

Big-Data-Technologien

Kapitel 1: Einführung

Hochschule Trier
Prof. Dr. Christoph Schmitz

Überblick

- Organisatorisches
- Was ist Big Data?
- Anwendungsfälle
- Herausforderungen
- Verarbeitungsmodelle

Organisatorisches

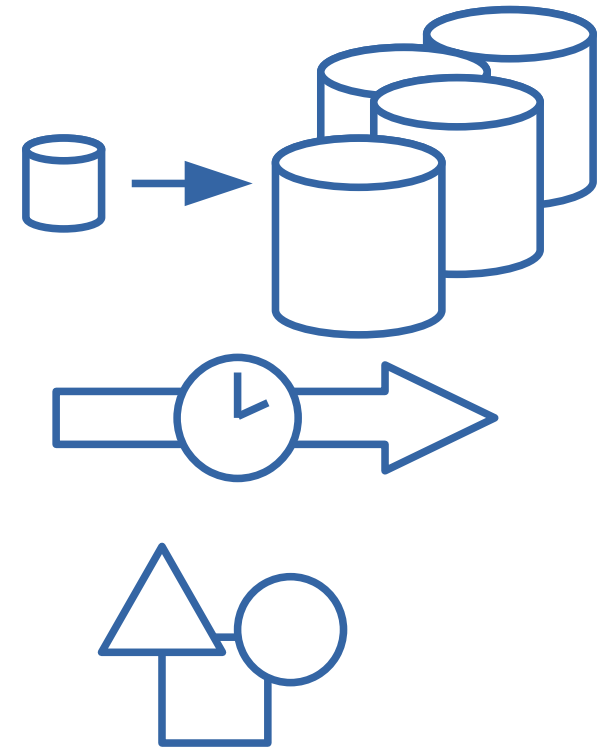
Organisatorisches

- Übung am Rechner
 - SSH auf Big-Data-Systeme
 - Java-Entwicklungsumgebung (Eclipse/IntelliJ/...)
 - evtl. Skriptsprachen
- Prüfungsvorleistung: 2/3 der Übungsaufgaben bearbeiten

Was ist Big Data?

Was ist Big Data? – 3 Vs

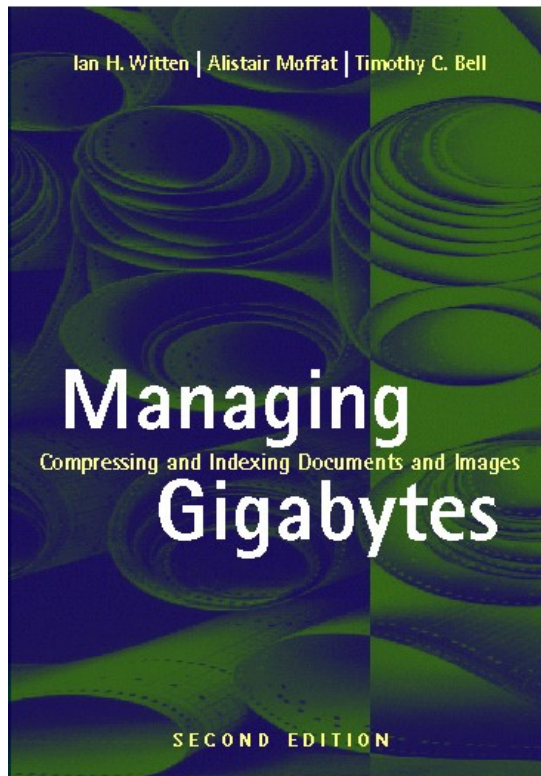
- **Volume:** viele Daten
- **Velocity:** schnelle Verarbeitung
- **Variety:** vielfältige Struktur



(Gartner, 2011)

Was ist Big Data?

- "Data whose size forces us **to look beyond the tried-and-true methods** that are prevalent at that time."

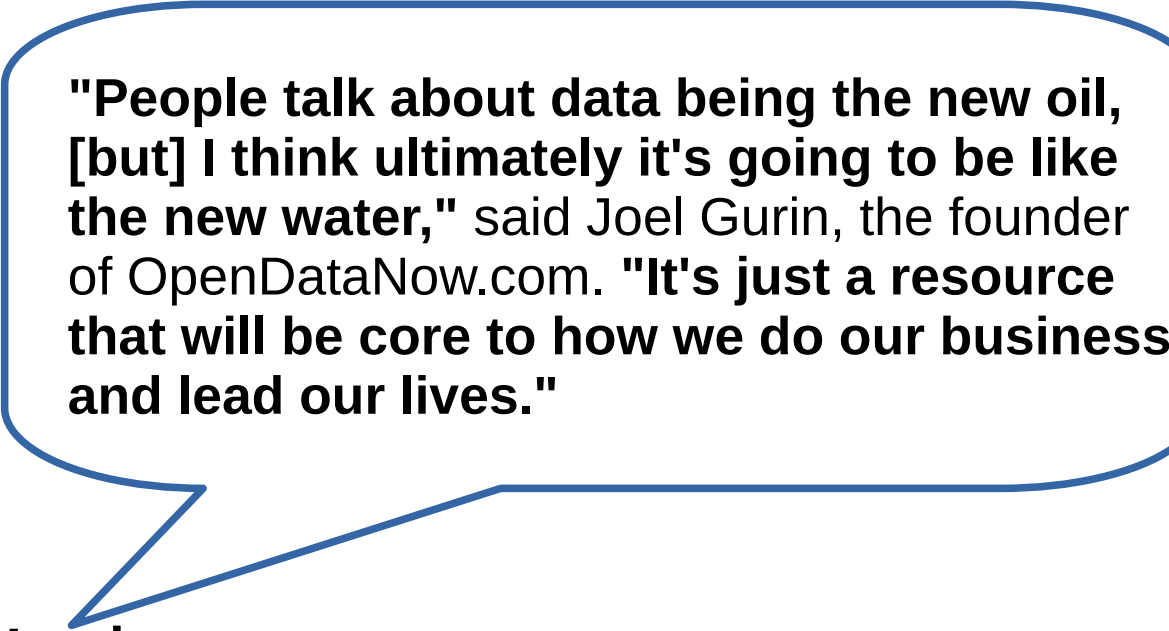


(Adam Jacobs, CACM Vol. 52 No. 8,
August 2009)

← 1999

Woher kommen diese Datenmengen?

