6. Reduktion für VF

Eine Projektion auf einem Fragment ist nutzlos wenn die Attribute der Projektion nicht im Fragment enthalten sind.

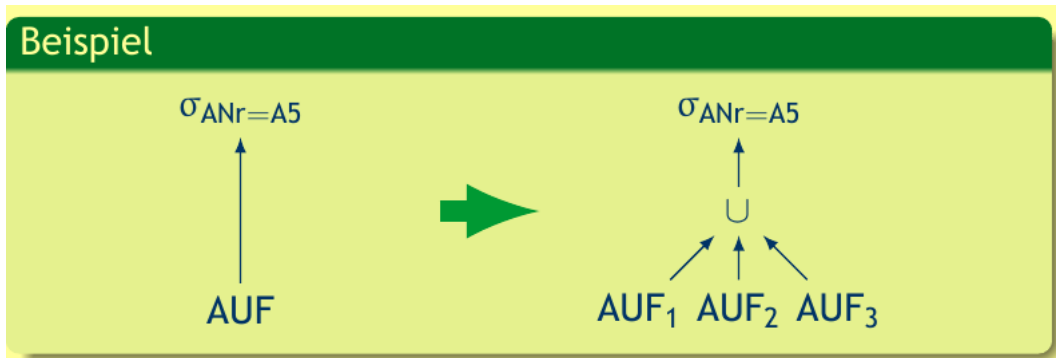


Abbildung 5.5.: figure

$ANr = A5$ widerspricht $ANr \leq A3$ (Fragment AUF_1) und $ANr > A6$ (Fragment AUF_3)

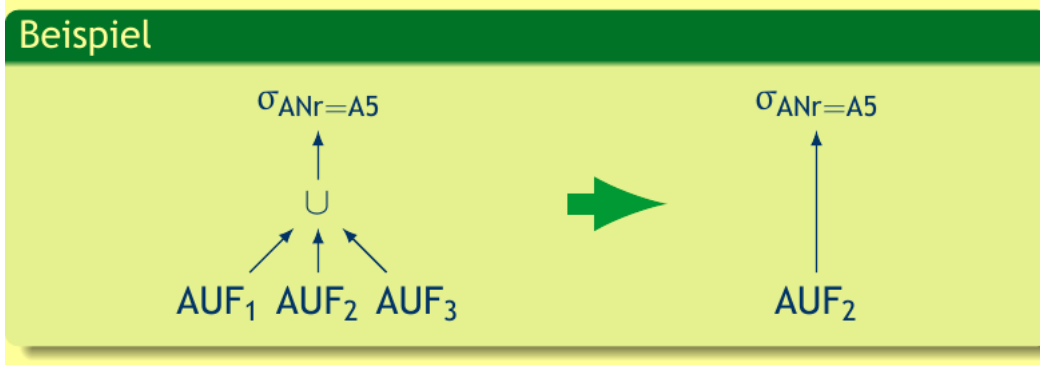


Abbildung 5.6.: figure

Beispiel

AUFTRAEGE sei fragmentiert wie vorher
 APOSTEN(ANr , BNr , Menge) ist fragmentiert:

$APO_1 = \sigma_{ANr \leq A3}(APO)$
 $APO_2 = \sigma_{ANr > A3}(APO)$

Rekonstruktion mit:

$APO = APO_1 \cup APO_2$

Abbildung 5.7.: figure

Beispiel

- SQL:
`SELECT *`
`FROM auftraege JOIN aposten USING (anr)`
- Relationale Algebra:
 $AUF \bowtie APO$

Abbildung 5.8.: figure

- Schritt 1: globaler Query Baum erstellen
- Schritt 2: globale Relation mit Fragmenten ersetzen

Beispiel

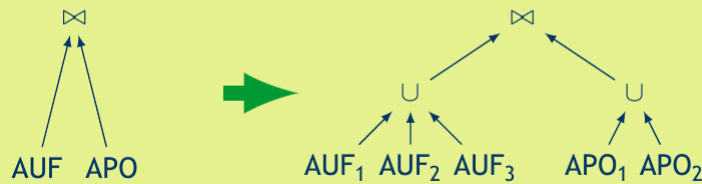


Abbildung 5.9.: figure

- Schritt 3: „Ausmultiplizieren“ und Reduzieren

$ANr \leq A3$ (AUF₁) widerspricht $ANr > A3$ (APO₂)

$A3 < ANr \leq A6$ (AUF₂) widerspricht $ANr \leq A3$ (APO₁)

$ANr > A6$ (AUF₃) widerspricht $ANr \leq A3$ (APO₁)

Beispiel

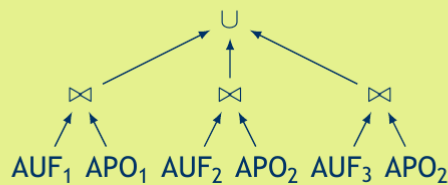


Abbildung 5.10.: figure

Beispiel

KUNDEN(KNr, KName, Ort) ist fragmentiert:

$KUN_1 = \pi_{KNr, KName}(KUN)$

$KUN_2 = \pi_{KNr, Ort}(KUN)$

Rekonstruktion mit:

$KUN = KUN_1 \bowtie KUN_2$

Abbildung 5.11.: figure

- Schritt 1: globaler Query Baum erstellen
- Schritt 2: globale Relation mit Fragmenten ersetzen
- Schritt 3: Reduktion

Beispiel

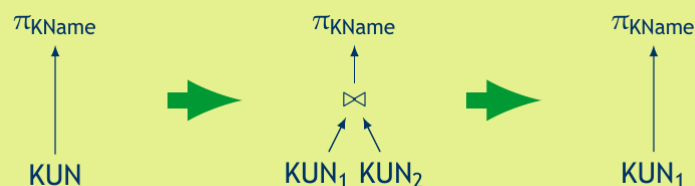


Abbildung 5.12.: figure

Beispiel

KUNDEN(KNr, KName, Ort) ist fragmentiert:

$$KUN_1 = \sigma_{\text{Ort}=\text{Basel}}(KUN)$$

$$KUN_2 = \sigma_{\text{Ort} \neq \text{Basel}}(KUN)$$

AUFTRAEGE(ANr, Datum, KNr) ist abgeleitet fragmentiert:

$$AUF_1 = AUF \triangleright KUN_1$$

$$AUF_2 = AUF \triangleright KUN_2$$

Rekonstruktion mit:

$$KUN = KUN_1 \cup KUN_2$$

$$AUF = AUF_1 \cup AUF_2$$

Abbildung 5.13.: figure

Anwenden der Join Reduktion**Beispiel**

- SQL:

```
SELECT *  
FROM kunden JOIN auftraege USING (knr)  
WHERE ort = 'Olten'
```

- Relationale Algebra:

$$AUF \bowtie (\sigma_{\text{Ort}=\text{Olten}}(KUN))$$

Abbildung 5.14.: figure

6. Concurrency

6.1. Notation + Definitionen für Concurrency

Notation	Bedeutung
$r_i[x]$	Read(x) durch Transaktion T_i
$w_i[x]$	Write(x) durch Transaktion T_i
$o_i[x]$	Read(x) oder Write(x) durch T_i
c_i	Commit durch Transaktion T_i
a_i	Commit durch Transaktion T_i
e_i	Commit oder Abort durch T_i

Transaktion Folge von Read und Write Operationen auf beliebigen Datenelementen x. Letzte Operation entweder Commit oder Abort.

Schedule, History ist eine Folge von Datenbank Scheduler auf der Datenbank ausgeführten Lese und Schreiboperationen. Es besteht aus Operationen verschiedener Transaktionen. Ordnung der Operationen innerhalb einer Transaktion muss beibehalten werden.

Serielle History Eine History H heisst seriell, wenn jeweils alle Operationen einer Transaktion direkt hintereinander ausgeführt werden (keine überlappte Ausführung der Transaktionen).

Konflikt Operation Zwei Datenbank-Operationen o_i und o_k heissen Konflikt-Operationen wenn :

- beide auf dasselbe Element x zugreifen.
- sie zu verschiedenen Transaktionen gehören.
- mindestens eine der beiden eine Schreiboperation ist.

Konflikt Serialisierbarkeit Eine History heisst konflikt-serialisierbar , wenn es eine serielle History mit gleicher Reihenfolge der Konflikt-Operationen gibt.

Recoverable History Eine History heisst Recoverable falls alle Commits der Transaktionen, von denen eine Transaktion T_i liest vor dem c_i erfolgen.

Vermeiden von cascading Aborts Eine History vermeidet cascading Aborts falls alle Commits der Transaktionen, von denen eine Transaktion T_i liest vor dem $r_i[x]$ erfolgen.

Strikte History Eine History heisst strikt falls jedes $r_i[x]$ oder $w_i[x]$ einer Transaktion T_i nach den Commits aller anderen Transaktionen erfolgt, die auf x schreiben haben.

2 Phasen Sperrprotokoll Two Phase Lock : Eine Transaktion folgt dem 2 Phasen Sperrprotokoll

- vor einem Read bzw Write muss ein share Lock bzw exclusive lock verlangt und erlangt werden.
- die Sperre darf erst freigegeben werden wenn die zugehörige Operation vollständig durchgeführt worden ist.
- nach der ersten Freigabe einer Sperre darf keine neue Sperre verlangt werden.

Striktes 2 Phasen Sperrprotokoll • Alle Sperren müssen auf einmal mit dem Commit atomar freigegeben werden.

6.2. Einführung

Atomicity Alles oder Nichts

Consistency keine Verletzung von Integritätsregeln

Isolation keine Nebeneffekte bei Nebenläufigkeit

Durability Bestätigte Änderungen sind persistent

Transaktionen bieten :

- unteilbare und zuverlässige Ausführung auch bei Ausfällen
- korrekte Ausführung auch bei gleichzeitigem Zugriff durch mehrere Benutzer
- korrekte Verwaltung von Replikaten (falls unterstützt)

6.2.1. Aspekte der Transaktionsverarbeitung

- Transaktionsstruktur (Transaktionsmodell)
 - flache Transaktion, verschachtelte Transaktion
- Konsistenzerhaltung der Datenbank
- Zuverlässigkeitsprotokolle
 - Unteilbarkeit, Dauerhaftigkeit
 - lokale Wiederherstellung
 - globale Commit Protokolle
- Nebenläufigkeitskontrolle
 - Ausführung nebenläufiger Transaktionen (Korrektheit?)
 - Konsistenzerhaltung zwischen Transaktionen, Isolation
- Kontrolle der Replikate
 - Kontrolle der gegenseitigen Konsistenz von Replikaten
- One-Copy Equivalence, ROWA

6.3. Architektur

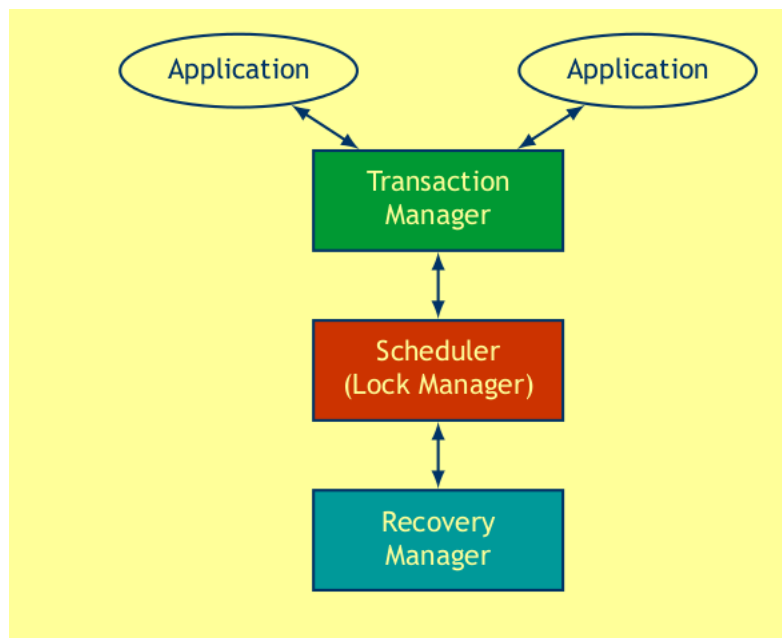


Abbildung 6.1.: figure

6.4. Nebenläufigkeit

Nebenläufigkeitskontrolle wie in zentralisierten DBS

- Ablaufplan, History, Schedule

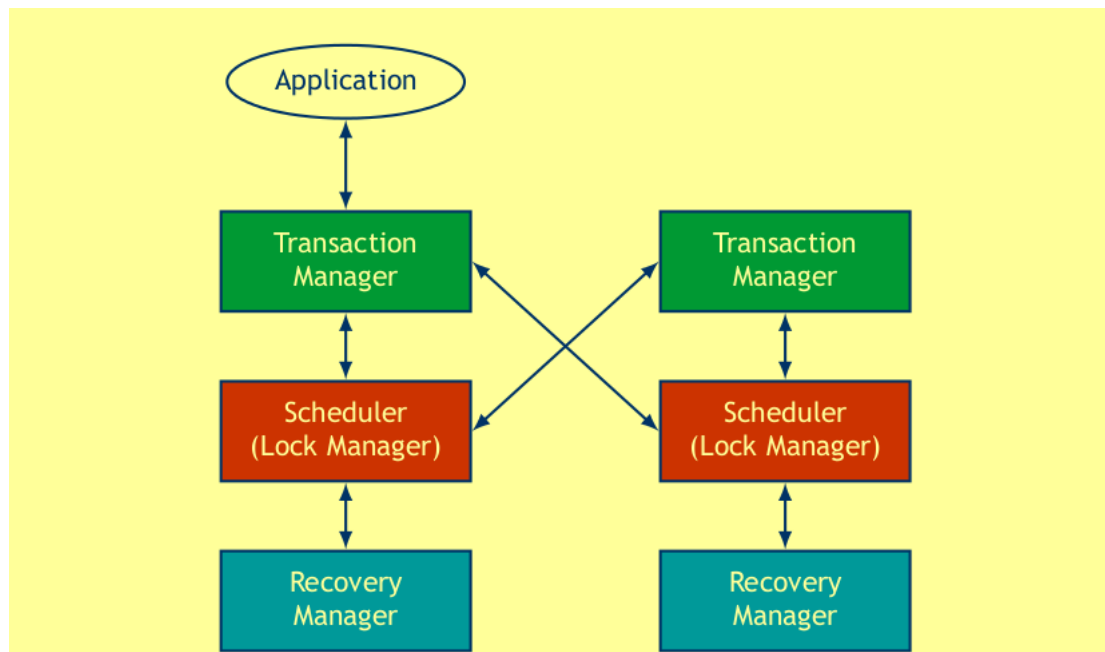


Abbildung 6.2.: figure

- serielle Ausführung, serielle History
- Äquivalenz von Histories
 - Konflikt Äquivalenz
 - Konflikt Operationen
- Serialisierbarkeit
- zwei verschiedene Arten Histories berücksichtigen
 - lokale Ablaufpläne
 - globaler Ablaufplan
- für Serialisierbarkeit des globalen Ablaufplans sind zwei Bedingungen nötig
 - jeder lokale Ablaufplan muss serialisierbar sein
 - zwei Konflikt Operationen müssen in der gleichen Reihenfolge auftreten in allen lokalen Ablaufplänen, in denen sie zusammen auftreten

6.4.1. Beispiel : Global nicht serialisierbar

serialisierbar auf Knoten aber global nicht wegen verschiedene Reihenfolge. Wir wissen im Zentral wie man es macht, aber

6.5. Realisierung

- Basierend auf 2 Phasen Sperrprotokoll
 - Primary Copy 2PL.
 - Verteiltes 2PL
- häufig mit Snapshot verfahren.
 - Read Constancy bei Oracle
 - verwenden der SCN Synchronization bei verteilten DBS.

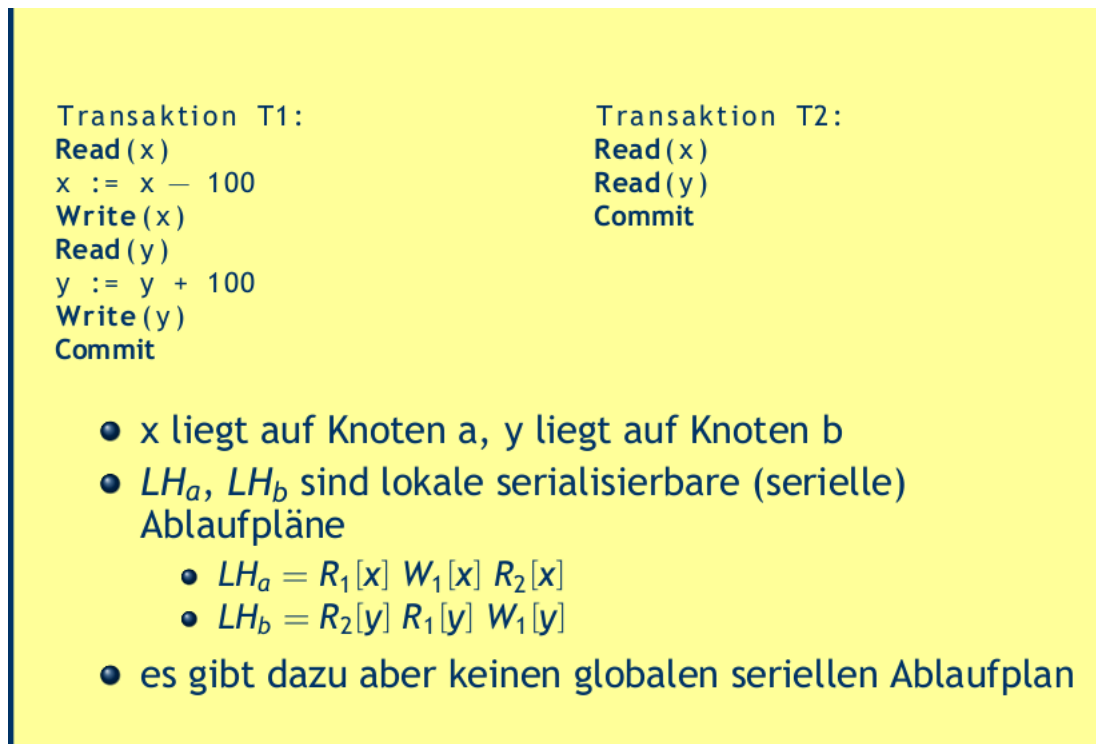


Abbildung 6.3.: figure

SCN (System Change Number) is a primary mechanism to maintain data consistency in Oracle database. SCN is used primarily in the following areas, of course, this is not a complete list.

