Big Data & Analytics

Nationaler Akademietag, Fulda

20.04.2018

Referent: Meinhard Lingo

E-Mail: meinhard.lingo@bs1in.de

Big Data & Analytics

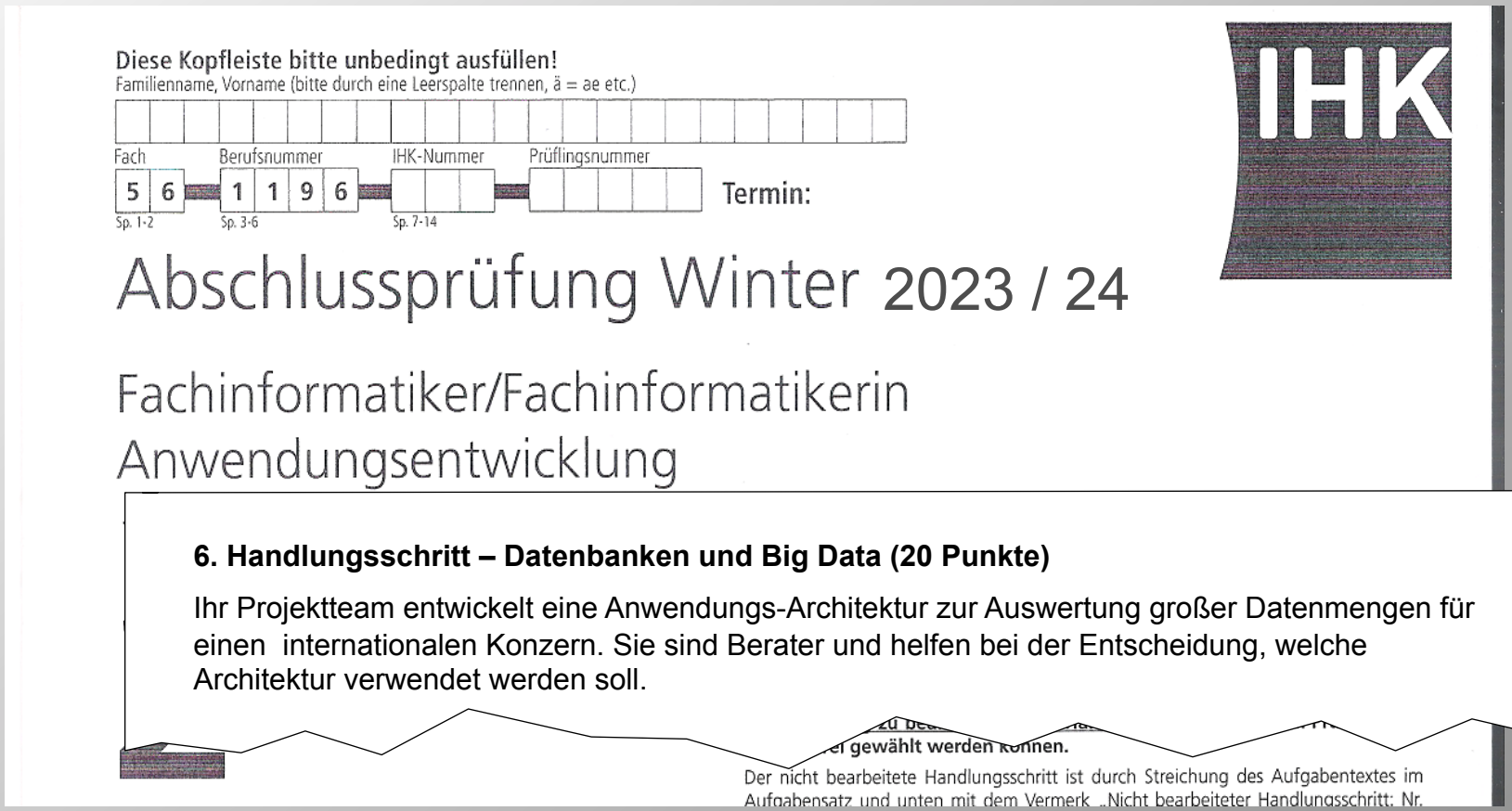
Big Data-Anwendungen: Ein Paradigmenwechsel.

Kompetenzen?

mögliche Unterrichtsinhalte?

Werkzeuge für den Unterricht?

Eine Vision …



Diese Kopfleiste bitte unbedingt ausfüllen!

Familienname, Vorname (bitte durch eine Leerspalte trennen, ä = ae etc.)

Fach	Berufsnummer	IHK-Nummer	Prüflingsnummer	
5 6	1 1 9 6			Termin:
Sp. 1-2	Sp. 3-6	Sp. 7-14		

IHK

Abschlussprüfung Winter 2023 / 24

Fachinformatiker/Fachinformatikerin
Anwendungsentwicklung

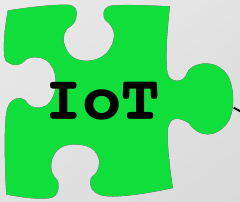
6. Handlungsschritt – Datenbanken und Big Data (20 Punkte)

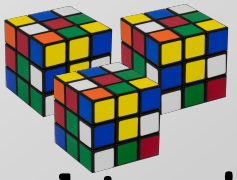
Ihr Projektteam entwickelt eine Anwendungs-Architektur zur Auswertung großer Datenmengen für einen internationalen Konzern. Sie sind Berater und helfen bei der Entscheidung, welche Architektur verwendet werden soll.

…l gewählt werden können.

Der nicht bearbeitete Handlungsschritt ist durch Streichung des Aufgabentextes im Aufgabensatz und unten mit dem Vermerk „Nicht bearbeiteter Handlungsschritt: Nr.

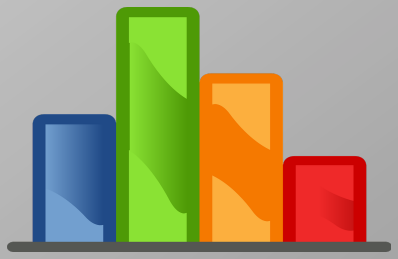
From Data To Sense...

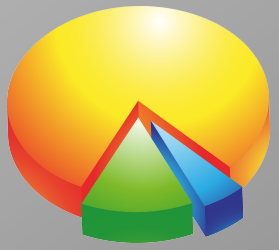


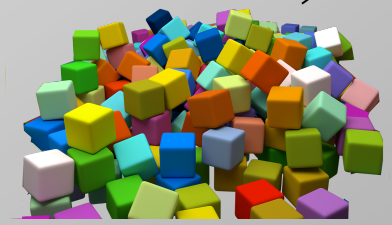


strukturiert

Big Data

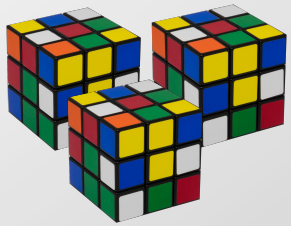




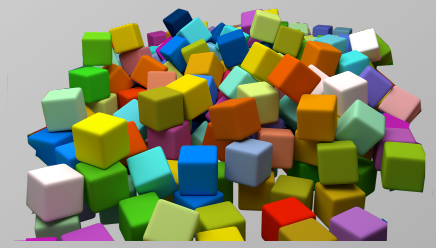


nicht strukturiert

Informationen, strukturiert und unstrukturiert



- Datenbank-Anwendungen
- Excel-Anwendungen,
- Formulare,
- CSV, XML, ...
- ...

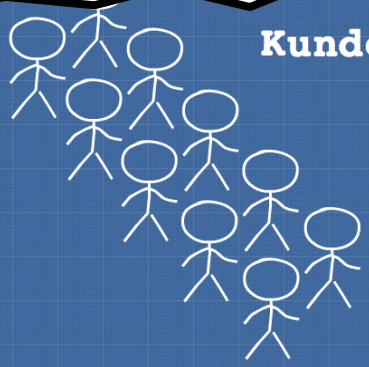


- Dokumente
- Web 2.0 (facebook, Twitter, Blogs, ...)
- Fotos, Musik, Videos
- ‘OK Google‘, Siri, Alexa, Google home ...
- ...

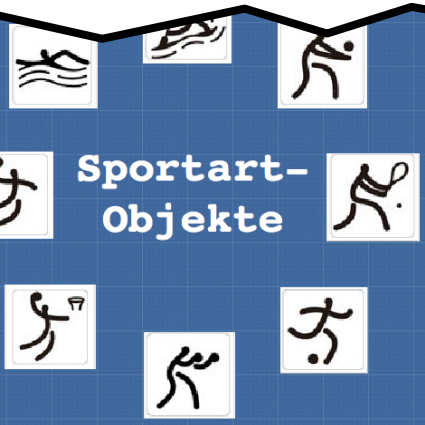
Teil 1:

Relationale Datenbanken

Erklären Sie den Zusammenhang zwischen einem semantischen Modell und dem Datenmodell in der 3NF.



