NoSQL - Einstieg

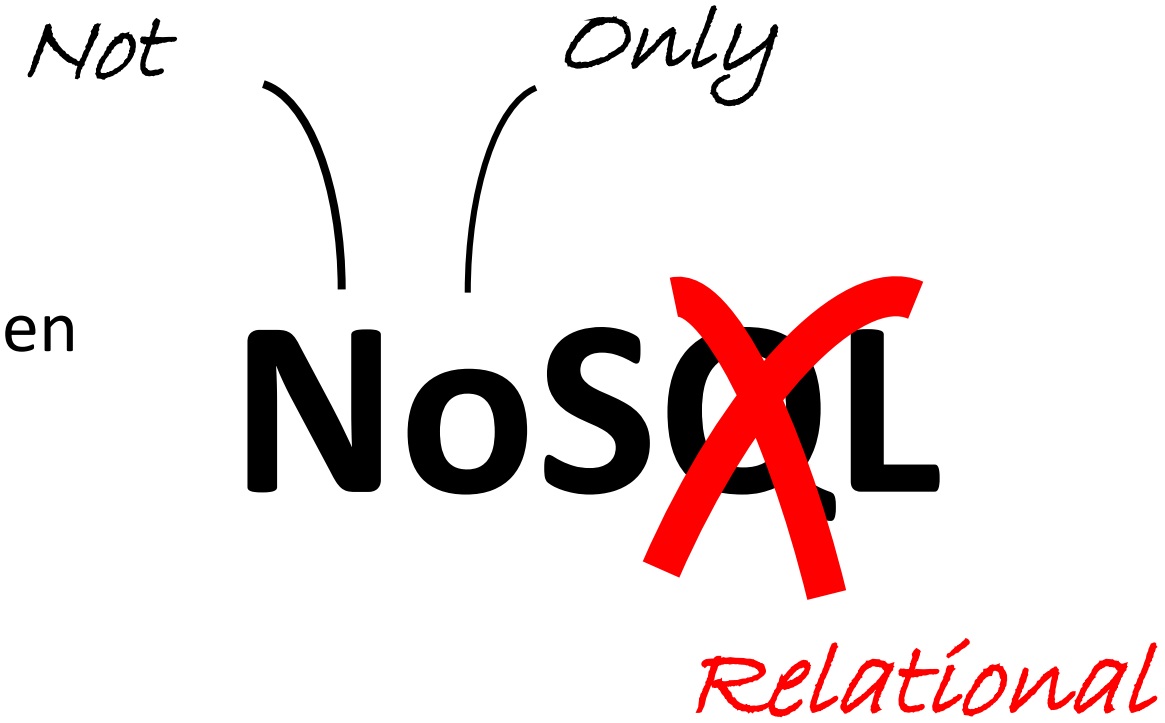
Thomas Grabetz, BsC MA

Lernziele

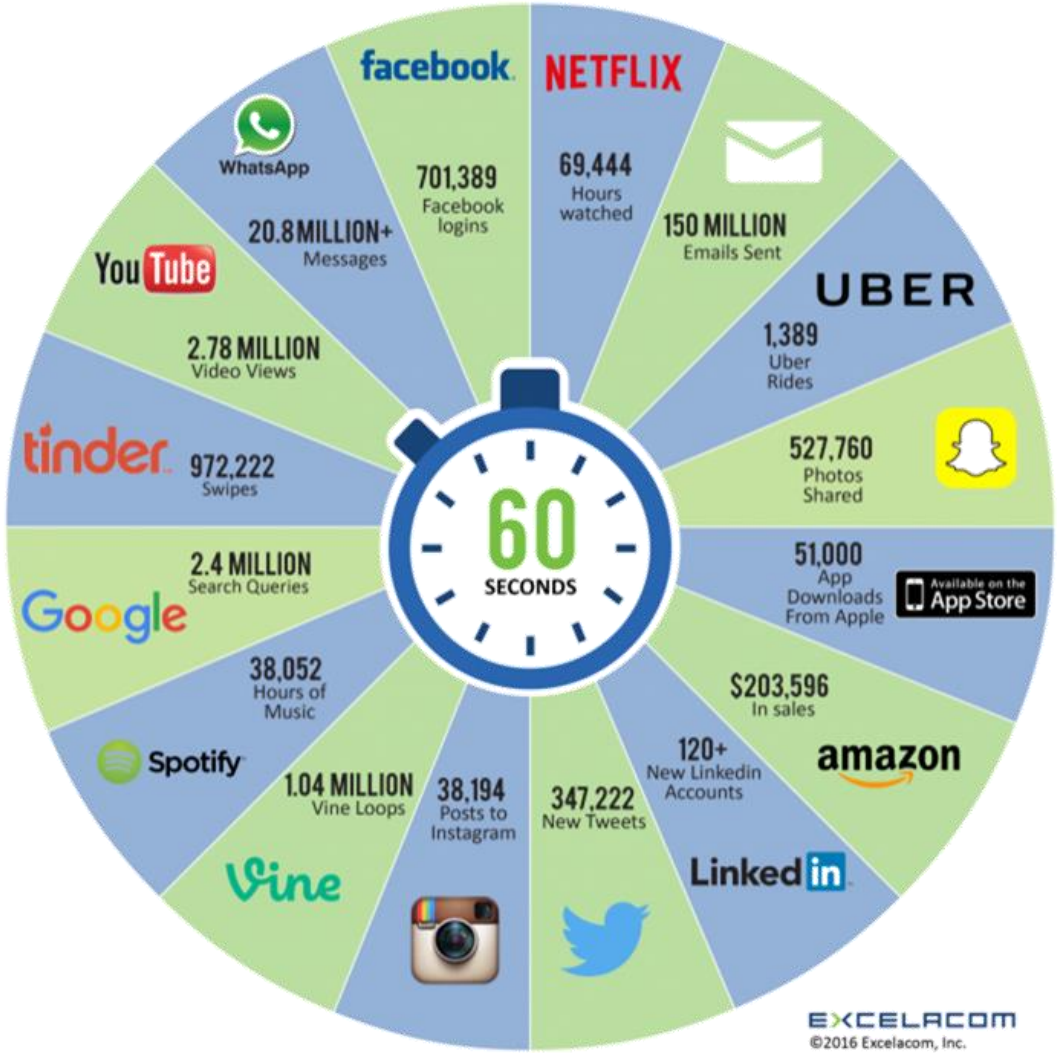
- Begriff „NoSQL“
- Wann sind NoSQL Datenbanken sinnvoll?
- Überblick über verschiedene Arten von NoSQL Datenbanken
- Konzeptioneller Unterschied: NoSQL vs. RDMS
- Ausblick: Dokumentenorientierte NoSQL Datenbank

Was bedeutet NoSQL?

- Ist nicht als Imperativ zu verstehen!
- Entsprechend der Anforderung die richtige Datenbank verwenden



Exkurs Big Data → was passiert in 60 Sek.?