- World Wide Web
- Mobilfunknetze
- Soziale Netzwerke
- Internet of Things
- Industrie 4.0



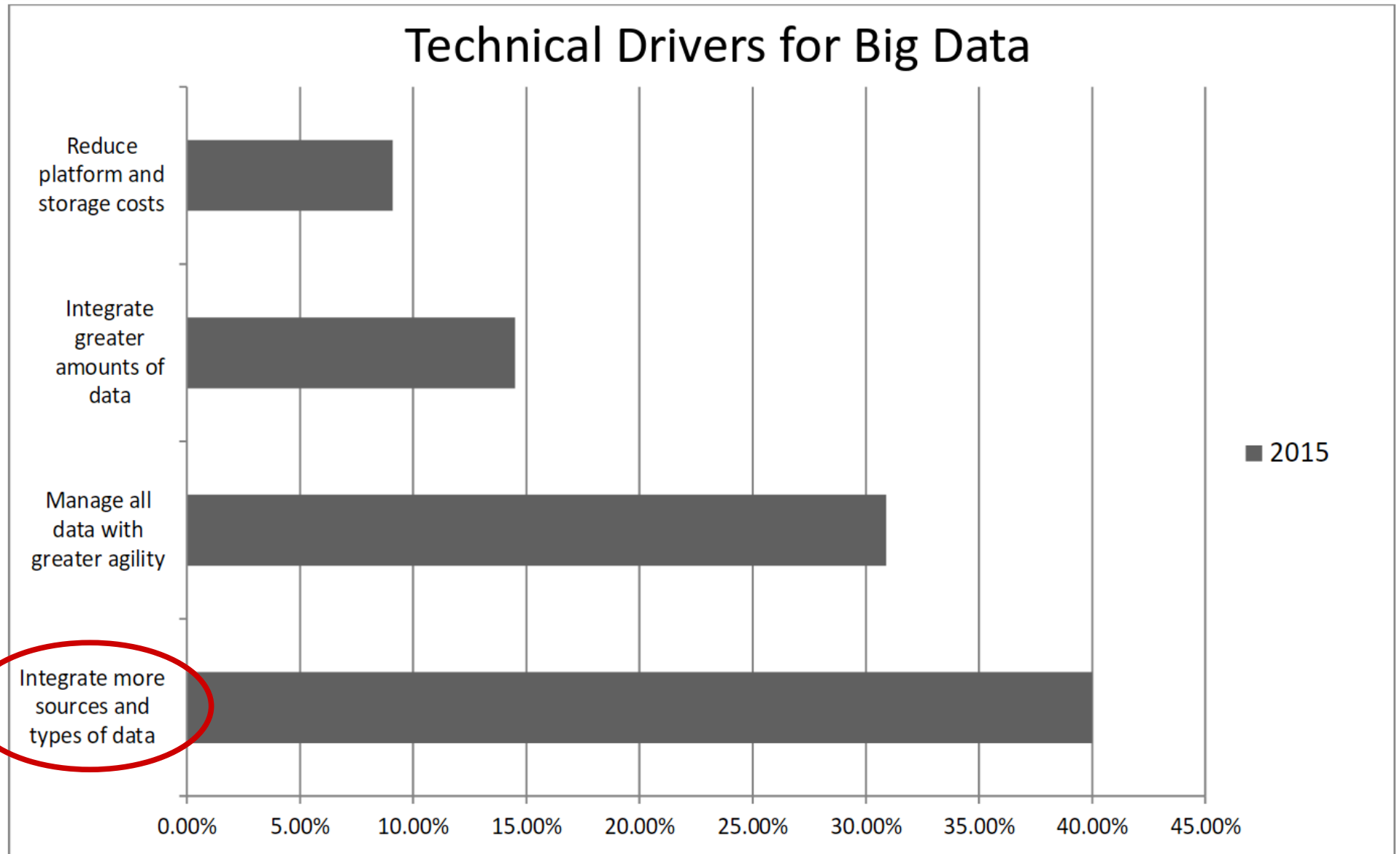
"People talk about data being the new oil, [but] I think ultimately it's going to be like the new water," said Joel Gurin, the founder of OpenDataNow.com. **"It's just a resource that will be core to how we do our business and lead our lives."**

- ... <Buzzword der Woche>...

Beispiele

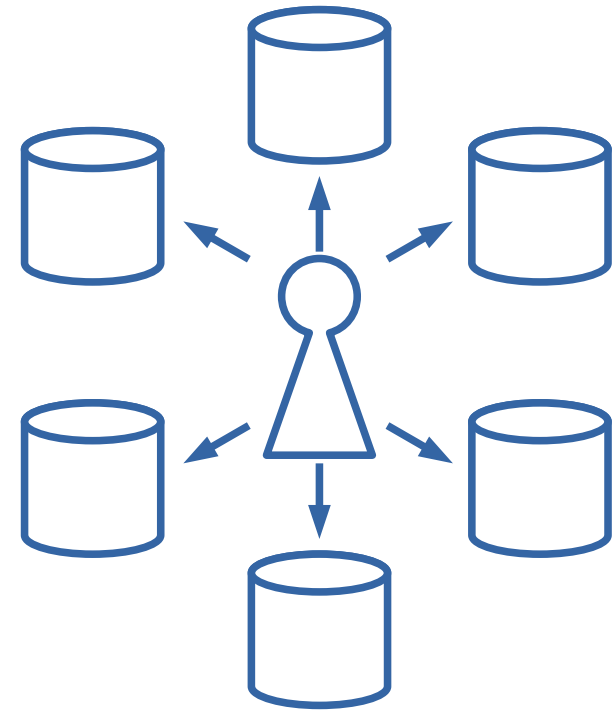
- **Velocity:** 40.000 Suchanfragen pro Sekunde, 1.2 Billionen pro Jahr bei Google
- **Volume:** 2.3 Mrd. aktive Benutzer, täglich 350 Mio. neue Fotos bei Facebook, 350 Mrd. Photos insgesamt
- **Volume:** 1 Mrd. aktive Benutzer, täglich 95 Mio. Posts bei Instagram
- **Volume:** ca. 1.5 Mio. Rechner bei Amazon

Was ist mit Variety?