Kunden-Objekte

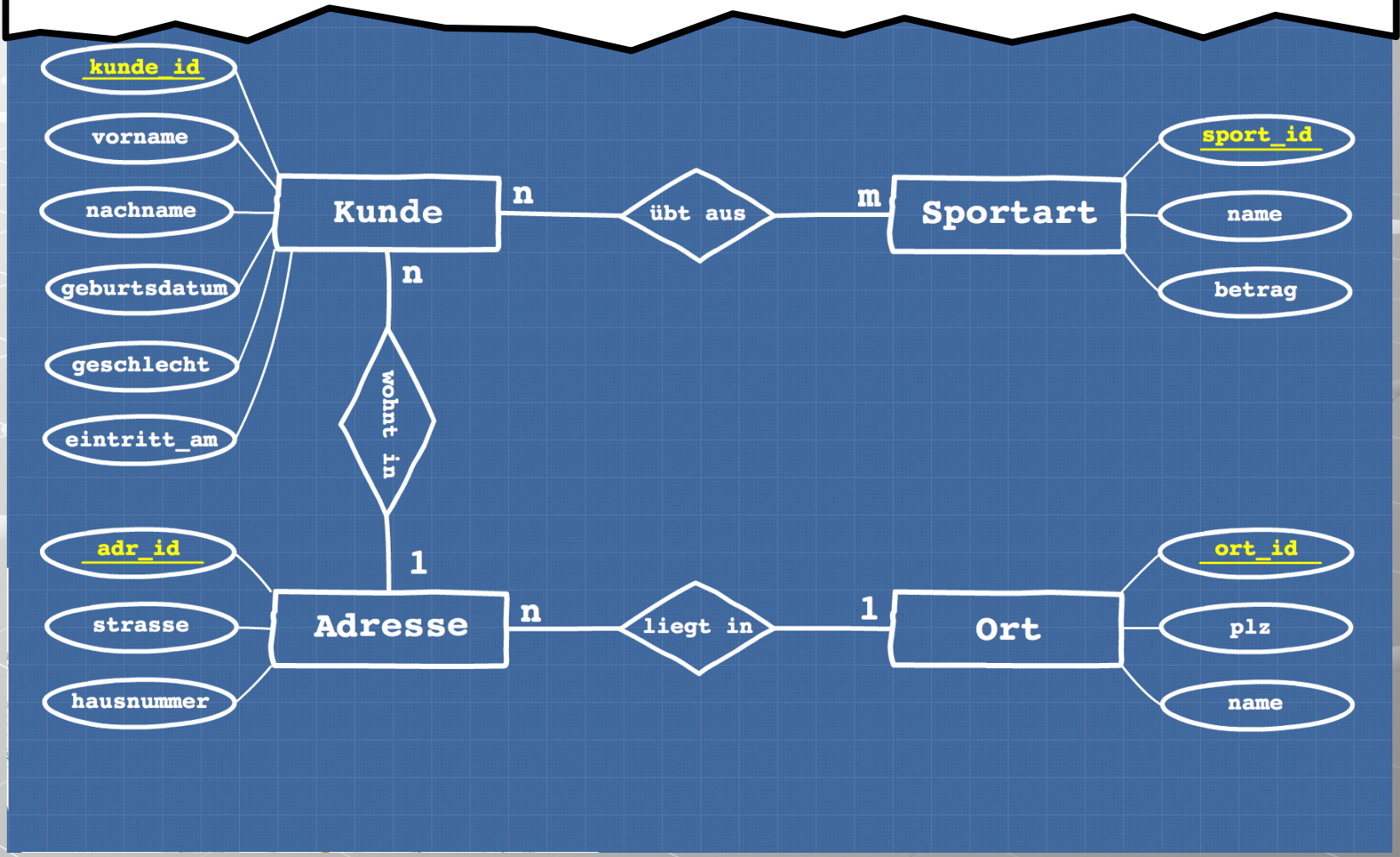


Sportart-Objekte

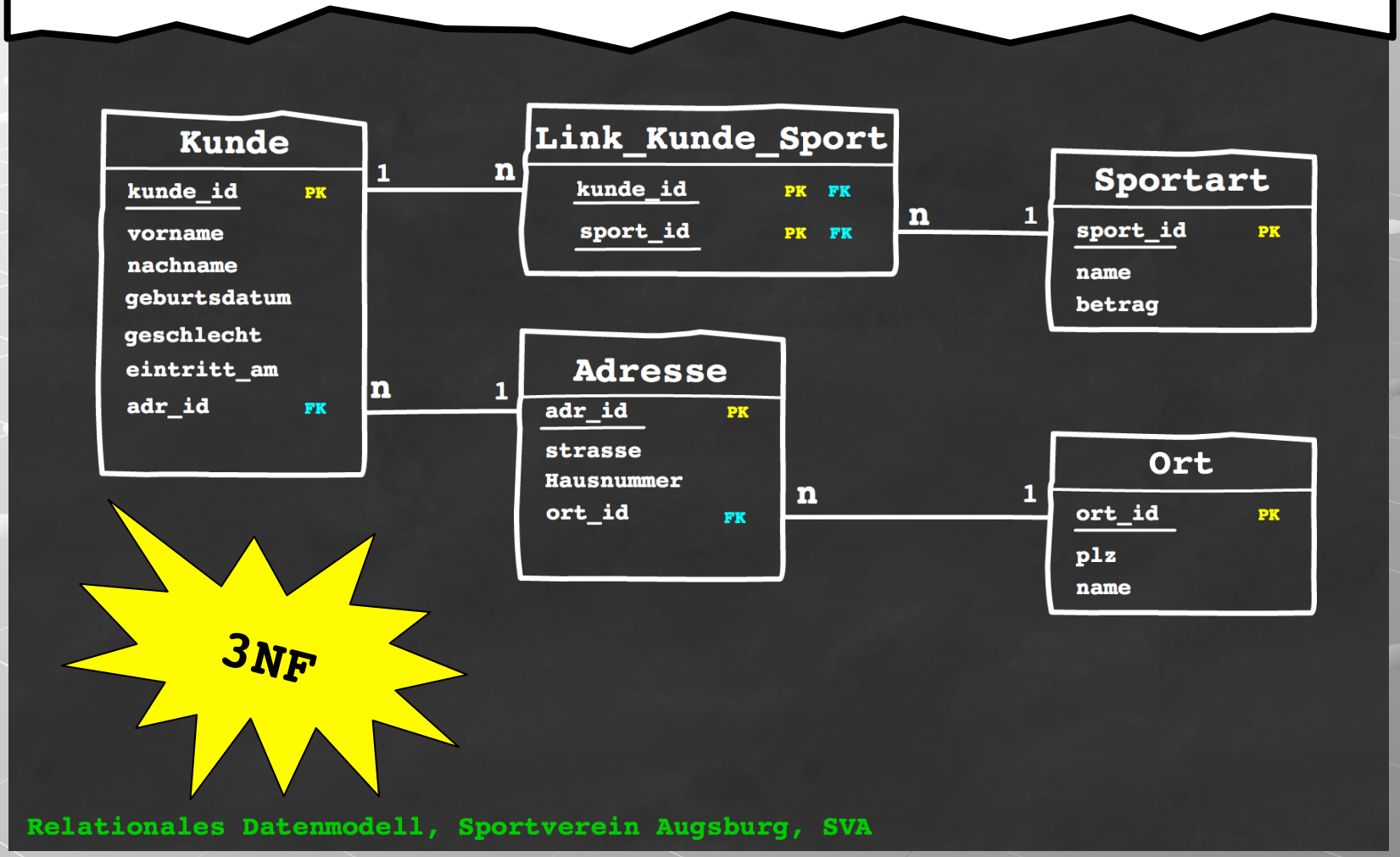
Adress-Objekte

Ort-Objekte

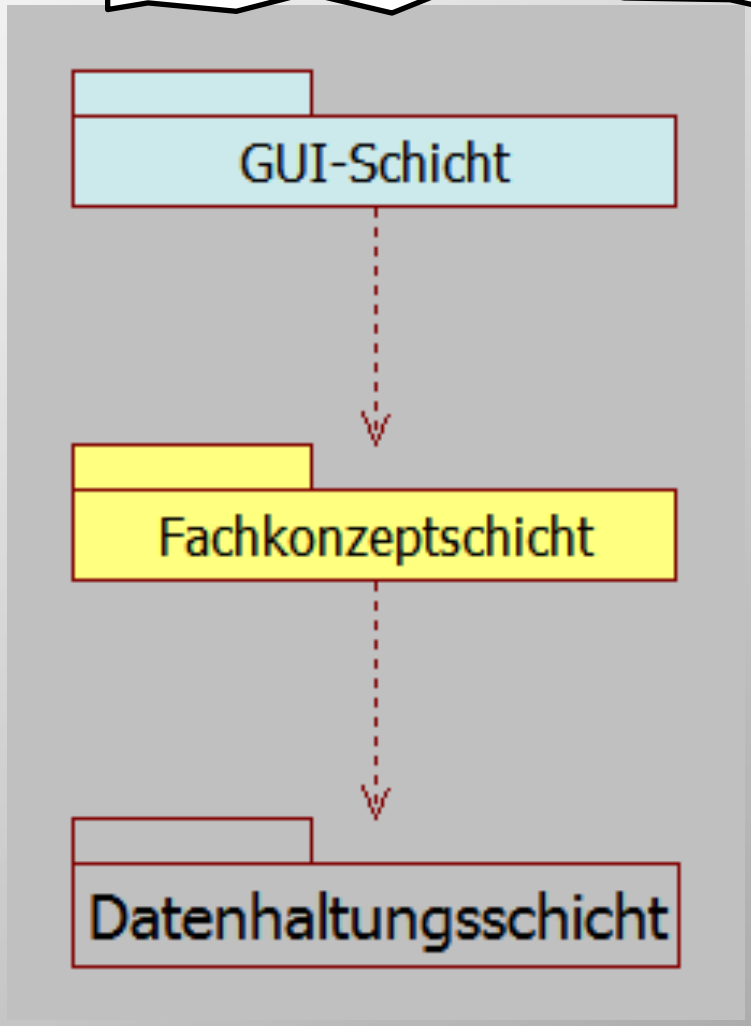
Erklären Sie den Zusammenhang zwischen einem semantischen Modell und dem Datenmodell in der 3NF.

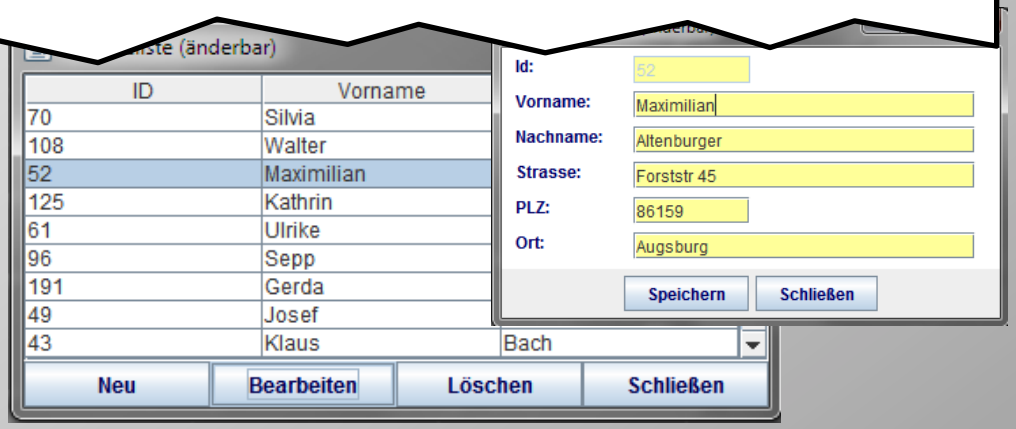


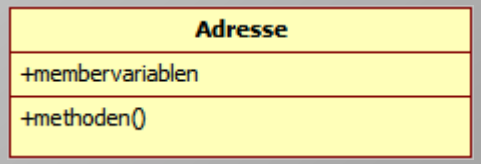
Erklären Sie den Zusammenhang zwischen einem semantischen Modell und dem Datenmodell in der 3NF.

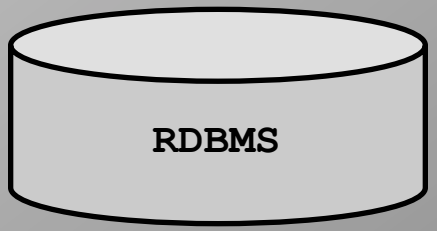


Skizzieren Sie die Architektur einer 3-Schicht-Datenbankanwendung.









Für viele Datenbank-Anwendungen werden relationale Datenbanken verwendet.
Welche wichtigen RDBMS kennen Sie?

342 Systeme im Ranking

Rang Apr 2018	Rang Mär 2018	Rang Apr 2017	DBMS	Datenbankmodell	Punkte Apr 2018	Punkte Mär 2018	Punkte Apr 2017
1.	1.	1.	Oracle	Relational DBMS	1289,79	+0,18	-112,21
2.	2.	2.	MySQL	Relational DBMS	1226,40	-2,46	-138,22
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1095,51	-9,28	-109,26
4.	4.	4.	PostgreSQL	Relational DBMS	395,47	-3,88	+33,69
5.	5.	5.	MongoDB	Document Store	341,41	+0,89	+15,98
6.	6.	6.	DB2	Relational DBMS	188,95	+2,28	+2,29
7.	7.	7.	Microsoft Access	Relational DBMS	132,22	+0,27	+4,04
8.	↑ 9.	↑ 11.	Elasticsearch	Suchmaschine	131,36	+2,81	+25,69
9.	↓ 8.	9.	Redis	Key-Value Store	130,11	-1,12	+15,75
10.	10.	↓ 8.	Cassandra	Wide Column Store	119,09	-4,40	-7,10
11.	11.	↓ 10.	SQLite	Relational DBMS	115,99	+1,17	+2,19
12.	12.	12.	Teradata	Relational DBMS	73,68	+1,21	-2,88
13.	13.	↑ 17.	Splunk	Suchmaschine	65,06	-0,61	+9,55
14.	↑ 15.	↑ 18.	MariaDB	Relational DBMS	64,56	+1,45	+15,83
15.	↓ 14.	↓ 14.	Solr	Suchmaschine	63,21	-1,60	-1,16
16.	16.	↓ 13.	SAP Adaptive Server	Relational DBMS	61,63	-0,99	-5,83
17.	17.	↓ 15.	HBase	Wide Column Store	59,69	-1,24	+1,22
18.	18.	↑ 20.	Hive	Relational DBMS	57,40	+0,39	+15,75
19.	19.	↓ 16.	FileMaker	Relational DBMS	55,00	-0,12	-2,17
20.	20.	↓ 19.	SAP HANA	Relational DBMS	48,90	+0,37	+0,75
21.	21.	↑ 22.	Amazon DynamoDB	Multi-Model	43,14	+0,69	+11,08
22.	22.	↓ 21.	Neo4j	Graph DBMS	40,90	-0,00	+5,99

Vergleichen Sie die Konsistenzmodelle ACID (RDBMS) und BASE (noSQL).

RDBMS

- **A:** atomicity
- **C:** consistency
- **I:** isolation
- **D:** durability

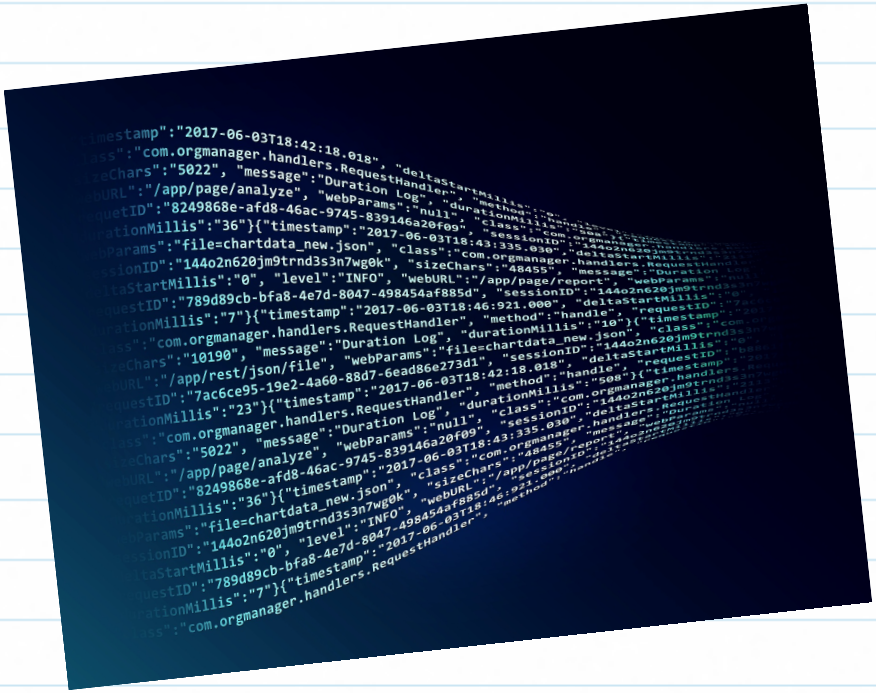
No SQL

- **B:** basically
- **A:** available
- **S:** soft state
- **E:** eventually consistent

Konsistenz (pessimistisch) — Verfügbarkeit (optimistisch)

