Google Spanner

Proseminar Ein-/Ausgabe Stand der Wissenschaft

Hanno Harte

Betreuer: Julian Kunkel

24.6.13

Gliederung

- Überblick
- Funktionsweise
    - True Time
    - Konsistenzsemantik
- Benchmarks
- Zusammenfassung

Überblick

- Projekt der GoogleInc. seit 2009

- Dezentralisiert

- Globale Datenbank
    - 13 Rechenzentren / 1 Million Server
    - Bald bis zu 10 Millionen Server ?

- Hohe Verfügbarkeit von Daten
    - Datenredundanzen
    - Verschiebung von Daten

- Geringe Latenz

Überblick

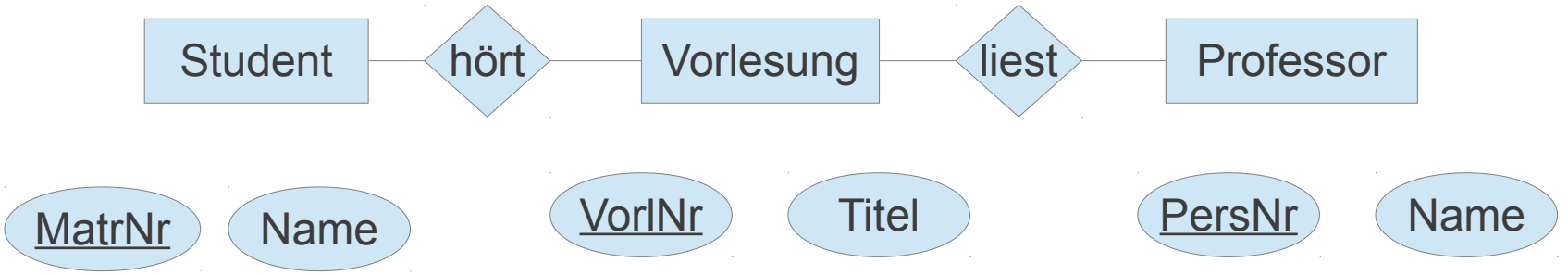
Exkurs SQL:
- Sprache in Datenbanken zur Bearbeitung und Abfrage von Datenbeständen