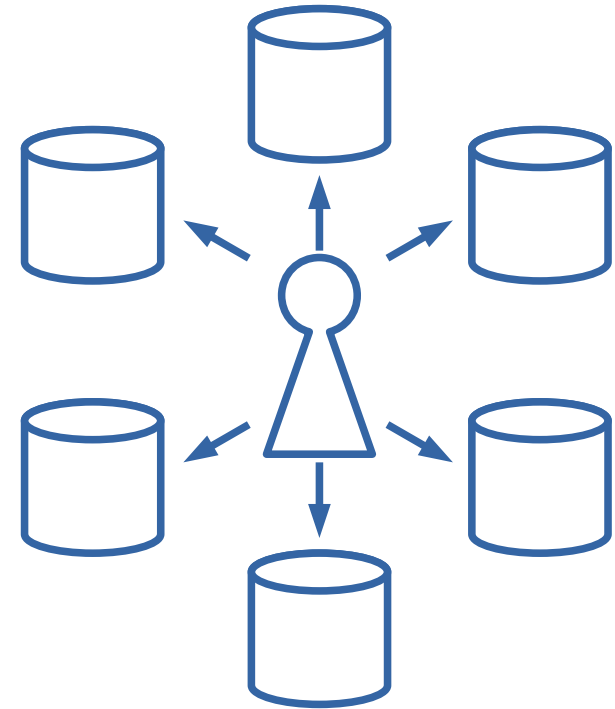
Variety

- **Informationsintegration**
 - Gewachsene Landschaften
 - Integration anderer Firmen(teile)
- **Stammdaten**
 - Name, Adresse, Vertragsdaten
- **Leistungsdaten**
 - Call Detail Records
 - Bestellungen
 - Abrechnungen
- **Customer Relationship Management (CRM)**
 - Kundenkontakte, Akquise, Hotline



Variety

- **Business Intelligence**
 - Analysen, Reports
- **Technische Überwachung**
 - Last, Ausfälle, Missbrauch
- **Zugekaufte Daten**
 - Marktforschung, Adresshandel



Anwendungsfälle

Anwendungsfälle

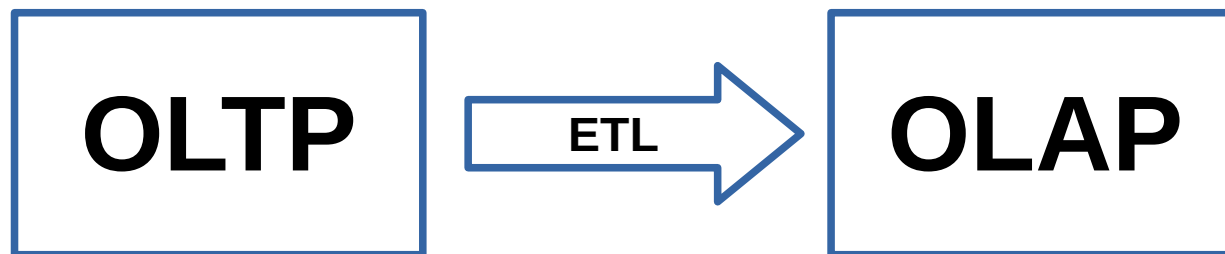
- **Entlasten/Ersetzen** bestehender Datenbanken
- Skalierbarkeit
- Durchsatz und Latenz
- Lizenzkosten

Anwendungsfälle

- **Transaktionsverarbeitung**
- **OLTP**: „On-Line Transaction Processing“
 - Suchen
 - Warenkörbe
 - Status-Updates

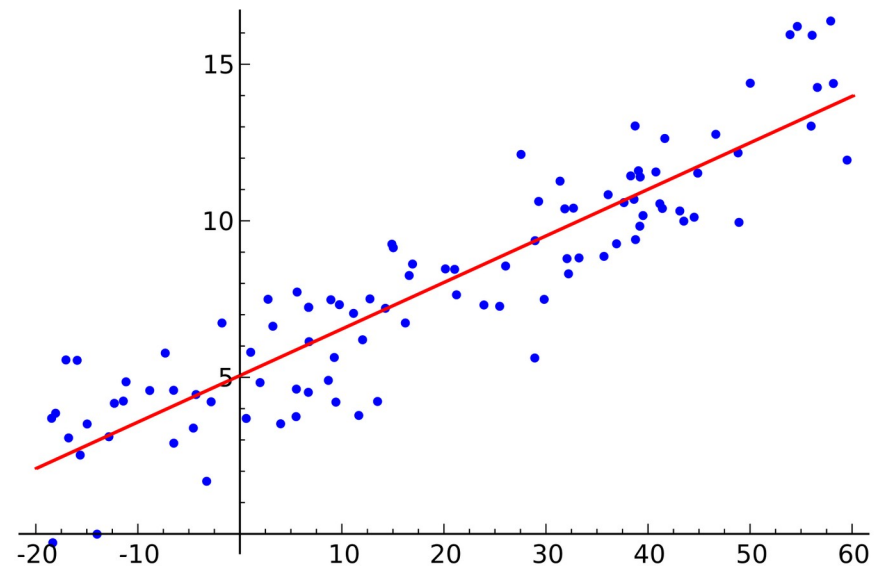
Anwendungsfälle

- Aufbereiten von Daten für **Business Intelligence** und Vorhersagen
- **OLAP**: „On-Line Analytical Processing“
- **ETL** – „Extract – Transform – Load“



Anwendungsfälle

- Interaktive **Analysen, Data Science**
 - Explorativer Umgang mit großen Datenmengen
 - Data Mining/Machine Learning
 - Generieren von neuen Erkenntnissen und Geschäftsideen



