Big-Data-Technologien

Kapitel 17: Zusammenfassung und Ausblick

Hochschule Trier
Prof. Dr. Christoph Schmitz

Big Data

- Verarbeiten von Daten, die wegen
  - **Umfang** (Volume)
  - **Geschwindigkeit** (Velocity)
  - **Vielfalt** (Variety)

  bisher nicht zu verarbeiten waren

Arten der Verarbeitung

- **Batch**
  - hoher Durchsatz
  - hohe Latenz
- **Stream**
  - geringe Latenz
  - geringerer Durchsatz
  - komplexe Verarbeitung problematisch
- **NoSQL**
  - Kompromisse gegenüber RDBMS
  - dafür: skalierbar, flexible Datenmodelle

Strategien

- **Verteilung**
  - ideal: n Maschinen = n-fache Leistung
  - Verteilung kostet: Aufwand, Ausdrucksmöglichkeiten, …
- **Vereinfachung**
  - Kompromisse bei Datenmodell und Algorithmen
  - z. B. MapReduce
- **Planung**
  - breite und schmale Transformationen
  - Medienbrüche vermeiden

Strategien

- **Partitionierung**
  - Welche Daten liegen wo?
  - Welche Berechnung läuft wo?
  - Welcher Zustand liegt wo?

- **Prinzipien**
  - Verteilten Zustand vermeiden
  - Code folgt den Daten
  - Daten im System belassen

Strategien

- Ideen aus der **funktionalen Programmierung**
  - **Unveränderliche Daten**
    replizierbar, Caching und Neuberechnung möglich
  - **Higher-Order Functions**
    trenne Berechnungsmodell von konkreter Berechnung (z. B. MapReduce)
  - **Vermeiden von veränderlichem Zustand**
    erschwert Parallelisierung, Robustheit, Wiederanlauf, …

Ausblick

- **Data Science:** kombiniert
  - **Fachlichkeit**
  - **Statistik/Machine Learning**
  - **Technologien/Big Data/…**

um bessere, schnellere Erkenntnisse basierend auf allen vorhandenen Daten zu gewinnen

Ausblick

- **Näherungsverfahren**

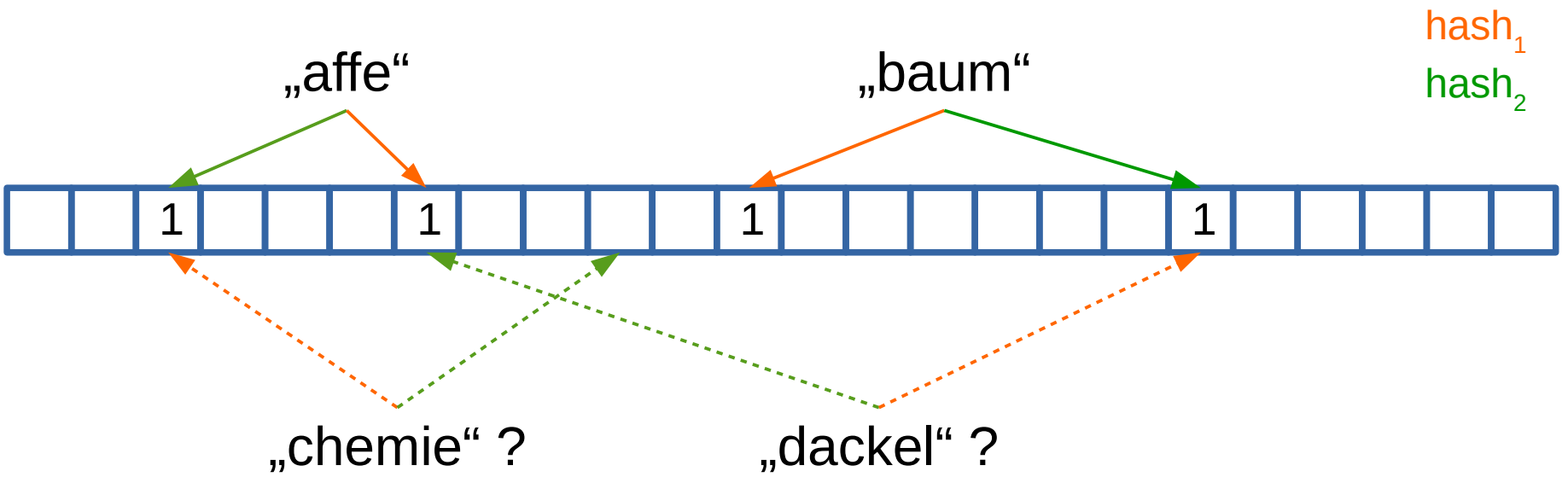
  Oft ist ein ungefähres Ergebnis gut genug!