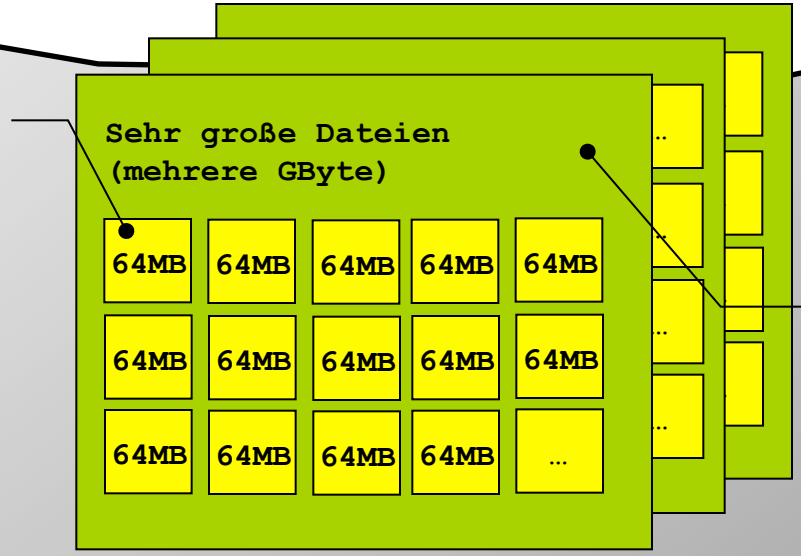
Teil 2:

Big Data – Konzepte

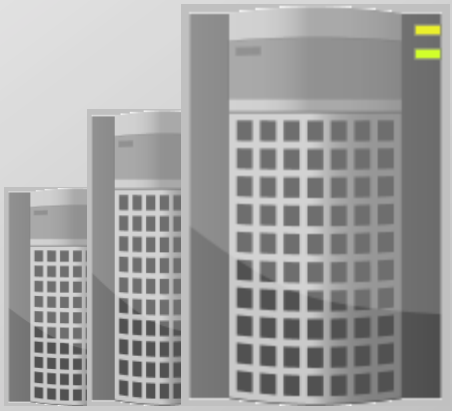


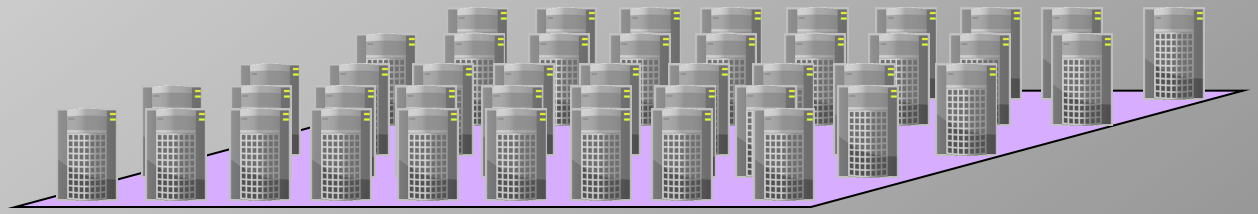
Erklären Sie den Begriff **horizontale Skalierung** am Beispiel des GFS.

64 MB Chunks mit 64 Bit Kennzeichnung



Datei wird mind. 3 mal pro Cluster gespeichert.



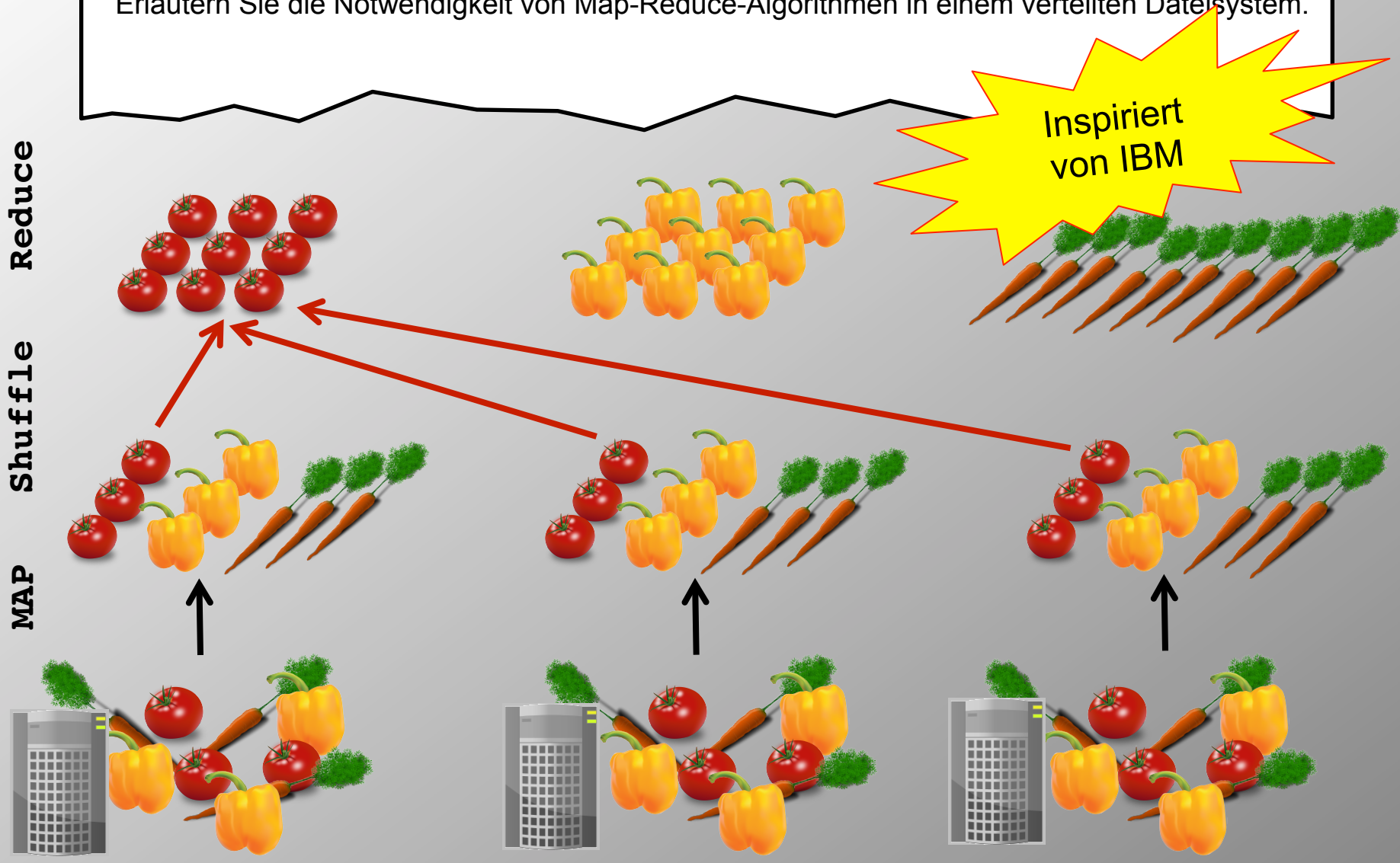


Cluster mit vielen Chunkserver

1 Master (Meta-Daten)

„heart-beat messages“

Erläutern Sie die Notwendigkeit von Map-Reduce-Algorithmen in einem verteilten Dateisystem.



Erläutern Sie den Map-Reduce-Algorithmen anhand eines Beispiels.

Daten

```
.........
081520171
123420183
081520181
123420182
471120184
.........
```

MAP

```
(....,.)
(2017,1)
(2018,3)
(2018,1)
(2018,2)
(2018,4)
(....,.)
```

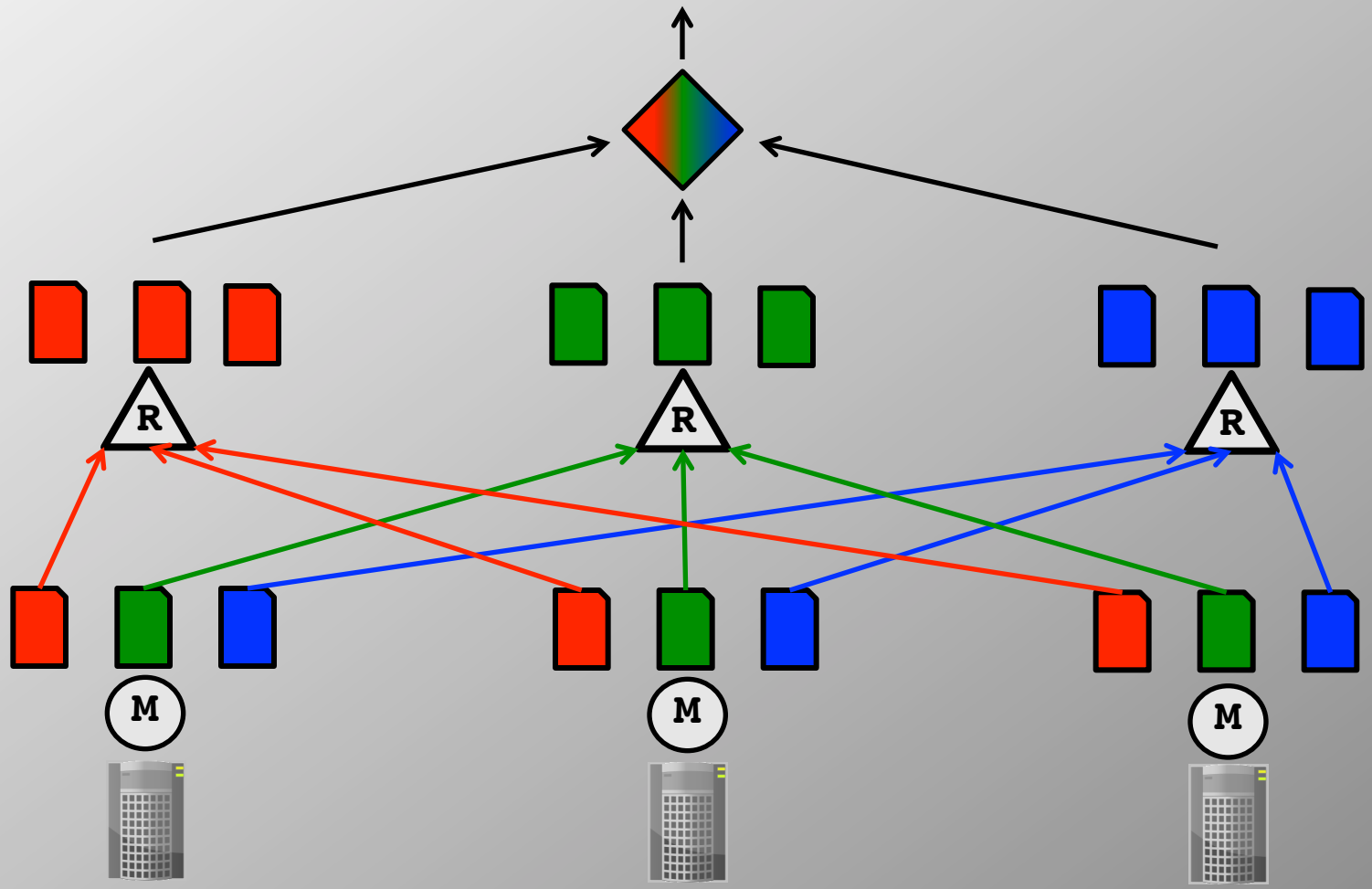
Shuffle

```
(....,.)
(2017,1)
(2018,[3,1,2,4])
(....,.)
```

Reduce

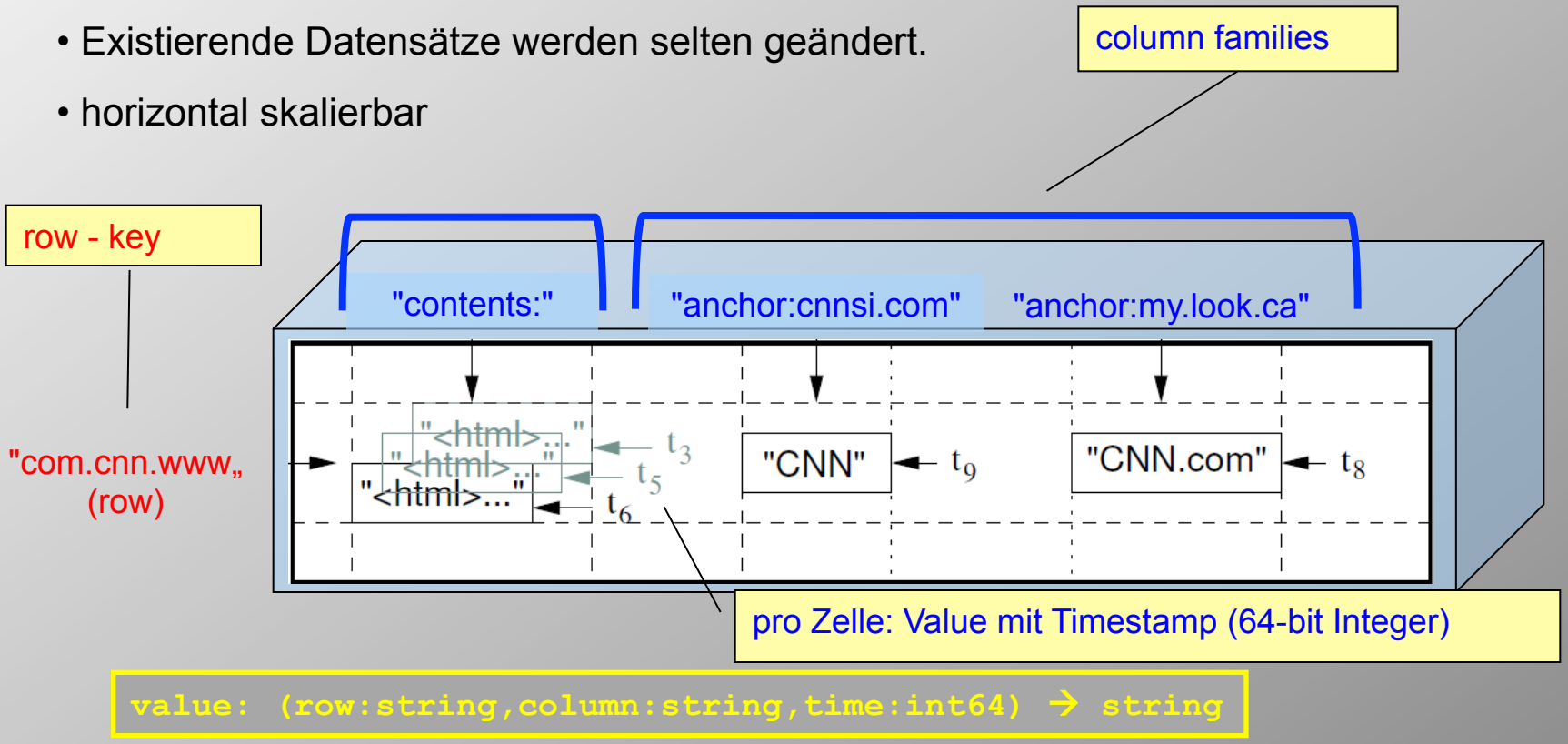
```
(....,.)
(2017,1.0)
(2018,2.5)
(....,.)
```

Map-Reduce–Algorithmen in verteilten Datei-systemen



Erläutern Sie das Konzept von Googles **Big Table**.