1. Every redo record has an SCN version of the redo record in the redo header (and redo records can have non-unique SCN). Given redo records from two threads (as in the case of recovery), recovery will order them in SCN order, essentially maintaining a strict sequential order. As explained in my [paper](#), every redo record has multiple change vectors too.
2. Every data block also has block SCN (aka block version). In addition to that, a change vector in a redo record also has expected block SCN. This means that a change vector can only be applied to one and only version of the block. Code checks if the target SCN in a change vector is matching with the block SCN before applying the redo record. If there is a mismatch, corrupt blocks are thrown.
3. Read consistency also uses SCN. Every query has query environment which includes an SCN at the start of the query. A session can see the transactional changes only if that transaction's commit SCN is lower than the query environment SCN.
4. Commit. Every commit will generate SCN, aka commit SCN, that marks a transaction boundary. Group commits are possible too.

Abbildung 6.4.: figure

6.5.1. Zentrales 2PL

- einzelner Knoten verwaltet alle Sperrinformation.
- ein Lock manager für das gesamte DDBMS
- jede Sperre muss beim zentralen Lock manager verlangt werden.
- einfach zu realisieren.
- Flächenhäufung, weniger zuverlässig.

6.5.2. Primary Copy 2pl

- Lock Manager auf einigen Knoten verteilt
- jeder Lock Manager verantwortlich für die Sperren einer Menge von Daten
- bei Replikaten wird eine Kopie als Primary Copy bestimmt, die anderen sind dann slave copies.
- Schreibsperre nur auf Primary Copy.
- nach Änderung der Primary Copy, Änderungen zu den slaves bringen.
- komplexeres Deadlock Handling
- weniger Kommunikation bessere Performance.

6.5.3. Verteiltes 2PL

- Jeder Lock manager auf allen Knoten verteilt.
- jeder Lock anager vrantwortlich für die Sperren siner Daten
- ohne Replikate gleich wie Primary Copy 2PL
- mit Replikaten wird Read One Write All Protokoll implementiert.
irgendeine Kopie kann zum Lesen verwendet werden
alle Kopien erhalten Schreibsperre vor Änderung.
- *komplexeres Daeadlock Handling*
- *höhere Kommunikationskosten als Primary Copy 2PL*

6.6. Deadlocks

Transaktionen sind im einem Deadlock wenn sie blockiert sind und **es bleiben bis das System eingreift**.

Deadlocks treten in 2PL auf zur Vermeidung von nicht serialisierbare Aufbaupläne. Um einen Deadlock aufzulösen muss eine Transaktion abgebrochen werden. In einem **Wait for Graph oder WFG** können Deadlocks erkannt werden.

WFG ist ein gerichtete Graph und stellt die "warte auf"-Beziehung zwischen Transaktionen dar. Transaktionen sind die Knoten im WFG.

Kante $T_i \leftarrow T_j$ bedeutet T_i wartet auf die Freigabe der Sperre auf T_j Zyklen in WFG sind dann deadlocks.

6.6.1. Deadlock Erkennung

- Ansatz automatisch Deadlocks erkennen und eine Betroffene Transaktion abbrehen.
- Bevorzugter Einsatz braucht viel Ressourcen ist aber ein einfaches Verfahren.
- Timeout: Transaktion die zu lange blockiert ist abbrehen.
einfach zu realisieren.
bricht Transaktionen unnötigerweise an.
Deadlocks können lange bestehen.

WFG kann benutzt werden um :

- Deadlock zu finden -
wird eine Transaktion blockiert, Kante in WFG einfügen
WFG periodisch für Zyklen durchsuchen.
- Wird ein Deadlock erkannt, eine Transaktion im Zyklus wählen und abbrehen.

6.6.2. Verteilte Deadlocks

- Wenn Deadlock zwei Knoten umfasst, kann ihn **keiner der Knoten erkennen**.
- Problem lokale und globaler WFG.
- Szenario : t_1 und T_2 auf Knoten 1 und T_3, T_4 auf Konten 2. $T_1 \rightarrow T_2 \rightarrow T_3 \rightarrow T_1$

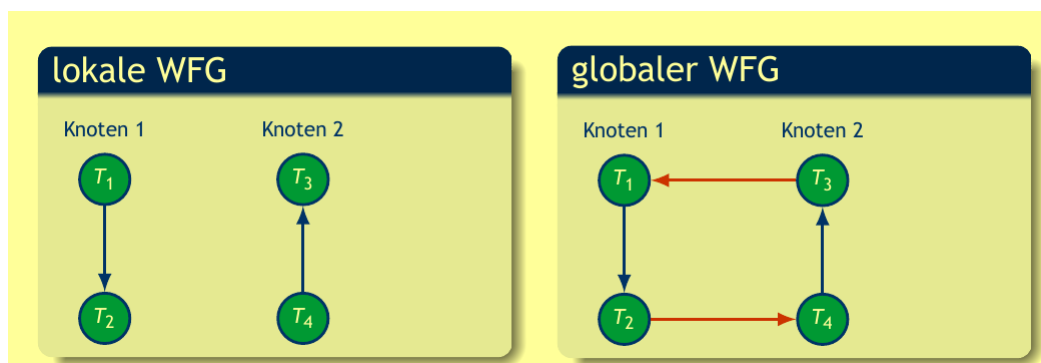


Abbildung 6.5.: figure

7. Verteilte Transaktionen 2

Eine Datenbank ist zuverlässig wenn sie robust und zur Wiederherstellung fähig ist. Diese Idee steckt auch in den ACID Eigenschaften :

Atomicity requires that each transaction is "all or nothing": if one part of the transaction fails, the entire transaction fails, and the database state is left unchanged. An atomic system must guarantee atomicity in each and every situation, including power failures, errors, and crashes. To the outside world, a committed transaction appears (by its effects on the database) to be indivisible ("atomic"), and an aborted transaction does not happen.

Consistency The consistency property ensures that any transaction will bring the database from one valid state to another. Any data written to the database must be valid according to all defined rules, including but not limited to constraints, cascades, triggers, and any combination thereof. This does not guarantee correctness of the transaction in all ways the application programmer might have wanted (that is the responsibility of application-level code) but merely that any programming errors do not violate any defined rules.

Isolation The isolation property ensures that the concurrent execution of transactions results in a system state that would be obtained if transactions were executed serially, i.e. one after the other. Providing isolation is the main goal of concurrency control. Depending on concurrency control method, the effects of an incomplete transaction might not even be visible to another transaction.[citation needed]

Durability Durability means that once a transaction has been committed, it will remain so, even in the event of power loss, crashes, or errors. In a relational database, for instance, once a group of SQL statements execute, the results need to be stored permanently (even if the database crashes immediately thereafter). To defend against power loss, transactions (or their effects) must be recorded in a non-volatile memory.

Wir werden auf Atomicity und Durability fokussieren.

Fehlerarten:

Transaktionsfehler Transaktionssabbruch, einseitig, Deadlock

System oder Kontenausfall Ausfall von CPU , Hauptspeicher , Stromversorgung - Hauptspeicherintervalle gehen verloren. Teil oder Totalausfall.

Plattenfehler Plattenfehler führen zu Verlust gespeicherte Daten. Headcrash, Ausfall des Controllers.

Verbindungsausfall Verlorene oder unzustellbare Nachrichten. Trennung des Netzwerks.

7.1. Lokale oder Verteilte Zuverlässigkeit

Zentrales DBS Behandelt Transaktionsfehler, System oder Knotenausfall und Plattenfehler. Bei Systemausfällen oder Plattenfehler handelt es immer um einen vollständigen Ausfall.

Verteiltes DBS Behandelt zusätzlich Koordination bei Transaktionsfehler, System oder Knotenausfall, Plattenfehler *Verbindungsausfall, erhöhte Verfügbarkeit.*

7.2. Lokale Wiederherstellung

Einzelne DBS sind fähig zur Wiederherstellung durch :

- angemessene Schreibstrategie - force oder no force.
- Angemessene Seitenersetzung (Paging Strategy) - Steal oder No Steal.
- Protokollierung von Änderungen - undo/redo.
- Write Ahead Logging (WAL)
 - Vor Commit (Redo) Log Einträge sichern.
 - Vor Änderung der permanenten Datenbank (Undo) Log Einträge sichern.

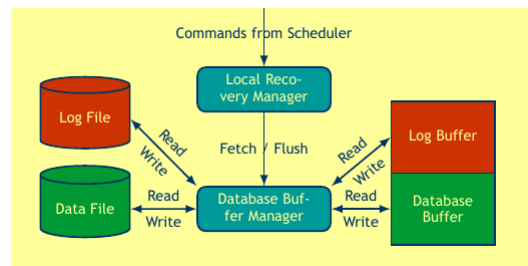


Abbildung 7.1.: figure

- Robustes verfahren benutzen
Undo, Redo idempotent.
ARIES

7.3. Zuverlässigkeit

- Sicherstellen von Atomicity und Durability verteilter Transaktionen.
- Jeder Knoten kann
lokaler Teil der Transaktion zuverlässig verarbeiten
nach Bedarf lokales Commit, Rollback, Recovery durchführen.
- Übergeordnete Protokolle müssen sich mit der Koordination der beteiligten Knoten befassen.
Start der Transaktion beim Ursprungsknoten verarbeiten
Read und Write an den Zielknoten verarbeiten, spezielle Vorkehrungen bei Replikation.
Abort, Commit, und Recover spezifisch für DDB.

7.3.1. Komponenten

Unterschieden in :

Coordinator Prozess beim Ursprungsknoten, steuert die Ausführung.

Participant Prozess bei den anderen Knoten die an der Transaktion beteiligt sind.

Protokolle:

Commit Protokolle Wie das Commit einer Transaktion verarbeiten wenn mehr als ein Knoten am Commit beteiligt ist? Regelt Atomicity und Durability einer verteilten Transaktion

Termination Protokolle Verwendet von Knoten, die vom Ausfall anderer Knoten betroffen sind. Regelt für einen Knoten das Commit/Abort, wenn andere Knoten ausfallen. Blocking/Non-Blocking: Müssen Knoten auf Recovery anderer Knoten warten, um Transaktion abzuschliessen?

Recovery Protokolle Verwendet von Knoten, die ausgefallen sind. Regelt das Recoveryverfahren beim Wiederanlauf nach einem Ausfall Unabhängig: Ein ausgefallener Knoten kann das Ergebnis einer Transaktion bestimmen, ohne Information von anderen Knoten zu haben.

7.4. 2 Phase Commit

1. Choice: Der Koordinator fragt alle Teilnehmer, ob sie bereit sind für ein Commit: Prepare. Jeder Teilnehmer teilt dem Koordinator seine Entscheidung mit: Vote-Commit oder Vote-Abort.
2. Descision: Der Koordinator trifft endgültige Entscheidung: Commit: Falls alle Teilnehmer bereit zum Commit Abort: In allen anderen Fällen und teilt sie den Teilnehmern mit: Global-Commit oder Global-Abort. Diese müssen sich entsprechend verhalten und bestätigen: Acknowledge.

Beobachtungen

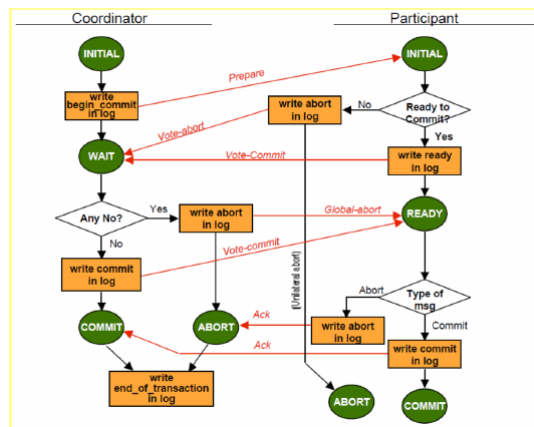


Abbildung 7.2.: figure

1. Ein Teilnehmer kann einseitig Abort durchführen, vorausgesetzt, er hat nicht mit Vote-Commit geantwortet.
2. Nachdem ein Teilnehmer mit Vote-Commit geantwortet hat, muss er zum Commit bereit sein und kann nicht mehr umentscheiden.
3. Ein Teilnehmer im Ready-Zustand muss bereit sein zum Commit oder Abort, je nach Entscheid des Koordinators.
4. Die endgültige Entscheidung ist Commit, falls alle Teilnehmer Vote-Commit geantwortet haben, oder Abort, falls irgendein Teilnehmer mit Vote-Abort geantwortet hat.
5. Koordinator und Teilnehmer verwenden Time-Out Methoden, um möglichen Wartezustand zu verlassen.

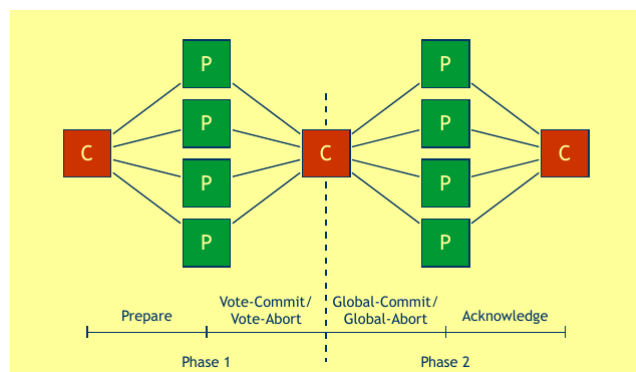


Abbildung 7.3.: Zentralisiertes 2pc

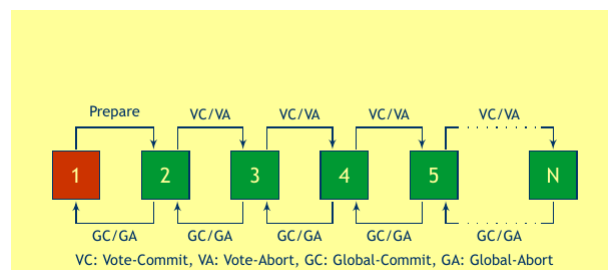


Abbildung 7.4.: Lineares 2PC

7.4.1. Zustandsübergänge

- Time out des Koordinates im Zustand wait, abort, commit möglich.
- wait Zustand :
Koordinator wartet auf Entscheid der Teilnehmer

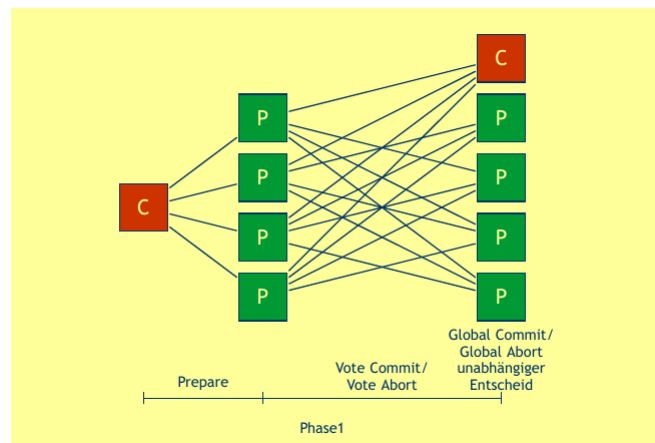


Abbildung 7.5.: Verteiltes 2PC

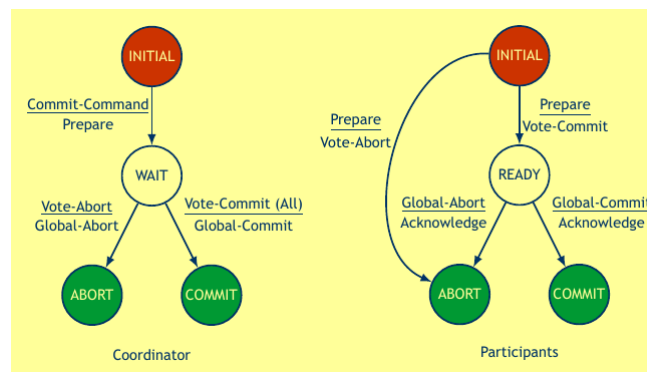


Abbildung 7.6.: Zustandsübergänge 2PC

Lösung: Koordinator entscheidet Global-Abort der Transaktion.

- commit oder abort :

Koordinator ist nicht sicher ob das Commit oder Abort verfahren durch die LRMs¹ alle Teilnehmer durchgeführt wurde.

Lösung: Global commit oder Global Abort an alle Teilnehmer die nicht geantwortet haben, erneut senden.

- Time-Out eines Teilnehmers im Zustand initial oder ready möglich.
- initial Zustand :

Teilnehmer wartet auf das Prepare des Koordinators.

Lösung: Teilnehmer führt einseitig Abort durch. Falls Prepare später ankommt, mit Vote Abort antworten oder ignorieren.

- ready Zustand :

Teilnehmer darf nach dem Vote Commit sich nicht umentscheiden und einseitig Abort durchführen.

Lösung: Teilnehmer ist blockiert und muss auf den endgültigen Entscheid des Koordinators warten.

7.4.2. Recovery Protocols

Protokolle die ein Koordinator oder Teilnehmer nach einem Ausfall verwenden für die Wiederherstellung beim Neustart. Annahmen:

1. Logbeschreibungen und Versenden von Nachrichten sind atomare Aktionen.
2. Zustandsübergänge erfolgen nach Versenden der Nachricht.

¹Local Recovery Manager

7.4.3. Ausfälle

Koordinator

- Koordinator im Zustand `initial` fällt aus - neustart der Transaktion
- `wait` - Neustart des Commit Verfahrens mit nochmaligem Versenden der Prepare Nachricht.
- `commit` oder `abort`
Falls alle Acknowledge Meldungen erhalten sind nichts tun sonst Termination Protokolls anwenden.

Teilnehmer

- `initial` Nach der Wiederherstellung führt Teilnehmer einseitig ein Abort der Transaktion durch.
- `ready` Ablauf wie nach Time out im `ready` Zustand und Termination Protocols anwenden.
- `commit`, `abort` Nichts tun.