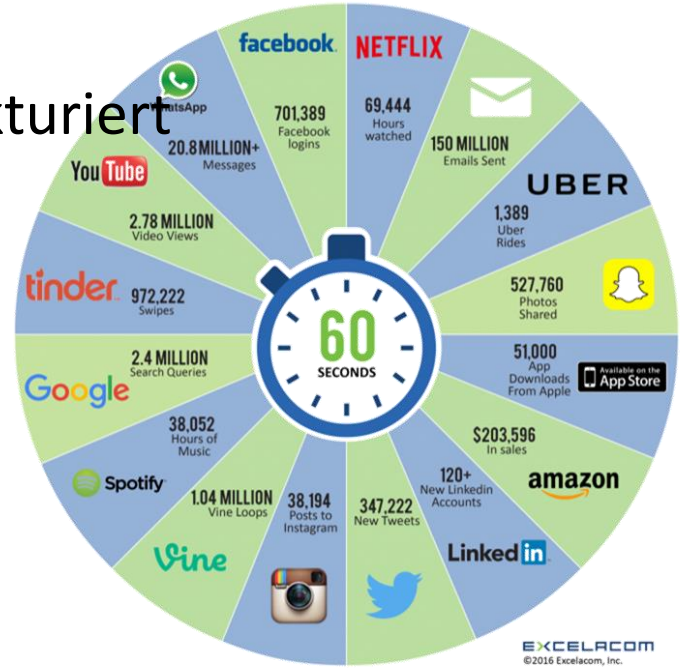


SPENGERGASSE
ausbildung mit zukunft

Exkurs Big Data

Definition:

- Grundlegende technische Herausforderungen: „3 V“
  - **Volume** → menge der Daten pro Zeiteinheit gespeichert
  - **Velocity** → Geschwindigkeit mit der Daten persistiert und verarbeitet werden
  - **Variety** → unterschiedliche Grade der Strukturiertheit
    unstrukturiert – semistrukturiert – stark strukturiert



Zusätzliche Aspekte:

- Kostenersparnis durch elastische Skalierung
- **Gewinnung neuer Informationen aus den Daten**

SPENGERGASSE
ausbildung mit zukunft

Exkurs Big Data
Gewinnung neuer Informationen aus Daten

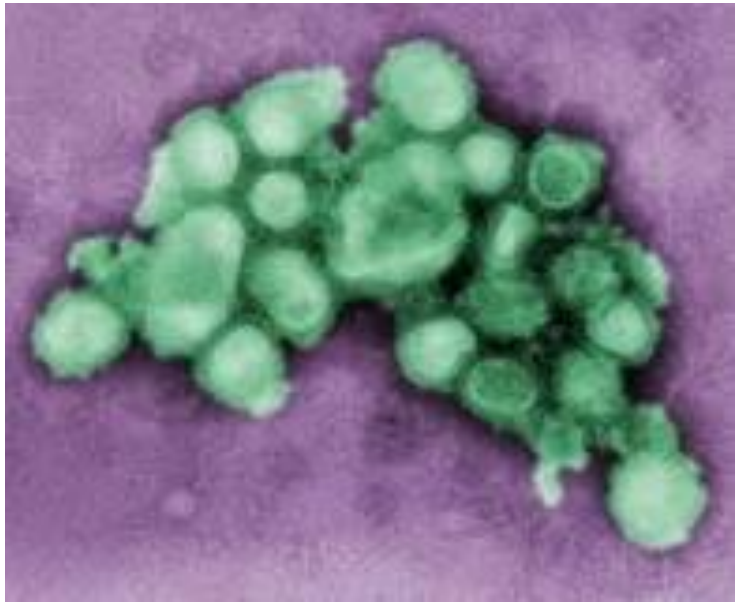
IEEE - Institute of Electrical and Electronics Engineers

Abstract: Predicting Flu Trends using Twitter data

Published in: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)

Vorhersage von Grippetrends

- Die Verringerung der Auswirkungen von saisonalen Grippeepidemien und anderen Pandemien wie dem H1N1 ist für die Gesundheitsbehörden von größter Bedeutung.
- Studien haben gezeigt, dass wirksame Maßnahmen zur Eindämmung der Epidemien ergriffen werden können, wenn eine frühzeitige Erkennung möglich ist.
- Der traditionelle Ansatz des Centers for Disease Control and Prevention (CDC) beinhaltet die Erhebung von Aktivitätsdaten zu grippeähnlichen Erkrankungen (ILI) aus "Wächterpraxen".
- Normalerweise gibt es eine Verzögerung von 1-2 Wochen zwischen der Diagnose eines Patienten und dem Zeitpunkt, zu dem der Datenpunkt in aggregierten ILI-Berichten verfügbar wird.
- In diesem Beitrag stellen wir den Rahmen des Social Network Enabled Flu Trends (SNEFT) vor, der die auf Twitter geposteten Nachrichten mit einer Erwähnung von Grippeindikatoren überwacht, um das Auftreten und die Ausbreitung einer Grippeepidemie in einer Bevölkerung zu verfolgen und vorherzusagen.
- Basierend auf den in den Jahren 2009 und 2010 gesammelten Daten stellen wir fest, dass das Volumen der Grippe-bezogenen Tweets stark mit der Anzahl der von CDC gemeldeten ILI-Fälle korreliert ist.
- Wir entwickeln weiter Auto-Regressionsmodelle, um das ILI-Aktivitätsniveau in einer Population vorherzusagen. Die Modelle prognostizieren Daten, die von CDC gesammelt und veröffentlicht werden, als den Prozentsatz der Besuche bei "Wächter"-Medizinern, die ILI in aufeinanderfolgenden Wochen zugeschrieben werden.
- Wir testen Modelle mit früheren CDC-Daten, mit und ohne Messungen von Twitter-Daten und zeigen, dass Twitter-Daten die Vorhersagegenauigkeit der Modelle erheblich verbessern können. Daher ermöglichen Twitter-Daten eine Echtzeitbewertung der ILI-Aktivität.