Anwendungsfälle

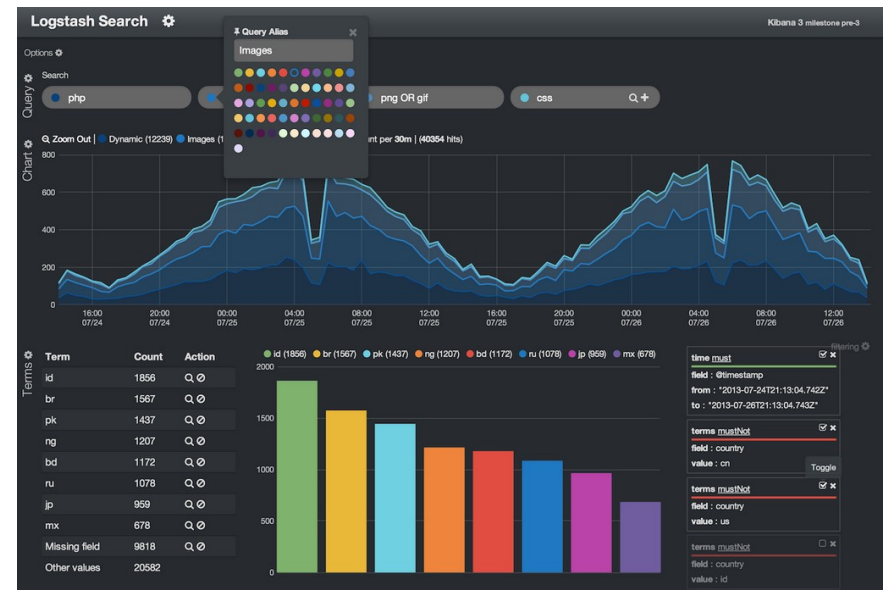
- **Recommender-Systeme**
 - Analyse von Benutzerverhalten
 - Aussprechen von Empfehlungen



Quelle: Netflix

Anwendungsfälle

- **Monitoring/Alerting**
 - Überwachung großer Systeme
 - Feststellen von Fehlerzuständen
 - Alarme

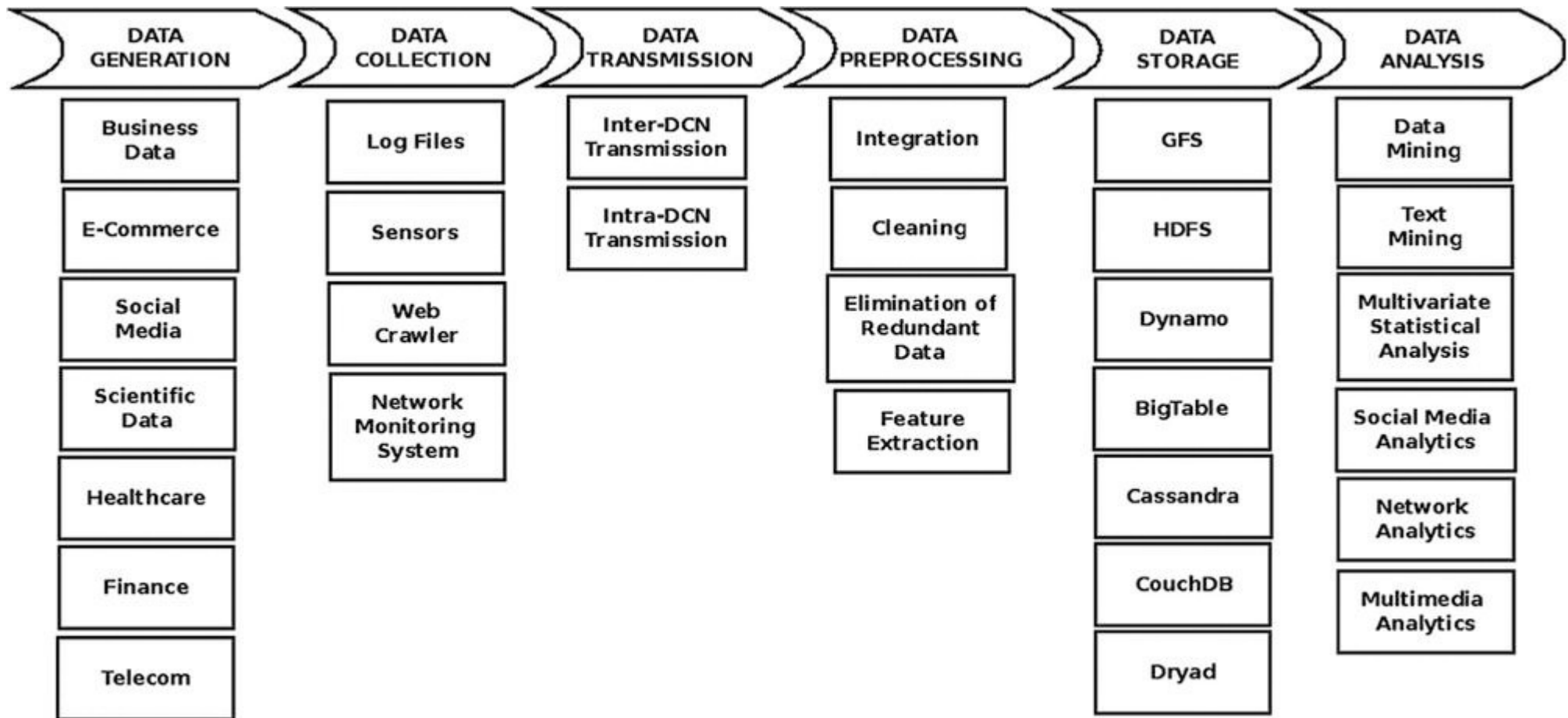


Quelle: elastic.co

Anwendungsfälle

- **Abuse**
 - personalisierte Spam-Filter
 - systematische Login-Versuche erkennen
- **Fraud Detection**
 - Kreditkartendaten prüfen
 - Zahlungswege vorgeben