7.4.4. Probleme mit 2pc

- Blockierend. Im Zustand `ready` muss Teilnehmer auf Entscheid des Koordinators warten. Fällt Koordinator aus, ist Teilnehmer blockiert. Verfügbarkeit reduziert.
- Recovery eines Teilnehmers nach Ausfall im Zustand `ready` kann nicht unabhängig erfolgen. Verfügbarkeit des Koordinators oder eines Teilnehmers, der die Rolle des Koordinators übernimmt ist nötig.

7.5. 3 Phase Commit

- Nicht Blockierend.
- Commit Protocol ist **nicht blockierend genau dann wenn:**
 - Synchron innerhalb Zustandsübergang
 - Keine Zustände hat die benachbart sind zu einem Commit und Abort Zustand. Oder die die zu einem Commit zustand *nicht* benachbart sind aber nicht committable sind.
 - Committable : Zustand in dem sichergestellt ist, dass alle Teilnehmer mit Vote-Commit geantwortet haben

8. Replikation I

8.1. Einleitung

8.1.1. Gründe für Replikation

Zuverlässigkeit vermeidet "single points of failure."

Performance Vermeidet Kommunikationskosten durch lokalen Zugriffe.

Skalierbarkeit Unterstützt Wachstum des Systems

Anforderungen durch Anwendungen Spezifikation??

Probleme

- Transparenz der Replikation - Abbilden logische Zugriffe in Physische Zugriffe auf Kopien.
- Fragen der Konsistenz : Kriterien und Sync der Kopien.

8.1.2. Ausführungsmodell

Es gibt physische Kopien der logischen Objekte im System. Zugriffe betreffen logische Objekte, werden aber in Zugriffe auf physische Objekte umgesetzt.:

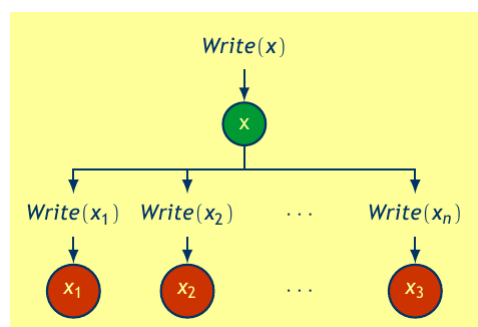


Abbildung 8.1.: figure

8.2. Konsistenzmodelle

Wie wird die Konsistenz des globalen DB Zustands definiert?

- Transaktionsbezogene Kriterien :
 - One Copy Serializability
 - Snapshot Isolation
- Datenbezogene Kriterien : Eventual Consistency
- Begrenzte Inkonsistenz : Epsilon Serializability

Mit welchen Verfahren wird gegenseitige Konsistenz der Replikate erreicht?

Transaktionsbezogene Kriterien (Strong Consistency)

- One Copy Equivalenz : Gegenseitige Konsistenz wenn alle Kopien identische Werte haben. Wirkung einer Transaktion auf replizierte Daten ist die gleiche, wenn sie auf einer einzelnen Datenmenge operierte.

- One Copy Serializability : Wirkung von Transaktionen auf replizierte Daten ist die gleiche wenn sie eine nach der anderen auf eine einzelnen Datenmenge operierten. Histories sind Äquivalenz zu einer seriellen Ausführung auf nicht replizierten Daten.

Datenbezogene Kriterien

- Angeschwächte Formen der Konsistenz:
- Eventual Consistency : letztendlich oder später erreicht. Wenn weitere Updates ausblieben konvergieren die Replikate zu identischen Kopien.
- Nur die Ausbreitung der Updates muss garantiert sein. Das ist kein Problem solange die Nutzer immer auf das gleiche Replikat zugreifen :
 Übergang zu Client-Centric Consistency. Garantiert einem einzelnen Nutzer konsistenten Zugriff auf Daten.

Client Centric Consistency

- Monotonic Reads - (gleichbleibend) Liest ein Prozess Datenelement x , dann gibt jedes nachfolgende Lesen von x durch diesen Prozess denselben oder neueren Wert zurück.
- Monotonic Writes - Schreiboperation auf Datenelement x ist abgeschlossen, bevor derselbe Prozess nachfolgende Schreiboperationen auf x ausführt.
- Read your Writes - Die Wirkung von Schreiboperation auf Datenelement x wird immer in nachfolgenden Leseoperationen auf x durch denselben Prozess gesehen.
- Writes follow Reads - Schreiboperationen auf Datenelement x , die auf ein Lesen durch denselben Prozess folgen, überschreiben immer denselben oder neueren Wert von x .

8.3. Update Propagation Strategies

Können entweder Eager(sync) oder Lazy(async) im Kombination mit Primary Copy (Master) oder Update Everywhere (Group).

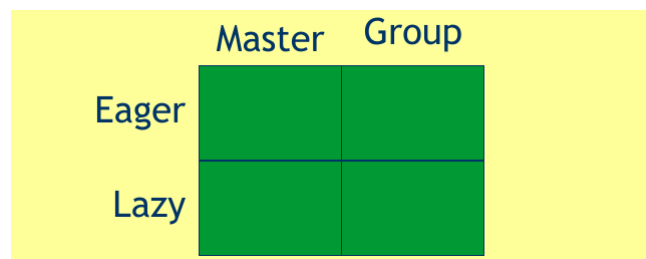


Abbildung 8.2.: figure

Eager(Sync Replikation) Jede Änderung wird sofort zu allen Kopien übertragen. Übertragung der Änderungen erfolgt innerhalb der Grenzen der Transaktion. ACID Eigenschaften gelten für alle Kopien.

Lazy(Async Replikation) Zuerst werden die lokalen Kopien geändert. Anschliessend werden die Änderungen zu allen anderen Kopien übertragen *Push/Pull*. Während der Übertragung sind Kopien inkonsistent. Zeitraum der Inkonsistenz kann in Abhängigkeit der Anwendung angepasst werden.

Master/Primary Copy Replication Es gibt eine einzige Kopie auf der Änderungen ausgeführt werden können (Master). Alle anderen Kopien (Slaves) übernehmen die Änderungen vom Master. Für verschiedene Datenelemente können verschiedene Knoten Master sein.

Group/Update everywhere Replication Änderungen können auf jeder Kopie ausgeführt werden. D.h. jeder Knoten der eine Kopie besitzt, kann auf ihr Änderungen ausführen.



Abbildung 8.3.: figure

8.4. Replikationsstrategien

Eager Master (Synchronous Primary Copy) Replication

- Primary Copy
 - Read** Lokales lesen (eigene Kopie), Resultat zurückgeben.
 - Write** lokales Schreiben, **Write an alle slaves weiterleiten (FIFO/Timestamp reihenfolge)**, Kontrolle sofort dem Nutzer zurückgeben.
 - Commit** verwende 2PC als Koordinator
 - Abort** lokales Abort, slaves informieren.
- Slave:
 - Read** Lokales lesen (eigene Kopie), Resultat zurückgeben.
 - Write Write von Master** : ausführen der writes in richtiger Reihenfolge, **Writes von Client**: zurückweisen oder an Master weiterleiten.
 - Slave ist Teilnehmer an 2PC der Transaktionen vom Master.
- Vorteile :
 - Änderungen müssen nicht koordiniert werden.
 - Keine Inkonsistenzen.
- Nachteile:
 - Längste Antwortzeit
 - Nur bei wenigen Updates sinnvoll (Master ist Flaschenhals)
 - Lokale Kopien sind beinahe nutzlos
 - selten eingesetzt

Eager Group (Synchronous Update Everywhere) Replication

- Read One Write All (ROWA)
 - jeder Knoten verwendet 2 Phasen Sperrprotokoll
 - Leseoperationen werden lokal durchgeführt
 - Schreiboperationen werden auf allen Knoten ausgeführt mithilfe eines verteilten Sperrprotokolls
- Vorteile:
 - Keine Inkonsistenzen
 - elegante Lösung (symmetrisch)
- Nachteile:
 - Vielzahl von Nachrichten
 - Antwortzeiten der Transaktionen sind sehr lang
 - beschränkte Skalierbarkeit (Deadlock Wahrscheinlichkeit wächst mit Anzahl Knoten)

Lazy Master (Asynchronous Primary Copy) Replication

- **Primary Copy**
 - Read:** lokales Lesen (eigene Kopie), Resultat zurückgeben
 - Write:** lokales Schreiben Kontrolle dem Nutzer zurückgeben
 - Commit, Abort:** Transaktion lokal beenden Irgendwann nach dem Commit: an alle Knoten die Änderungen in einer einzigen Nachricht übermitteln (FIFO oder Timestamp Reihenfolge)
- **Slave**
 - Read:** lokales Lesen (eigene Kopie), Resultat zurückgeben
 - Nachricht von Master: Änderungen in richtiger Reihenfolge (FIFO oder Timestamp) anwenden
 - Write von Client:** zurückweisen oder an Master weiterleiten
 - Commit, Abort:** nur für lokale Read-only Transaktionen
- **Vorteile:**
 - Keine Koordination nötig
 - kurze Antwortzeiten (Transaktionen sind lokal)
- **Nachteile**
 - lokale Kopien sind nicht aktuell
 - Inkonsistenzen (verschiedene Knoten haben unterschiedliche Werte für das gleiche Datenelement)

Lazy Group (Asynchronous Update Everywhere) Replication jeder Knoten:

- **textbfRead:** lokales Lesen (eigene Kopie), Resultat zurückgeben
- **Write:** lokales Schreiben Kontrolle dem Nutzer zurückgeben
- **Commit, Abort:** Transaktion lokal beenden Irgendwann nach dem Commit: an alle Knoten die Änderungen in einer einzigen Nachricht übermitteln (FIFO oder Timestamp Reihenfolge)
- **Nachricht von anderem Knoten:**
 - Erkennen von Konflikten
 - Änderungen anwenden
 - Reconciliation (Abgleichen)
- **Vorteile:**
 - Keine Koordination nötig
 - kürzeste Antwortzeiten
- **Nachteile**
 - Inkonsistenzen
 - Änderungen könne verloren gehen (Abgleich)

9. Replikation II (Mat Views)

Eine Updatable Materialized View Replication entspricht einer Lazy Group (asynchronous, update everywhere) Replikation. Eine Materialized View ist ein Replikat einer Master Tabelle bezogen auf einen bestimmten Zeitpunkt. Durch sog. Refreshs werden die Materialized View und bei einer Updatable Materialized View Replication auch die Master Tabelle auf den neusten Stand gebracht, indem alle Änderungen seit dem letzten Refresh angewendet werden. Die Refreshs werden mit einem Pull/Push Verfahren auf der Seite der Materialized View durchgeführt.

9.1. Komponenten

- Zur Replikationsumgebung gehören verschiedene Komponenten:
- Replication Objects
 - Tables
 - Materialized Views
- Replication Groups
 - Master Groups
 - Materialized View Groups
- Replication Sites
 - Master Sites
 - Materialized View Sites

9.2. Komponenten

Replication Objects sind Datenbankobjekte, die auf mehreren Knoten eines verteilten Datenbanksystems vorhanden sind. Änderungen an einem Objekt auf einem Knoten werden auf die Kopien an allen anderen Knoten angewendet. In einer Updatable Materialized View Replication Umgebung werden Tables und Materialized Views als Replication Objects unterschieden.

Replication Groups dienen der Verwaltung von Replication Objects. Eine Replication Group ist eine Sammlung von logisch zusammengehörenden Replication Objects. Typischerweise wird eine Replication Group erzeugt, um die Schema Objekte einer bestimmten Anwendung zu organisieren.

Replication Sites sind die Instances, die an einer oder mehreren Replication Groups beteiligt sind. Es werden zwei Grundtypen von Replication Sites unterschieden: Master Sites und Materialized View Sites. Eine Replication Group auf einem Master Site wird als Master Group bezeichnet. Eine Replication Group auf einem Materialized View Site basiert auf einer Master Group und wird als Materialized View Group bezeichnet. Zudem hat jede Master Group einen Master Definition Site. Von dort aus werden die Replication Group und die Replication Objects verwaltet.

9.3. Benutzer und Rollen

Die verschiedenen Aufgaben für den Betrieb einer Replikationsumgebung werden durch bestimmte Rollen wahrgenommen. Für eine Updatable Materialized View Replication Umgebung sind dies:

- Master Site
 - Replication Administrator
 - Proxy Materialized View Administrator
 - Proxy Refresher

- Materialized View Site
 - MView Administrator
 - MView Propagator
 - Refresher

9.4. Master Site Strukturen

Die Master Table ist die Grundlage der Materialized View. Änderungen an einer Master Table werden im Materialized View Log aufgezeichnet, um bei einem Refresh zu der Materialized View übertragen zu werden. Ein interner After-Row-Trigger fügt bei jedem Insert, Update oder Delete der Master Table eine Zeile in das Materialized View Log ein. Im Materialized View Log werden die Änderungen an der Master Table aufgezeichnet. Das Log muss explizit erstellt werden, dabei wird eine zugrundeliegende Log-Tabelle erzeugt mit dem Namen:

```
$_Master_Table_Name
```

9.5. Materilized View Site Strukturen

Wird eine Materialized View erstellt, werden weitere Strukturen auf dem Materialized View Site angelegt. Zu jeder Materialized View wird ein Index mit Namen Materialized_View_Name_PK angelegt, der dem Primärschlüssel der Master Table entspricht. Ein Updatable Materialized View Log:

```
( USLOG$_Materialized_View_Name)
```

wird verwendet um zu bestimmen, welche Zeilen der Materialized View überschrieben oder entfernt werden müssen. Ein interner Trigger zeichnet die Änderungen an der Updatable Materialized View im Log auf.

9.6. Konfliktauflösung

Werden Änderungen sowohl auf der Master Site als auch auf der Materialized View Site durchgeführt, kann es zu Konflikten kommen. Oracle bietet die Möglichkeit, für verschiedene Anforderungen spezielle Regeln zur Konfliktauflösung zu hinterlegen. Siehe dazu das Kapitel 6 in Oracle Database Advanced Replication Management API Reference:

http://download.oracle.com/docs/cd/E11882_01/server.112/e10707/toc.htm

10. Nosql

10.1. Einführung

A blogger, often referred to as having made the term popular is Rackspace employee Eric Evans who later described the ambition of the NoSQL movement as “the whole point of seeking alternatives is that you need to solve a problem that relational databases are a bad fit for” (cf. [Eva09b]). This section will discuss rationales of practitioners for developing and using nonrelational databases and display theoretical work in this field. Furthermore, it will treat the origins and main drivers of the NoSQL movement.

10.1.1. Motivation

Avoidance of unneeded complexity: databases provide a variety of features and strict data consistency. But this rich feature set and the ACID properties implemented by RDBMSs might be more than necessary for particular applications and use cases. As an example, Adobe’s ConnectNow holds three copies of user session data; these replicas do not have to undergo all consistency checks of a relational database management systems nor do they have to be persisted. Hence, it is fully sufficient to hold them in memory (cf. [Com09b]).

High Throughput Some NoSQL databases provide a significantly higher data throughput than traditional RDBMSs. For instance, the column-store Hypertable which pursues Google’s Bigtable approach allows the local search engine Zvent to store one billion data cells per day [Jud09]. To give another example, Google is able to process 20 petabyte a day stored in Bigtable via its MapReduce approach [Com09b].

Horizontal Scalability and Running on Commodity Hardware NoSql trying to address problems such as Data Scaling, Performance of single servers, Rigid schema design. Nosql databases are designed to scale well in the horizontal direction and not rely on highly available hardware. Sharding is easier.

Avoidance of expensive Object Relational Mapping Most of the NoSQL databases are designed to store data structures that are either simple or more similar to the ones of object-oriented programming languages compared to relational data structures. They do not make expensive object-relational mapping necessary (such as Key/Value-Stores or Document-Stores). This is particularly important for applications with data structures of low complexity that can hardly benefit from the features of a relational database. Dare Obasanjo claims a little provokingly that “all you really need [as a web developer] is a key<->value or tuple store that supports some level of query functionality and has decent persistence semantics.” (cf. [Oba09a]). The blogger and database-analyst Curt Monash iterates on this aspect: “SQL is an awkward fit for procedural code, and almost all code is procedural. [For data upon which users expect to do heavy, repeated manipulations, the cost of mapping data into SQL is] well worth paying [. . .] But when your database structure is very, very simple, SQL may not seem that beneficial.” Jon Travis, an engineer at SpringSource agrees with that: “Relational databases give you too much. They force you to twist your object data to fit a RDBMS.” (cited in [Com09a]).

Complexity and Cost of Setting up Database Clusters See easier sharding.

Compromising Reliability for better performance “different scenarios where applications would be willing to compromise reliability for better performance.” As an example of such a scenario favoring performance over reliability, he mentions HTTP session data which “needs to be shared between various web servers but since the data is transient in nature (it goes away when the user logs off) there is no need to store it in persistent storage.”

Problems with one size fits all thinking

1. Continuous growth of data
2. Process large amounts of data in short amounts of time.
3. Social media example - unrelated islands of data, low user/transaction value - no need for string data integrity.

Myth of Effortless Dist and Partitioning of centralised data models and nor work correct any longer. The professionals of Ajatus agree with this in a blog post stating that if a database grows, at first, replication is configured. In addition, as the amount of data grows further, the database is sharded by expensive system admins requiring large financial sums or a fortune worth of money for commercial DBMS-vendors are needed to operate the sharded database (cf. [Aja09]). Shalom concludes that in his opinion relational database management

systems will not disappear soon. However, there is definitely a place for more specialized solutions as a "one size fits all" thinking was and is wrong with regards to databases.

Developments in Programming ORMs¹ hide SQL in many frameworks. The NoSQL databases react on this trend and try to provide data structures in their APIs that are closer to the ones of programming languages (e. g. key/value-structures, documents, graphs).

Cloud Computing 1. High -> inf scalability.