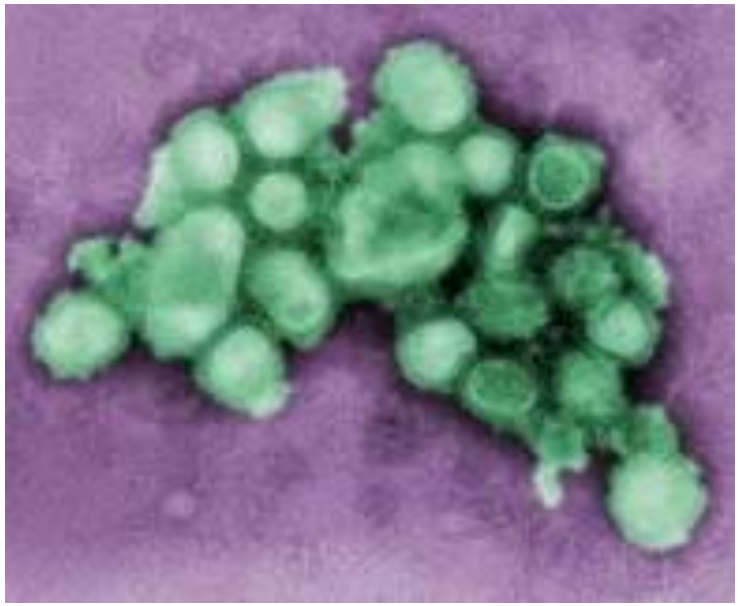
H1N1 Virus

SPENGERGASSE
ausbildung mit zukunft

Exkurs Big Data
Gewinnung neuer Informationen aus Daten

Vorhersage von Grippetrends mit Hilfe von Twitter-Daten

- Die Verringerung der Auswirkungen von saisonalen Grippeepidemien und anderen Pandemien wie dem H1N1 ist für die Gesundheitsbehörden von größter Bedeutung.
- Studien haben gezeigt, dass wirksame Maßnahmen zur Eindämmung der Epidemien ergriffen werden können, wenn eine frühzeitige Erkennung möglich ist.
- Der traditionelle Ansatz des Centers for Disease Control and Prevention (CDC) beinhaltet die Erhebung von Aktivitätsdaten zu grippeähnlichen Erkrankungen (ILI) aus "Wächterpraxen".
- Normalerweise gibt es eine Verzögerung von 1-2 Wochen zwischen der Diagnose eines Patienten und dem Zeitpunkt, zu dem der Datenpunkt in aggregierten ILI-Berichten verfügbar wird.
- In diesem Beitrag stellen wir den Rahmen des Social Network Enabled Flu Trends (SNEFT) vor, der die auf Twitter geposteten Nachrichten mit einer Erwähnung von Grippeindikatoren überwacht, um das Auftreten und die Ausbreitung einer Grippeepidemie in einer Bevölkerung zu verfolgen und vorherzusagen.
- Basierend auf den in den Jahren 2009 und 2010 gesammelten Daten stellen wir fest, dass das Volumen der Grippe-bezogenen Tweets stark mit der Anzahl der von CDC gemeldeten ILI-Fälle korreliert ist.
- Wir entwickeln weiter Auto-Regressionsmodelle, um das ILI-Aktivitätsniveau in einer Population vorherzusagen. Die Modelle prognostizieren Daten, die von CDC gesammelt und veröffentlicht werden, als den Prozentsatz der Besuche bei "Wächter"-Medizinern, die ILI in aufeinanderfolgenden Wochen zugeschrieben werden.
- Wir testen Modelle mit früheren CDC-Daten, mit und ohne Messungen von Twitter-Daten und zeigen, dass Twitter-Daten die Vorhersagegenauigkeit der Modelle erheblich verbessern können. Daher ermöglichen Twitter-Daten eine Echtzeitbewertung der ILI-Aktivität.



H1N1 Virus

Exkurs Big Data
Gewinnung neuer Informationen aus Daten

Auswertung von Tweets

Tweets welche die Worte:

- Schweinegrippe (**blau**)
- H1N1 (**rot**) oder
- beide Worte enthalten (**grün**)

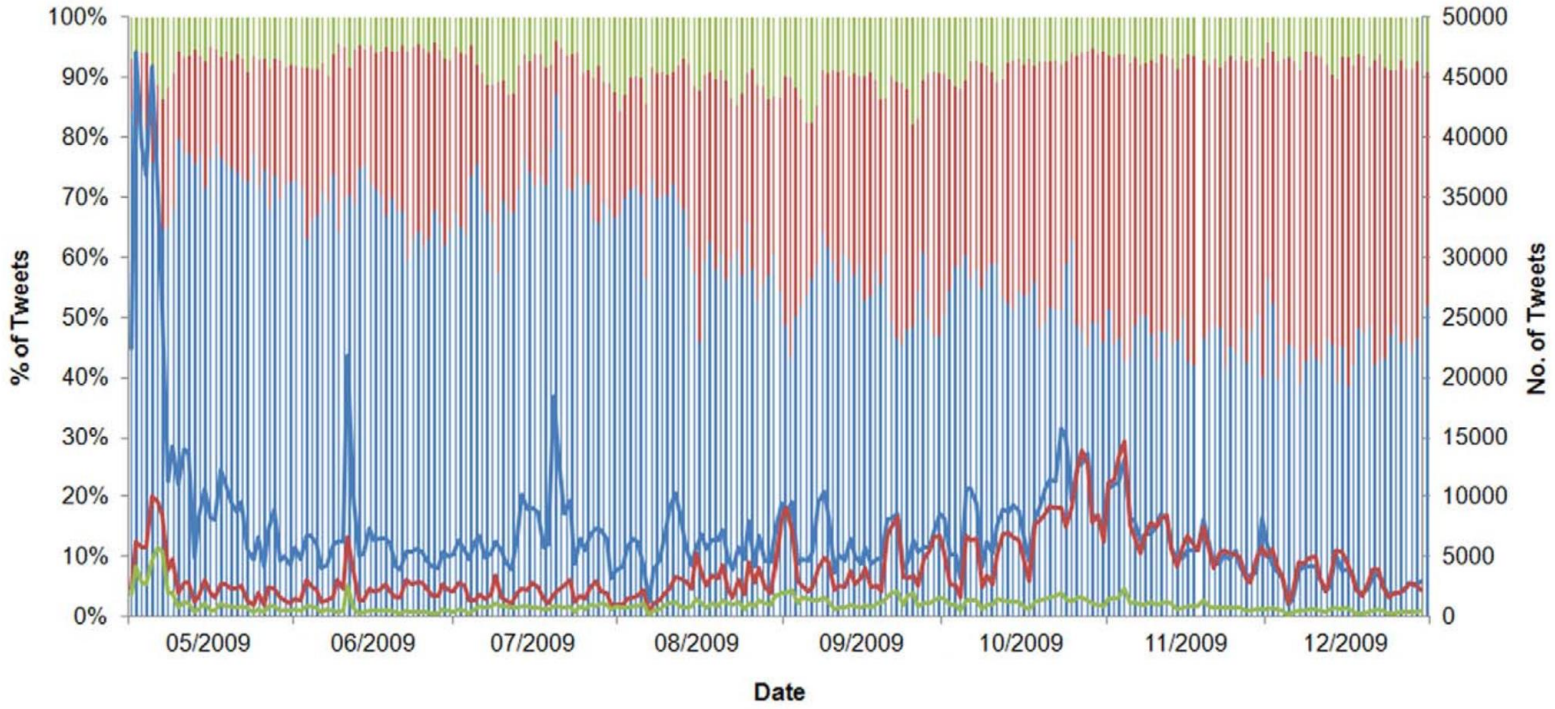


Figure 1. Tweets containing H1N1, swine flu, or both from May to December 2009. Lines = absolute number. Bars = relative percentage. Blue = "swine flu" or swineflu. Red = H1N1. Green = ("swine flu" or swineflu) AND... Continue Reading