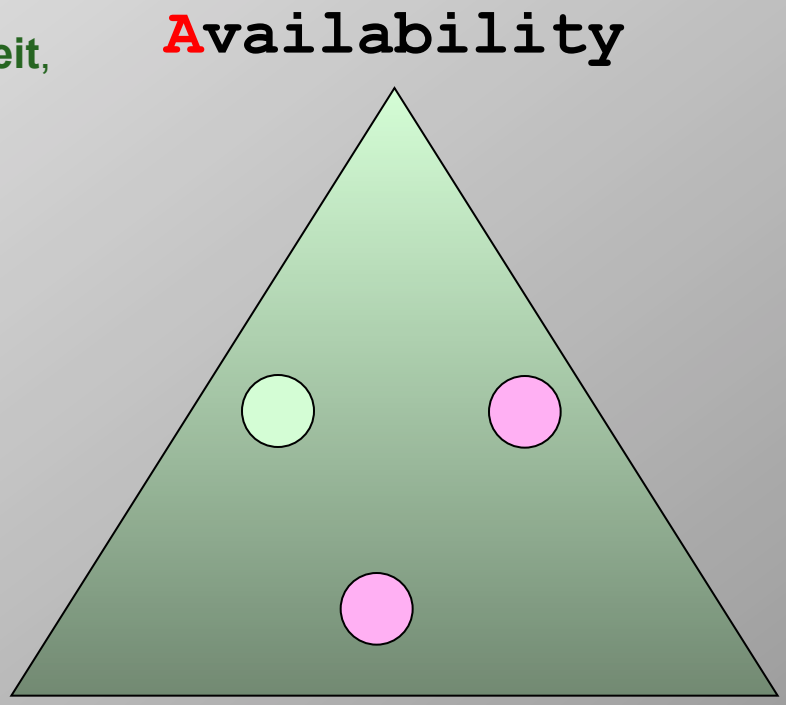
- Datensätze werden oft hinzugefügt.
- Existierende Datensätze werden selten geändert.
- horizontal skalierbar




```
value: (row:string,column:string,time:int64) → string
```

Erklären Sie das CAP-Theorem von Eric Brewer im Zusammenhang mit relationalen und noSQL Datenbanken.

Das Ziel ist die optimale Kombination aus **Verfügbarkeit**, **Konsistenz** und **Toleranz** bei Ausfall einzelner Partitionen.

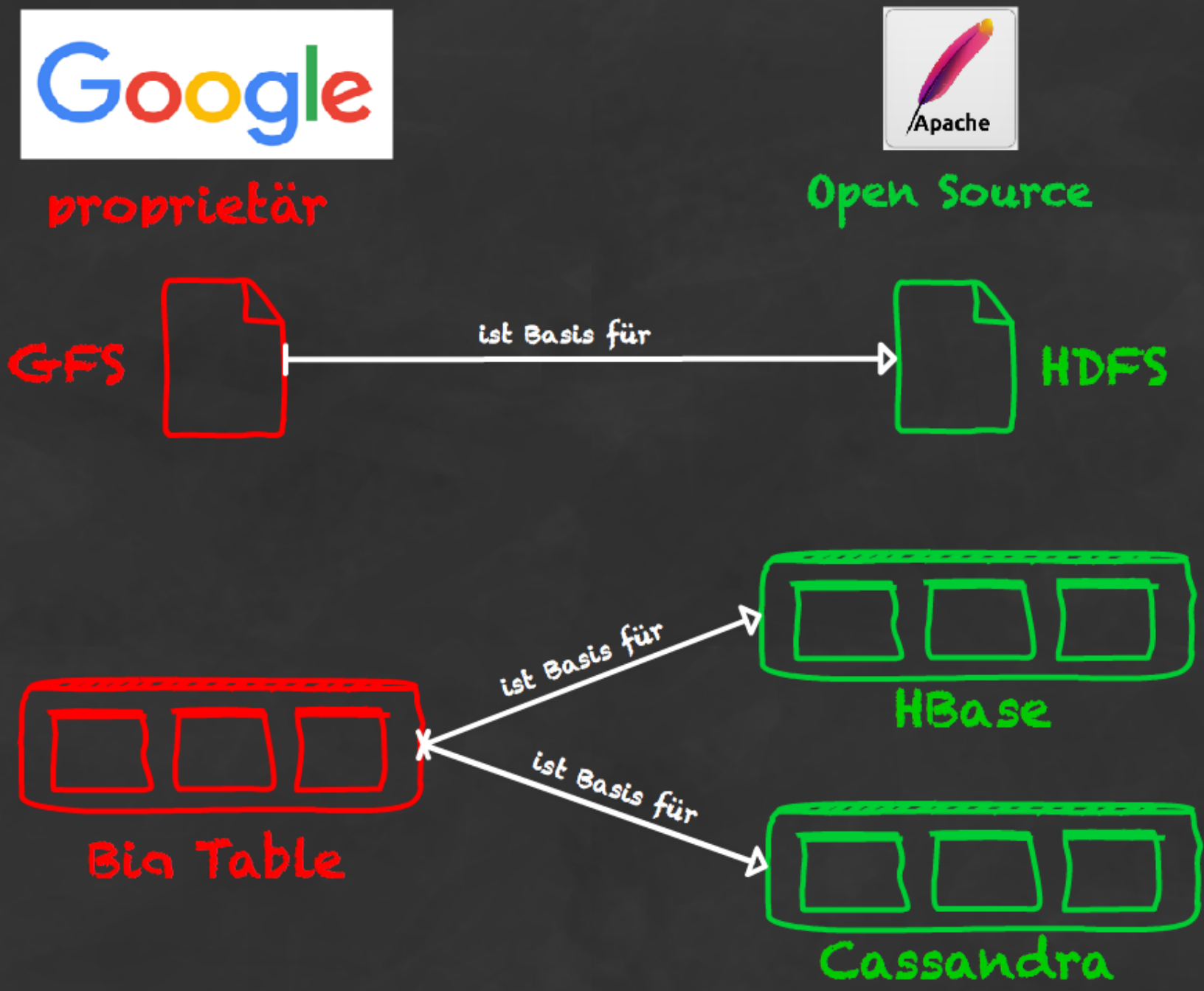


RDBMS

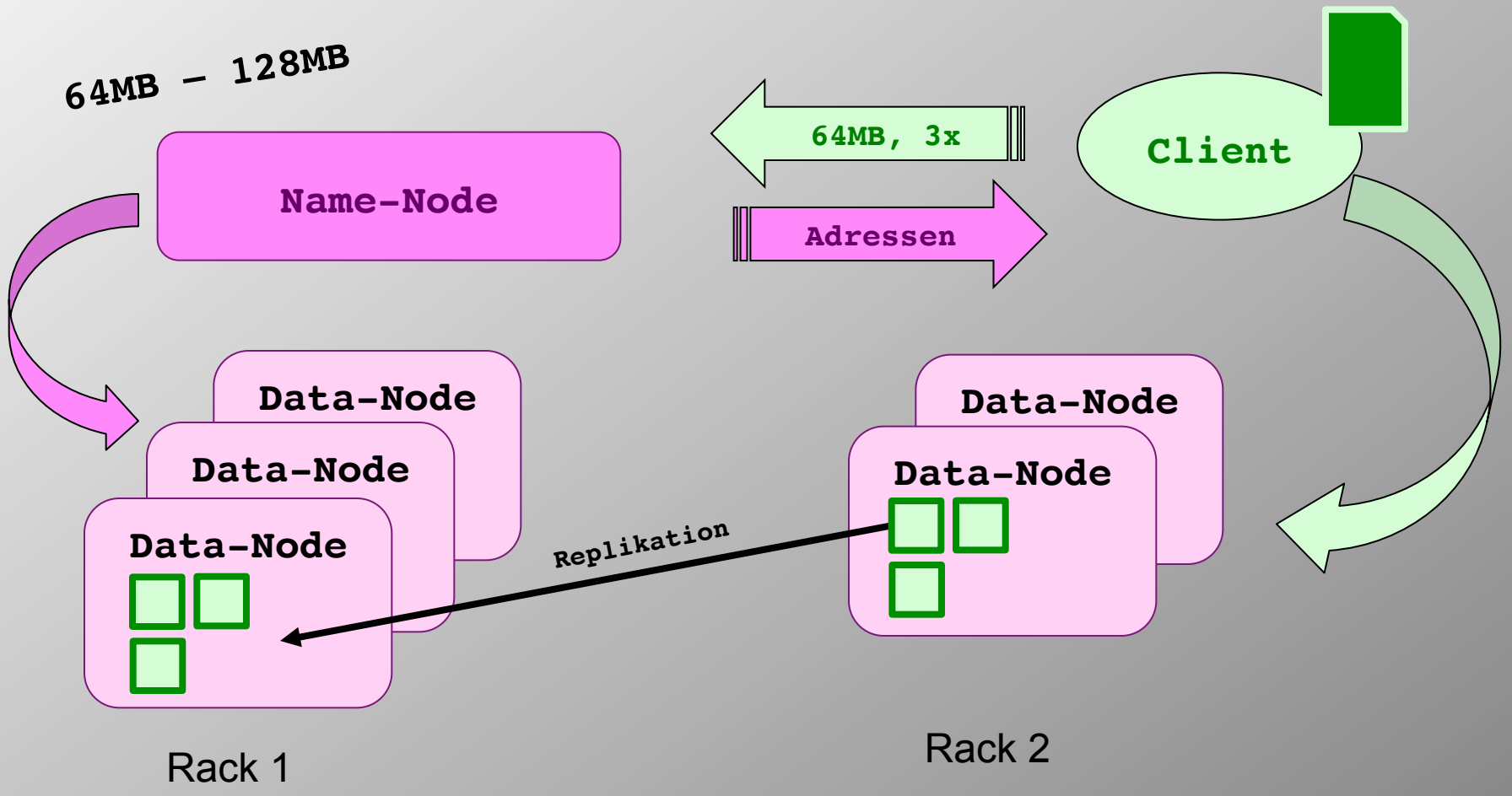
No SQL

Teil 3:

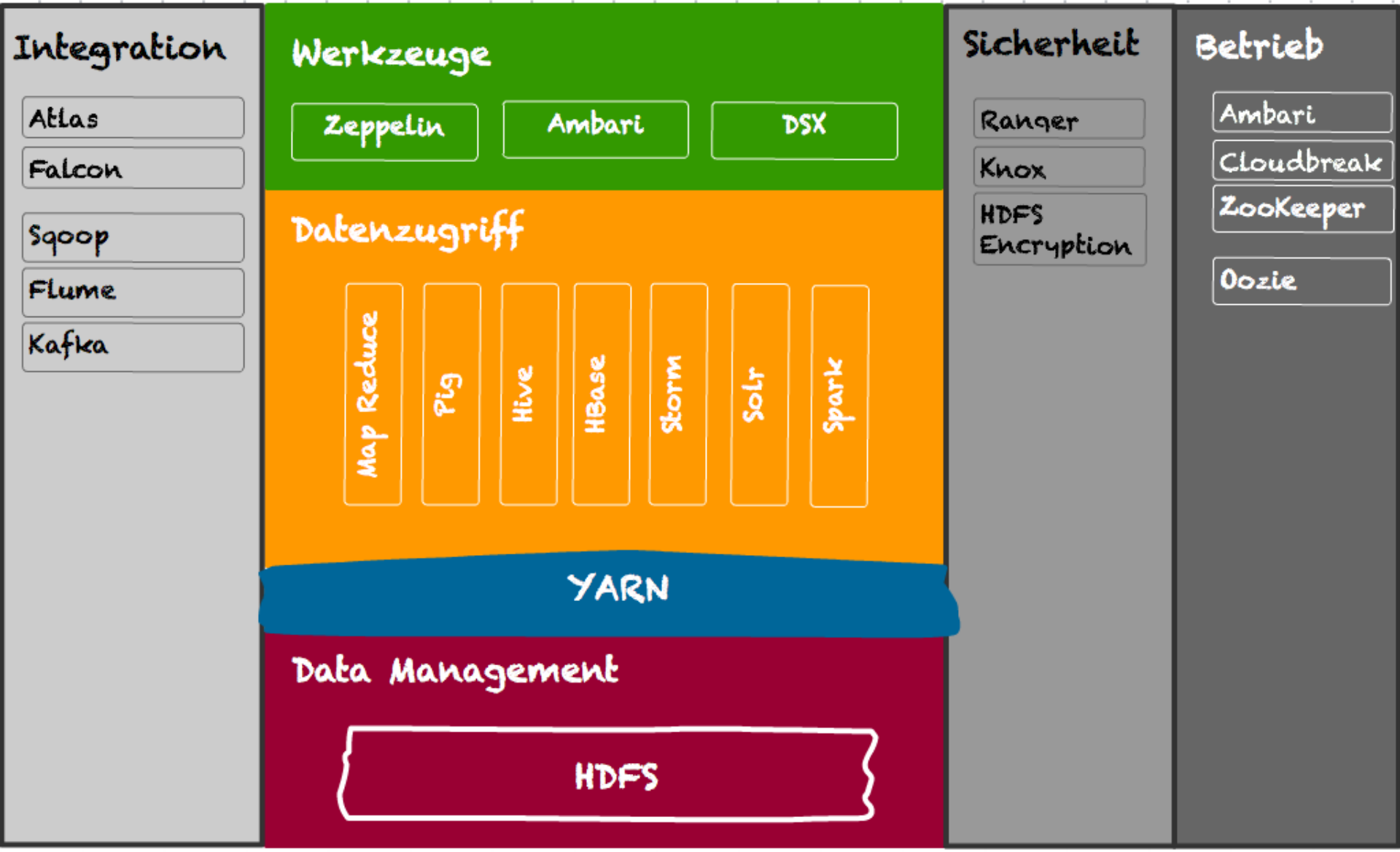
Big Data – Architektur


Google
Apache
proprietär
Open Source
GFS
ist Basis für
HDFS
ist Basis für
ist Basis für
HBase
Big Table
Cassandra

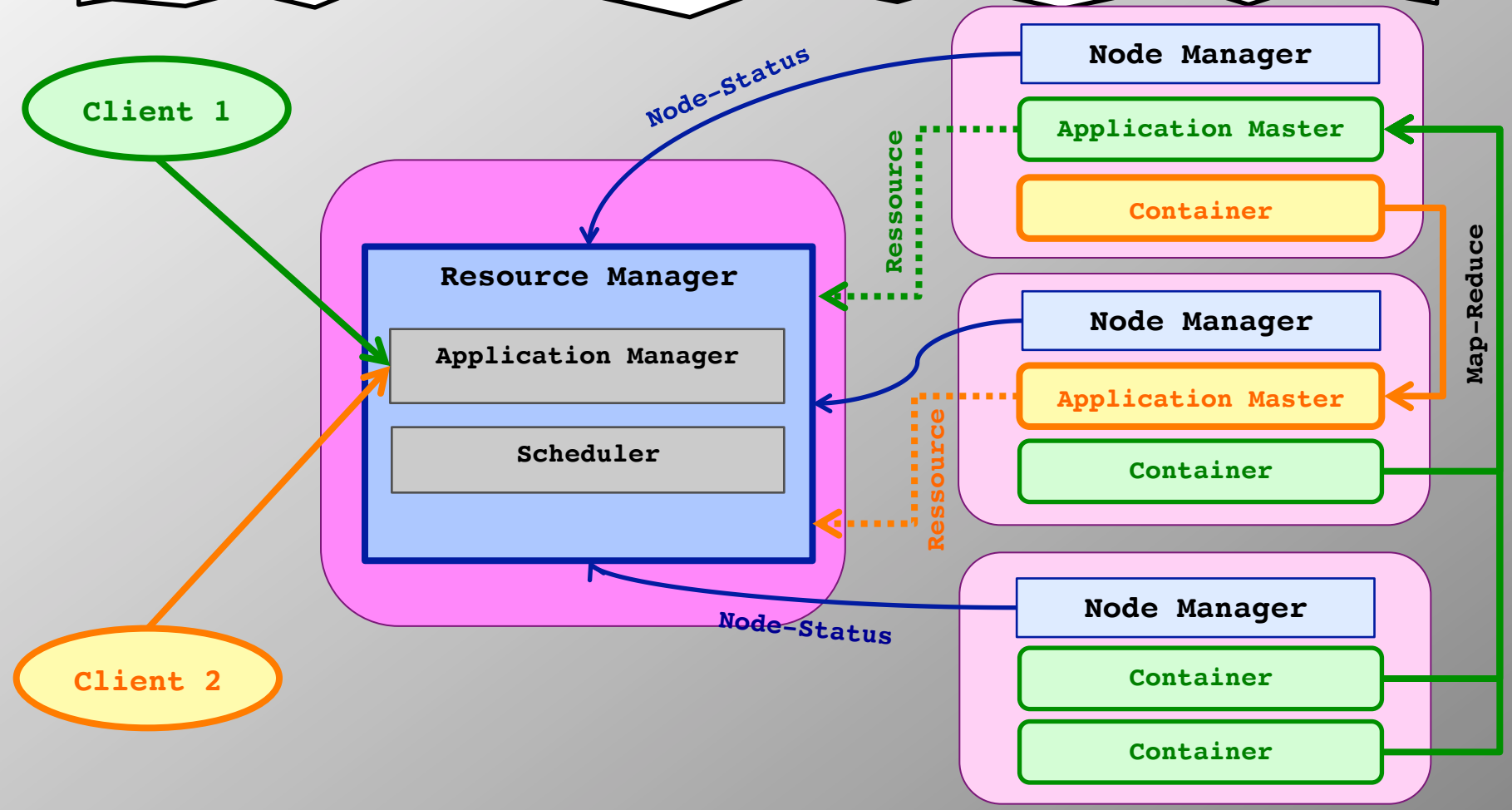
Wie wird die Idee des GFS im **Hadoop-Distributed-File-System** umgesetzt?



HADOOP - Eco-System

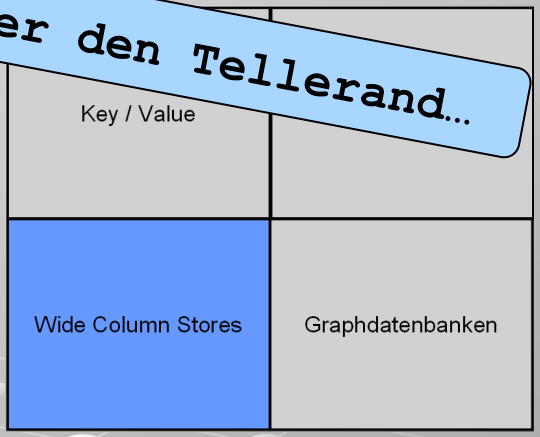


Erläutern Sie die Aufgabe von **YARN** als Bindeglied zwischender Daten-Management-Schicht und der Datenzugriffsschicht einer HDFS-Anwendung.



Cassandra

über den Tellerand...

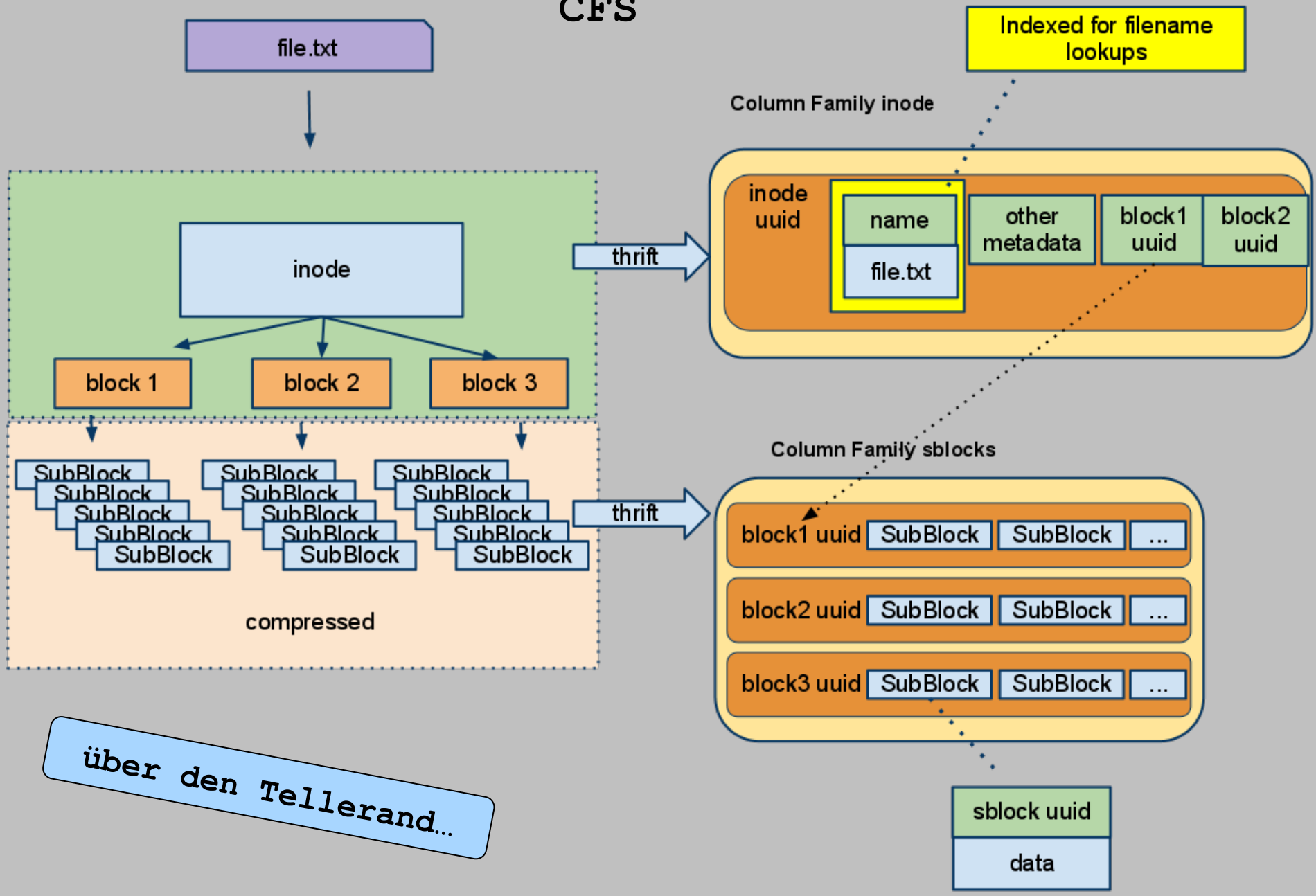


- Flexibilität und Skalierbarkeit durch Key / Value
- vertraute Schemasicherheit
- verteilte Architektur

Keyspace:WildWest				
CF:Users				
	"Jim"	„reactivity":"2"		
	"Joe"	„reactivity":"9"	„nick_name":"Little Joe"	
CF:Horses				
	"Kelly"	"color":"white"	"owner":"Jim"	
CF:Weapons				
	"Colt"	"ammo":"15"	"owner":"Joe"	
CF:Duels				
	"high_noon"	"player_1":"Jim"	"player_2":"Joe"	"winner":"Joe"
CF:Messages				
	"high_noon_1"	"from":"Joe"	"to":"Jim"	"text":"Looooser!"