2. Low admin/overhead.

Following databases work well:

- Data warehousing specific dbs
- Simple scalable and fast key value stores.
-

The RDBMS plus Caching-Layer Pattern/Workaround vs. Systems Built from Scratch with Scalability in Mind : As scalability requirements grow and these technologies are less and less capable to suit with them. In addition, as NoSQL datastores are arising Hoff comes to the conclusion that "[with] a little perspective, it's clear the MySQL + memcached era is passing. It will stick around for a while. Old technologies seldom fade away completely." (cf. [Hof10c]). As examples, he cites big websites and players that have moved towards non-relational datastores including LinkedIn, Amazon, Digg and Twitter. Hoff mentions the following reasons for using NoSQL solutions which have been explained earlier in this paper:

- Relational databases place computation on reads, which is considered wrong for large-scale web applications such as Digg. NoSQL databases therefore do not offer or avoid complex read operations.
- The serial nature of applications¹ often waiting for I/O from the data store which does no good to scalability and low response times.
- Huge amounts of data and a high growth factor lead Twitter towards facilitating Cassandra, which is designed to operate with large scale data.
- Furthermore, operational costs of running and maintaining systems like Twitter escalate. Web applications of this size therefore "need a system that can grow in a more automated fashion and be highly available." (cited in [Hof10c]).

Yesterday vs Today by Stonebraker, see below). In the 1960s and 1970s databases have been designed for single, large high- end machines. In contrast to this, today, many large (web) companies use commodity hardware which will predictably fail. Applications are consequently designed to handle such failures which are considered the "standard mode of operation", as Amazon refers to it (cf. [DHJ+ 07, p. 205]). Furthermore, relational databases fit well for data that is rigidly structured with relations and allows for dynamic queries expressed in a sophisticated language. Lehnhardt and Lang point out that today, particularly in the web sector, data is neither rigidly structured nor are dynamic queries needed as most applications already use prepared statements or stored procedures. Therefore, it is sufficient to predefine queries within the database and assign values to their variables dynamically (cf. [PLL09]). Furthermore, relational databases were initially designed for centralized deployments and not for distribution. Although enhancements for clustering have been added on top of them it still leaks through that traditional were not designed having distribution concepts in mind at the beginning (like the issues adverted by the "fallacies of network computing" quoted below). As an example, synchronization is often not implemented efficiently but requires expensive protocols like two or three phase commit. Another difficulty Lehnhardt and Lang see is that clusters of relational databases try to be "transparent" towards applications. This means that the application should not contain any notion if talking to a singly machine or a cluster since all distribution aspects are tried to be hidden from the application. They question this approach to keep the application unaware of all consequences of distribution that are e. g. stated in the famous eight fallacies of distributed computing (cf. [Gos07]² :

1. The network is reliable
2. Latency is zero
3. Bandwidth is infinite
4. The network is secure
5. Topology doesn't change
6. There is one administrator
7. Transport cost is zero
8. The network is homogeneous

¹Object Relational Mappers

10.2. Cassandra

1. Welches sind die wichtigsten Gründe, die zur Entwicklung Ihres NoSQL Systems führten? Welche Anforderungen erfüllen diese Systeme?

- Verarbeitung von grossen Datenmengen
- Availability
- Reliability
- Kein Single Point of Failure
- Läuft auf Cheap Comodity Hardware
- High Write Output
- Inbox Search Support (Suche durch die Facebook Inbox)
- Geographisch verteilt

2. Welches „logische“ Datenmodell bietet das System? Welches API bietet das System? Wie werden Daten manipuliert und abgefragt?

Datenmodell: Eine Instanz von Cassadra besteht aus nur 1 Tabelle, die eine verteilte, multidimensionale Map bezeichnet, die durch einen Schlüssel indexiert ist. Schlüssel: 16-36 Bytes, Value: Object. Struktur:

Zeilen mit String als Schlüssel

Spalten: haben Namen und speichern eine Anzahl Werte, die durch einen Zeitstempel identifiziert sind. Jede Zeile kann eine unterschiedliche Zahl von Spalten haben.

Spaltenfamilien: willkürliche Zahl Spalten pro Zeile können zu Familien zusammengefasst werden. Es können beliebig viele Familien spezifiziert werden, aber effektiv nur wenige. Familien können dynamisch zur Laufzeit mit Spalten und Superspalten ergänzt werden.

-Superspalten: weitere Gruppierung: Name und willkürliche Anzahl Spalten Ansprache von Daten über Tripel (Zeilenschlüssel, Spaltenschlüssel, Zeitstempel). Zeilenschlüssel über column-family:supercolumn:column Sortierung nach Zeitstempel oder Name möglich

API: drei Operationen:

- get(table, key, columnName)

- insert(table, key, rowMutation)

- delete(table, key, columnName) à Familie, Superzeile, Zeile

Operationen sind atomar per replica, Anzahl betroffene Spalten ist egal.

-read: Konsistenzgarantie kann spezifiziert werden (nächster Knoten, Antworten verschiedener Knoten)

-insert/update: „a quorum of replica nodes has to acknowledge the completion of the writes“. Weiterleitung zu willkürlichem Server eines Cassandra Clusters, so dass Partitionierung und Redistribution ermöglicht werden.

3. Verteilungsarchitektur

Ringarchitektur mit allen Knoten. Jedem Knoten ist ein bestimmter Range der Hashkeys zugeteilt. Replikation

SimpleStrategy: Replikation auf den N-1 nächsten Knoten

NetworkTopologyArchitetur: definierbare Anzahl Replikate pro Datacenter

4. Wie werden Daten repliziert? Wie wird die Konsistenz der Daten sichergestellt?

- Auf jeder Instanz ist ein Replikationsfaktor N definiert, wobei alle Daten auf N Hosts repliziert werden
- Jeder Schlüssler wird an einen "Coordinator Node" zugewiesen, welcher einerseits die Daten dazu lokal speichert und andererseits die Daten auf N-1 Knoten verteilt
- Daten können nach verschiedenen Richtlinien verteilt werden z.B. Rack Unaware, Rack Aware und Datacenter Aware um Risiken besser zu vermindern
- Cassandra bestimmt mit Hilfe von Zookeeper einen Leader, welcher die Verantwortlichkeit für Datenbereiche an die Knoten verteilt
- Metadaten werden lokal und fehlertolerant in Zookeeper gespeichert
- Die Daten werden über mehrere Datacenter repliziert und sollten auch bei Ausfall eines ganzen Datacenters verfügbar sein.

- Konsistenz wird sichergestellt indem ein Quorum festgelegt wird, wieviel Knoten einen Schreibvorgang bestätigen müssen, auch für Lesevorgänge kann ein Quorum festgelegt werden, muss aber nicht

5. Welche Vorkehrungen werden für die Verfügbarkeit und Ausfallsicherheit getroffen?

- Daten bzw. Nodes sind geographisch verteilt und werden über mehrere Datacenter repliziert. Dadurch wird gewährleistet, dass die Latenzzeiten klein gehalten werden und der Ausfall eines Knotenpunktes keinen Einfluss auf das System hat. Durch das abschwächen des Quorum-System, der die Konsistenz aufrecht erhalten soll, wird die Beständigkeit des Systems garantiert.
- Fehlerdetektion durch Accrual Failure Detector: Jedem Knoten wird ein Wert zugeteilt, anhand dessen man abschätzen kann, wie hoch die Chance ist, dass dieser Knoten ausfällt, statt diesem Knoten einen boolean-Wert zuzuteilen.
- Der Knoten wählt beim Starten ein Random-Token, um seine Position auf dem Ring zu ermitteln. Beim joinen in ein Cluster, wird der Node dessen Config-File lesen. Bei einem Ausfall eines Teilsystems, wird nun aber kein Re-Balancing vorgenommen, da angenommen wird, dass es wieder online kommen wird. Um vorzubeugen, dass nicht erreichbare Nodes anderen Instanzen beitreten, wird beim Kommunizieren der Cluster-Name der Cassandra-Instanz mitgeschickt, um seinen Quellort zu bestimmen. Das hinzufügen oder entfernen eines Knoten aufgrund eines manuellen Fehlers im Konfiguration-File muss ebenfalls manuell über ein Command Line Tool erfolgen.

6. Wie werden die Daten lokal auf den einzelnen Knoten gespeichert? Eine Schreiboperation schreiben in ein Commit-Log. Auf jeder Maschine befindet sich eine Disk welche nur dem Commit-Log zugeteilt ist (dedicated). Das Write-Log ist dabei sequentiell. Konnten die Daten erfolgreich in's Commit-Log geschrieben werden, werden die Daten zusätzlich in eine In-Memory Struktur geladen. Ab einer gewissen Speichergröße, wird die In-Memory-Datenstruktur auf eine Disk geschrieben. Alle Schreiboperationen sind dabei sequentiell und die Daten werden gleichzeitig für einen schnelleren Zugriff indexiert. Mit der Zeit entstehen somit mehrere Files mit den entsprechenden Dump Daten. Ein Merge Prozess fügt die verschiedenen Dump-Files wieder zu einer Datei zusammen.

Ein Query auf den Daten arbeitet Primär auf den In-Daten-Memory. Ein Lookup auf den Files geht von neueren Files über zu den älteren Files. Die Daten werden dabei mit Hilfe eines keys gesucht. Um zu verhindern, dass Files durchsucht werden, welche den Filter nicht enthalten wird ein Bloom Filter über die vorhandenen Keys verwendet. Eine Tabelle der Bloom Filter wird sowohl auf jedem Dump-File als auch In-Memory gehalten.

10.3. Neo4J

1. Welches sind die wichtigsten Gründe, die zur Entwicklung Ihres NoSQL Systems führten? Welche Anforderungen erfüllen diese Systeme?

Neo4j gehört in die Kategorie der Graphendatenbanken und wurde ursprünglich für die Echtzeitsuche von verschlagworteten(indexierten) Dokumenten in einem Online Dokumentationssystem entwickelt. „Viele aktuelle Anwendungen müssen stark vernetzte, rekursive Daten behandeln, welche sich nur unzureichend auf relationalen Datenbanken abbilden lassen. Ein gutes Beispiel dafür sind soziale Netzwerke. Auch ist es schwierig, das exakte Schema der Daten bereits zur Entwurfszeit zu kennen, speziell im Umgang mit Benutzergesteuertem Inhalt.“ Frei übersetzt aus: <http://highscalability.com/neo4j-graph-database-kicks-butt> Neo4j ist in Java implementiert und sowohl als Bibliothek in die JVM eingebettet als auch als Server-Version verfügbar. Die Datenbank nutzt einen eigenen, optimierten Persistenzmechanismus für die Speicherung und Verwaltung der Graphen. Neo4j ist als open source und kommerzielle Version verfügbar. Die Datenbank ist gut dokumentiert und verfügt über einen breiten Community Support. Verfügt über Schnittstellen zu diversen Programmiersprachen, ein REST Interface ist der empfohlene Zugriffs-weg. Hoch skalierbar. Das .jar ist kleiner als 500kb und hat nur eine dependency. Einfaches API.

2. Welches „logische“ Datenmodell bietet das System? Welches API bietet das System? Wie werden Daten manipuliert und abgefragt?

Das Datenmodell hinter neo4j ist ein Graph Datenbank System. Dabei gibt es Knoten und Kanten welche in einem Property Graphen das System ausmachen. Das heisst, zwischen den Knoten gibt es Relationships und die Kanten und Knoten haben verschiedene Properties : Um die Daten zu erhalten, spielt die Traversierung eine wesentliche Rolle. Jeder Knoten und jede Kante speichert einen "Mini-Index" mit den eigenen verbundenen Objekten. Globale Indizes werden als Startpunkte für die Traversierung festgelegt. Weiter werden Indizes benötigt um schnell Knoten mit bestimmten Werten zu erhalten. Danach wird nach bestimmten Mustern im Graphen gesucht.

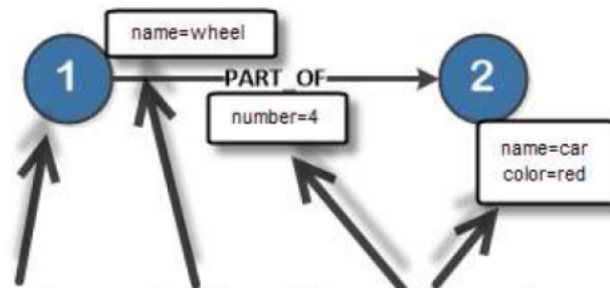


Abbildung 10.1.: figure

API:**API:**

- Neo4j's Java API (Knoten und Beziehungen als Java-Objekte)
- REST interface (JSON-Objekte)
- DSL language
 - Abfrage Sprache Cypher ist eine deklarative Graph Query Sprache von SQL inspiriert. (ASCII)

```

(m:Movie {title: "The Matrix"})
<-[:ACTED_IN {role:"Neo"}]-
(a:Actor {name:"Keanu Reeves"})
  
```

Quelle: Neo4jDE_ebook.pdf

- Gremlin: Verwendet Groovy und Pipes um Graphen zu traversieren.

Abbildung 10.2.: figure

3. Wie werden Daten über mehrere Knoten verteilt (Verteilungsarchitektur)?**4. Wie werden Daten repliziert? Wie wird die Konsistenz der Daten sichergestellt?**

Replikationen sind nur mit Neo4j Enterprise möglich. Dabei werden Cluster eingesetzt, die aus einem Master und mehreren Slaves bestehen. Die Daten werden automatisch auf die Slaves repliziert. Es sollte dabei aber darauf geachtet werden, dass Schreibvorgänge nur auf dem Master passieren. Die Replikation kann auf zwei Arten geschehen:

- Automatisch durch Abruf der Schreibtransaktionen seitens des Slaves auf dem Master (Millisekunden - paar Minuten)
- Automatisch durch Propagieren der Schreibtransaktionen seitens des Master auf eine vorgegebene Anzahl an Slaves (sofort). Dies wird über den push-factor definiert. Jeder Slave besitzt die ganze Datenbank. Die Skalierbarkeit ist somit nahezu linear (Read-Scaling), da Lesevorgänge beliebig auf die Slaves verteilt werden können. Bei Graphen Datenbanken ist nie die ganze Datenbank im Arbeitsspeicher, sondern nur der Teil, welcher bei der letzten Abfrage geladen wurde. Dies bietet nun die Möglichkeit über ein intelligentes und konsistentes Routing dieselben Klassen von Anfragen auf die gleichen Empfänger weiterzuleiten. So müssen nicht nach jeder Anfrage neue Teile geladen werden (Cache Sharding).

5. Welche Vorkehrungen werden für die Verfügbarkeit und Ausfallsicherheit getroffen?

Dies ist nur bei einem Cluster möglich. Es sind zwei Arten des Ausfalls möglich. Fällt ein Slave aus, wird dies von den anderen Datenbankinstanzen im Netzwerk erkannt und er wird als vorübergehend nicht verfügbar gekennzeichnet. Sobald er wieder verfügbar ist, sorgt er selbst dafür, dass er wieder auf dem aktuellen Stand des Masters ist.

Fällt der Master aus, sieht das ganze etwas anders aus. Der Cluster entscheidet, welcher Slave zum neuen Master wird. Bedingung ist lediglich, dass der Slave zum Master werden darf. Somit ist ein Failover kein Problem. Sobald der Master nicht mehr verfügbar ist, werden alle laufenden Schreibtransaktionen zurück gerollt. Der Wechsel vom Slave zum Master geschieht dann im Normalfall innert Sekunden. Während dieser Zeit sind keine

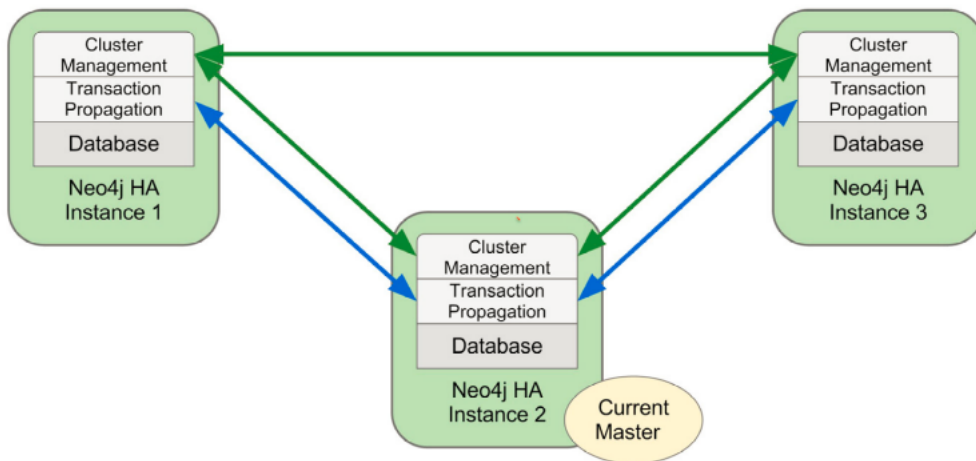


Abbildung 10.3.: figure

Schreibtransaktionen möglich. Die einzige Ausnahme ist, wenn auf dem alten Master noch Schreibvorgänge vorhanden sind, die noch auf keine andere Datenbankinstanz repliziert werden konnten und bereits Änderungen am neuen Master gemacht wurden. In diesem Fall wird der alte Master zwei „Branches“ führen. Der erste „Branch“ ist seine alte Datenbank. Dieser „Branch“ wird abgezogen, da er nicht mehr aktuell ist, jedoch noch nicht replizierte Änderungen enthält. Dann lädt er sich für den zweiten „Branch“ eine komplette Kopie des aktuellen Zustands herunter und wird dann selbst zum Slave.

6. **Wie werden die Daten lokal auf den einzelnen Knoten gespeichert?** Nodes, Properties und Relationships werden separat voneinander gespeichert. (*.nodestore.db, *.relationshipstore.db, *.propertystore.db)

10.4. Bigtable

1. **Welches sind die wichtigsten Gründe, die zur Entwicklung Ihres NoSQL Systems führten?**

Google brauchte ein unterliegendes System für Ihre Dienste. Diese Dienste arbeiten mit grossen Datenmengen (Petabytes), sind auf viele tausende Server im Cluster verteilt und generieren viele neue Einträge, aber wenige Änderungen an vorhandenen Einträgen. Eingesetzt u. a. für folgende Produkte: Google Suche (web indexing) Google Earth Google Finance

Welche Anforderungen erfüllen diese Systeme?

breite Anwendungspalette
Skalierbarkeit
Hochverfügbarkeit
Hochleistungsfähigkeit

2. **Speicherung und Verteilungsarchitektur:**

Speicherung im Dateisystem: A Bigtable is a sparse, distributed, persistent multidimensional sorted map.