Published in PloS one 2009

Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak

Cynthia Chew, Gunther Eysenbach

SPENGERGASSE
ausbildung mit zukunft

Exkurs Big Data
Gewinnung neuer Informationen aus Daten

Verleihung des Big Brother Awards 2019

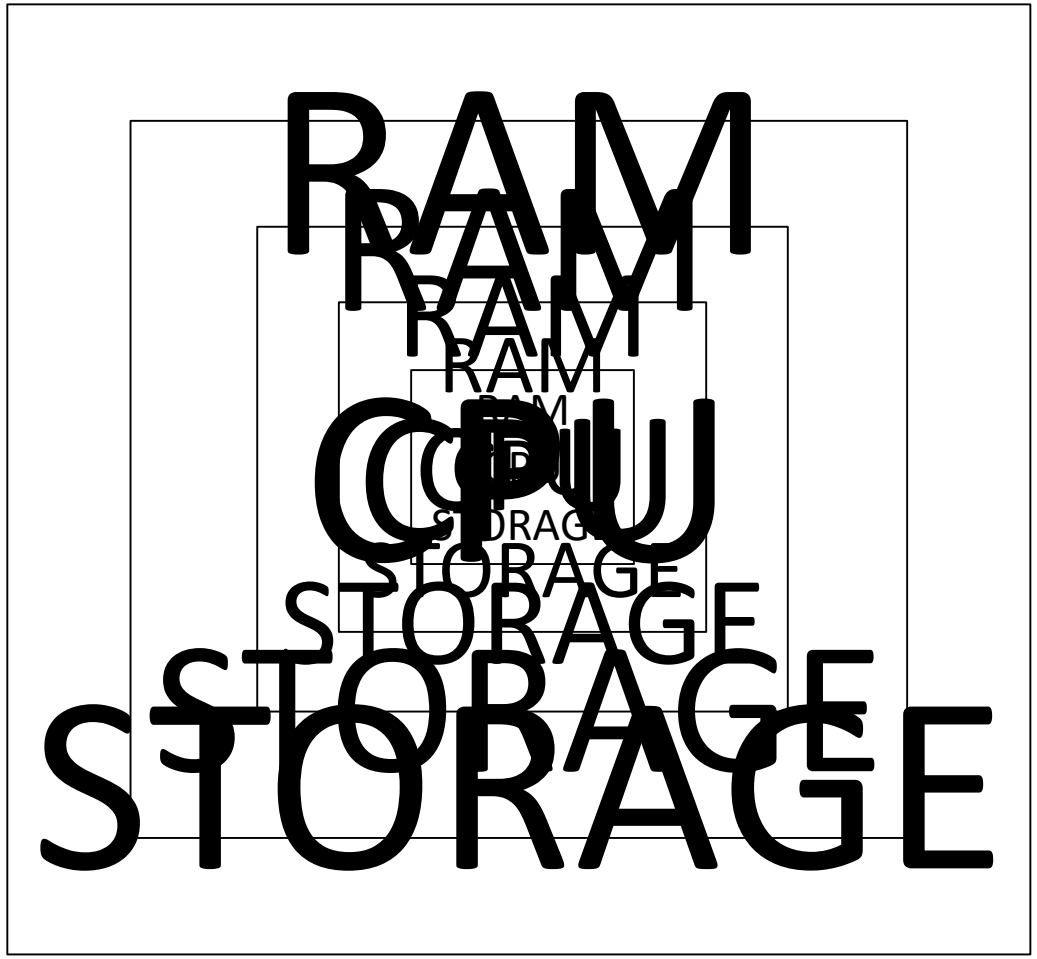
- **JÖ-Bonusclub: Profiling und schwierige Kündigung**
  (In der Kategorie "Kommunikation und Marketing„)
  Das neue Bonusprogramm vereint viele Unternehmen, auch über die Konzerngrenzen der Rewe-Gruppe hinaus (OMV, Merkur, Libro, interio, Verbund, Pagro, Billa, Zgonc, Billa Reisen, Penny, Bipa, Adeg, Bawag PSK, Pearle).
  Wenn alle Kundentransaktionen zusammen ausgewertet werden können, seien dem Profiling Tür und Tor geöffnet, so die Begründung der Jury. (Quelle „Konsument.at“)

- **Post-Algorithmus: Unzulässige Datenverarbeitung**
  (Die Kategorie "Business und Finanzen„)
  Es gewann der Post-Algorithmus, mit dem die Parteivorlieben von Bürgern errechnet wurden. Anfang des Jahres sorgte der Datenskandal um die Speicherung dieser Vorlieben von Millionen Post-Kunden und der mutmaßliche Verkauf dieser Daten an wahlwerbende Parteien für Aufregung.
  Die Post wurde – nicht rechtskräftig – im Oktober 2019 zu € 18 Mio Verwaltungsstrafe verurteilt