Abbildung 3.2.1 Cassandra-Zustand nach ersten Inserts

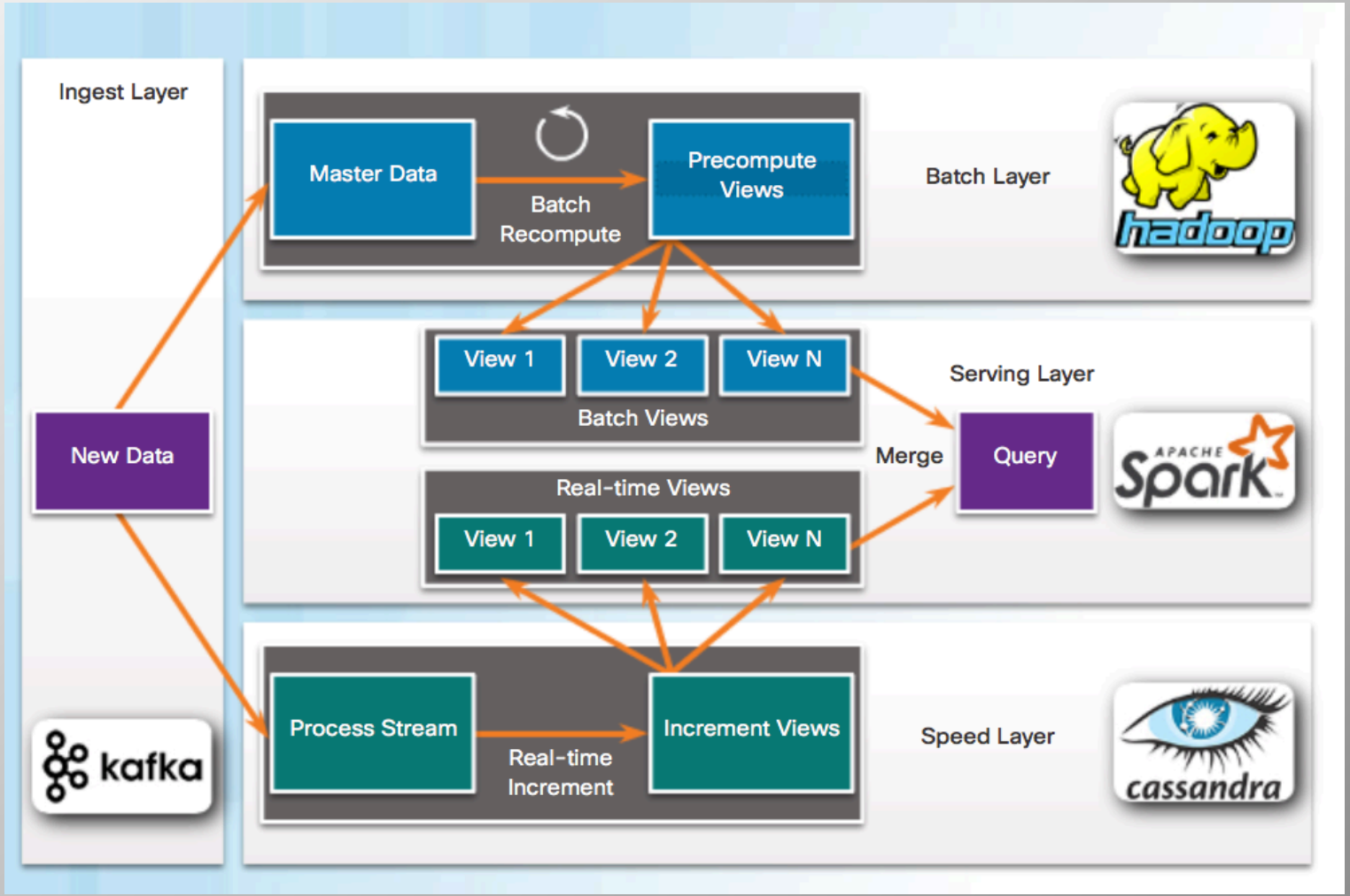
CFS



über den Tellerand...

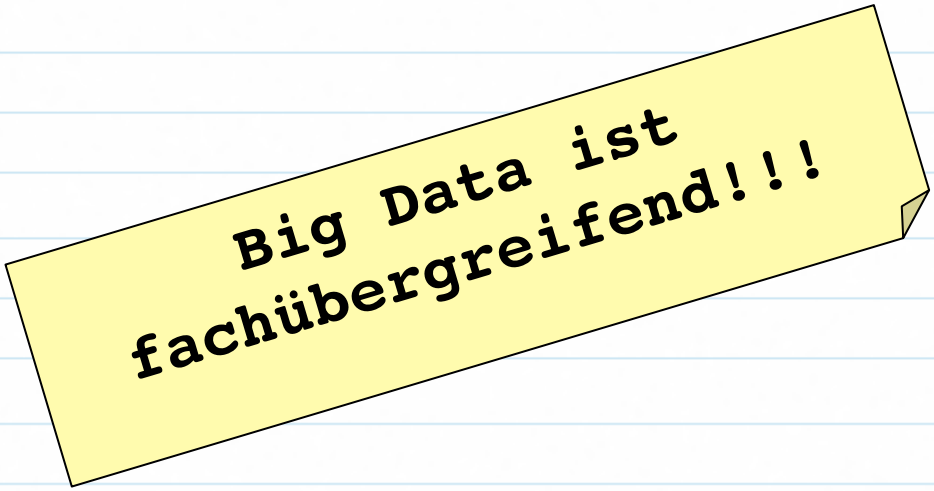
Quelle: https://www.datastax.com/dev/blog/cassandra-file-system-design

Aus welchen Komponenten besteht eine **`Lambda-Architektur`**? Skizzieren Sie die Komponenten...



Teil 4:

Big Data im Unterricht



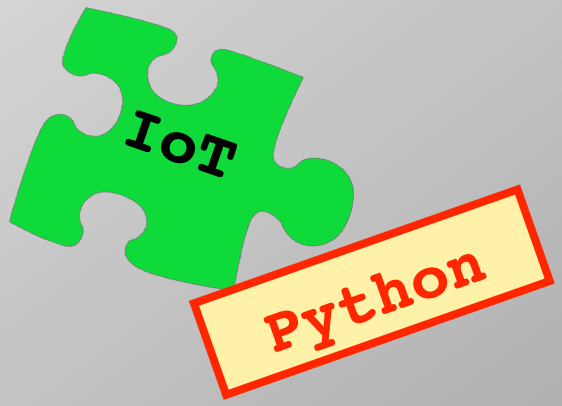
1. AJ	2. AJ	3. AJ
JAVA Basics	JAVA OOP, Android	RDBMS – SQL
		Big Data Basics
	UML	IoT + PT
	RDBMS	Python Basics
Packet Tracer, OSI Modell, vernetzte Systeme	Packet Tracer, Protokolle, CCNA Sem. 1, Raspberry Pi, Sensoren, Virtualisierung	

Geführtes Selbststudium & Aufgaben

WIE?

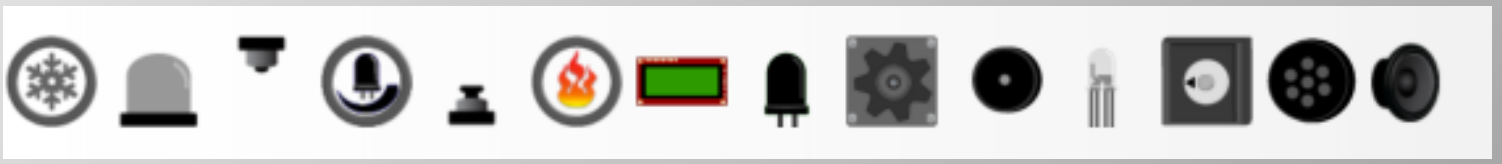
8 Blöcke mit je 3 Std.

Ein Packet Tracer, viele Möglichkeiten...



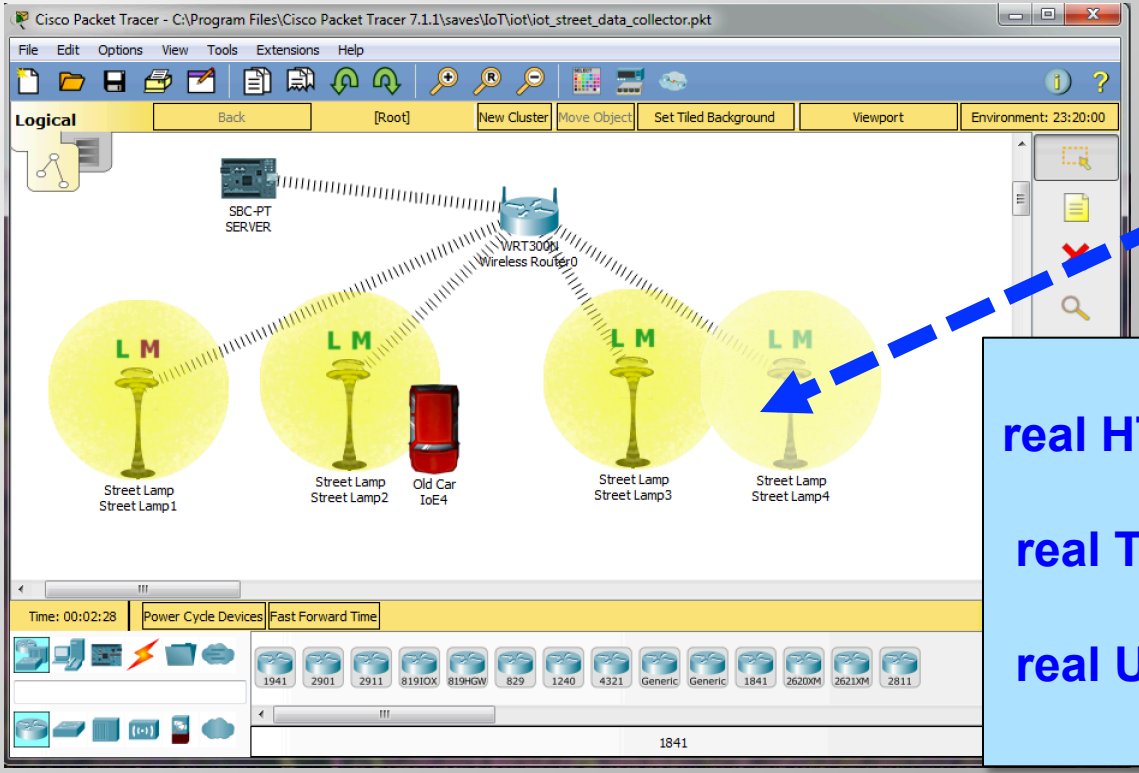
Boards

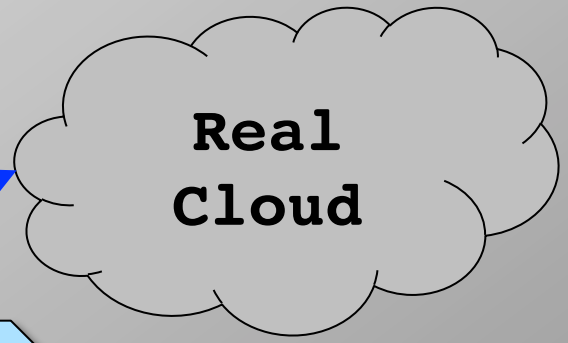
Aktoren



Sensoren

Packet Tracer mit Cloud-Anbindung





real HTTP,

real TCP,

real UDP

Projekt: Smart Home & Android App in Zusammenarbeit mit 2.AJ

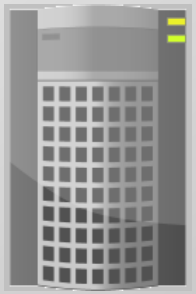
Real Cloud

WIE?

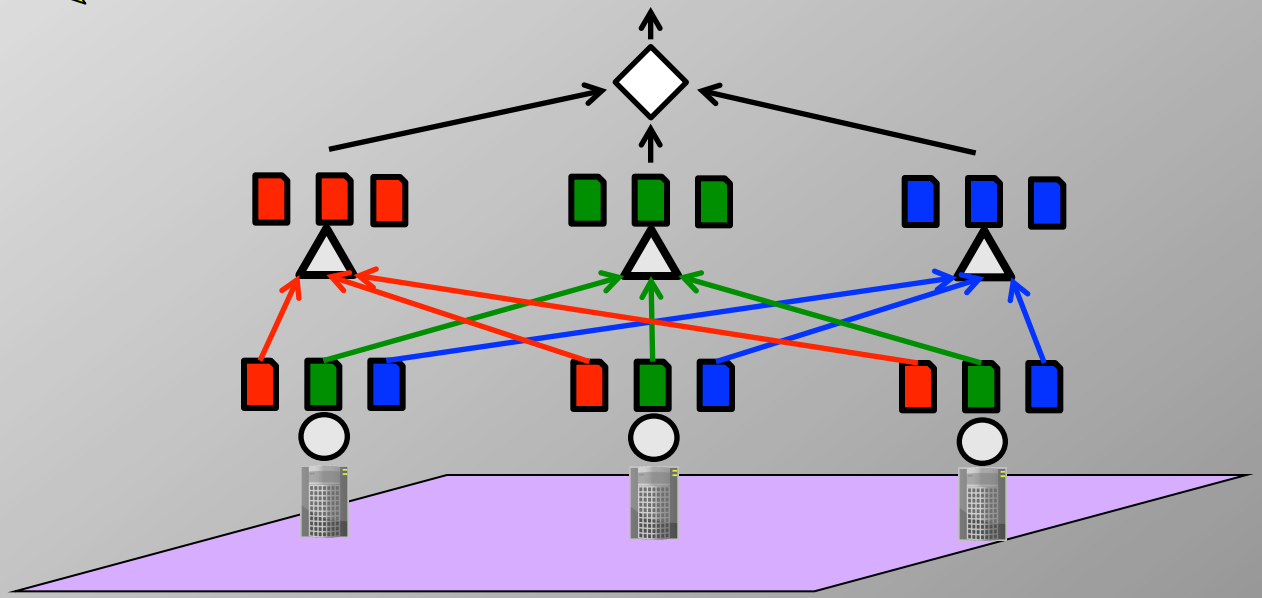
Staatliche BS1 Ingolstadt

Virtueller Hadoop-Cluster mit Yarn, Map-Reduce im Klassenraum

Primzahlen
Zwischen
1 und 300.000



Hadoop
Master



Virtuelle Hadoop-Knoten

Auszug aus dem Lehrplan...