Der Schlüssel für den Index besteht aus:

row key
column key
timestamp

Jeder Wert ist ein Array aus beliebigen Bytes (String).

(row:string, column:string, time:int64) -> string

Speicherung der Tabellen: Tabellen werden in „tablets“ unterteilt und so gespeichert. Die Grösse eines solchen Segments ist etwa 200 MB. Dies ist optimiert für das Google File System (GFS). Wenn viele Daten vorhanden sind werden diese komprimiert. Es wird das „SSTable file format“ verwendet.

Zugriff auf die Daten Pro Tabelle existiert ein Index, der komplett(?) in den Speicher geladen wird. Das Auffinden der Daten im Dateisystem ist als Binär-Suche möglich. Der Index bezieht sich direkt auf den Block auf der HDD, so dass Zugriffe auf HDD optimiert werden können. Die Daten sind grundsätzlich unveränderbar (SSTable ist immutable). Somit ist Zugriff immer Thread-Safe. Es werden Serien von

SSTables verwendet, so dass Änderungen möglich sind. Änderungen kommen aber zuerst ins änderbare „memtable“. Darin sind also immer die neuesten Änderungen gespeichert. Zugriff ist hier immer noch parallel möglich dank Copy-On-Write. Ein Tablet-Server verwaltet etwa 10-1000 Tablets.

Garbage Collection Das GFS hat eine GC (Mark-And-Sweep). Der Master-Server muss sich darum kümmern.

Verteilung der Daten: Speziell ist hier dass das GFS bereits verteilt ist. BigTable benötigt ein Cluster-Management-System zum Scheduling, Ressourcen-Management und für Erkennung und Verarbeitung von Fehlern. Es gibt einen Master-Server und mehrere Tablet-Server. Solche Tablet-Server können dynamisch entfernt oder hinzugefügt werden. Chubby wird verwendet damit Tablet-Server sich im Cluster anmelden können.

Datenmodell: persistent multidimensional sorted map.

Zeilen werden lexographisch sortiert. (Row Keys 64KB)

Tabelle wird unterteilt in Tablets. Mehrere Rows ergeben ein Tablet. Diese werden auf verschiedene Server verteilt.

Keine Limitierung in der Anzahl Spalten

Gruppierung der Spalten -> Column Families

Auf den Column Families werden die Access Control Rules definiert

Lesbarer Name

best Practice: wenige Column Families. Keine Änderungen an diesen.

Eine Gruppe -> Ein file

Timestamps Jede Zelle kann mehrer Versionen der selben Daten beinhalten. Daher wird jeder Eintrag mit einem 64bit timestamp versehen.

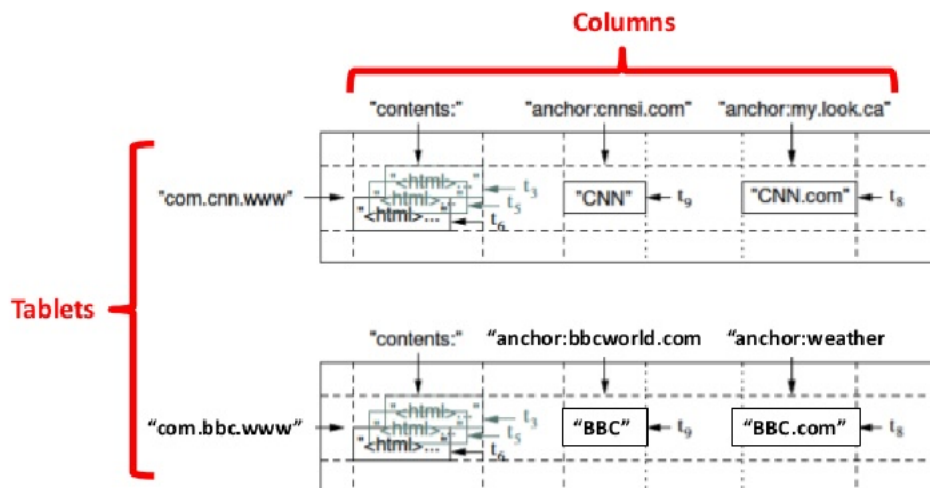


Abbildung 10.4.: figure

API

Read Operations: Row by Key, Limitation of Column Families + TimeStamps, Column iterators

Write for Rows: create, delete values of column

Write for Columns Families: create, delete

Administration: cluster, table, column family metadata -> Access Control jede read/write operation in einer Zeile ist atomar

3. **Wie werden Daten repliziert? Wie wird die Konsistenz der Daten sichergestellt?** BigTable verwendet ein Cluster-Management-System (Chubby) um mit verteilten Knoten zu interagieren. Jede Chubby Instanz besitzt einen Cluster von 5 aktiven Kopien, von welchem jeweils einer der Master ist und Anfragen bearbeitet. Um die Kopien konsistent zu halten wird der Paxos Algorithmus verwendet. Die gesplitteten Tablets werden auf die verschiedenen Knoten verteilt.

Chubby ist ein verstreuter Lock-Service, der ein BigTable Cluster mit mehrere tausenden Knoten koordiniert und 5 Replikationen hat. Der Klient verbindet sich jeweils zu Chubby um mit den Daten zu arbeiten. Falls für eine bestimmte Zeit Chubby unerreichbar wird, wird BigTable auch unerreichbar.

Für Bigtable sind zurzeit zwei Replikationsmechanismen verfügbar. Einerseits die Master/Slave Replikation, bei der ein Datacenter als Master ausgewählt wird, und sämtliche Schreibzugriffe über dieses Center abgewickelt werden. Hierbei werden die Änderungen asynchron auf andere Datacenter repliziert.

Andererseits wird der High Replication Datastore angeboten, der mittels des sog. axos-Algorithmus die Schreibfragen auf verschiedene Datacenter verteilt und diese dann synchron auf andere Datacenter repliziert. Bigtable besitzt eigene Algorithmen, die eine automatische Lastanpassung (Skalierung) durchführen, daher ist kein manueller Eingriff notwendig.

HBase selbst kennt keine Replikation. Dort wird mit Hilfe von Hadoop die Replikation durchgeführt.

Verfügbarkeit Jeder TabletServer schreibt seine CommitLogs durch 2 Threads, jeder Thread schreibt ins eigene LogFile. Zu 1 Zeitpunkt ist nur 1 Thread aktiv. Wenn Writes durch den aktiven Thread zu langsam werden, wird der andere Thread als der aktive gesetzt. Alle Commits auf der Queue, die auf Write beim langsamen Thread gewartet haben, werden vom neuen Thread übernommen. · Performance: Caching! The Scan Cache is a higher-level cache that caches the key-value pairs returned by the SST able Interface to the tablet server code (bestens geeignet fürs Lesen der gleichen Daten). The Block Cache is a lower-level cache that caches SST ables blocks that were read from GFS (bestens geeignet für Lesen der Daten, die in der Nähe der neulich gelesenen Daten sind, d.h. sequentielles Lesen oder Lesen im gleichen Block).

Ausfallsicherheit Wenn ein Tablet Server stirbt: Seine Tablets werden auf die andere Server verteilt. Um den State jedes Tablets nachzubauen, muss der CommitLog durchgesucht werden auf betreffende Commits, um diese auf den Tablets auszuführen. D.h. 100 Server – 100 Reads auf 1 File. Da es schlechtes Vorgehen ist: Der CommitLog wird sortiert nach Table-Key. CommitLog wird zuerst auf 64 MB grosse Blocks aufgeteilt, das Sortieren der Blöcke passiert parallel. Das sortierte Output ermöglicht blockweises sequentielles Lesen der nötigen Daten.

10.5. MongoDB

Bemerkung : Entschuldigung für CAPS LOCK.

1. **Welches sind die wichtigsten Gründe, die zur Entwicklung Ihres NoSQL Systems führten? Welche Anforderungen erfüllen diese Systeme?**

DOKUMENTORIENTIERT: Verwendet BSON (Binary JSON), da sehr effizient - Bessere Interaktion mit modernen (objektorientierten) Programmiersprachen als Relationale Datenbanken

FLEXIBILITÄT: Schemaloses Datenmodell.

- Dokumente können jederzeit erweitert werden ohne dass hohe Aufwände entstehen (Einfaches hinzufügen von Feldern und Listen, kein Fehler beim Lesen von noch nicht existierenden Datenbanken).
- Iterative Entwicklung möglich

MÄCHTIGKEIT - Hoher Funktionsumfang

- Komplexe Datentypen möglich (Strukturen mit dynamischen Attributen, Arrays, usw.)

GESCHWINDIGKEIT/SKLAIERBARKEIT Schneller als SQL-Systeme durch: - Key-Value Prinzip: Zugriff direkt auf den Key anstatt diesen via Indexierung zu suchen - Vermeidung von Joins. Alle Daten können direkt über die aktuelle Liste im Dokument abgefragt werden - In-Memory lesen/schreiben

2. **Welches „logische“ Datenmodell bietet das System?**

- MongoDB-Server besteht aus mehrere MongoDB Datenbanken
- MongoDB Datenbank besteht aus einer oder mehreren Sammlungen eine Sammlung (hat einen Namen) besteht aus einem oder mehreren Dokumenten.
- Es ist möglich eine hierarchische Struktur der Sammlung zu bilden, indem ein Punkt eingefügt wird. Dies wird jedoch nicht so logisch abgebildet.
- ein Dokument besteht aus mehreren Feldern (Werte, Arrays, binär Daten, sub-Dokumente)
- Jedes Dokument in der gleichen Sammlung kann verschiedene Felder haben.
- Das Format wird BSON genannt und ist ähnlich wie JSON, jedoch in binärer Form damit es schneller Verarbeitet werden kann.
-
-

Beispiel Dokument:

```
{
  title : " MongoDB ",
  last_editor : "172.5.123.91",
  last_modified : new Date ("9/23/2010"),
  body : " MongoDB is a ...",
  categories : [" Database ", " NoSQL ", " Document Database "],
  reviewed : false
}
```

Vergleich zu SQL:

Sammlung <-> Tabelle

Dokument <-> Zeile

Feld <-> Spalte

MongoDB ist schema frei

SQL: fünf Tabellen <-> MongoDB: zwei Sammlungen

Abbildung 10.5.: figure

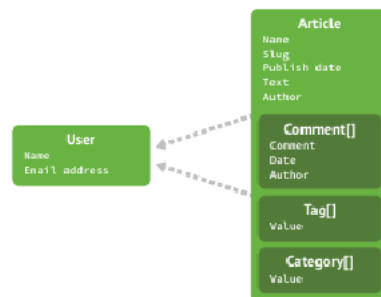


FIGURE 2 // Example document data model for a blogging application.

Abbildung 10.6.: figure

-
-

3. Wie werden Daten repliziert? Wie wird die Konsistenz der Daten sichergestellt?**a) MASTER-SLAVE-REPLIKATION (Alte Variante)**

MongoDB unterstützt die Master-Slave-Replikation. Ein Master kann Lese- („Reads“) und Schreibzugriffe („Writes“) ausführen. Ein Slave kopiert die Daten vom Master und kann nur für Lesezugriffe oder Backups eingesetzt werden, nicht aber für Schreibzugriffe. Fällt der Masterknoten aus, so kann einer der Slaveknoten zum neuen Masterknoten umgewandelt werden (muss manuell erfolgen, Administrationsaufwand).

b) REPLICA SET (Neue Variante)

Ein Replica Set ist eine Gruppe von Knoten, die Kopien voneinander sind. Einer dieser Gruppe wird als PrimärerKnoten gewählt. Nur über diesen werden Schreiboperationen abgewickelt. Der Primärknoten reicht diese an die anderen Knoten der Gruppe weiter. Wenn der Primärknoten ausfällt, übernimmt ein beliebiger anderer Knoten aus der Gruppe dessen Rolle (dies kann automatisch erfolgen, kein Administrationsaufwand).

Vorteile der neuen Variante sind z.B.: - Geringerer Administrationsaufwand

- Konsistenzbedingungen können einfach gesetzt werden (z.B. Schreiben ist erfolgreich wenn mehr als die Hälfte des Replica Sets die Änderungen übernommen hat)

- Grosse Skalierbarkeit, da nahezu jeder Knoten aus einem Replica Set für Leseoperationen verwendet werden kann

4. Welche Vorkehrungen werden für die Verfügbarkeit und Ausfallsicherheit getroffen?

REPLICATION Operationen welche die Datenbank auf dem PRIMARY verändern, werden in einem Log namens oplog gespeichert. Der 'oplog' enthält eine geordnete Menge von idempotenten Operationen, welche auf den "Secondaries" repliziert werden. Die Grösse des oplogs beträgt standardmässig 5des freien

SELECT

```
db . < collection > . findOne ( { _id : 921394 } );
db . < collection > . find ( { categories : [ " NoSQL " , " Document Databases " ] } );
db . < collection > . find ( { < selection criteria > } , { < field_1 > : 1 , ... } );
< field_1 > : 1 die Felder
< field_1 > : 0 ohne diese Felder
```

INSERT

```
db . < collection > . insert ( { title : " MongoDB " , last_editor : ... } );
db . < collection > . save ( { ... } ); # wenn kein PK
```

UPDATE

```
db . < collection > . save ( { ... } ); # wenn PK
db . < collection > . update ( < criteria > , < new document > , < upsert > , < multi > );
```

DELETE

```
db . < collection > . remove ( { < criteria > } );
```

Abbildung 10.7.: figure

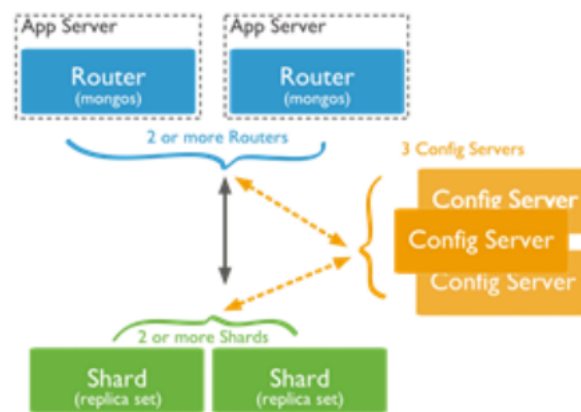


Abbildung 10.8.: figure

Festplattenspeichers. Der oplog enthält Änderungen einiger Stunden. Sollte ein Secondary ausfallen, kann er so wieder "aufholen". Sollte ein Secondary länger als die Periode vom oplog beträgt ausfallen, dann werden mittels "initial synchronization" alle Datenbanken usw. komplett repliziert.

ELECTIONS AND FAILOVER Falls der Primary aus irgendeinem Grund ausfallen sollte, dann bestimmen die Secondaries unter sich einen neuen Primary. Dieser Prozess wird "Election" genannt. Sobald der neue Primary bestimmt wurde, konfigurieren sich die anderen Secondaries selber, so dass sie nun Updates vom neuen Primary erhalten. Sollte der ausgefallene Primary wieder online kommen, erkennt er, dass er nicht mehr der Primary ist und wird zu einem Secondary.

ELECTION PRIORITY Der "Election" Prozess kann mit einer Prioritäts-Konfiguration beeinflusst werden. Der Secondary mit der höchsten Priorität wird dann Primary. Zum Beispiel könnte man alle Instanzen in einem "Secondary Data Center" so konfigurieren, dass nur wenn das komplette "Primary Data Center" ausfällt, einer der Secondaries dann Primary wird.

CONFIGURABLE WRITE AVAILABILITY MongoDB besitzt die Möglichkeit, dass ein Client bei jeder Abfrage einen "write concern" mitteilen kann. Dieser gibt an, wann der Client eine Antwort erhält. Der Client wartet dann solange auf eine Antwort bis der gewünschte "write concern" eingetreten ist und kann so z.B. sicherstellen, dass die Daten definitiv geschrieben wurden.

Folgende Optionen sind z.B. möglich:

- Client braucht keine Antwort, (One-Way) -> Keine Garantie, dass Daten geschrieben werden
- Client will eine Bestätigung sobald die Änderungen auf mehreren Replicas vollzogen worden sind.
- Client will eine Bestätigung sobald die Änderungen auf der Mehrheit der Replicas vollzogen worden sind.
- Client will eine Bestätigung sobald die Änderungen auf allen Replicas vollzogen worden sind.
- Und weitere, einstellbare, Möglichkeiten wären z.B: Mind. 2 Replicas enthalten die Änderungen im

Primary-Datcenter und min. 1 Replica im Secondary Datacenter.

5. Abspeicherung Lokal

DATA AND EXTENTS

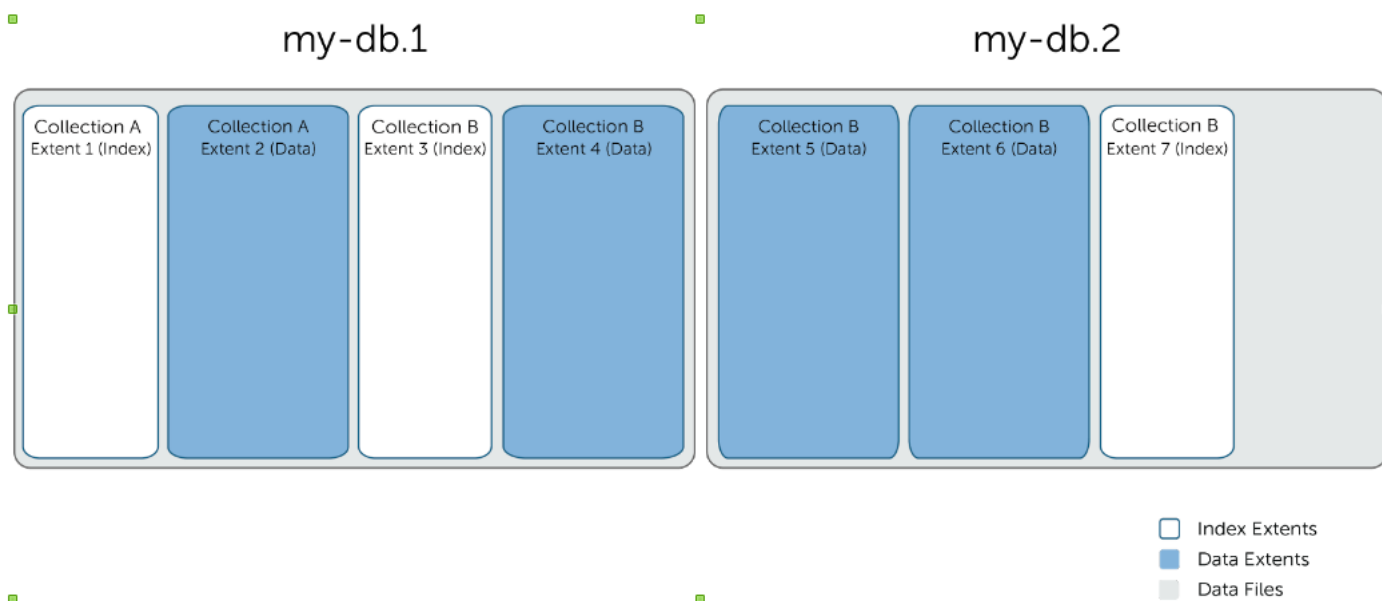


Abbildung 10.9.: figure

11. Semantic Web

11.1. World Wide Web

- HTTP als Protokoll, mit dem der Browser Informationen vom Webserver anfordern kann.
- HTML als Dokumentenbeschreibungssprache, die festlegt, wie die Information gegliedert ist und wie die Dokumente verknüpft sind (Hyperlinks).
- URLs als eindeutige Adresse bzw. Bezeichnung einer Ressource (z. B. einer Webseite), die in Hyperlinks verwendet wird.