Skalierung großer Datenbanken

Große **Relationale Datenbanken**

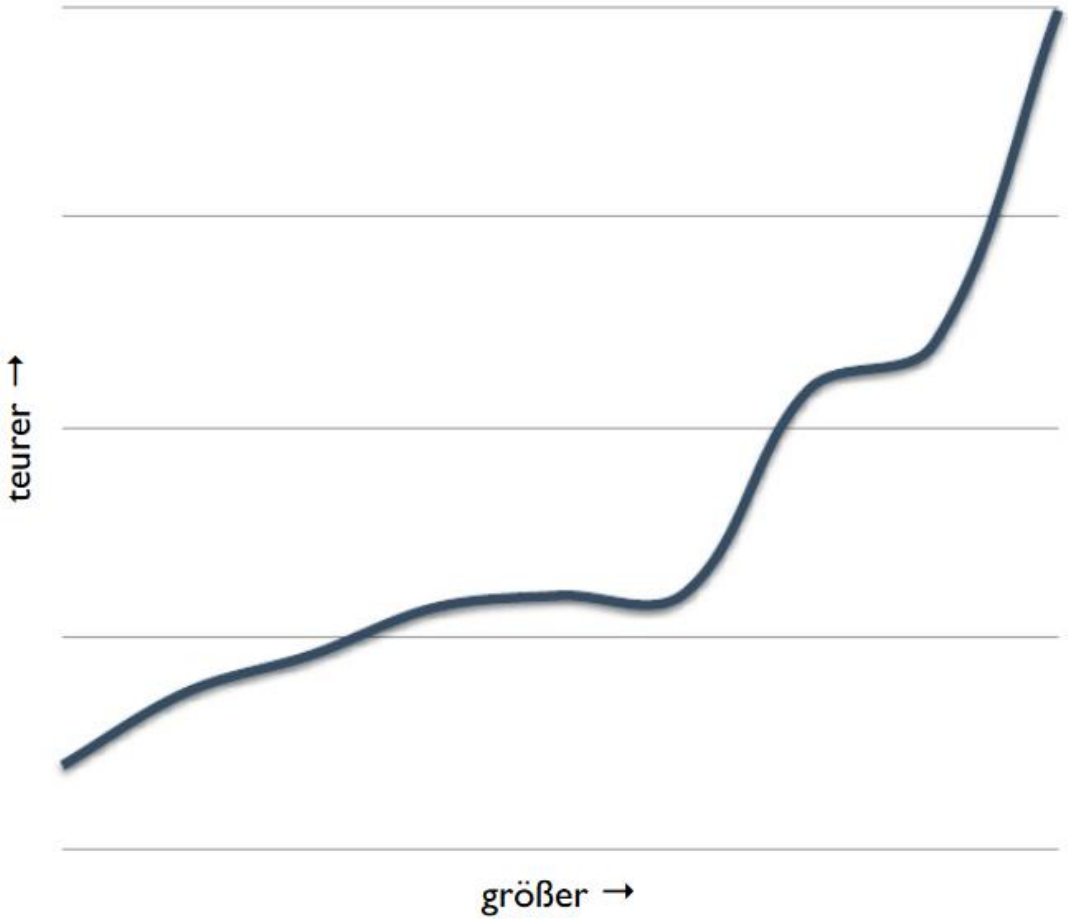
Skalieren vertikal



Skalierung großer Datenbanken

Große **Relationale Datenbanken**

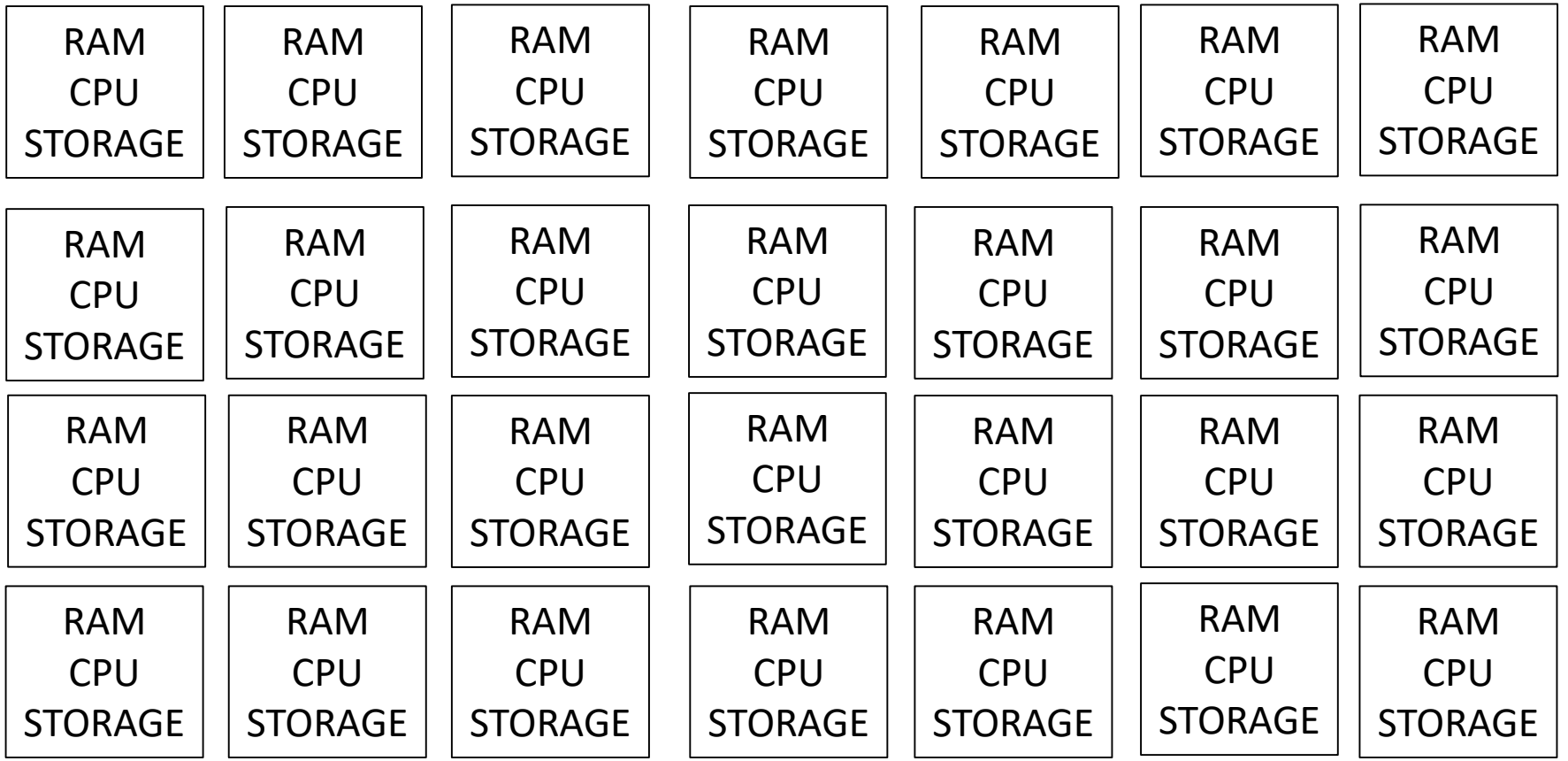
Skalieren vertikal



Skalierung großer Datenbanken

Große **NoSQL Datenbanken**

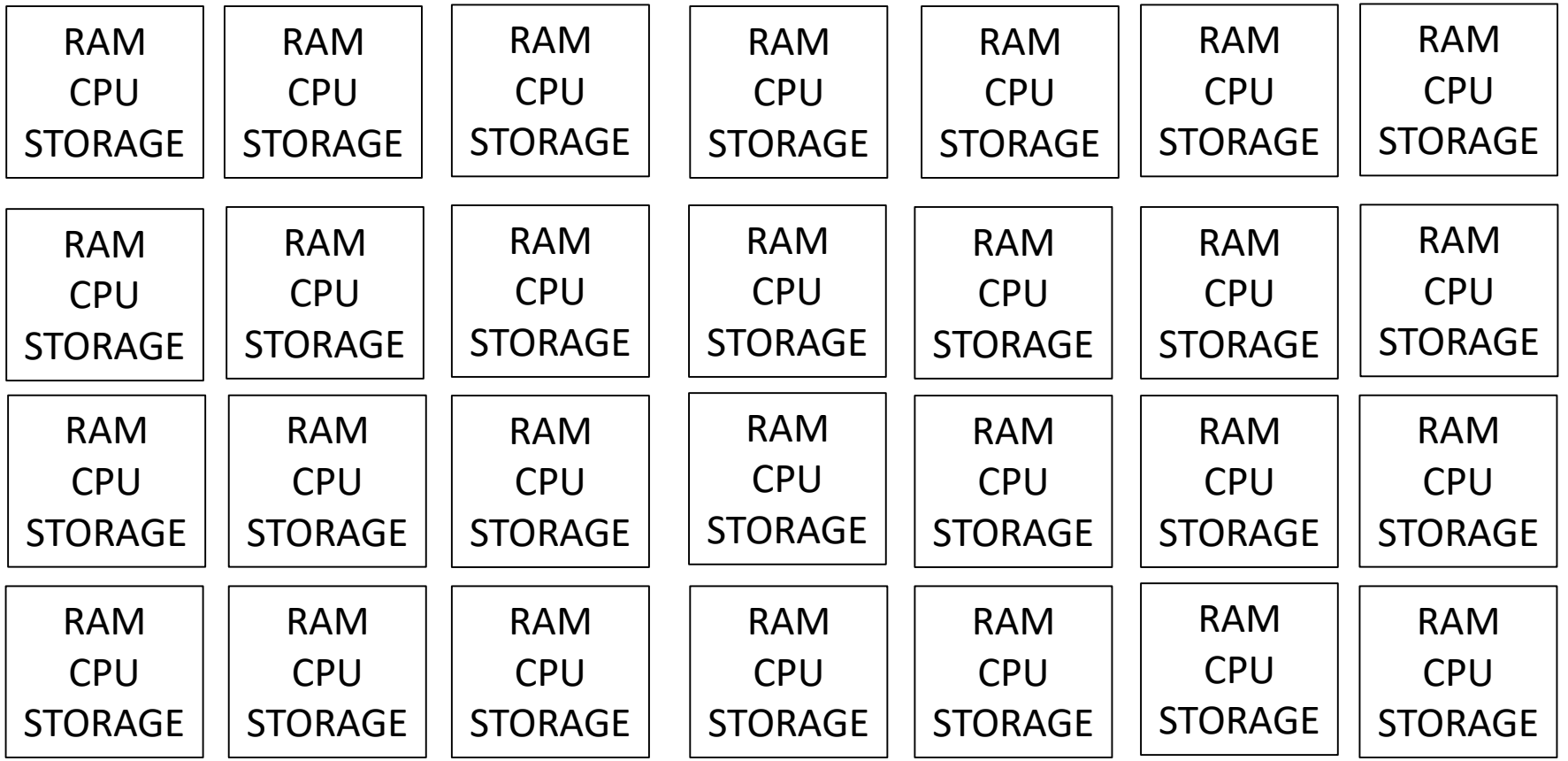
Skalieren horizontal



Skalierung großer Datenbanken

Beispiel Facebook

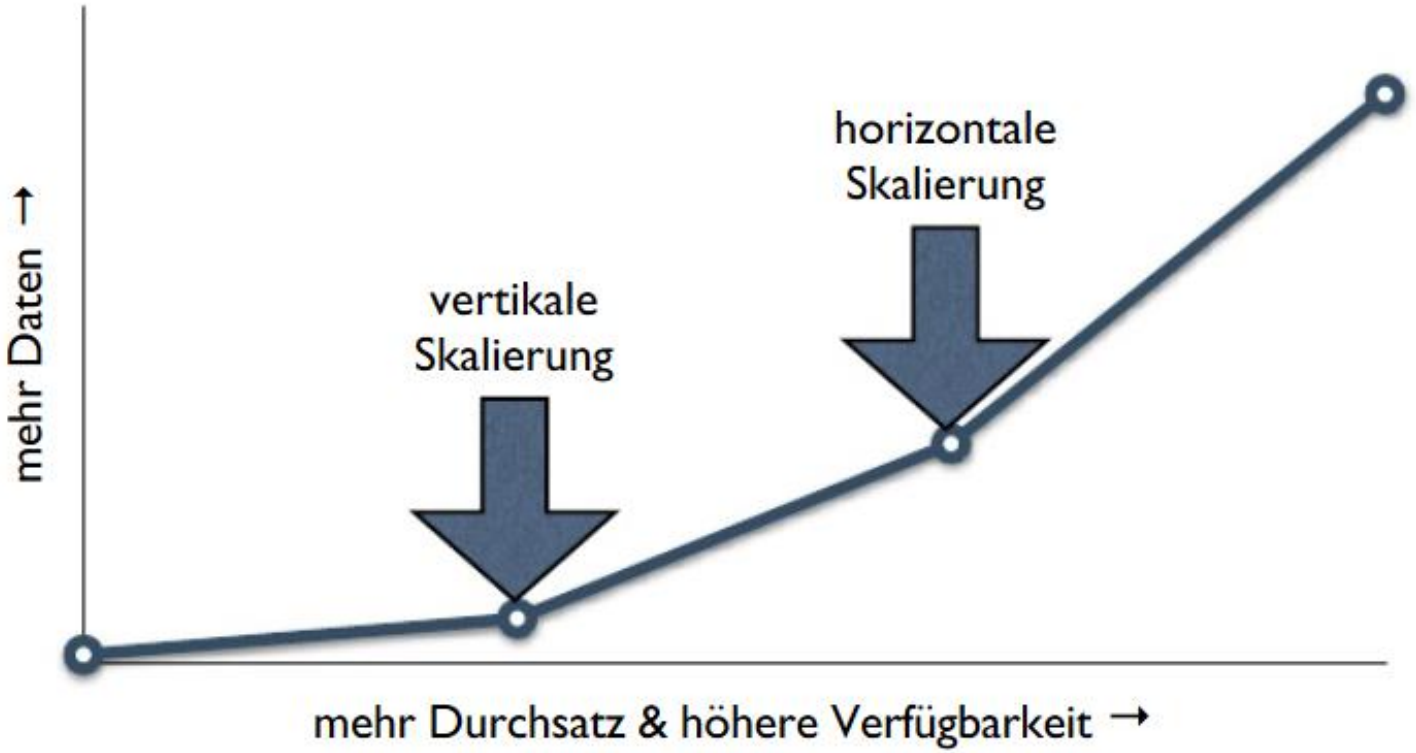
- 30.000 Server
- 25 Terabyte Logdaten täglich
- 300.000.000 Nutzer
- 230 Ingenieure



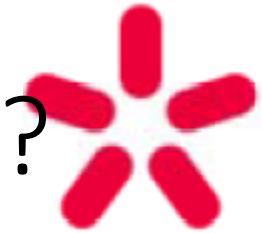
SPENGERGASSE
ausbildung mit zukunft

Skalierung großer Datenbanken

Horizontale Skalierung führt zu verteilten Systemen



Wann verwendet man NoSQL Datenbanken?



- Probleme bei der Anzahl schreibender Operationen / Sekunde?
- Wie gut passt das logische Datenmodell zum Relationen-Modell?
- Wie schnell oder wie oft müssen Reports erstellt werden können?
- Müssen sehr viele Daten sehr schnell gespeichert und abgerufen werden?
- Wie skalierbar muss die Datenbank sein?
- Wie viele Nutzer arbeiten (werden arbeiten) mit der Datenbank (Lizenzproblem)