- Auf meiner Webseite waren 15.324.432 unterschiedliche Benutzer. ← **5h Rechenzeit**

- Auf meiner Webseite waren ungefähr 15 Mio. unterschiedliche Benutzer. ← **5s Rechenzeit**

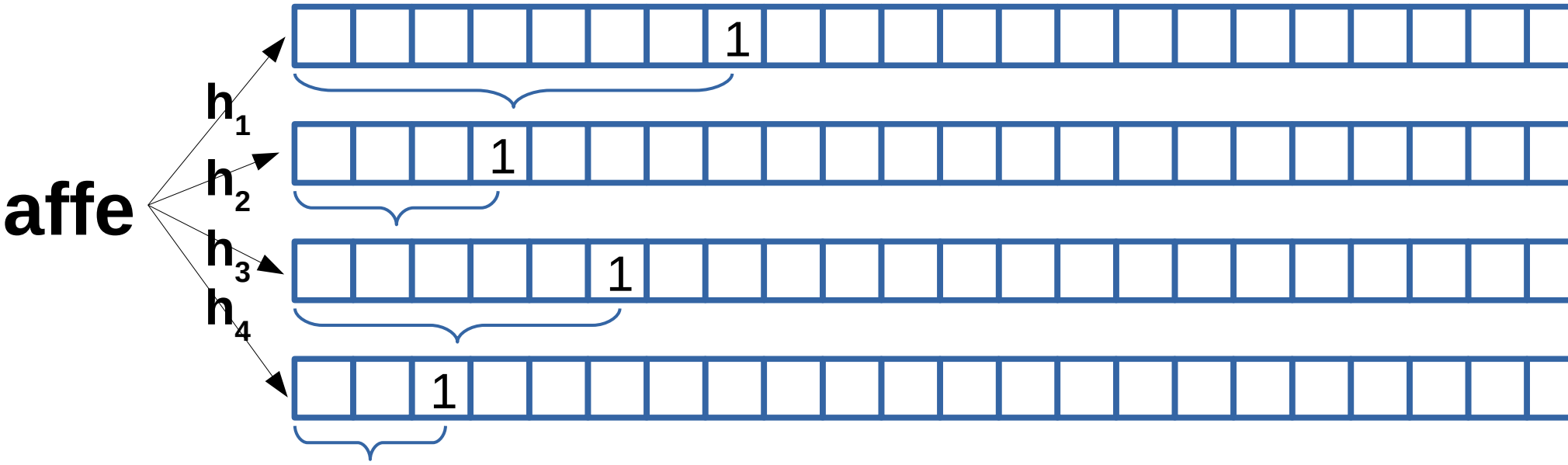
Näherungsverfahren

- **Bloom-Filter:** Element in Menge oder nicht?