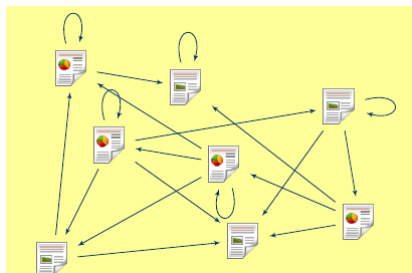


Abbildung 11.1.: figure

11.2. Idee des Semantic Webs

Das Semantic Web unterstützt ein verteiltes Netz auf der Ebene der Daten und der Beschreibung der Daten statt nur auf der Ebene der Dokumente. Dies wird u. a. erreicht, indem nicht mehr zwischen den Daten und ihrer Beschreibung (Metadaten) unterschieden wird. Aber beide bleiben wichtig: Ohne Metadaten gibt es nichts, worüber man reden kann und ohne Daten gibt es nichts zu sagen!

Subjekt	Prädikat	Objekt
Umberto Eco	schrieb	Das Foucaultsche Pendel
Umberto Eco	schrieb	Der Name der Rose
Umberto Eco	lebt in	Bologna
Renate Ramge	heiratete	Umberto Eco
Renate Ramge	geboren in	Frankfurt
Bologna	ist in	Italien
Frankfurt	ist in	Deutschland
Italien	gehört zu	Europäische Union
Deutschland	gehört zu	Europäische Union

Abbildung 11.2.: figure

11.3. Ziele des Semantic Webs

- Informationen im Web sollen für Maschinen verständlich sein Bereitstellen eines semantischen Modells, um das
- Wissen eines Fachbereichs auszudrücken Informationsaustausch zwischen Fachbereichen ermöglichen.

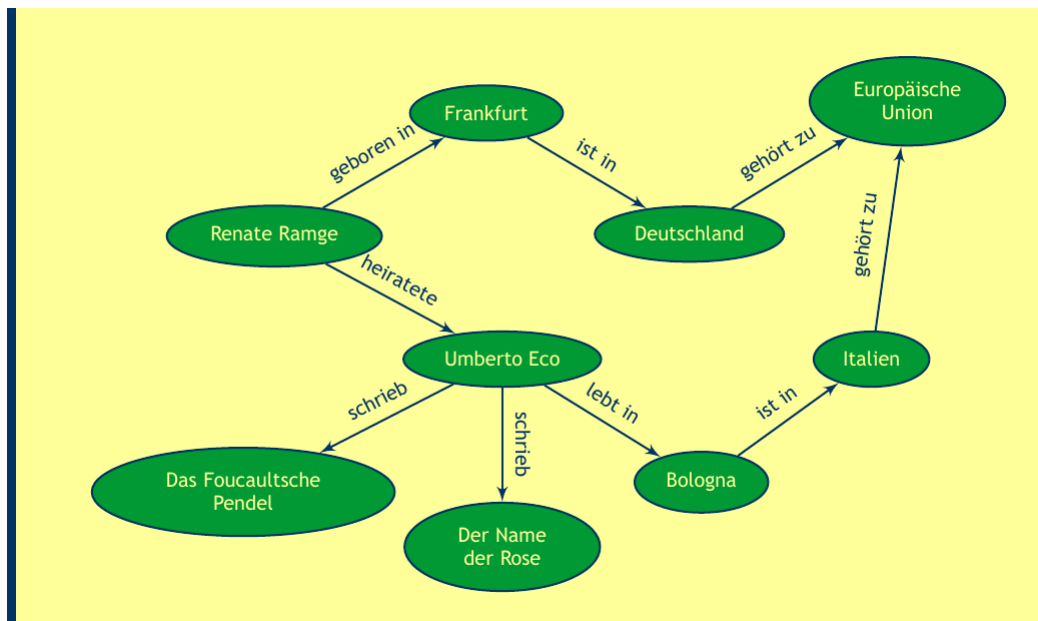


Abbildung 11.3.: figure

11.4. Datenmodell WWW und semantic web

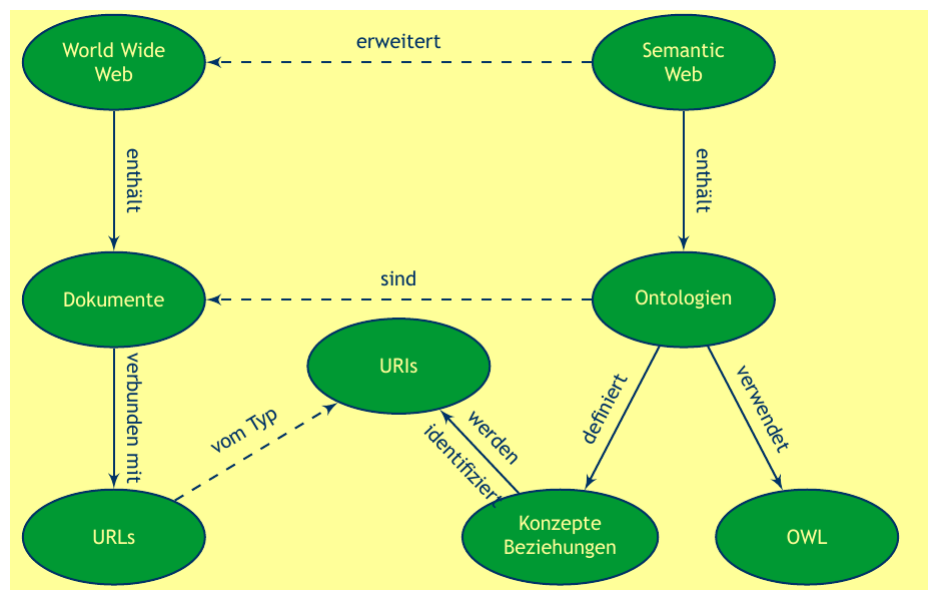


Abbildung 11.4.: figure

11.5. Technologien

11.5.1. RDF - Resource Description Framework

- einfaches Datenmodell
- formale Semantik und Schlussfolgerung
- erweiterbares Vokabular mit URI
- XML Syntax
- unterstützt XML Schema Datentypen

- Jeder kann Aussagen über beliebige Ressourcen machen
- Alles, was benannt werden kann, ist eine Ressource
- Global eindeutige Namen durch URI
- Konstruktion von Aussagen mit Subjekt, Prädikat und Objekt
- Subjekt und Prädikat sind immer Ressourcen.
- Das Objekt kann entweder eine Ressource oder ein Literal sein
- Dargestellt als gerichteter Graph
Knoten: Subjekt, Objekt
Kanten: Prädikat

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

  <rdf:Description about = "URI-Referenz">
    <propName>string </propName>
  </rdf:Description>
</rdf:RDF>
```

Abbildung 11.5.: RDF Beispiel

Weitere Elemente:

- rdf:type
- rdf:ID
- rdf:Statement
- rdf:Property
- rdf:subject
- rdf:predicate
- rdf:object

11.6. RDF und OWL

RDFS: RDF Schema

OWL: Web Ontology Language

RDF bietet allein keine Möglichkeit anzugeben, was die Aussagen bedeuten. Es braucht: ein gemeinsames Vokabular von Begriffen, eine hierarchische Organisation der Begriffe (Taxonomie), Regeln und Beschreibungen, wie Klassen und Eigenschaften verwendet werden und wie sie mit einander in Beziehung stehen können.

Elemente:

- rdfs:Class
- rdfs:domain
- rdfs:range
- rdfs:subPropertyOf
- rdfs:subClassOf
- owl:inverseOf
- owl:SymmetricProperty
- owl:TransitiveProperty
- owl:FunctionalProperty

Teil II.

Labs

12. Lab 1 - Trigger

12.1. Event Logging

12.1.1. Aufgabenstellung

Manipulationen auf der Tabelle REGISTRIERUNGEN loggen. Trigger und Log-Tabelle befinden sich auf dem gleichen System wie REGISTRIERUNGEN.

12.1.2. Lösung

Log-Tabelle

```
1 -- Erste Möglichkeit
2 CREATE TABLE registrierungen_log (
3   username VARCHAR2(15) NOT NULL,
4   systemdate DATE NOT NULL,
5   maniptype VARCHAR2(3) NOT NULL,
6   oldval VARCHAR(15),
7   newval VARCHAR(15)
8 );
9 -- Zweite Möglichkeit
10 CREATE TABLE registrierungen_log (
11   username VARCHAR2(15) NOT NULL,
12   systemdate DATE NOT NULL,
13   maniptype VARCHAR2(3) NOT NULL,
14   oldmnr VARCHAR2(4) NULL,
15   newmnr VARCHAR2(4) NULL,
16   oldfnr VARCHAR2(3) NULL,
17   newfnr VARCHAR2(3) NULL,
18   oldanr VARCHAR2(4) NULL,
19   newanr VARCHAR2(4) NULL,
20   olddatum DATE NULL,
21   newdatum DATE NULL
22 );
```

Trigger

```
1 -- Erste Möglichkeit
2 CREATE OR REPLACE TRIGGER registrierungen_change
3 BEFORE
4   INSERT OR
5   UPDATE OR
6   DELETE
7 ON registrierungen
8 FOR EACH ROW
9
10 DECLARE
11   username VARCHAR(15);
12   maniptype VARCHAR2(3);
13   systemdatum DATE := SYSDATE;
14 BEGIN
15   SELECT USER INTO username FROM DUAL;
16   CASE
17     WHEN INSERTING THEN
18       maniptype := 'INS';
19       INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
20         VALUES(username, systemdatum, maniptype, NULL, :NEW.mnr);
21       INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
22         VALUES(username, systemdatum, maniptype, NULL, :NEW.fnr);
23       INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
24         VALUES(username, systemdatum, maniptype, NULL, :NEW.anr);
```

```

22     INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
        VALUES(username, systemdatum, maniptype, NULL, TO_CHAR(:NEW.datum, 'DD.MM.YYYY'
        ));
23     WHEN DELETING THEN
24         maniptype := 'DEL';
25         INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
            VALUES(username, systemdatum, maniptype, :OLD.mnr, NULL);
26         INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
            VALUES(username, systemdatum, maniptype, :OLD.fnr, NULL);
27         INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
            VALUES(username, systemdatum, maniptype, :OLD.anr, NULL);
28         INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
            VALUES(username, systemdatum, maniptype, TO_CHAR(:OLD.datum, 'DD.MM.YYYY'),
            NULL);
29     WHEN UPDATING THEN
30         maniptype := 'UPD';
31         CASE
32             WHEN :OLD.mnr != :NEW.mnr THEN
33                 INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
                    VALUES(username, systemdatum, maniptype, :OLD.mnr, :NEW.mnr);
34             WHEN :OLD.fnr != :NEW.fnr THEN
35                 INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
                    VALUES(username, systemdatum, maniptype, :OLD.fnr, :NEW.fnr);
36             WHEN :OLD.anr != :NEW.anr THEN
37                 INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
                    VALUES(username, systemdatum, maniptype, :OLD.anr, :NEW.anr);
38             WHEN :OLD.datum != :NEW.datum THEN
39                 INSERT INTO registrierungen_log(username, systemdate, maniptype, oldval, newval)
                    VALUES(username, systemdatum, maniptype, TO_CHAR(:OLD.datum, 'DD.MM.YYYY'
                    ), TO_CHAR(:NEW.datum, 'DD.MM.YYYY'));
40         END CASE;
41     END CASE;
42 END;
43
44 -- Zweite Möglichkeit
45 CREATE OR REPLACE TRIGGER registrierungen_change
46 AFTER
47     INSERT OR
48     UPDATE OR
49     DELETE
50 ON registrierungen
51 FOR EACH ROW
52 DECLARE
53     log_action VARCHAR2(3);
54 BEGIN
55     IF INSERTING THEN
56         log_action := 'INS';
57     ELSIF UPDATING THEN
58         log_action := 'UPD';
59     ELSIF DELETING THEN
60         log_action := 'DEL';
61     ELSE
62         DBMS_OUTPUT.PUT_LINE('This code is not reachable.');

```

12.1.3. Lösungsbeschreibung

Code	Zweck	Kommentar
BEFORE, AFTER, INSTEAD OF	Das Timing des Triggers, wobei INSTEAD OF die auslösende Operation auf eine View mit Operationen des Triggers ersetzt.	
FOR EACH ROW	Der Trigger soll für jede Reihe einzeln ausgelöst werden, statt für ein Statement.	Für diesen Trigger sehr wichtig, da Manipulationen auf der Granularität von Zeilen geloggt werden müssen. Dadurch hat man auch Zugriff auf alte und neue Werte.
systemdatum DATE := SYSDATE; SELECT USER INTO username FROM DUAL;	Zwei verschiedene Möglichkeiten Systemvariablen in einer Variable zu speichern.	FROM DUAL muss verwendet werden, da ein SELECT nicht ohne FROM verwendet werden darf. Dabei wird DUAL als eine DUMMY Tabelle verwendet und enthält keine Informationen.
:NEW.mnr :OLD.mnr	Falls die Granularität des Triggers auf Zeilenebene fest gelegt ist, kann durch :NEW und :OLD auf die Werte zugegriffen werden, die vor und nach der Manipulation in der Tabelle stehen würden.	
INSERTING, DELETING, UPDATING	Können in einem WHEN-CASE verwendet werden, um abzufragen, bei welcher Operation der Trigger ausgelöst wurde.	Ist nur nötig, falls der Trigger bei mehr als eine Operation triggert.

12.2. Referential Integrity

12.2.1. Aufgabenstellung

In einer zentralisierten Datenbank (sämtliche Tabellen auf einem System) wird die referentielle Integrität vom Datenbanksystem überprüft. In einer verteilten Datenbank ist dies nicht mehr möglich, da die Datenbanksysteme voneinander getrennt sind. Die referentielle Integrität muss somit manuell über Trigger überprüft werden.

Für die Aufgabe werden die Tabellen Filialen[PK:FNR] und DVDKopien[PK:DVDNr, FK1:FNR] behandelt.

12.2.2. Lösung

Link ananke <

```

1 CREATE DATABASE LINK orion.helios.fhnw.ch
2 CONNECT TO mvdb00 IDENTIFIED BY mvdb00
3 USING 'orion';
4
5 -- Mit Views, filialen-Trigger auf ananke
6 CREATE OR REPLACE VIEW filialen AS
7 SELECT * FROM filialen@orion.helios.fhnw.ch;
8
9 -- Ohne Views, filialen-Trigger auf helios
10 CREATE SYNONYM filialen FOR filialen@orion.helios.fhnw.ch;
11
12 CREATE SYNONYM dvdkopien FOR dvdkopien@ananke.hades.fhnw.ch;
```

FILIALEN-Trigger auf ananke

```

1 CREATE OR REPLACE TRIGGER dvdkopien_ref_check_ins_upd
2 BEFORE
3   INSERT OR
4   UPDATE
5   ON dvdkopien
6   FOR EACH ROW
```



```

7  DECLARE
8      msg VARCHAR2(100) := 'Integritäts-Constraint verletzt - übergeordneter Schlüssel nicht
                              gefunden';
9      cnt NUMBER;
10 BEGIN
11     SELECT COUNT(*) INTO cnt FROM filialen WHERE filialen.fnr = :NEW.fnr;
12     IF (cnt = 0) THEN RAISE_APPLICATION_ERROR(-20000, msg); END IF;
13 END;

```

DVDKOPIEN-Trigger auf helios

```

1  CREATE OR REPLACE TRIGGER filialen_del_upd
2  BEFORE DELETE OR UPDATE ON filialen
3  FOR EACH ROW
4  DECLARE
5      msg VARCHAR2(100) := 'Integritäts-Constraint verletzt - untergeordneter Datensatz
                              gefunden';
6      cnt NUMBER;
7  BEGIN
8      SELECT COUNT(*) INTO cnt FROM dvdkopien WHERE dvdkopien.fnr = :OLD.fnr;
9      IF (cnt > 0) THEN RAISE_APPLICATION_ERROR(-20000, msg);
10     END IF;
11 END;

```

12.2.3. Lösungsbeschreibung

Voraussetzungen	Synonym von FILIALEN auf ananke	Auf dem ursprünglichen System muss eine ein Synonym (inkl. Datenbanklink) zu FILIALEN auf helios bestehen, damit auf dessen Daten zugegriffen werden kann, um die Integrität zu überprüfen.
Trigger-Lokalität	Die Trigger sollten auf den Systemen installiert werden, auf denen sich ebenfalls die Tabellen befinden. Sämtliche Trigger können aber auch auf einem System installiert werden. Dann müssen die Trigger jedoch auf Views der entfernten Tabellen triggern und die Operation durch INSTEAD OF ersetzen, damit zuerst die Integrität überprüft und danach die ursprüngliche Operation ausgeführt werden kann.	
RAISE_APPLICATION_ERROR(-20000, msg);	Hier wird eine Fehlermeldung erzeugt.	Der Code ist dabei entscheidend, was angezeigt wird. -20000 ist nicht obligatorisch für dieses Lab, ein Besseres konnte jedoch nicht auf die Schnelle gefunden werden.
BEFORE	Logik des Triggers wird vor der Ausführung der Operation ausgeführt	In diesem Lab ist ein BEFORE-Trigger nötig, da durch das Werfen eines Fehlers die folgende Operationen abgebrochen wird.

13. Lab 3 - Verteilter Datenbankentwurf

13.1. Aufgabenstellung

Anhand von Anwendungen und deren Ausführungshäufigkeiten eine verteilte Datenbank entwerfen.

Anwendungen