- Definition von Datenstrukturen

- Für relationale Datenbanken

- Angelehnt an englische Umgangssprache

Überblick



Beispiel Sql Datenbank

Student

MatrNr	Name
26120	Fichte
25403	Jonas
27103	Fauler

Hört

MatrNr	VorlNr
25403	5001
26120	5001
26120	5045

Vorlesung

VorlNr	Titel	PersNr
5001	ET	15
5022	IT	12
5045	DB	12

Professor

PersNr	Name
12	Wirth
15	Tesla
20	Urlauber

Überblick

- Multi-versionale Datenbank mit Zeitstempel

- SQL ähnliches relationales Datenbanksystem
    - nicht vollständig relational

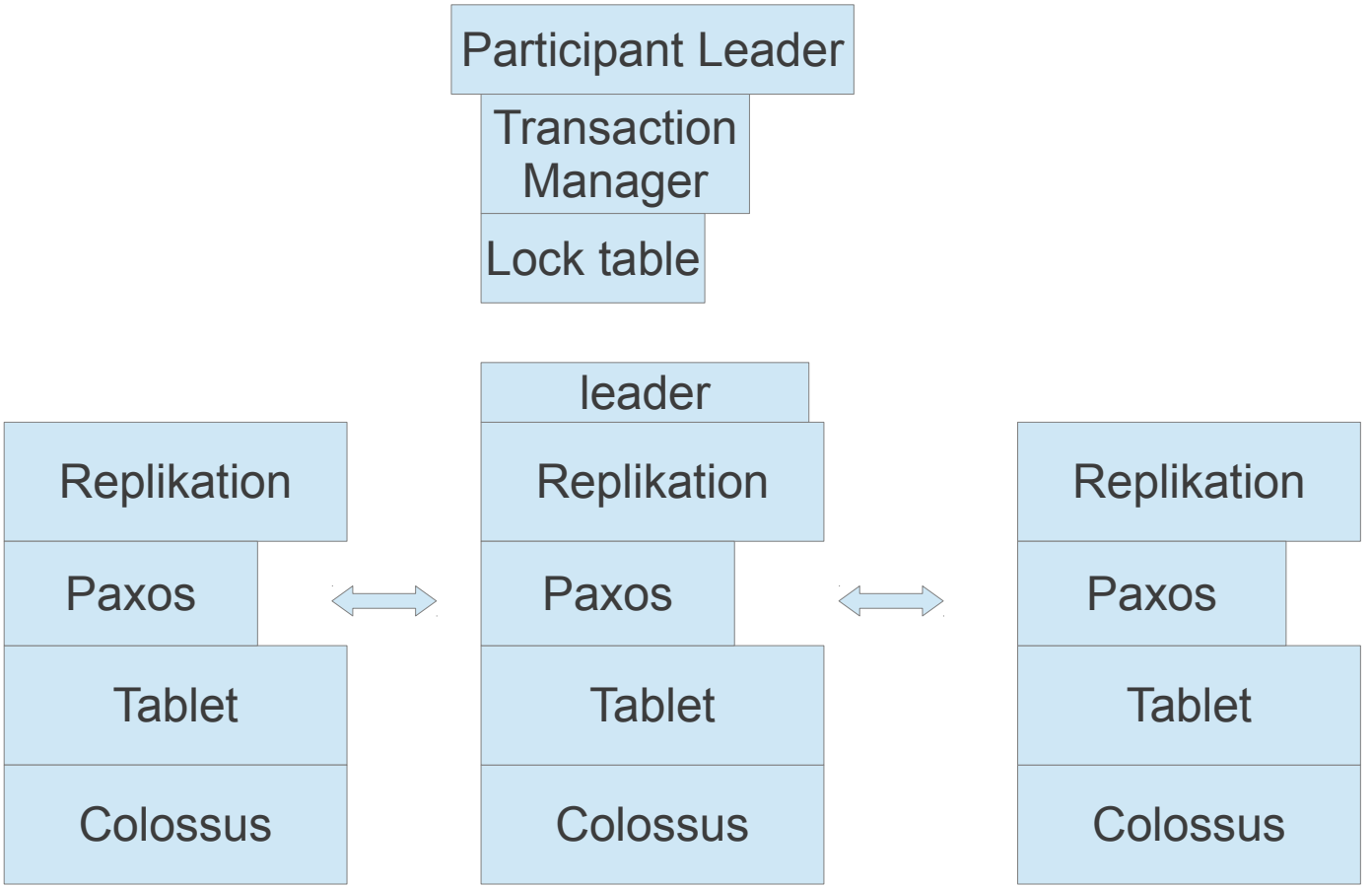
Funktionsweise

- Zugriff auf eine Information in einer Google Datenbank
- Information in einer Tabelle
- SQL ähnliche relationale Datenbank
- Tabellen auf einem „Spannserver“
    - B-Trees auf „Colossus“

Funktionsweise

- Background-Anwendung „Movedir“
    - erzeugt und löscht Replikationen
    - nicht blockierend
- 100-1000 Datenstrukturen auf einem Spannserver
    - genannt Tablet
- So genannter Paxos über dem Tablet

Funktionsweise



Exkurs Paxos:

- Protokolle zur Vereinfachung der Interaktion zwischen Prozessoren

- Anfrage an mehrere Prozessoren

- Eine Replikation pro Prozessor

Exkurs Paxos:

Aufgaben der Prozessoren:

- Client
- Acceptor
- Proposer
- Leader
- Learner

Exkurs Paxos:

- Long lived leader
- tt.after(smax) = true; smax = maximaler Zeitstempel
- Gleiche Zustände übermittelt
- Vereinfachung der Verschiebung zwischen Servern
- Viele Anfragen gleichzeitig möglich
- Korrekte Reihenfolge garantiert