Wertschöpfung durch Big Data



Herausforderungen

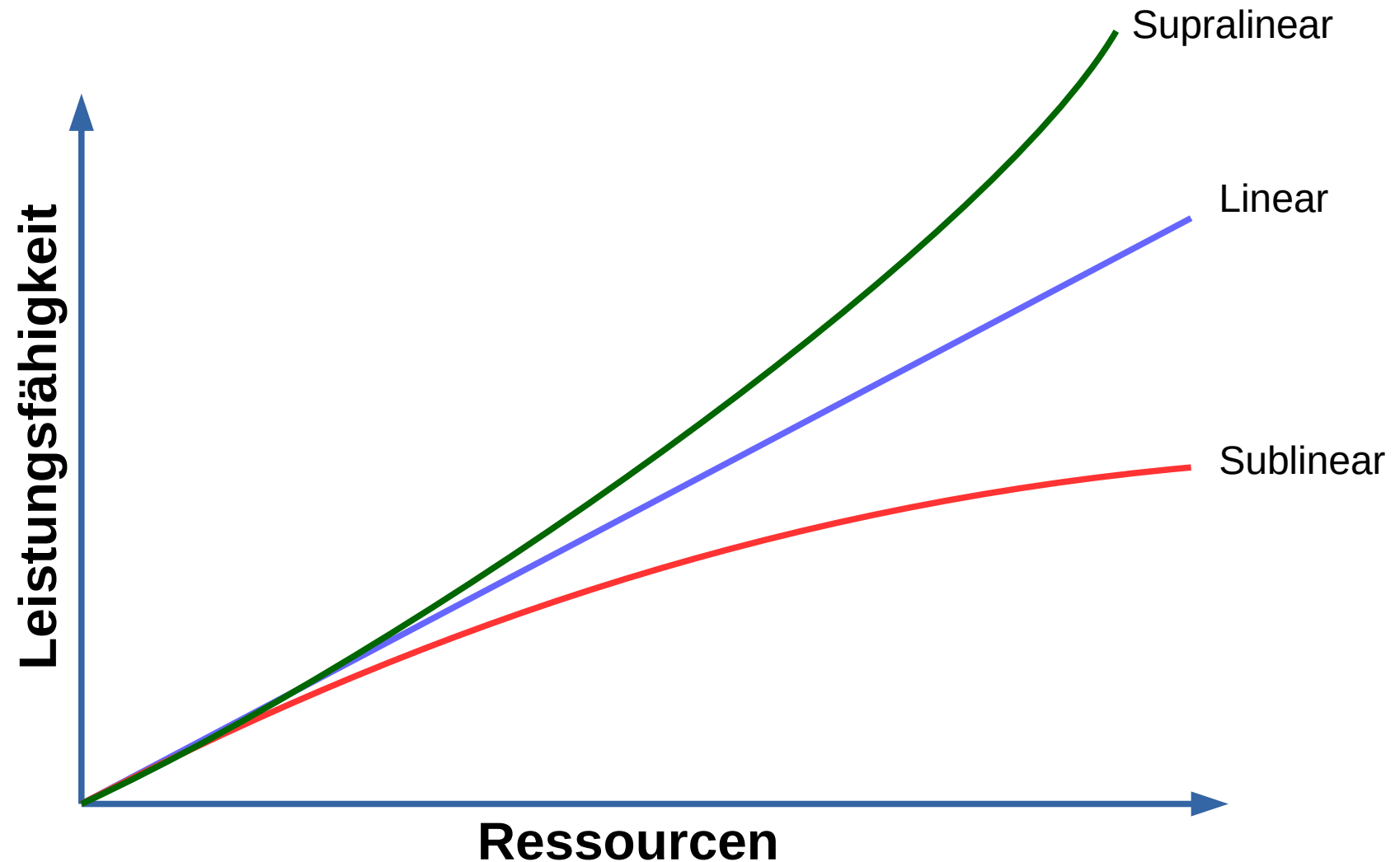
Herausforderungen

- Integration
- Skalierbarkeit
- Ausgewogenheit
- Verteilung
- Konsistenz
- Durchsatz
- Latenz

Herausforderungen: Integration

- Vielfalt gewachsener Systeme
- Unterschiede bei
 - Business-Verständnis
 - Datenmodellen
 - nonfunktionalen Anforderungen
 - Schlüsseln
- **Single Source of Truth?**

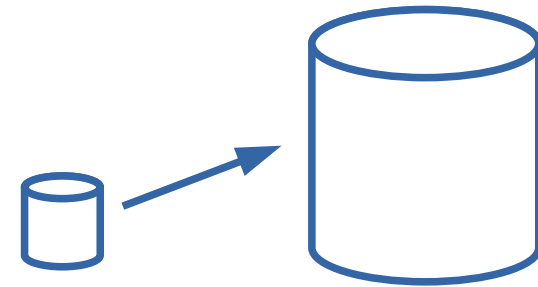
Herausforderungen: Skalierbarkeit



Horizontale und vertikale Skalierung

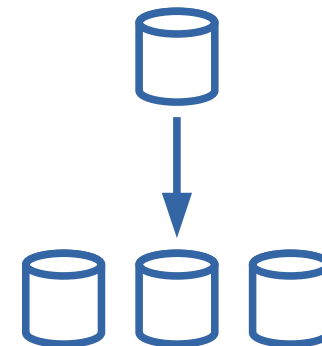
- **Vertikale Skalierung („scale up“)**

- schnellere CPU
- mehr Speicher
- schnellere/größere Festplatte
- ...