Kategorien von NoSQL Datenbanken

Die wichtigsten Vertreter von NoSQL Datenbanken lassen sich in 4 Kategorien einteilen:

1. **Key-Value-Datenbanken:**
   verwalten Tupel bestehend aus einem Schlüssel und einem Wert Abfragen nur über den Schlüssel möglich

2. **Spaltenorientierte Datenbanken *(„Wide Column Stores“)***
   verwenden Tabellen, bei denen ein Datensatz allerdings eine dynamische Anzahl an Spalten haben kann. Können als „Verallgemeinerung“ von Key Value Datenbanken gesehen werden. Indizes sind frei definierbar und ermöglichen die Abfrage über beliebige Spalten

Kategorien von NoSQL Datenbanken

Die wichtigsten Vertreter von NoSQL Datenbanken lassen sich in 4 Kategorien einteilen:

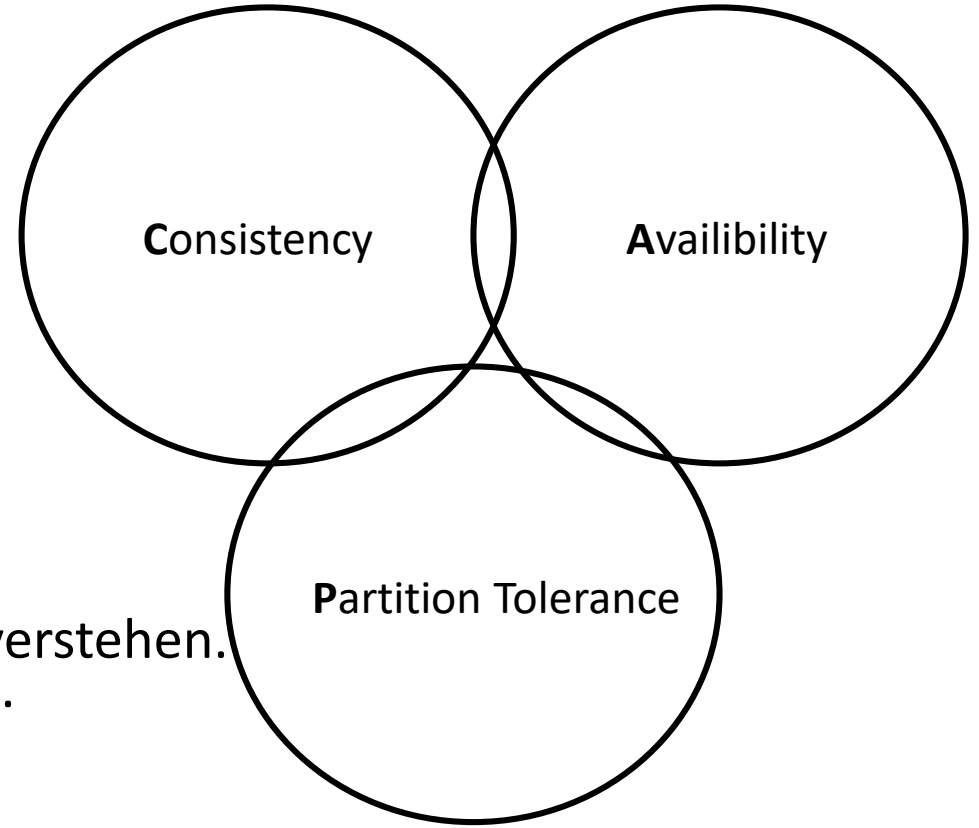
3. **Graphen Datenbanken**
   spezialisieren sich auf de Verwaltung von
   Knoten und Kanten zwischen diesen Knoten.
   Abfragen ermöglichen u.a. das Traversieren des Graphen

4. **Dokumentenorientierte Datenbanken**
   ein Dokument ist ein einzelner Datensatz, der im Prinzip aus einer geordneten Liste von Key-Value-Paaren besteht und als Werte auch Arrays und eingebettete Dokumente zulässt.