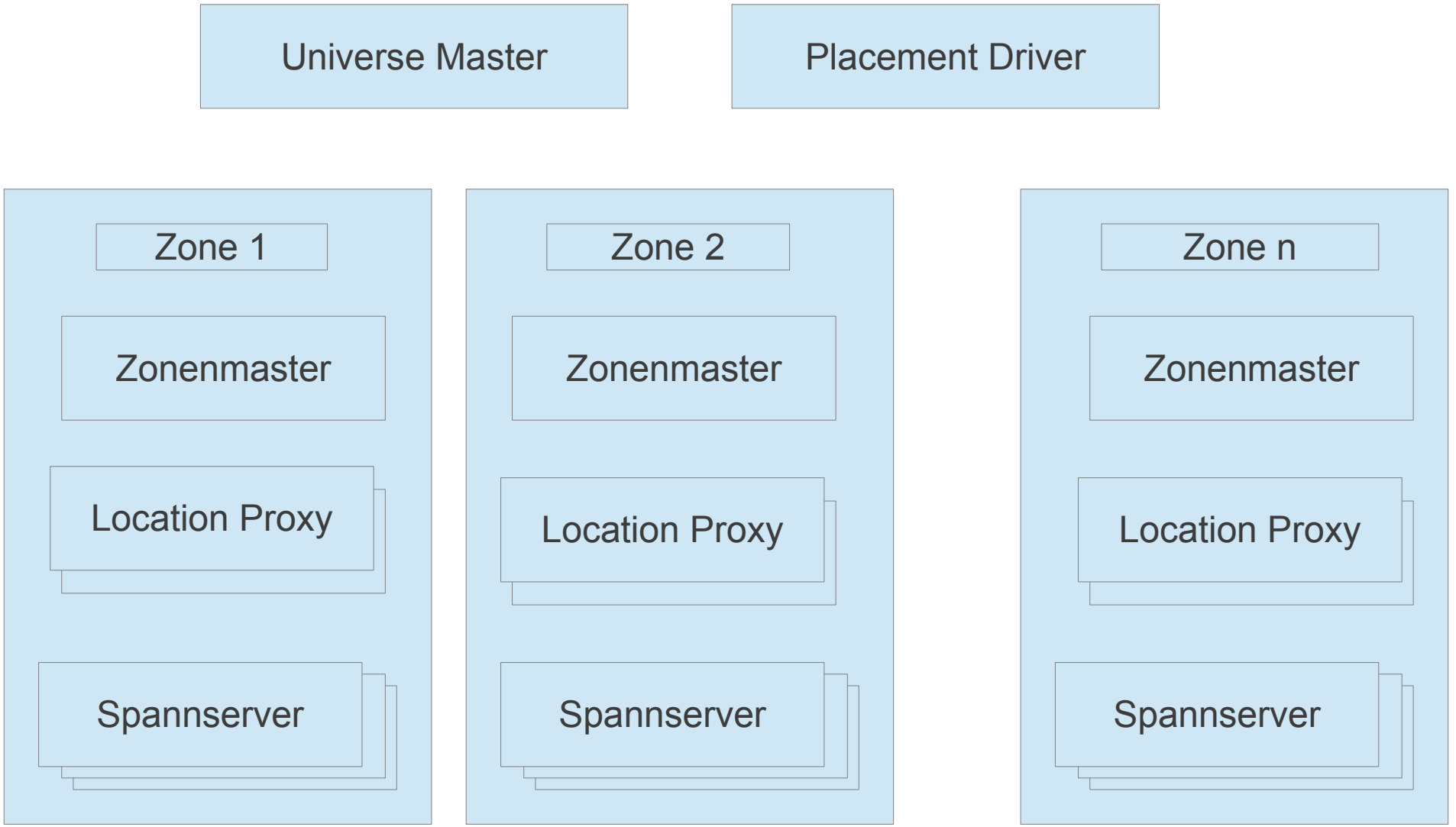
Funktionsweise

- Schreibt in 2 logs
- Paxos Group
- Schreibende Zugriffe müssen Paxos Protokoll initiieren
- Lesende Zugriffe vom Tablet verwaltet

Funktionsweise

- Lock table für:
- 2 Phase locking
    - Expending, Shrinking Phase
- Transaction Manager für Koordination

Funktionsweise



Funktionsweise

- Location Proxies für Position des Spannservers
- Zonen
    - 1 Zonenmaster, bis zu einigen tausend Spannservern
- ZonenMaster als Verwaltungseinheit
- Zonen nicht lokal gebunden