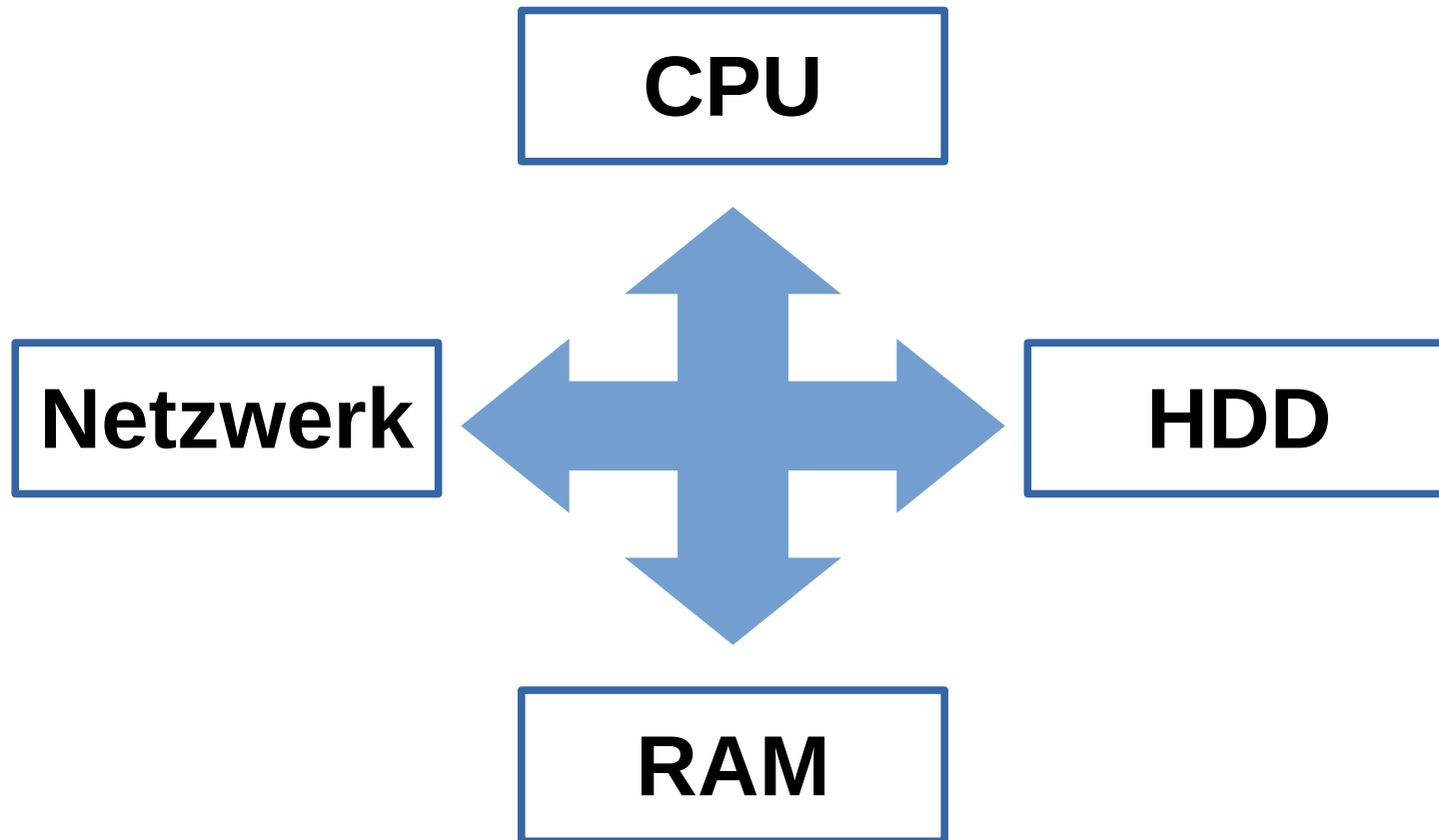


- **Horizontale Skalierung („scale out“)**

- mehrere Kerne
- mehrere Festplatten
- mehrere Rechner



Ausgewogenheit



Herausforderung: Verteilung

- **„My First Law of Distributed Object Design: Don't distribute your objects.“**

(Martin Fowler, PoEAA)

- Verteilung bringt Probleme mit sich:
 - Programmiermodell
 - Verteilter Zustand → Konsistenz
 - Synchronisation
 - Zeitbegriff
 - Ausfälle ↔ Robustheit

Herausforderungen: Durchsatz und Latenz

- Durchsatz/Bandbreite:

Anzahl / Zeiteinheit bzw.

Menge / Zeiteinheit

- Transaktionen/s
- Datendurchsatz MB/s

Herausforderungen: Durchsatz und Latenz

- Latenz:

Zeit zwischen Auslöser und Reaktion

- Antwortzeit eines Datenbanksystems
 - Reaktionszeit auf Ping
- Durchsatz und Latenz sind oft widerstrebende Optimierungsziele!

Durchsatz und Latenz: Sneakernet

"Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway."



Durchsatz und Latenz

- MicroSD-Karte:
 - 0.5 g, 165 mm³, 512 GB
- Container:
 - 33 m³, 21 t Nutzlast, 2.4 t
- Containerschiff OOCL Hong Kong:
 - 21.413 Container, 197.000 t Nutzlast
- Latenz Hamburg – New York:
 - 6 Tage

4 Exabit/s

- **Durchsatz in Bytes/s?**



Quelle: Wikipedia

Soll das ein Witz sein?!

AWS Snowball: Accelerating Large-Scale Data Ingest Into the AWS Cloud | AWS Public Sector Summit 2016

AWS Snowball—data transport service for large datasets



Bilder: amazon.com

Durchsatz und Latenz

- Beispiel: 10-Gb-Ethernet
- Query-Response:
 - Roundtrip: **ca. 50 μ s, 20.000/s**
 - MTU: 1.500 Bytes
 - $20.000/\text{s} * 1.500 \text{ Bytes} = \mathbf{30 \text{ MB/s}}$
- Batch:
 - ca. 10 Gb/s = **1 GB/s**
 - ca. **600.000 Nachrichten/s** bei Nachrichtengröße 1.500 Bytes
 - Latenz: **Sekunden**

Exkurs: Überschlagsrechnungen

- Wie schnell ist...?
- Wie lang dauert...?
- Wie groß ist...?
- Wie zuverlässig ist...?
- **Geht das?**
- **Was kostet es?**

Überschlagsrechnungen: Beispiele

- **Hardware**

- HDD: 200 MB/s, ~10 ms Seek, 130 IOps/s
- HDD Annual Failure Rate: z. B. 0.7%
- SSD: 500 MB/s, ~0.05 ms Seek, 1000–10000 IOps/s

- **Kommandozeile**

- AWK durchsucht 100 Mio. Zeilen in 20 s
- grep durchsucht 100 Mio. Zeilen in 5 s
- GNU sort 100 Mio. Zeilen in 220 s

Überschlagsrechnungen: Beispiele

- **Netzwerk**

- Lichtgeschwindigkeit: 300.000 km/s
- Ethernet-Roundtrip (10Gb): $\sim 50 \mu\text{s}$

- **Kosten**

- Betrieb eines Servers: 500 €/Monat

Überschlagsrechnungen: Fragestellungen

- Wie viel Hardware brauchen wir, um Logfiles von 5.000 Servern zu analysieren?
- Was würde es kosten, 500 Mio. Events pro Tag zusätzlich zu verarbeiten?
- Sollen wir Server mit 1 HE und 8 Kernen kaufen, oder die mit 2 HE und 20 Kernen?

Organisatorische und rechtliche Herausforderungen

- **Datenschutz**

- Welche Daten sammeln?
- Welche Daten verknüpfen?
- Wie lange aufbewahren?
- Pseudonymisieren, Anonymisieren

- **Datensicherheit**

- Authentifizierung/Autorisierung
- Verschlüsselung

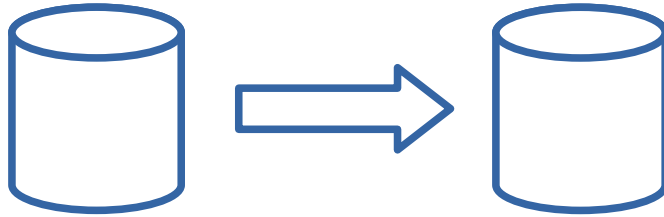
Organisatorische und rechtliche Herausforderungen

- Mangel an **Fachkräften**
 - Anwendungsentwicklung
 - Betriebserfahrung in der IT
 - Data Mining und Machine Learning
 - Data Science
- Mangelndes **Verständnis** in der Organisation
 - „Big-Data-Abteilung“?