CAP –Theorem & Verteilte Systeme

In verteilen Systemen können maximal zwei der folgenden Eigenschafen gleichzeitig gelten:

1. **Consistency (C)**
   Alle Knoten haben jederzeit den gleichen Datenbestand
2. **Availability (A)**
   Das System steht für Lese- und Schreibzugriffe zur Verfügung.
3. **Partition Tolerance (P)**
   Toleranz gegenüber dem Ausfall einzelner Knoten und/oder Netzwerkstrecken.

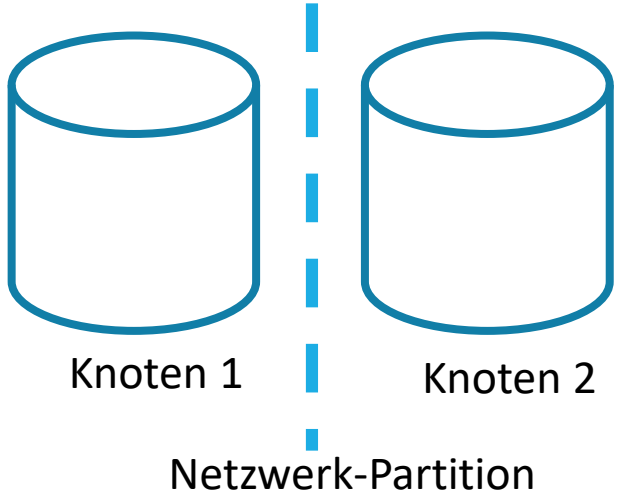


Diese Eigenschaften sind dabei als graduelle Größen zu verstehen.
D.H. Sie können zwischen gar nicht und voll erfüllt liegen.

Auswirkungen des CAP Theorems

Beispiel:

Das Netzwerk zwischen diesen beiden Knoten ist gestört. Es liegt eine Partition des Netzwerkes in zwei Teilen vor – die Knoten können einander nicht erreichen.



Fall 1: Ein Knoten kann seinen Zustand ändern (im Fall eines DB Systems schreibende Operation)

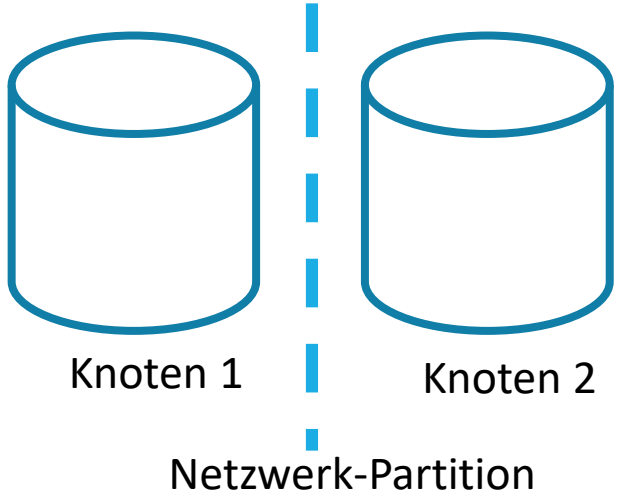
→ Das Gesamtsystem wird inkonsistent = man gibt „C“ auf

1. Consistency (C)
2. Availabliity (A)
3. Partition Tolerance (P)

Auswirkungen des CAP Theorems

Beispiel:

Das Netzwerk zwischen diesen beiden Knoten ist gestört. Es liegt eine Partition des Netzwerkes in zwei Teilen vor – die Knoten können einander nicht erreichen.



Fall 2: Die Konsistenz soll erhalten bleiben; Es muss sich der Knoten, dessen Zustand sich nicht ändert, als nicht verfügbar „abmelden“ da er sonst, gegenüber den Clients einen nicht aktuellen Datenstand ausliefern würde.

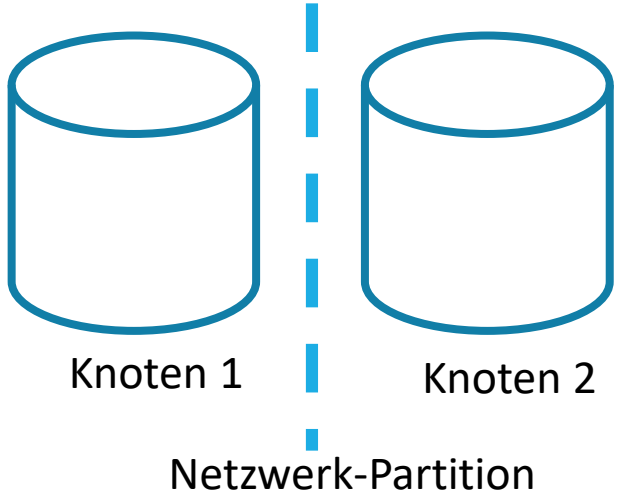
→ Die Verfügbarkeit wäre nur auf einen Knoten reduziert
= „A“ wäre nicht gegeben

1. Consistency (C)
2. Availabliity (A)
3. Partition Tolerance (P)

Auswirkungen des CAP Theorems

Beispiel:

Das Netzwerk zwischen diesen beiden Knoten ist gestört. Es liegt eine Partition des Netzwerkes in zwei Teilen vor – die Knoten können einander nicht erreichen.



Fall 3: Nur wenn alle Knoten stets miteinander Kommunizieren können, kann man „C“ und „A“ dauerhaft aufrecht erhalten.

→ Steht im Wiederspruch zu „P“

1. Consistency (C)
2. Availabliity (A)
3. Partition Tolerance (P)

CAP –Theorem → RDBMS vs. NoSQL DB

1. Consistency (C)
2. Availabliity (A)
3. Partition Tolerance (P)

RDBMS:

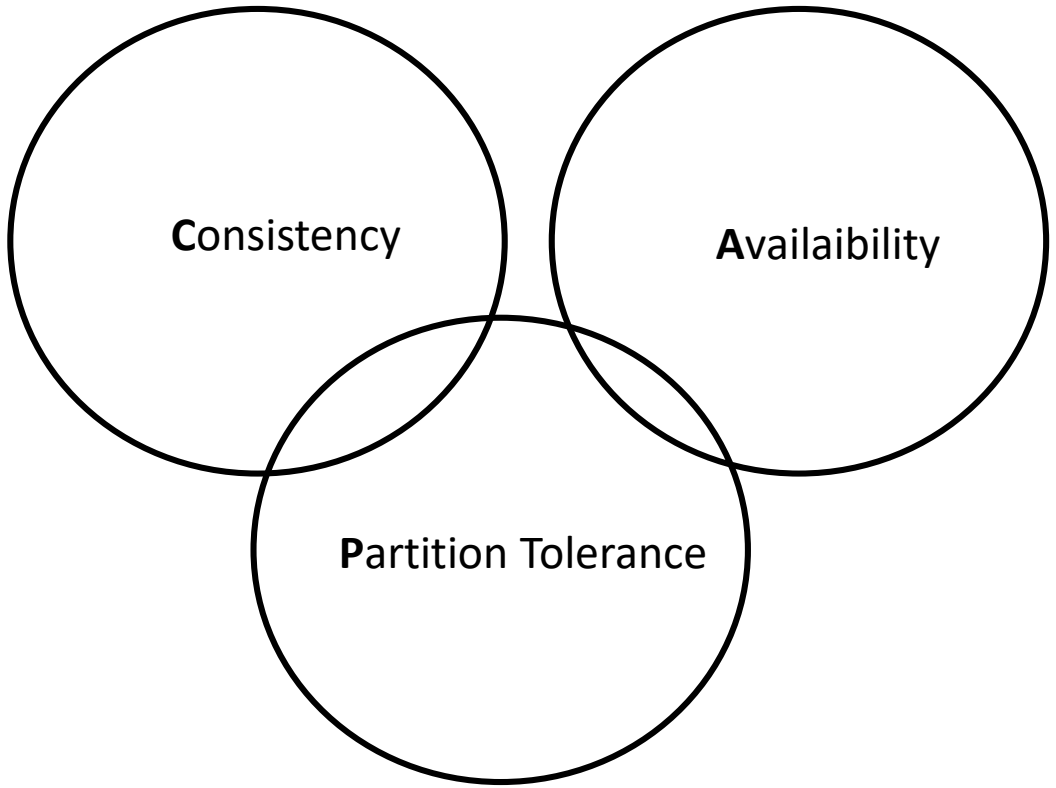
- Legen großen Wert auf die Konsistenz und garantieren hohe Verfügbarkeit („CA“-Kategorie)
- Bei Datenbanken die nur auf einem Knoten laufen, kann die Ausfallsicherheit durch Kauf teurer HW auf nahezu 100% gebracht werden. → Fällt der Knoten jedoch aus, ist das System nicht mehr verfügbar
- Bei Datenbank-Clustern verringert sich die Verfügbarkeit, da die Transaktionen zur Sicherstellung der Konsistenz auf mehreren Knoten eine längere Laufzeit haben.