Funktionsweise

- Universe Master für interaktives Debugging
- Placement Driver
    - Verschieben von Daten bei Überfüllung
    - Abgleichen der Replikationen
- Gesamter Einsatzbereich als Universum bezeichnet

True Time

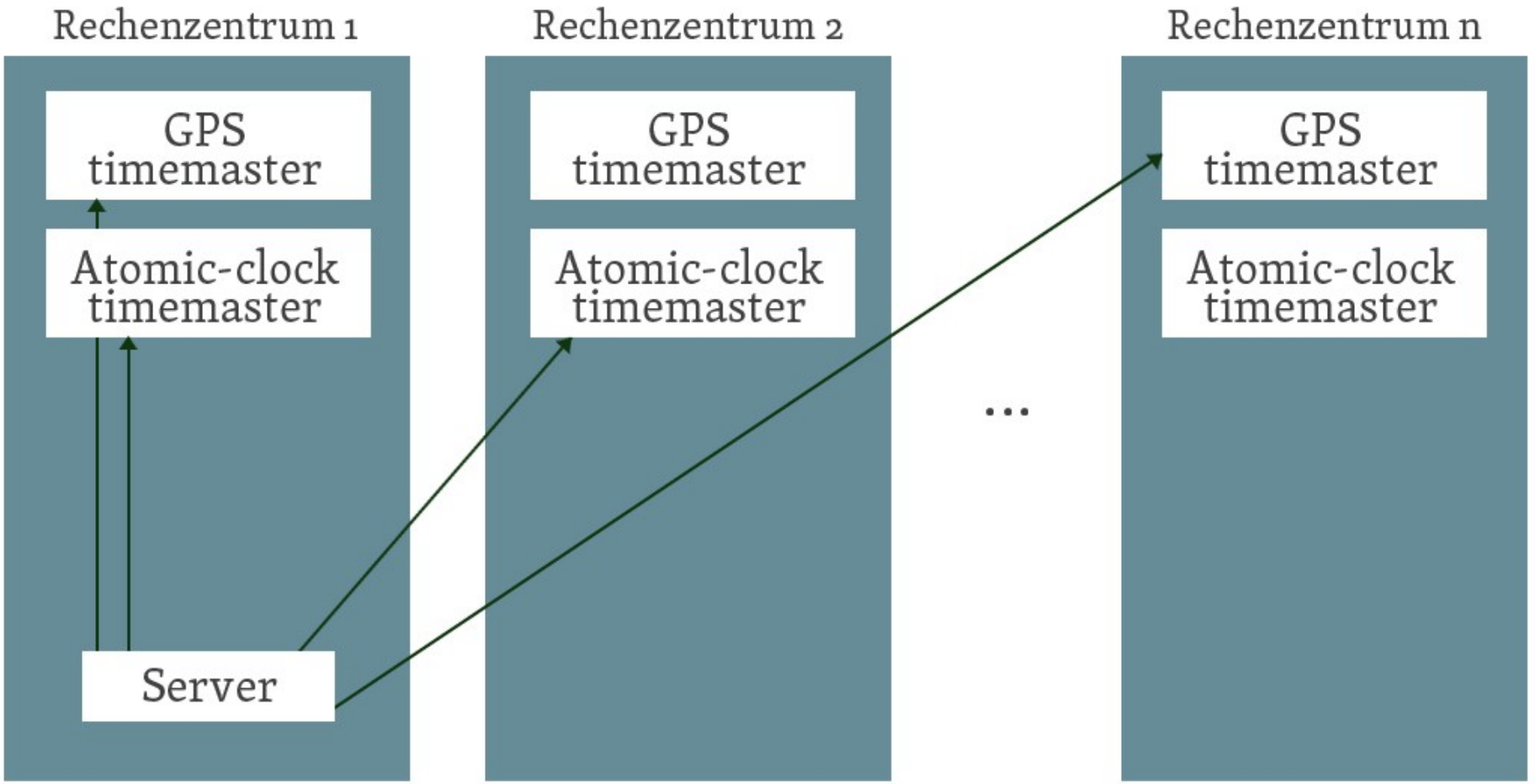
- NTP (Network Time Protocoll)