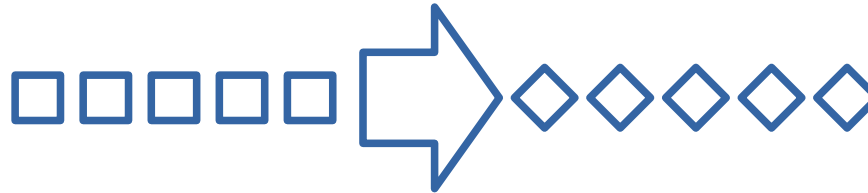
Verarbeitungsmodelle

Verarbeitungsmodelle

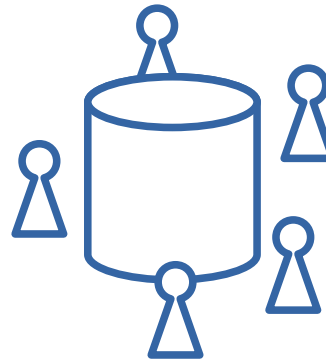
- Batch



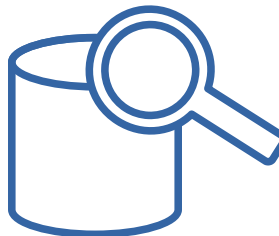
- Stream



- Transaktionen (OLTP)

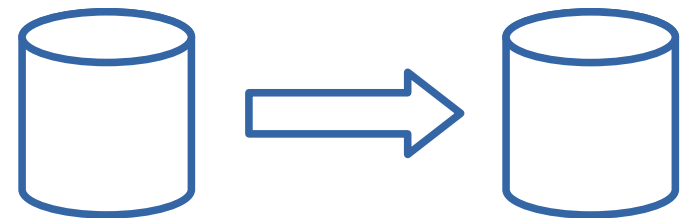


- Analysen (OLAP)



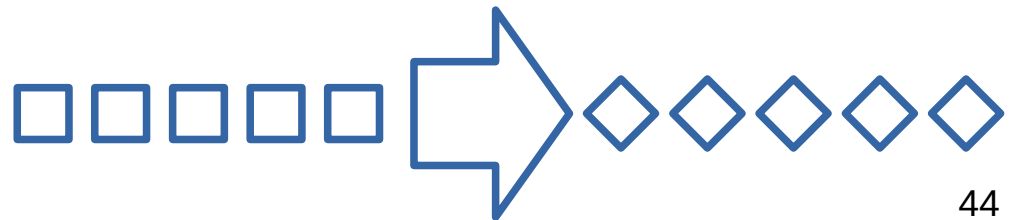
Verarbeitungsmodell: Batch

- Verarbeitung großer Datenmengen in einem Durchgang
- Optimiert auf Durchsatz
- Latenz: Minuten, Stunden, Tage
- Durchsatz: Gigabytes/s
- Beispiele
 - Abrechnungsläufe
 - Vorberechnung von Aggregaten
 - Aufwendige Analysen



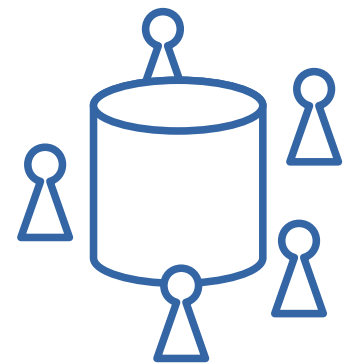
Verarbeitungsmodell: Stream

- Verarbeitung von Einzelereignissen in (weicher) Echtzeit
- Latenz: Millisekunden, Sekunden
- Durchsatz: Tausende/Sekunde
- Beispiele
 - Autovervollständigung bei Suchen
 - Retargeting
 - Monitoring
 - Erkennen von Anomalien



Verarbeitungsmodell: Transaktionen

- Viele gleichzeitige Lese- und Schreiboperationen auf persistenten Daten
- Durchsatz: Tausende/s
- Latenz: Millisekunden
- Beispiele
 - Warenkörbe
 - Buchungen
 - Geschäftssysteme