CAP –Theorem & Verteilte Systeme

„…in larger distributed-scale systems, network partitions are given.

Therefore:
Consistency and **Availaiblity** cannot be achieved at the same time…“

Werner Vogls, Amazon.com



CAP –Theorem → RDBMS vs. NoSQL DB

NoSQL:

- Bei vielen NoSQL Datenbanken kann man aufgrund der hohen Verteilung nicht auf die Partition Tolerance (P) verzichten.

- Die Verfügbarkeit nimmt ebenfalls einen hohen Stellenwert ein, da man grundsätzlich alle Anfragen an das System beantworten können möchte.

- Somit kann nurmehr an der Consistenz (C) „gespart“ werden (ist eine bewusst Design Entscheidung).

ACID vs. BASE

Das ACID-Prinzip von relationalen Datenbanken stellt, in Zusammenhang mit Transaktionen, folgendes sicher:

- **Atomicity (Atomarität):**
  Eine Transaktion wird nach dem „Alles-oder-nichts“ Prinzip entweder vollständig oder gar nicht ausgeführt. Wird eine atomare Transaktion abgebrochen, ist das System in einem unverändertem Zustand.

- **Consistency (Konsistenz)**

In einem konsistenten Datenbanksystem führt eine Folge von Datenbankoperationen wieder zu einem konsistenten Zustand

ACID vs. BASE

Das ACID-Prinzip von relationalen Datenbanken stellt, in Zusammenhang mit Transaktionen, folgendes sicher:

- **Isolation**
  Parallel ausgeführte Transaktionen beeinflussen sich nicht gegenseitig

- **Durability (Dauerhaftigkeit)**

Die Auswirkungen von Transaktionen müssen dauerhaft im System gespeichert werden – insbesondere bei Systemabstürzen.

ACID vs. BASE

NoSQL – Datenbanken → BASE

- **B**asically Available
- **S**oft State
- **E**ventual Consistency

Das Akronym „BASE“ ist gekünstelt um einen einprägsamen Gegensatz zu „ACID“ darzustellen (engl. Lauge / Säure). Für „Basically Available“ und „Soft State“ gibt es keine präzise Definition.

Vielmehr steht „BASE“ für ein Design Prinzip, das das Konzept der „absoluten Konsistenz“ aufgibt, statt dessen die Verfügbarkeit des Systems erhöht und dadurch zwischenzeitlich in ***einem etwas undefinierten Zustand*** sein kann.

ACID vs. BASE → 2 verschiedene Konzepte

BASE → Konsistenz in verteilten Systemen

ACID → Integrität im Zusammenhang mit Transaktionen

Somit betrachten wir zwei Seiten einer Medaille…. Aber nicht der selben Medaille.

SQL vs. NoSQL (1) - Allgemein

SQL Datenbanken

- sind mächtig und können für die meisten Datenbankprobleme herangezogen werden
- Aufgrund langjähriger (Weiter-)Entwicklung wird die Arbeit mit SQL-Datenbanken immer einfacher
- Vielzahl an Anbieter, Schulungen, Support, Manpower etc…
- Einheitliche, sehr mächtige Abfragesprache → SQL

SQL vs. NoSQL (2) - Skalierung

NoSQL Datenbanken

- Horizontale Skalierung → DIE Stärke von NoSQL Datenbanken
  - Bei SQL Datenbanken ist horizontale Skalierung zwar möglich, jedoch nur mit wesentlich höherem Verwaltungsaufwand und nur begrenzt (ab einem gewissen Punkt führen die Vorteile von SQL ins Negative)
  - Aufgrund des einfachen Schemas (bzw. keines Schemas) sind NoSQL Datenbanken für hohe Skalierbarkeit geschaffen.
  - Die Skalierbarkeit bleib auch bei sehr hohen Datenvolumina erhalten

SQL vs. NoSQL (3) - Performance

NoSQL Datenbanken

- Sind performanter als Relationale Datenbanken → besonders bei hohen Datenvolumen bemerkbar
- Relationale Datenbanken stoßen bei sehr großen Datenvolumina an ihre Grenzen (ist ein Grund weshalb NoSQL Datenbanken erfunden wurden)
- Unabhängig ob Lese- oder Schreibzugriff, NoSQL Datenbanken sind den SQL Datenbanken voraus

SQL vs. NoSQL (4) - Konsistenz

SQL Datenbanken

- Aufgrund der ACID-Eigenschaften besitzen Relationale Datenbanken eine bessere Konsistenz → eigentlich eine absolute Konsistenz
- NoSQL Ansatz „eventually consistency“ („schlussendlich konsistent“) → es wird nicht garantiert, dass nach einem Update immer derselbe Wert zurückgegeben wird. Eine Reihe von Bedingungen müssen erfüllt sein, bis alle denselben Wert bekommen.
- Grundsätzlich muss der Nutzer entscheiden ob Performance im Vordergrund steht, oder ob eine gute Konsistenz notwendig ist.

SQL vs. NoSQL (5) - Beziehungen

- Beziehungen sind eine der schwierigsten und ressourcen-intensivsten Dinge, die man mit Hilfe von SQL-Datenbanken erstellen kann.
- Das Speichern von vernetzten Informationen oder zusammenhängenden Objekten kann man bei SQL-Datenbanken sehr schwer realisieren (bzw. nur mit sehr hohem Aufwand).
- Graphen-Datenbanken sind NoSQL Datenbanken, welche darauf spezialisiert sind, vernetzte Informationen zu speichern.

SQL vs. NoSQL (6) - Fazit

- Aufgrund des zunehmenden Datenvolumens, der Notwendigkeit von steigender Performance und der zunehmenden Wichtigkeit, Beziehungen in Datenbanken zu definieren, werden NoSQL Datenbanken immer beliebter.

- Sie werden SQL Datenbanken jedoch **nicht** Ablösen. Beide Systeme werden parallel existieren und **einander ergänzen**.

- Je nach Verwendungszweck muss zwischen den beiden Systemen gewählt werden.