- TrueTime Googles eigene API
  → Eigene Atomuhren und GPS Reciever

- „Timeslave deamon“
  → Abgleich mit Master Server

- Abweichungen einberechnet

True Time


Rechenzentrum 1
GPS timemaster
Atomic-clock timemaster
Server
Rechenzentrum 2
GPS timemaster
Atomic-clock timemaster
...
Rechenzentrum n
GPS timemaster
Atomic-clock timemaster

[4]

True time

Method	Returns
TT.now()	TTinterval [earliest, latest]
TT.after(t)	True if t has definitely passed
TT.before(t)	True if t has definitely not arrived

[1]

- TT.now() erzeugt den Zeitstempel
- TTinterval absolute Zeit in der TT.now() benutzt wurde

Konsistenzsemantik

- Read Only:

- Vorteile von Snapshot isolation

- Muss vordeklariert sein als read only

- Blockiert nicht

- Read only als snapshot reads implementiert

    s-read = tt.now().latest

Konsistenzsemantik

- Snapshot reads

- Scope Ausdruck

- Zeitstempel in monoton steigender Reihenfolge übergeben

Konsistenzsemantik

Read / Write:

- Writes buffern auf dem Client

- Wound/wait bei den reads innerhalb von read/write

- Read locks

Konsistenzsemantik

2 Phase Commit:

- Protokoll zur Sicherung der Transaktionen bei Fehlern

- 2 Phasen:
    - 1. Phase
        - Nachricht des Koordinators an Ausführende Stellen
        - Führt Transaktion aus und schreibt in 2 logs undo/redo
    - 2. Phase:
        - Antwort : Ja
            - Transaktion wird fertiggestellt und locks freigegeben
        - Antwort : Nein
            - Zurücksetzen der Transaktion mit undo log

Konsistenzsemantik

- Write locks
- Prepare Zeitstempel
- Einen Zeitstempel für die gesamte Transaktion
- Wartet nun auf TT.after(s)
    - Garantie das der Zeitstempel in der Vergangenheit liegt
- Locks werden gelöst

Konsistenzsemantik

- read t <= t-safe
- t-safe = min(t-paxos-safe, t-TM-safe)
- t-paxos- safe
- t-TM-safe

Konsistenzsemantik

- Schema Change Transaktion
    - Als „normale“ Transaktion
    - Nicht blockierend
    - Zeitstempel in der Zukunft

Benchmarks

- 4GB Ram, AMD Barcelona 2200MHz
- Netzwerkdistanz <= 1 MS
- 50 Paxos Groups
- Reads and Writes mit 4 KB größe
- Paxos Latenz bei ca 9 ms und Wartezeit ca 5 ms

Benchmarks

	Latenz (ms)	
participants	mean	99th percentile
1	17.0 + - 1.4	75.0 + - 34.9
2	24.5 + - 2.5	87.6 + - 35.9
5	31.5 + - 6.2	104.5 + - 52.2
10	30.0 + - 3.7	95.6 + - 25.4
25	35.5 + - 5.6	100.4 + - 42.7
50	42.7 + - 4.1	93.7 + - 22.9
100	71.4+ - 7.6	131.2 + - 17.6
200	150.5 + - 11	320.3 + - 35.1

- 2 phase commit getestet an 3 Zonen á 25 Spannservern

Zusammenfassung

- Globale Vernetzung aller Rechenzentren
- Minimierung von Latenzen
- Hohe Fehlertoleranz
- Eindeutiger Zeitstempel möglich
- Nicht blockierende Reads

Quellen

[1] http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en//archive/spanner-osdi2012.pdf
[2] http://business.chip.de/news/Google-Spanner-Redundanz-fuer-10-Millionen-Server_42369620.html
[3] http://www.wired.com/wiredenterprise/2012/11/google-spanner-time/
[4] Vorlage: http://www.neogrid.de/Bilder-Lexikon.php?Bild_Nr=8&Feld=Google-Spanner&Serie=no
[5] http://de.wikipedia.org/wiki/SQL