```
1 -- Anwendung q1
2 SELECT name, typ, maxmiete
3 FROM kunde;
4 -- Anwendung q2
5 SELECT telefonnummer, name, typ
6 FROM kunde
7 WHERE maxmiete <= 2000;
8 -- Anwendung q3
9 SELECT telefonnummer, maxmiete
10 FROM kunde;
11 -- Anwendung q4
12 SELECT DISTINCT(typ)
13 FROM kunde;
14 -- Anwendung q5
15 SELECT telefonnummer, name
16 FROM kunde
17 WHERE typ = 'Wohnung';
```

Ausführungshäufigkeiten

Anwendung	System	Zugriffshäufigkeit
q1	ananke	10
q1	telesto	20
q2	ananke	5
q2	calypso	10
q3	telesto	35
q3	calypso	5
q4	telesto	10
q5	telesto	15

13.2. Lösung

13.3. Lösungsbeschreibung

13.3.1. Weg zur globalen Affinität

Als Grundlage für die Usage Matrix dienen die Anwendung/Queries. Bei jedem Attribut, auf die ein Query zugreift, wird das Flag[1=true,0=false] in der Matrix gesetzt. Beim Zusammentragen der Attribute muss man sich darauf achten, dass nicht nur der Select berücksichtigt wird, sondern auch andere Zugriffe, wie zum Beispiel im Where.

Die Zugriffshäufigkeitsmatrix wird durch das Zusammentragen der Informationen aus der Ausführungshäufigkeit der einzelnen Queries auf den einzelnen Systemen gebildet. Es ist ratsam für weitere Anwendung gleich eine Summe zu bilden, wie viel mal auf ein Query zugegriffen wird.

Die Attributs-Affinität Matrix bildet nun sämtliche Attribute ab. Diese Matrix kann wie folgt ausgefüllt werden:

- Parameter Zeile / Spalte => Name / Telefon
- In der Usage Matrix prüfen welche Queries auf diese Attribute zugreifen, also eine 1 in beiden
- Feldern stehen haben => q2, q5
- Zugriffshäufigkeit dieser Queries aufsummieren => 15 + 15 = 30

Usage Matrix U

	Name	Telefon	Typ	MaxMiete
q1	1	0	1	1
q2	1	1	1	1
q3	0	1	0	1
q4	0	0	1	0
q5	1	1	1	0

Zugriffshäufigkeits Matrix Acc

	Ananke	Telesto	Calpyso	Orion	Sum
q1	10	20	0	0	30
q2	5	0	10	0	15
q3	0	35	5	0	40
q4	0	10	0	0	10
q5	0	15	0	0	15

Attributs-Affinitäten Matrix AA

	Name	Telefon	Typ	MaxMiete
Name	60	30	60	45
Telefon	30	70	30	55
Typ	60	30	70	45
MaxMiete	45	55	45	85

Globale Affinität ohne Optimierung

	Name	Telefon	Typ	MaxMiete
Name	60	30	60	45
Telefon	30	70	30	55
Typ	60	30	70	45
MaxMiete	45	55	45	85

1800	1800	2700
2100	2100	1650
1800	2100	3150
2475	2475	3825
8175	8475	11325
		27975

Abbildung 13.1.: Globale Affinität

Bond Energy Algorithmus
Schritt 1

	Typ	Name	Telefon
Name	60	60	30
Telefon	30	30	70
Typ	70	60	30
MaxMiete	45	45	55

	Name	Typ	Telefon
Name	60	60	30
Telefon	30	30	70
Typ	60	70	30
MaxMiete	45	45	55

	Name	Telefon	Typ
Name	60	30	60
Telefon	30	70	30
Typ	60	30	70
MaxMiete	45	55	45

3600	1800
900	2100
4200	1800
2025	2475
10725	8175
	18900

3600	1800
900	2100
4200	2100
2025	2475
10725	8475
	19200

1800	1800
2100	2100
1800	2100
2475	2475
8175	8475
	16650

Abbildung 13.2.: Bond Energy Algorithm

- In der AA-Matrix unter Zeile / Spalte und Spalte / Zeile eintragen

Da die AA-Matrix symmetrisch an der Diagonale aufgebaut ist, können meistens Zeile und Spalte vertauscht werden, um das gleiche Resultat in der Matrix einzutragen.

Achtung! Der Bond Energy Algorithmus ist hier vereinfacht dargestellt, da mit Excel gearbeitet wurde. Beim korrekten Vorgehen muss der Beitrag einer Änderung berechnet werden. Dabei kann wie folgt überlegt werden:

- Die "lokale Affinität" zweier benachbarter Spalten kann berechnet werden indem Werte der benachbarten Spalten in einer Zeile multipliziert und anschliessend zeilenweise aufsummiert werden. Beispiel Name Telefon: $60 \cdot 30 + 30 \cdot 70 + 60 \cdot 30 + 45 \cdot 55 = 8175$.
- Wird bei BEA nun eine Änderung durch das Hinzufügen einer weiteren Spalte bewirkt, wird der Beitrag wie folgt berechnet:
 - bestehende Nachbarschaften werden nicht aufgelöst: "lokale Affinität" berechnen; das ist der Beitrag; dies ist der Fall, wenn der neue Nachbar links oder rechts der Matrix hinzugefügt wird
 - bestehende Nachbarschaften werden aufgelöst: "lokale Affinität" der aufgelösten Nachbarn berechnen (z.B. Name-Telefon), dieser Wert wird vom Beitrag abgezogen; "lokale Affinität" des neuen Nachbarn jeweils mit der rechten und linken Spalte berechnen (z.B. Name-Typ, Typ-Telefon); diese Werte werden zum Beitrag addiert
- Die Änderung mit dem grössten Beitrag wird durchgeführt und die nächste Iteration beginnt.
- Sind alle Iterationen durch, können die Zeilen entsprechend den Spalten angeordnet werden.

Es wurde die Matrix berechnet, die die grösste globale Affinität besitzt. Nun muss noch berechnet werden, wie die Fragmente dieser Tabelle gebildet werden sollen. Dazu wird entlang der Diagonalen der Matrix der Trennpunkt verschoben. Es muss nun die Fragmentierung mit der grössten Trennqualität gefunden werden. Der Algorithmus acc berechnet sich wie folgt:

- ein Parameter: falls in einer Zeile/Query in der Usage Matrix in irgend einer Spalte des Parameters/des Fragments eine 1 steht und in allen Spalten, die nicht zum Fragments gehören, eine 0, dann wird die Zugriffshäufigkeit des Queries zum Resultat addiert

Resultat BEA

	Name	Typ	MaxMiete	Telefon
Name	60	60	45	30
Telefon	30	30	55	70
Typ	60	70	45	30
MaxMiete	45	45	85	55

3600	2700	1350	
900	1650	3850	
4200	3150	1350	
2025	3825	4675	
10725	11325	11225	33275

Abbildung 13.3.: Bond Energy Algorithm

Anordnen der Zeilen entsprechend den Spalten

	Name	Typ	MaxMiete	Telefon
Name	60	60	45	30
Typ	60	70	45	30
MaxMiete	45	45	85	55
Telefon	30	30	55	70

Abbildung 13.4.: Bond Energy Algorithm

- zwei Parameter: gleich wie bei einem Parameter, nur muss mindestens die Bedingung für Spalten des ersten UND des zweiten Fragments zutreffen

Für das Splitting wählt man ebenfalls die Variante mit dem höchsten Wert aus und drückt es anschliessend in relationaler Algebra aus.

Hilfsmatrix U nach der Reihenfolge der Attribute geordnet

	Name	Typ	MaxMiete	Telefon
q1	1	1	1	0
q2	1	1	1	1
q3	0	0	1	1
q4	0	1	0	0
q5	1	1	0	1

Hilfsmatrix Acc

	Ananke	Telesto	Calpyso	Orion	Sum	
q1	10	20	0	0	30	q1
q2	5	0	10	0	15	q2
q3	0	35	5	0	40	q3
q4	0	10	0	0	10	q4
q5	0	15	0	0	15	q5
					110	

Abbildung 13.5.: Hilfsmatrizen

Splitting

Variante 1

Reihenfolge Attribute

Name, Typ, MaxMiete, Telefon	Telefon, Name, Typ, MaxMiete	MaxMiete, Telefon, Name, Typ	Typ, MaxMiete, Telefon, Name
acc(VF1) 0	acc(VF1) 0	acc(VF1) 0	acc(VF1) 10
acc(VF2) 50	acc(VF2) 40	acc(VF2) 25	acc(VF2) 40
acc(VF1,VF2) 60	acc(VF1,VF2) 70	acc(VF1,VF2) 85	acc(VF1,VF2) 60
sq -3600	sq -4900	sq -7225	sq -3200

Variante 2

Reihenfolge Attribute

Name, Typ, MaxMiete, Telefon	Telefon, Name, Typ, MaxMiete	MaxMiete, Telefon, Name, Typ	Typ, MaxMiete, Telefon, Name
acc(VF1) 10	acc(VF1) 0	acc(VF1) 40	acc(VF1) 10
acc(VF2) 40	acc(VF2) 10	acc(VF2) 10	acc(VF2) 0
acc(VF1,VF2) 60	acc(VF1,VF2) 100	acc(VF1,VF2) 60	acc(VF1,VF2) 100
sq -3200	sq -10000	sq -3200	sq -10000

Variante 3

Reihenfolge Attribute

Name, Typ, MaxMiete, Telefon	Telefon, Name, Typ, MaxMiete	MaxMiete, Telefon, Name, Typ	Typ, MaxMiete, Telefon, Name
acc(VF1) 40	acc(VF1) 25	acc(VF1) 40	acc(VF1) 50
acc(VF2) 0	acc(VF2) 0	acc(VF2) 10	acc(VF2) 0
acc(VF1,VF2) 70	acc(VF1,VF2) 85	acc(VF1,VF2) 60	acc(VF1,VF2) 60
sq -4900	sq -7225	sq -3200	sq -3600

Fragmente mit bester Trennqualität

Kunde1: $\pi_{Name,Typ}(Kunde)$ Kunde2: $\pi_{MaxMiete,Telefon}(Kunde)$

Abbildung 13.6.: Splitting und Fragmentierung

Korrektheit:

vollständigkeit

Die Datenelemente der Spalte KNr sind in beiden Fragmenten enthalten. Die Spalten Name und Typ sind in der Fragmente Kunde1 enthalten. Die Spalten MaxMiete und Telefon sind in Kunde2 enthalten. FNR muss laut Aufgabenstellung nicht berücksichtigt

rekonstruierbar
disjunktheit

SELECT KNr, Name, Telefon, Typ, MaxMiete FROM Kunde1 JOIN Kunde2 USING KNr; Dieses Statement ergibt die Tabelle Kunde.
Kunde1 und Kunde2 haben nur die Spalte KNr gemeinsam.
SELECT Name, Typ FROM Kunde1; SELECT MaxMiete, Telefon FROM Kunde2; Diese Statements sind nicht möglich.

Abbildung 13.7.: Korrektheit

14. Lab 8 - Replikation und Konflikte

14.1. Setup Master Site

1. Connect as System at Master Site
2. Replication Administrator anlegen
3. Rechte an Admin geben
4. Proxy User angeben

Bemerkung Diese Schritte sind schon durchgeführt worden! in der Übung aber könnten relevant sein

Master Site Setup

```
1  CONNECT system@telesto
2  CREATE USER repadmin IDENTIFIED BY &password
3  DEFAULT TABLESPACE studspace
4  TEMPORARY TABLESPACE temp;
5  BEGIN
6  DBMS_REPCAT_ADMIN.GRANT_ADMIN_ANY_SCHEMA (
7  username => 'repadmin');
8  END;
9
10 BEGIN
11 DBMS_REPCAT_ADMIN.REGISTER_USER_REPGROUP (
12 username => 'repadmin',
13 privilege_type => 'receiver',
14 list_of_gnames => NULL);
15 END;
16
17 CREATE USER proxy_mviewadmin IDENTIFIED BY &password
18 DEFAULT TABLESPACE studspace
19 TEMPORARY TABLESPACE temp;
20 -- any GRANTS?
21 BEGIN
22 DBMS_REPCAT_ADMIN.REGISTER_USER_REPGROUP (
23 username => 'proxy_mviewadmin',
24 privilege_type => 'proxy_snapadmin',
25 list_of_gnames => NULL);
26 END;
27
28 CREATE USER proxy_refresher IDENTIFIED BY &password
29 DEFAULT TABLESPACE studspace
30
31 TEMPORARY TABLESPACE temp;
32
33 GRANT CREATE SESSION TO proxy_refresher;
34 GRANT SELECT ANY TABLE TO proxy_refresher;
```

14.2. Setup Materialized View Site

1. Connect as System at Materialized View Site
2. Users anlegen
3. Database Links erstellen
4. Schedule Purge

Bemerkung Diese Schritte sind schon durchgeführt worden! aber könnten relevant sein

SQL-Skript

```

1 CONNECT system@calypso
2 CREATE USER mviewadmin IDENTIFIED BY &password
3 DEFAULT TABLESPACE studspace
4 TEMPORARY TABLESPACE temp;
5
6 BEGIN
7 DBMS_REPCAT_ADMIN.GRANT_ADMIN_ANY_SCHEMA (
8 username => 'mviewadmin');
9 END;
10
11 GRANT COMMENT ANY TABLE TO mviewadmin;
12 GRANT LOCK ANY TABLE TO mviewadmin;
13
14
15 CREATE USER propagator IDENTIFIED BY &password
16 DEFAULT TABLESPACE studspace
17 TEMPORARY TABLESPACE temp;
18
19 BEGIN
20 DBMS_DEFER_SYS.REGISTER_PROPAGATOR (
21 username => 'propagator');
22 END;
23
24 --Create DB Links: as SYSTEM on calypso
25 CREATE PUBLIC DATABASE LINK telesto.janus.fhnw.ch
26 USING 'telesto';
27
28 -- for mviewadmin db link to proxy_mviewadmin on master site
29 CONNECT mviewadmin@calypso;
30
31 CREATE DATABASE LINK telesto.janus.fhnw.ch
32 CONNECT TO proxy_mviewadmin IDENTIFIED BY &password;
33
34 -- for propagator db link to repadmin on master site
35 CONNECT propagator@calypso
36
37 CREATE DATABASE LINK telesto.janus.fhnw.ch
38 CONNECT TO repadmin IDENTIFIED BY &password;
39 -- as mviewadmin on calypso
40 CONNECT mviewadmin@calypso;
41
42 BEGIN
43 DBMS_DEFER_SYS.SCHEDULE_PURGE (
44 next_date => SYSDATE,
45 interval => 'SYSDATE + 1/24',
46 delay_seconds => 0,
47 rollback_segment => '');
48 END;

```

14.3. Erstellen von Master Group

1. Schema an der Master Site erstellen
2. Master Group erstellen
3. Objekte zur Master Group hinzufügen
4. Replication Support einrichten
5. Replikation starten

SQL-Skript

```

1 -- Diese Schritte müssen für neue Situation durchgeführt werden!
2 CONNECT repadmin@telesto
3
4 BEGIN
5 DBMS_REPCAT.CREATE_MASTER_REPGROUP (
6 gname => 'mvdbsx2_repg');

```

```

7  END;
8
9  BEGIN
10 DBMS_REPCAT.CREATE_MASTER_REPOBJECT (
11 gname => 'mvdbxsx2_repg',
12 type => 'TABLE',
13 oname => 'filialen',
14 sname => 'mvdbxsx2',
15 use_existing_object => TRUE,
16 Copy_rows => FALSE);
17 END;
18
19 BEGIN
20 DBMS_REPCAT.GENERATE_REPLICATION_SUPPORT (
21 sname => 'mvdbxsx2',
22 oname => 'filialen',
23 type => 'TABLE',
24 min_communication => TRUE);
25 END;
26
27 BEGIN
28 DBMS_REPCAT.RESUME_MASTER_ACTIVITY (
29 gname => 'mvdbxsx2_repg');
30 END;

```

14.4. Erstellen von Master View Group

1. Materialized View Logs an der Master Site erstellen
2. Repliziertes Schema und Database Links erstellen
3. Materialized View Group erstellen
4. Refresh Group erstellen
5. Objekte zur Materialized View Group hinzufügen
6. Objekte zur Refresh Group hinzufügen
7. Manuell Refresh durchführen

SQL-Skript

```

1  -- Bemerkung Diese Schritte müssen für neue Situation durchgeführt werden!
2  -- use dvd rental sample user mvdbxsx2
3  CONNECT mvdbxsx2@telesto
4
5  CREATE MATERIALIZED VIEW LOG ON mvdbxsx2.filialen;
6  -- for all owner db link to proxy_refresher on master site
7  CONNECT mvdbxsx2@calypso
8
9  CREATE DATABASE LINK telesto.janus.fhnw.ch
10 CONNECT TO proxy_refresher IDENTIFIED BY &password;
11 -- create materialized view group
12 CONNECT mviewadmin@calypso
13
14 BEGIN
15 DBMS_REPCAT.CREATE_MVIEW_REPGROUP (
16 gname => 'mvdbxsx2_repg',
17 master => 'telesto.janus.fhnw.ch',
18 propagation_mode => 'ASYNCHRONOUS');
19 END;
20 /
21 -- create refresh group
22 -- with scheduled refresh
23 BEGIN
24 DBMS_REFRESH.MAKE (
25 name => 'mviewadmin.mvdbxsx2_refg',
26 list => '',
27 next_date => SYSDATE,
28 interval => 'SYSDATE + 1/24',

```