ANWENDUNGSENTWICKLUNG/PROGRAMMIERUNG
Jahrgangsstufe 12
Fachrichtung: Systemintegration

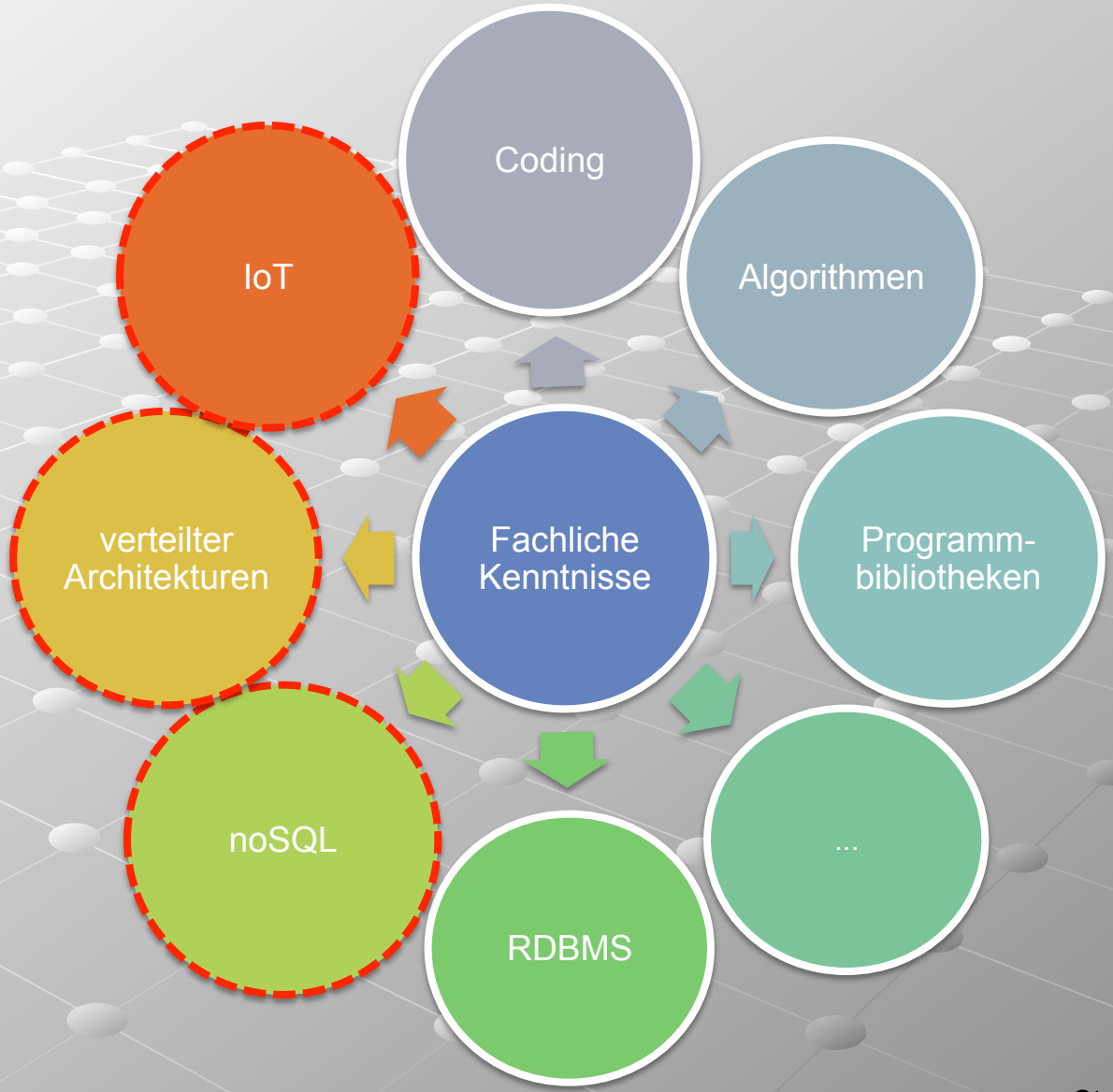
Lernfeld	**77 Std.**
Entwickeln und Bereitstellen von Anwendungssystemen – Schwerpunkt: Datenbankanwendungen	

ANWENDUNGSENTWICKLUNG/PROGRAMMIERUNG
Jahrgangsstufe 12
Fachrichtung: Anwendungsentwicklung

Lernfeld	**165 Std.**
Entwickeln und Bereitstellen von Anwendungssystemen – Schwerpunkt: Programmentwicklungsmethoden, Programmierung und Datenbankkonzepte	

Ziele

Die Schülerinnen und Schüler wenden für das Entwickeln von Anwendungssystemen eine Programmentwicklungsmethode an und erstellen die (Anwendungs-)Programme auf der



Appendix A:

Big Data Analyse, Daten-auswertung,Visualisierung

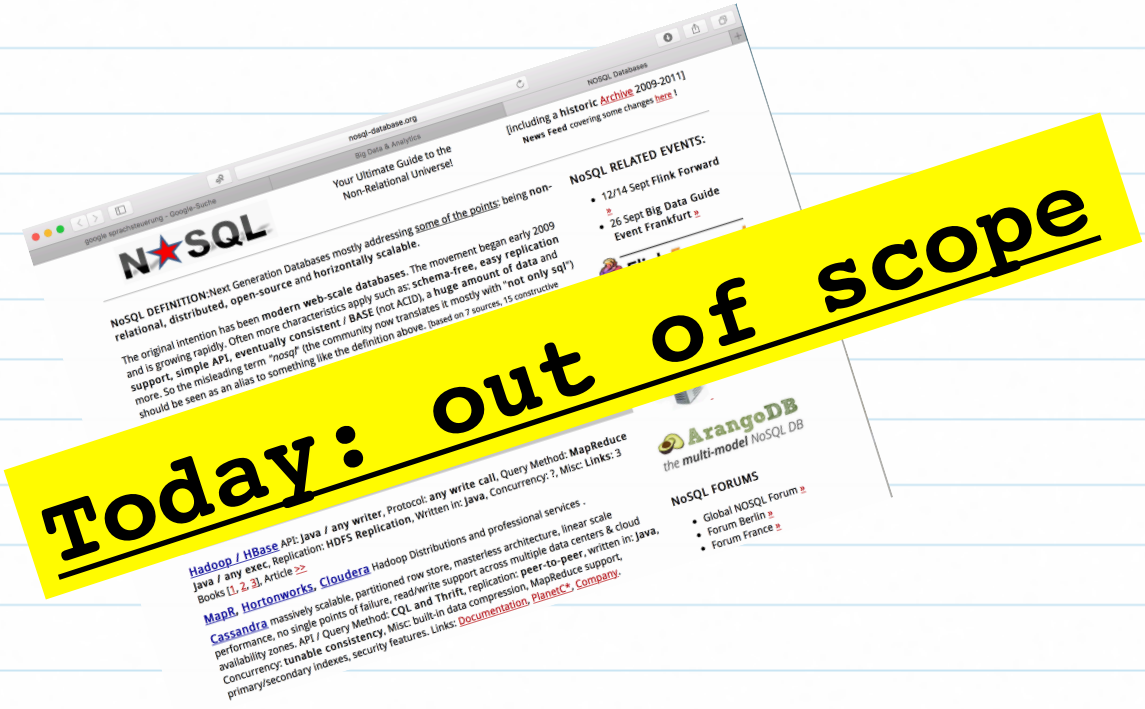
Today: out of scope

Datenanalyse und Visualisierung

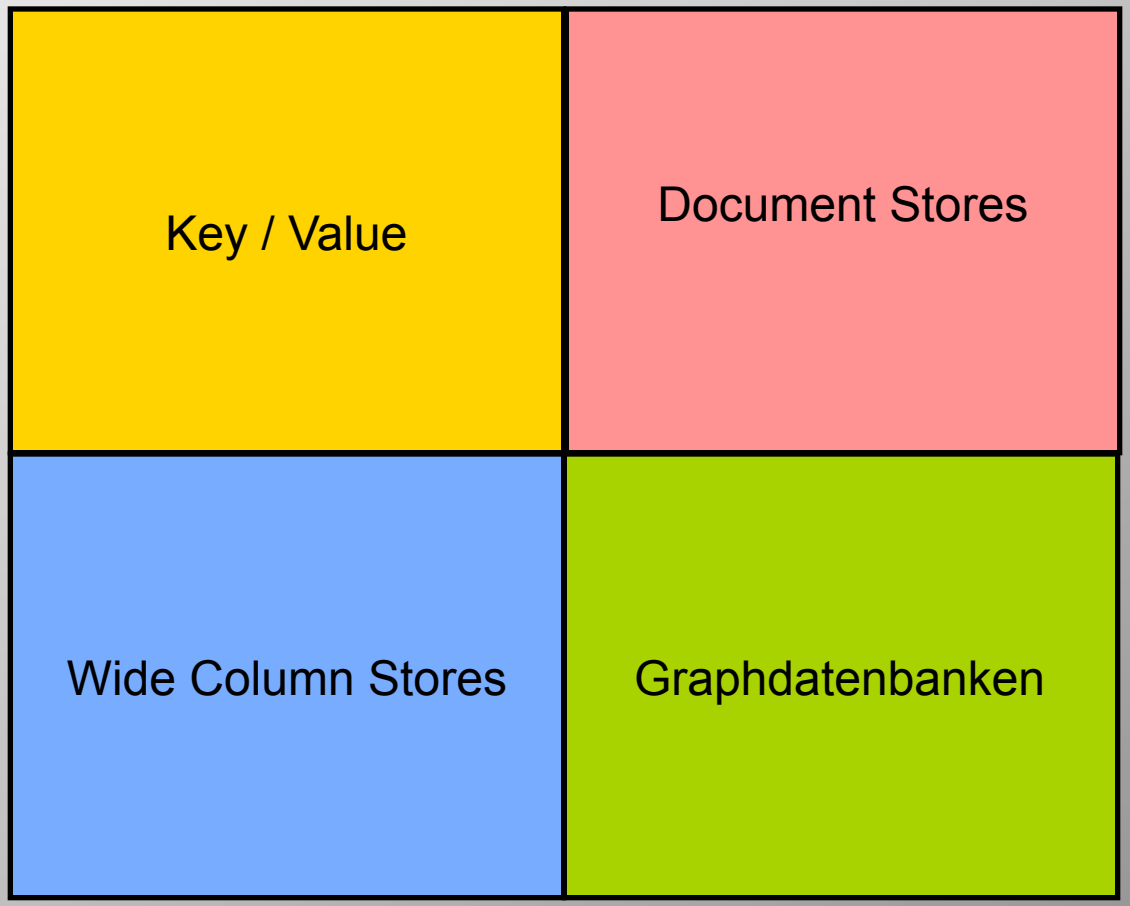
Grundlagen	• Was ist Datenanalyse • Big Data verarbeiten • unterschiedliche Datenquellen • Daten verarbeiten • ethische Verantwortung und Bedenken
Datenanalyse	• Statistik • Verteilungen und Merkmale von Daten • beschreibende Statistik • Korrelationen
Machine Learning	• Vorhersagen • Modell-Abschätzungen
Storytelling mit Daten	• Hypothesen und Nachweise • Visualisierungs-Werkzeuge • geeignete Visualisierungen

Appendix B:

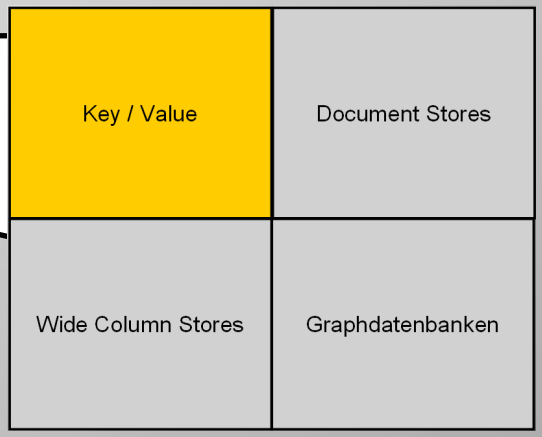
noSQL – Datenbanken



In welche vier Kategorien werden NoSQL – Kernsysteme eingeteilt?



Beschreiben Sie die genannten NoSQL-Kategorien.



- Daten: Schlüssel / Wert - Paare
- Schlüssel sind: Datenbanken, Namensräume, Attribute
- Werte : Zeichenketten aber auch Hashes, Listen oder Mengen
- Vorteil: einfaches Datenmodell
- Nachteil: Mächtigkeit der Abfragesprache oft gering (hier muss man sich auf die API verlassen)

Beschreiben Sie die genannten NoSQL-Kategorien.

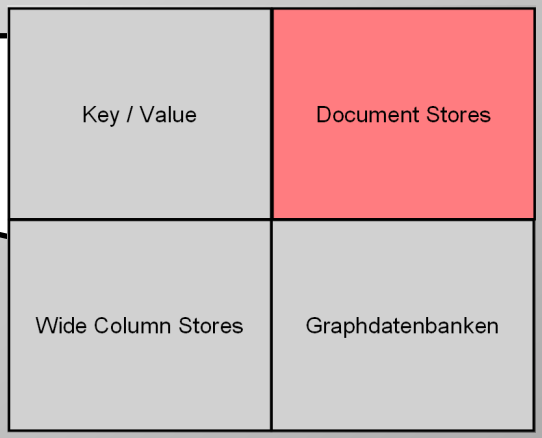
Key / Value	Document Stores
Wide Column Stores	Graphdatenbanken

```
SET server:name "bs1in"
GET server:name => "bs1in"
```

…ein kleines API Beispiel…

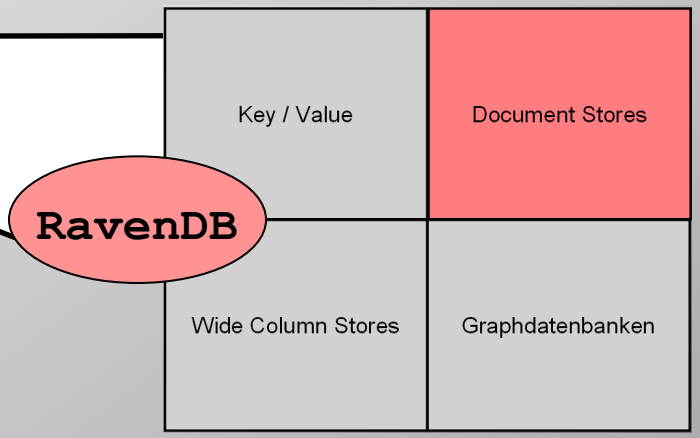
```
RPUSH friends "Tom"
RPUSH friends "Bob"
LPUSH friends "Sam"
LRANGE riends 1 2 => ["Tom","Bob"]
LLEN friends => 3
LPOP friends => "Sam"
RPOP friends => "Bob"
LLEN friends => 1
```

Beschreiben Sie die genannten NoSQL-Kategorien.