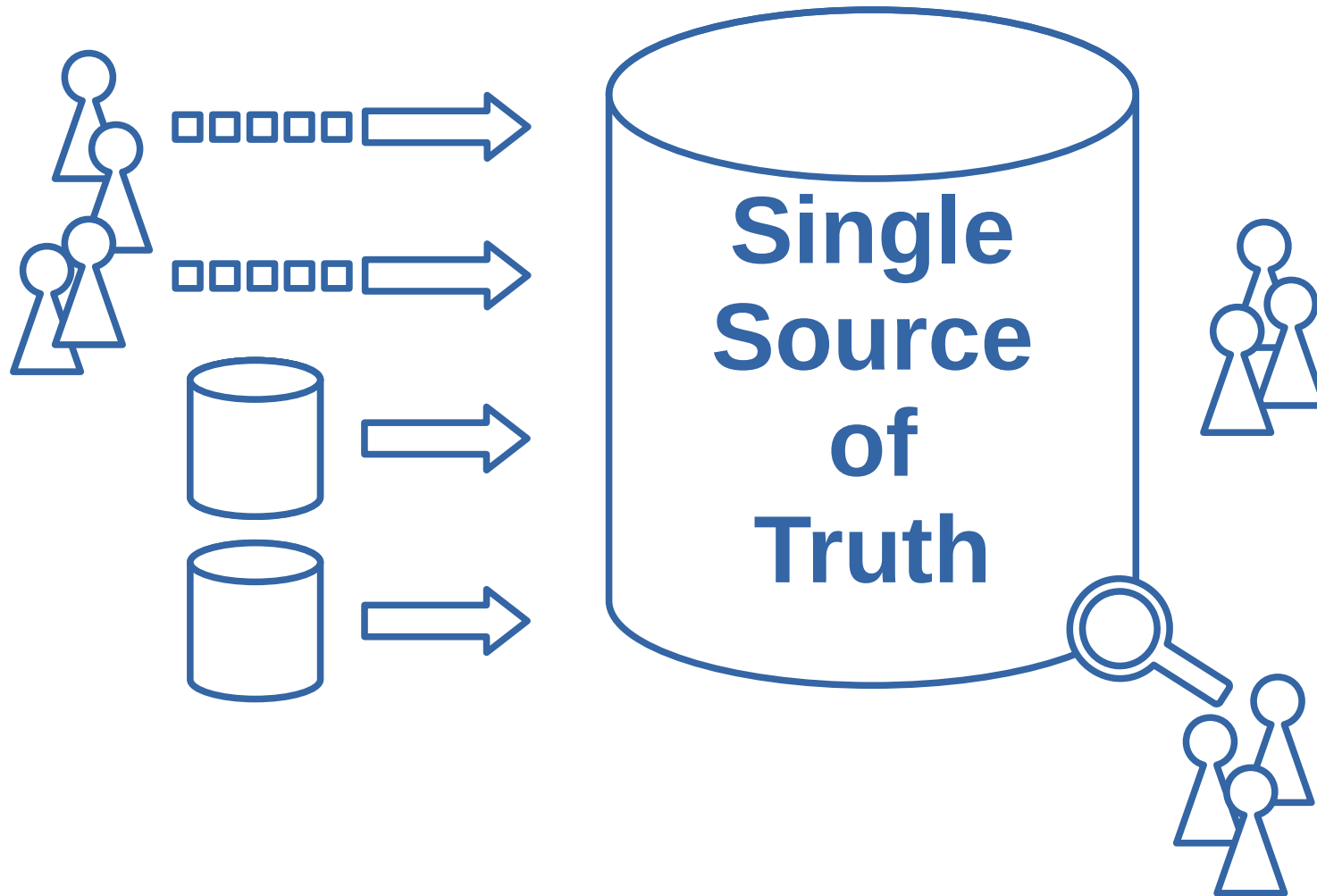


Verarbeitungsmodell: OLAP

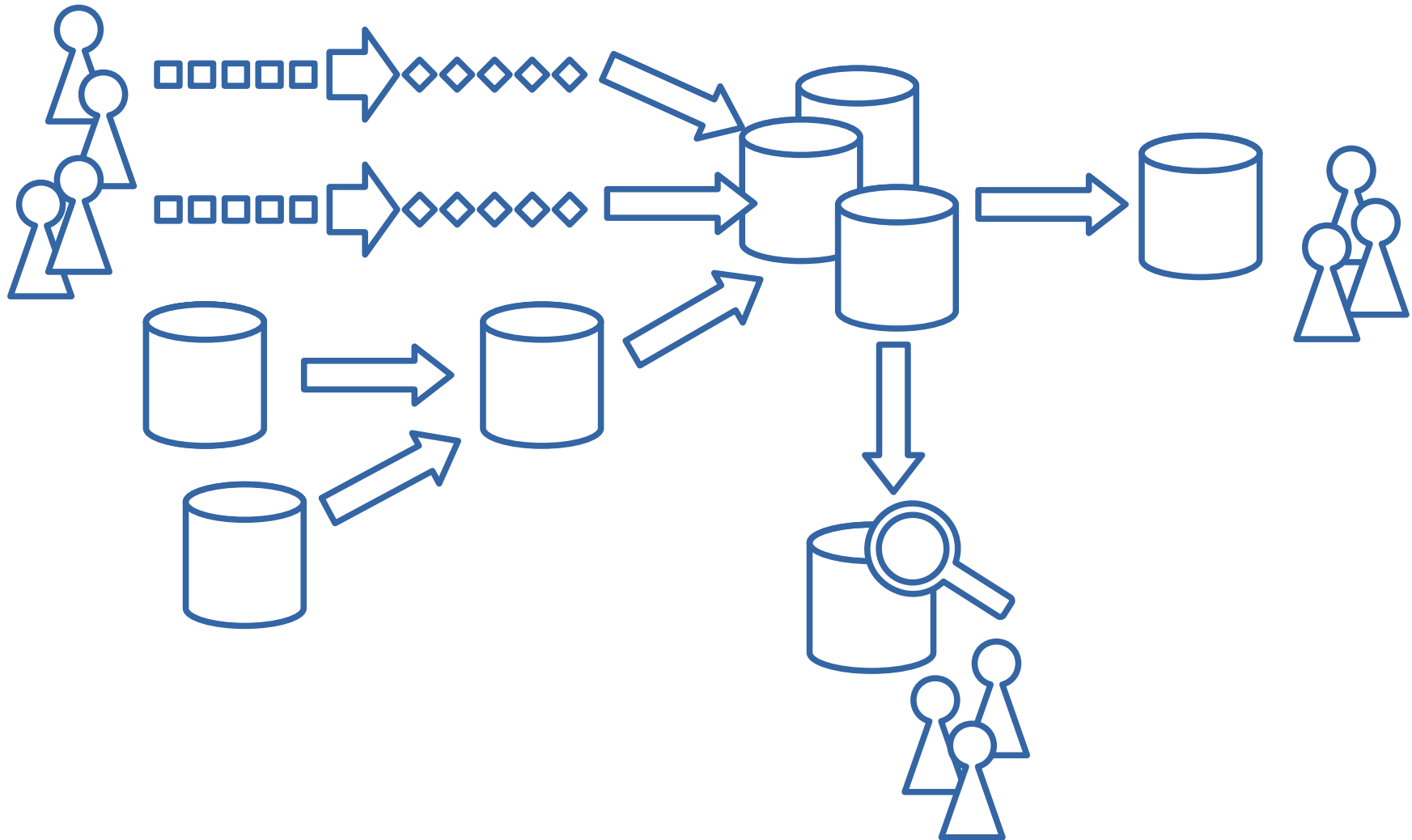
- Berechnung von Aggregaten auf historisierten Daten
- Aufbereitete Datenbestände
- Interaktive Nutzung oder periodische Berichte
- Latenz: Sekunden, Minuten
- Beispiele
 - Reporting
 - Business Intelligence
 - Data Science



Kombination: Big-Data-Architektur



Kombination: Big-Data-Architektur



Zusammenfassung

- Definitionen von Big Data
- Anwendungsfälle
- Herausforderungen
- Verarbeitungsmodelle

Ausblick

- Batch-Verarbeitung
- NoSQL-Datenbanken
- Stream-Verarbeitung
- Big-Data-Architekturen
- Näherungsverfahren