```

29 implicit_destroy => FALSE,
30 rollback_seg => '',
31 push_deferred_rpc => TRUE,
32 refresh_after_errors => FALSE);
33 END;
34 /
35 -- Add objects to the materialized view group
36 -- as mvdbxs2
37 CONNECT mvdbxs2@calypso
38
39 CREATE MATERIALIZED VIEW mvdbxs2.filialen
40 REFRESH FAST WITH PRIMARY KEY FOR UPDATE
41 AS SELECT * FROM mvdbxs2.filialen@telesto.janus.fhnw.ch;
42
43 -- add to rep group
44 -- as mviewadmin
45 CONNECT mviewadmin@calypso
46
47 BEGIN
48 DBMS_REPCAT.CREATE_MVIEW_REPOBJECT (
49 gname => 'mvdbxs2_repg',
50 sname => 'mvdbxs2',
51 oname => 'filialen',
52 type => 'SNAPSHOT',
53 min_communication => TRUE);
54 END;
55 /
56 -- as mviewadmin
57 BEGIN
58 DBMS_REFRESH.ADD(
59 name => 'mviewadmin.mvdbxs2_refg',
60 list => 'mvdbxs2.filialen',
61 lax => TRUE);
62 END;
63 /
64 -- refresh on demand
65 -- as mviewadmin
66 BEGIN
67 DBMS_REFRESH.REFRESH (
68 name => 'mviewadmin.mvdbxs2_refg' );
69 end;
70 /

```

14.5. Lösungen der Konflikte

14.5.1. Update Konflikt Lösung

Im folgenden soll eine Update-Konflikt-Lösung definiert werden, damit bei einem Update, ein allfälliger Konflikt automatisch behoben wird.

Zusätzlich wird überprüft, dass bei einem INSERT- oder DELETE-Statement die Lösungsstrategie nicht angewendet wird.

Definieren der Update-Konflikt-Lösung als repadmin auf telesto

```

1 BEGIN
2   DBMS_REPCAT.SUSPEND_MASTER_ACTIVITY (
3     gname => 'mvdbxs02_repg');
4 END;
5 /
6
7 BEGIN
8   DBMS_REPCAT.MAKE_COLUMN_GROUP (
9     sname => 'mvdbxs02',
10    oname => 'filialen',
11    column_group => 'fil_cg',
12    list_of_column_names => 'fnr,ort,plz');
13 END;
14 /
15 BEGIN
16   DBMS_REPCAT.ADD_UPDATE_RESOLUTION (
17     sname => 'mvdbxs02',
18     oname => 'filialen',

```

```

19     column_group => 'fil_cg',
20     sequence_no => 1,
21     method => 'DISCARD',
22     parameter_column_name => 'fnr,ort,plz');
23 END;
24 /
25 BEGIN
26     DBMS_REPCAT.GENERATE_REPLICATION_SUPPORT (
27         sname => 'mvdba02',
28         oname => 'filialen',
29         type => 'TABLE',
30         min_communication => TRUE);
31 END;
32 /
33 BEGIN
34     DBMS_REPCAT.RESUME_MASTER_ACTIVITY (
35         gname => 'mvdba02_repg');
36 END;
37 /

```

14.5.2. Uniqueness Konflikt Lösung

Definition der Uniqueness-Konflikt-Lösung als repadmin auf telesto

```

1 BEGIN
2     DBMS_REPCAT.SUSPEND_MASTER_ACTIVITY (
3         gname => 'mvdba02_repg');
4 END;
5 /
6
7 BEGIN
8     DBMS_REPCAT.ADD_UNIQUE_RESOLUTION (
9         sname => 'mvdba02',
10        oname => 'filialen',
11        constraint_name => 'fl_pk',
12        sequence_no => 1,
13        method => 'DISCARD',
14        parameter_column_name => 'fnr');
15 END;
16 /
17 BEGIN
18     DBMS_REPCAT.GENERATE_REPLICATION_SUPPORT (
19         sname => 'mvdba02',
20         oname => 'filialen',
21         type => 'TABLE',
22         min_communication => TRUE);
23 END;
24 /
25 BEGIN
26     DBMS_REPCAT.RESUME_MASTER_ACTIVITY (
27         gname => 'mvdba02_repg');
28 END;
29 /

```

14.5.3. Delete Konflikt

Im folgenden soll eine Delete-Konflikt-Lösung definiert werden, damit bei einem DELETE-Statement, welches einen Konflikt auslöst, diesen automatisch behebt. Bei der Delete-Konflikt-Lösung wird eine User-Funktion¹ benötigt, die den Konflikt löst.

Definieren des Konflikt-Lösungs-Handler als repadmin auf telesto

```

1 CREATE OR REPLACE FUNCTION mvdba02_fl_delete_handler (
2     old_fnr IN OUT VARCHAR2,
3     old_ort IN OUT VARCHAR2,
4     old_plz IN OUT VARCHAR2,
5     ignore_discard_flag OUT BOOLEAN ) RETURN BOOLEAN IS
6

```

¹http://docstore.mik.ua/oreilly/oracle/bipack/ch16_05.htm

```
7 BEGIN
8 DELETE FROM mvdb02.filialen
9 WHERE fnr = old_fnr;
10
11 ignore_discard_flag := TRUE;
12
13 RETURN TRUE;
14 END mvdb02_fl_delete_handler;
```

Definieren der Delete-Konflikt-Lösung als repadmin auf telesto

```
1 BEGIN
2 DBMS_REPCAT.SUSPEND_MASTER_ACTIVITY (
3 gname => 'mvdb02_repg');
4 END;
5 /
6 BEGIN
7 DBMS_REPCAT.ADD_DELETE_RESOLUTION(
8 sname => 'mvdb02',
9 oname => 'filialen',
10 sequence_no => 1,
11 parameter_column_name => 'fnr,ort,plz',
12 function_name => 'mvdb02_fl_delete_handler');
13 END;
14 /
15 BEGIN
16 DBMS_REPCAT.GENERATE_REPLICATION_SUPPORT (
17 sname => 'mvdb02',
18 oname => 'filialen',
19 type => 'TABLE',
20 min_communication => TRUE);
21 END;
22 /
23 BEGIN
24 DBMS_REPCAT.RESUME_MASTER_ACTIVITY (
25 gname => 'mvdb02_repg');
26 END;
27 /
```

15. Lab 11 - Relationale Operationen mit Map Reduce

15.1. Aufgabenstellung

Tabellen werden im csv-Format abgegeben. Nun können SQL-Befehle als Map/Reduce-Funktionen in einer Hadoop Umgebung realisiert werden.

15.2. Group by mit Count

SQL-Skript

```
1 SELECT mnr, COUNT(*)
2 FROM registrierungen
3 GROUP BY mnr;
```

```
1 /**
2  * Implementing a group by with count in a select statement with apache hadoop.
3  * This algorithm is similar to the wordcount example.
4  */
5 public class GroupByWithCount {
6
7     private static int index = 0;
8
9     /**
10    * Mapper which is able to process a csv-file.
11    * The first column must be the key in the group by statement.
12    * Stores the number one with every line it maps.
13    */
14    public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
15        private final static IntWritable one = new IntWritable(1);
16        private Text word = new Text();
17
18        @Override
19        protected void map(LongWritable key, Text value, Context context) throws IOException,
20            InterruptedException {
21            // one line in the csv
22            String line = value.toString();
23            // get columns by splitting String
24            String[] cols = line.split(";");
25            // first column = mnr
26            word.set(cols[index]);
27            context.write(word, one);
28        }
29    }
30
31    /**
32    * Reducer which sums up the integers stored by the mapper.
33    */
34    public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
35
36        @Override
37        protected void reduce(Text key, Iterable<IntWritable> values, Context context) throws
38            IOException, InterruptedException {
39            int sum = 0;
40            Iterator<IntWritable> it = values.iterator();
41            while (it.hasNext()) {
42                sum += it.next().get();
43            }
44            context.write(key, new IntWritable(sum));
45        }
46    }
47 }
```

```

45 }
46
47 public static void main(String[] args) throws Exception {
48     Configuration conf = new Configuration();
49     String[] otherArgs = new GenericOptionsParser(conf, args)
50         .getRemainingArgs();
51     if (otherArgs.length != 3) {
52         System.err.println("Usage: groupbywithcount.jar <csv-File> <output-Directory> <key>")
53         ;
54         System.exit(2);
55     }
56     index = Integer.parseInt(otherArgs[2]);
57     Job job = Job.getInstance(conf, "groupbywithcount");
58     job.setInputFormatClass(TextInputFormat.class);
59     job.setMapperClass(Map.class);
60     job.setReducerClass(Reduce.class);
61     job.setOutputKeyClass(Text.class);
62     job.setOutputValueClass(IntWritable.class);
63     job.setOutputFormatClass(TextOutputFormat.class);
64     FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
65     FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
66     System.exit(job.waitForCompletion(true) ? 0 : 1);
67 }
68 }
69 }
70 }
71 }
72 }
73 }
74 }

```

15.3. Join

SQL-Skript

```

1 SELECT *
2 FROM registrierungen JOIN mitglieder USING (mnr);

```

```

1 /**
2  * Implementing a map-reduce-join with apache hadoop.
3  * This algorithm joins two csv-files using the first index of each.
4  * Join-Type: Reduce-Side-Join
5  */
6 public class Join {
7
8     private static int leftIndex = 0;
9     private static int rightIndex = 0;
10
11     /**
12      * Input-Mapper for the first file. This file must be a csv.
13      * KeyOut is the join key. ValueOut is the according line in the csv without the join key
14      */
15     public static class RightInputMapper extends Mapper<LongWritable, Text, Text,
16         DatasetTagger> {
17         private Text word = new Text();
18
19         @Override
20         protected void map(LongWritable key, Text value, Context context) throws IOException,
21             InterruptedException {
22             // fetching one line in the csv
23             String line = value.toString();
24             // extracting columns by splitting the String
25             String[] cols = line.split(",");
26             // wrapping the value of the first column in a Text
27             word.set(cols[rightIndex]);
28             // Rest of the line will be stored in a custom datastructure with a tag from which
29             // file it comes
30             String rest = assembleLineWithoutIndex(cols, rightIndex);
31             DatasetTagger ckw = new DatasetTagger(rest, DatasetTagger.Side.RIGHT);

```

```

29     context.write(word, ckw);
30 }
31 }
32 }
33
34 /**
35  * Input-Mapper for the second file. This file must be a csv.
36  * KeyOut is the join key. ValueOut is the according line in the csv without the join key
37  */
38 public static class LeftInputMapper extends Mapper<LongWritable, Text, Text,
39     DatasetTagger> {
40     private Text word = new Text();
41
42     @Override
43     protected void map(LongWritable key, Text value, Context context) throws IOException,
44         InterruptedException {
45         // one line in the csv
46         String line = value.toString();
47         // get columns by splitting String
48         String[] cols = line.split(";");
49         // first column = mnr
50         word.set(cols[leftIndex]);
51         String rest = assembleLineWithoutIndex(cols, leftIndex);
52         DatasetTagger ckw = new DatasetTagger(rest, DatasetTagger.Side.LEFT);
53         context.write(word, ckw);
54     }
55 }
56
57 /**
58  * Assembles the splitted line of a csv-File excluding the joining index.
59  * @param line Line of a csv-File splitted over the semicolon
60  * @param index Joining Index
61  * @return Resulting String.
62  */
63 private static String assembleLineWithoutIndex(String[] line, int index) {
64     StringBuilder sb = new StringBuilder();
65     for (int i = 0; i < line.length; i++) {
66         if (i != index) {
67             sb.append(line[i] + ";");
68         }
69     }
70     sb.deleteCharAt(sb.length()-1);
71     return sb.toString();
72 }
73
74 /**
75  * KeyOut is the join key. ValueOut is the sorted lines concatenated to several strings.
76  */
77 public static class Reduce extends Reducer<Text, DatasetTagger, Text, Text> {
78
79     @Override
80     protected void reduce(Text key, Iterable<DatasetTagger> values, Context c) throws
81         IOException, InterruptedException {
82         // Values of the mapper which processes the file which is on the right side of the
83         // join
84         List<String> right = new ArrayList<String>();
85         // Values of the mapper which processes the file which is on the left side of the
86         // join
87         List<String> left = new ArrayList<String>();
88
89         // Fill the two Lists with the values of the Iterable
90         Iterator<DatasetTagger> it = values.iterator();
91         DatasetTagger v;
92         while(it.hasNext()) {
93             v = it.next();
94             if (v.getSide() == DatasetTagger.Side.LEFT) {
95                 left.add(v.getValue());
96             } else if (v.getSide() == DatasetTagger.Side.RIGHT) {
97                 right.add(v.getValue());
98             }
99         }
100
101         // Concatenate the output according to the join-rules
102         for (String l : left) {
103             for (String r : right) {

```

```

99         String res = rearrangeOutput(l, r);
100         Text val = new Text(res);
101         c.write(key, val);
102     }
103 }
104 }
105 }
106
107 /**
108  * Rearrange the output from the two sides so that it displays tabs instead of a
109  * semicolon between the columns
110  * @param right String from the right side of the join.
111  * @param left String from the left side of the join.
112  * @return String ready to be displayed.
113  */
114 private static String rearrangeOutput(String right, String left) {
115     StringBuilder res = new StringBuilder();
116     String[] r = right.split(";");
117     String[] l = left.split(";");
118     for (String string : r) {
119         res.append(string + "\t");
120     }
121     for (String string : l) {
122         res.append(string + "\t");
123     }
124     return res.toString();
125 }
126
127 public static void main(String[] args) throws Exception {
128     Configuration conf = new Configuration();
129     String[] otherArgs = new GenericOptionsParser(conf, args)
130         .getRemainingArgs();
131     if (otherArgs.length != 5) {
132         System.err.println("Usage: join.jar <left-csv-File> <right-csv-File> <output-Directory> <left-key> <right-key>");
133         System.exit(2);
134     }
135     Job job = Job.getInstance(conf, "join");
136
137     // Left csv
138     Path in1 = new Path(otherArgs[0]);
139     // Right csv
140     Path in2 = new Path(otherArgs[1]);
141     // Output
142     Path out = new Path(otherArgs[2]);
143     // Left joining index
144     leftIndex = Integer.parseInt(otherArgs[3]);
145     // Right joining index
146     rightIndex = Integer.parseInt(otherArgs[4]);
147
148     job.setReducerClass(Reduce.class);
149
150     // Setting these is necessary. Otherwise it does not recognize the custom class
151     job.setMapOutputKeyClass(Text.class);
152     job.setMapOutputValueClass(DatasetTagger.class);
153
154     job.setOutputKeyClass(Text.class);
155     job.setOutputValueClass(Text.class);
156
157     job.setOutputFormatClass(TextOutputFormat.class);
158
159     FileOutputFormat.setOutputPath(job, out);
160
161     MultipleInputs.addInputPath(job, in1, TextInputFormat.class, LeftInputMapper.class);
162     MultipleInputs.addInputPath(job, in2, TextInputFormat.class, RightInputMapper.class);
163
164     System.exit(job.waitForCompletion(true) ? 0 : 1);
165 }
166
167 }

```

```

1 /**
2  * Wraps and Tags a String in an object that can be processed by hadoop for a join-
   statement.

```

```

3  */
4  public class DatasetTagger implements Writable {
5      /**
6       * Enum to determine on which side the wrapped value is.
7       */
8      public static enum Side {RIGHT, LEFT};
9
10     private String value;
11     private Side side;
12
13     public DatasetTagger() {
14     }
15
16     public DatasetTagger(String value, Side side) {
17         this.value = value;
18         this.side = side;
19     }
20
21     public void readFields(DataInput dataInput) throws IOException {
22         value = WritableUtils.readString(dataInput);
23         side = WritableUtils.readEnum(dataInput, Side.class);
24     }
25
26     public void write(DataOutput dataOutput) throws IOException {
27         WritableUtils.writeString(dataOutput, value);
28         WritableUtils.writeEnum(dataOutput, side);
29     }
30
31     public String getValue() {
32         return value;
33     }
34
35     public void setValue(String value) {
36         this.value = value;
37     }
38
39     public Side getSide() {
40         return side;
41     }
42
43     public void setSide(Side side) {
44         this.side = side;
45     }
46
47     @Override
48     public String toString() {
49         return "DatasetTag [value=" + value + ", side=" + side + "]";
50     }
51
52
53 }

```

15.3.1. Lösungsbeschreibung

Group by mit Count

In diesem Algorithmus muss ausgerechnet werden, wie viel mal eine bestimmte Spalte im csv-File vorkommt.

Der Mapper bekommt nun jede Reihe vom csv-File einzeln. Somit kann diese Linie aufgesplittet und der gewünschte Index extrahiert werden. Dieser extrahierte Wert wird in ein Text-Objekt gewrappt und mit einer 1, gewrappt in ein IntWritable, in den Context geschrieben.

Der Reducer bekommt als key-value Paar ein Text, der den gewünschte Index erhält und eine List von Einsen, die der Mapper in den Context geschrieben hat. Aus dieser Liste kann nun ermittelt werden, wie viel mal das Wort im csv-File vorgekommen ist. Als Resultat wird wiederum der Text mit der Summer in den Context geschrieben. Dieser Wert wird dann auf der Konsole ausgegeben.

15.3.2. Join

Der Reduce-Side Join verlangt, dass der eigentliche Join im Reducer durchgeführt werden muss. Der Mapper ist somit für die Zuordnung der Datensätze zum Join-Key zuständig. Damit zusätzlich ein Left-Join erreicht werden kann, muss der ValueOut des Mappers gekennzeichnet werden, von welcher Seite, bzw. welchem File die Zeile stammt. Dies setzt eine eigene Datenstruktur voraus, die zusätzlich vom Hadoop erkannt werden muss. Als Attribute

muss ein Datenelement einen String, sowie eine Enumeration beinhalten. Die Enumeration definiert dabei von welcher Seite des Joins es stammt. Dadurch ist es möglich, im Reducer das Konkatenieren des Endresultats nach den Regeln eines Left-Joins durchzuführen. Damit Hadoop diese Datenstruktur als ValueOut des Mappers erkennt, muss dieser als MapOutputValue konfiguriert werden.