- Daten: Informationen werden als Dokument abgelegt.
- Dokument: Strukturierte Datensammlung JSON, BSON, YAML, RDF (**`Ressource Description Framework`**).
- Identifizierung: Jedes Dokument erhält eine ID.
- Vorteil: Die Verantwortung für das Dokumenten-Schema liegt in der Anwendung. Erweiterungen sind kein Problem.
- Nachteil: keine referenzielle Integrität, keine Normalisierung.

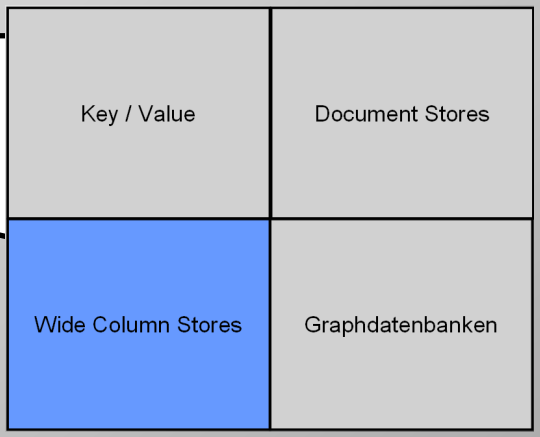
Beschreiben Sie die genannten NoSQL-Kategorien.



Java Script Object Notation

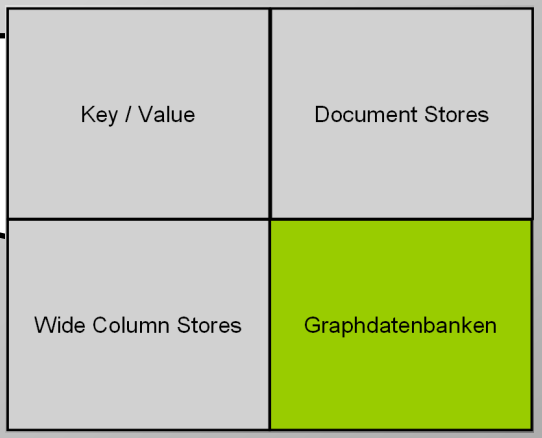
```
{
  "Performers": ["Rebekka Bakken"],
  "Composers": [],
  "Title": "If Only",
  "Album": "Art of How to Fall",
  "Duration": "00:03:42.5740000",
  "Genre": "Pop"
};
```

Beschreiben Sie die genannten NoSQL-Kategorien.



- Daten: spaltenorientiert, d.h. pro Attribut wird eine Tabelle verwendet.
- Vorteil: gute Analyse und einfache Aggregation der Daten, Verwendung in OLAP- und Data-Warehouse-Umgebungen.
- Nachteil: Suche und das Einfügen von Daten ist aufwendiger.

Beschreiben Sie die genannten NoSQL-Kategorien.



- Speicherung von Informationen erfolgt in Knoten und Kanten
- Property-Graphen bieten die Möglichkeit, Knoten und Kanten mit Eigenschaften zu versehen.
- Anwendungsbereiche: Graphen, semantische Netzwerke und Location Bases Services (Smartphones)

Beschreiben Sie die genannten NoSQL-Kategorien.

Key / Value	Document Stores
Wide Column Stores	Graphdatenbanken

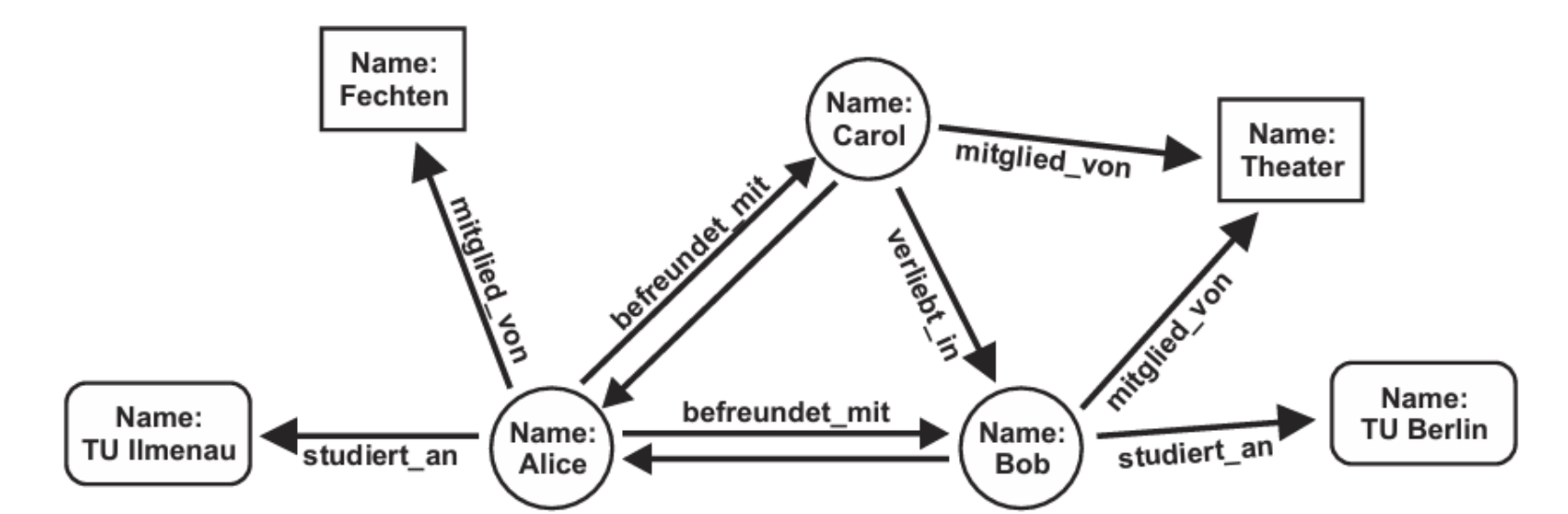
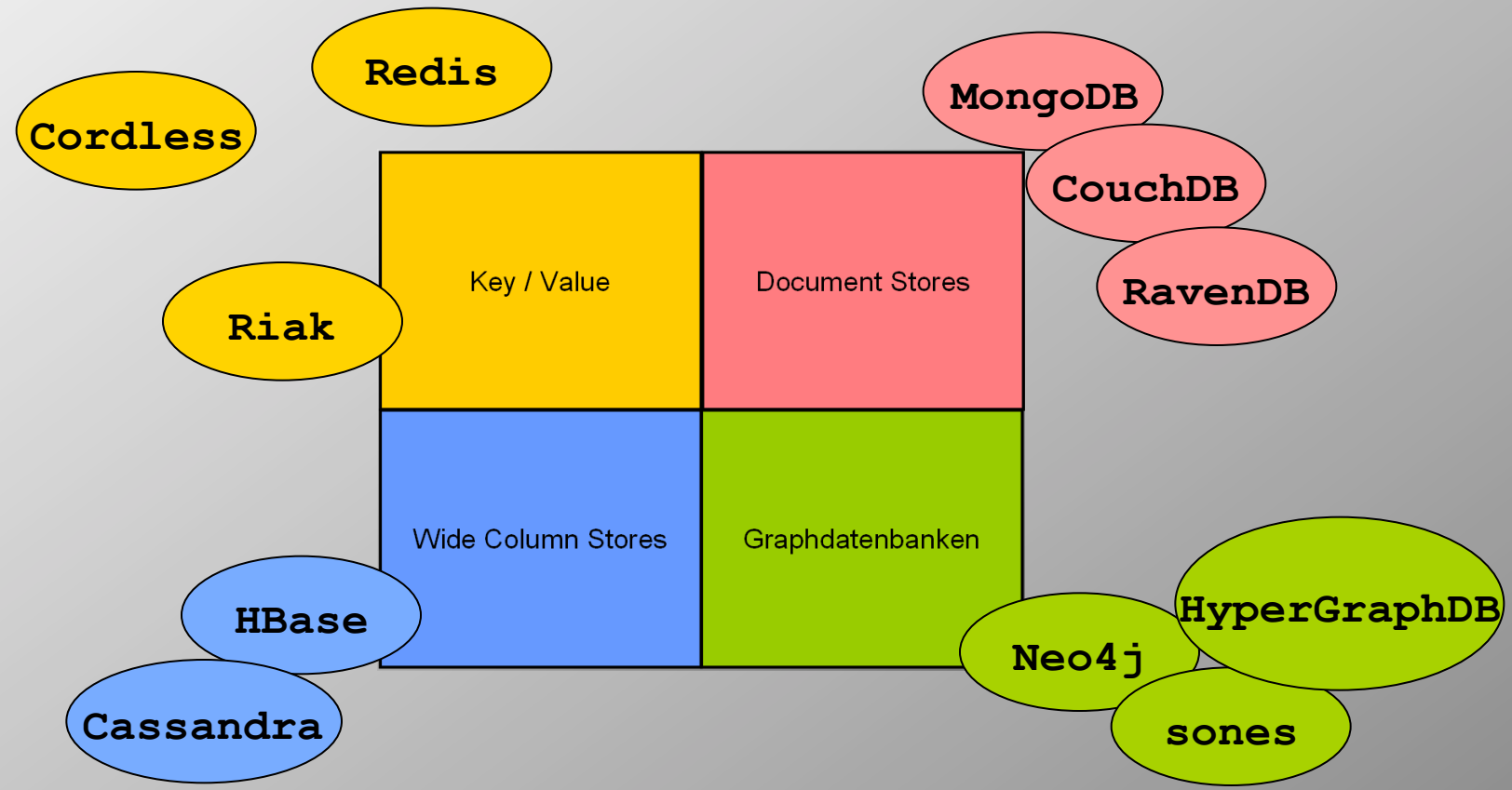


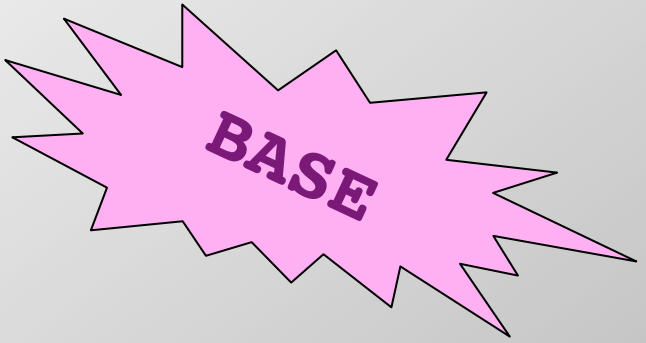
Abbildung 6.1.3 Beispiel eines Property-Graphen anhand eines „sozialen Netzes"

[sones GraphDB VisualGraph Tool]

Ordnen Sie wichtige NoSQL–Datenbanken den Kategorien zu.



Welche Konzepte stehen bei NoSQL-Datenbanksystemen im Vordergrund? Erklären Sie in diesem Zusammenhang den Begriff BASE.



Datenmodell ist nicht relational

Keine Speicherung in herkömmlichen Tabellen

Datenbank ist schemafrei

oder hat nur schwächere Schemarestriktionen

verteilte Systeme

einfache Replikationsmechanismen

horizontale Skalierbarkeit

Server-Cluster

einfache API

Open Source

Vergleichen Sie die Konsistenzmodelle ACID (RDBMS) und BASE (noSQL).

A: atomicity

C: consistency

I: isolation

D: durability

B: basically

A: available

S: soft state

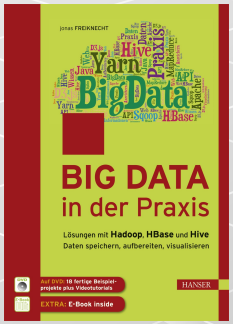
E: eventually consistent

Big Data

Entwicklung und Programmierung von Systemen für große Datenmengen und Einstaz der Lambda-Architektur

Nathan Marz, James Warren

mitp Professional



Big Data in der Praxis

Lösungen mit Hadoop, HBAse und Hive. Daten speichern, aufbereiten, visualisieren

Jonas Freiknecht

Hanser

NoSQL

Einstieg in die Welt nichtrelationaler Web2.0 Datenbanken

Stefan Edlich, Achim Friedland, Jens Hampe, Benjamin Brauer

Hanser

research.google.com

google bigTable

google file system

Big Data im Web

Big Data auf Wikipedia

Apache Hadoop

hadoop.apache.org

NoSQL im Web

www.nosql-database.org

Hadoop Ecosystem

de.hortonworks.com

DB-Engines

db-engines.com

Vielen Dank für Ihre Aufmerksamkeit