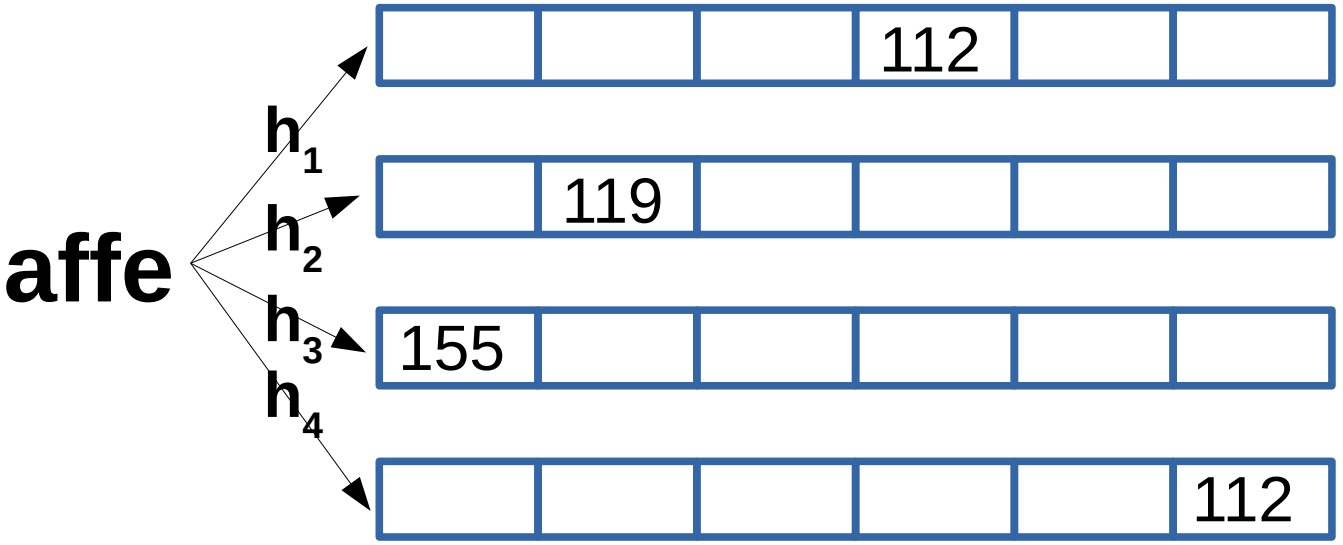
Näherungsverfahren

- **HyperLogLog:** Wie viele unterschiedliche Elemente? (Mengenkardinalität)



Näherungsverfahren

- **Count-Min-Sketch**: Welches Element ist wie häufig?

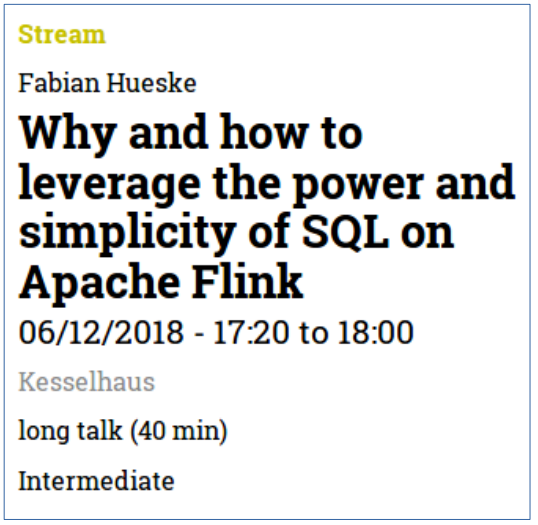


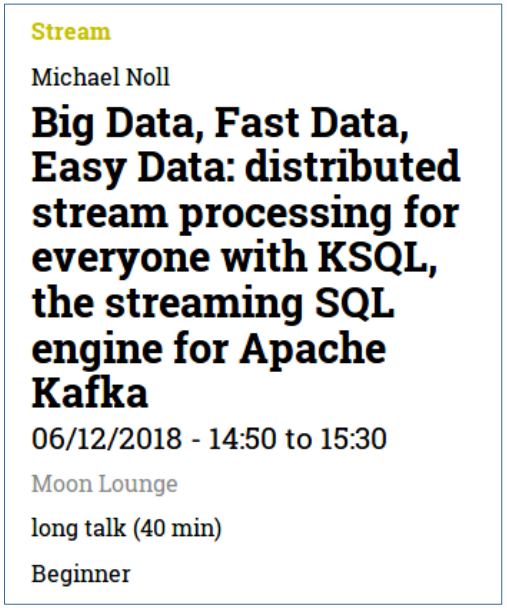
Näherungsverfahren

- Gemeinsamkeiten
  - Tausche **Speicherplatz**/**Rechenzeit** gegen **Genauigkeit**
  - **Günstiges** Verhältnis von Ersparnis zu Gewinn
  - "**Sketches**" haben günstige Eigenschaften (z. B. Additivität)
- Typische Aussage:
  - "Bei **Speicherplatz** $O(1/\varepsilon + 1/\ln \delta)$ Zählung mit **Fehler** kleiner ε mit **Wahrscheinlichkeit** $1-\delta$"

Aktuelle Themen

- **Stream**-Verarbeitung
- Big Data in der **Cloud**
- **Data Science**, Analytics, Machine Learning, neuronale Netze
- **Anwendungen:** Big Data für XYZ
- **SQL** ist wieder da!





Screenshots: berlinbuzzwords.de