DIHSÜD

FH JOANNEUM
University of Applied Sciences

Digital Innovation Hub Süd

Big Data Technologies

Wilhelm Zugaj, Joachim Schauer (IIT/FH JOANNEUM)

FFG
Forschung wirkt.

Das Land Steiermark

LAND KÄRNTEN

Agenda

- Einführung in Big Data & Distributed Storage (Hadoop)	40 Minuten
- Fragen & Pause	15 Minuten
- Verteilte Datenbanksysteme (NoSQL)	40 Minuten
- Fragen & Pause	15 Minuten
- Data Analytics & Machine Learning	40 Minuten
- Diskussion & Fragen	30 Minuten

Big Data: Einführung

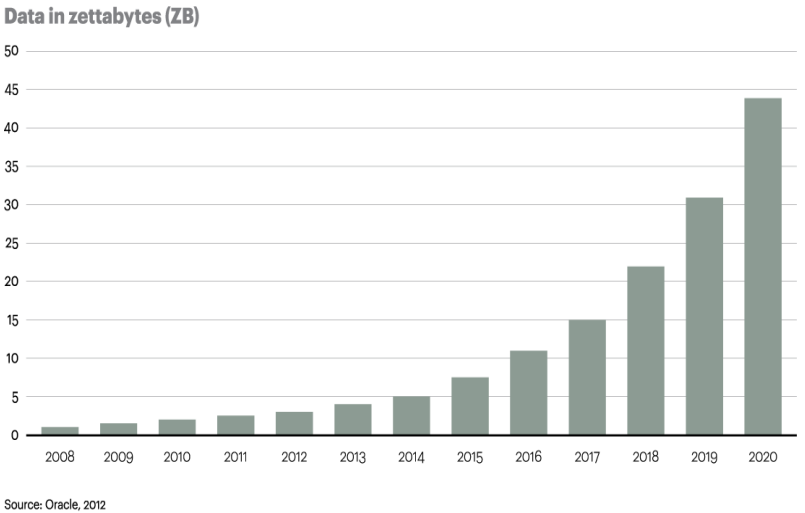
Globales Datenwachstum

Typische Treiber:

- Cloud
- IOT
- Sensoren
- Smart meters
- Social Networks
- Online Shopping
- Mobile

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020

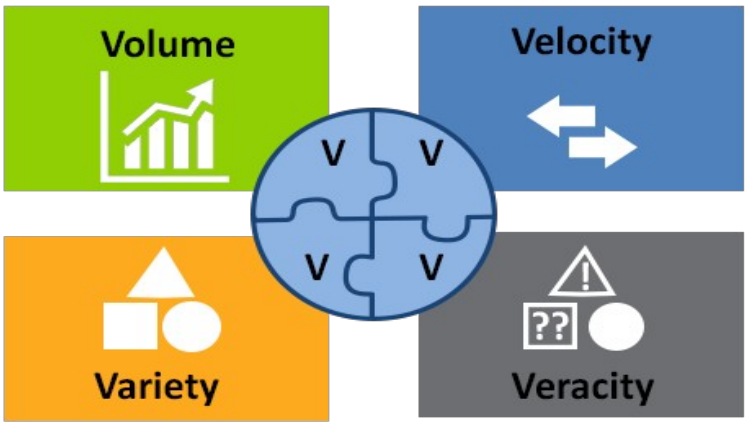


Source: Oracle, 2012

http://www.atkearney.com/strategic-it/ideas-insights/article/-/asset_publisher/LCcgOeS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192

Definition Big Data

- Es gibt keine einheitliche Definition
- Wird beschrieben durch die 3-9 V's:
  - Volume
  - Velocity
  - Variability



http://www.zarantech.com/blog/the-4-vs-of-big-data/

- Validity
- Value
- Variety
- Velocity
- Veracity
- Viability
- Visibility
- Volatility
- Volume

Definitionsversuche

SAS

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

NIST

Big data is data which "exceed(s) the capacity or capability of current or conventional methods and systems."

Ward/Barker University of St. Andrews

"Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning."

Big Data ist ein Paradigmenwechsel: Technologien & Methoden

- Storage Systems
  - alt: Centralized local Storage & Relational Database Systems
  - neu: Distributed remote Storage & NoSQL Database Systems
- Processing
  - alt: Single Transaction Processing
  - neu: Parallel Data Processing
- Analytics
  - alt: Statistische Datenanalyse
  - neu: Machine Learning Algorithms

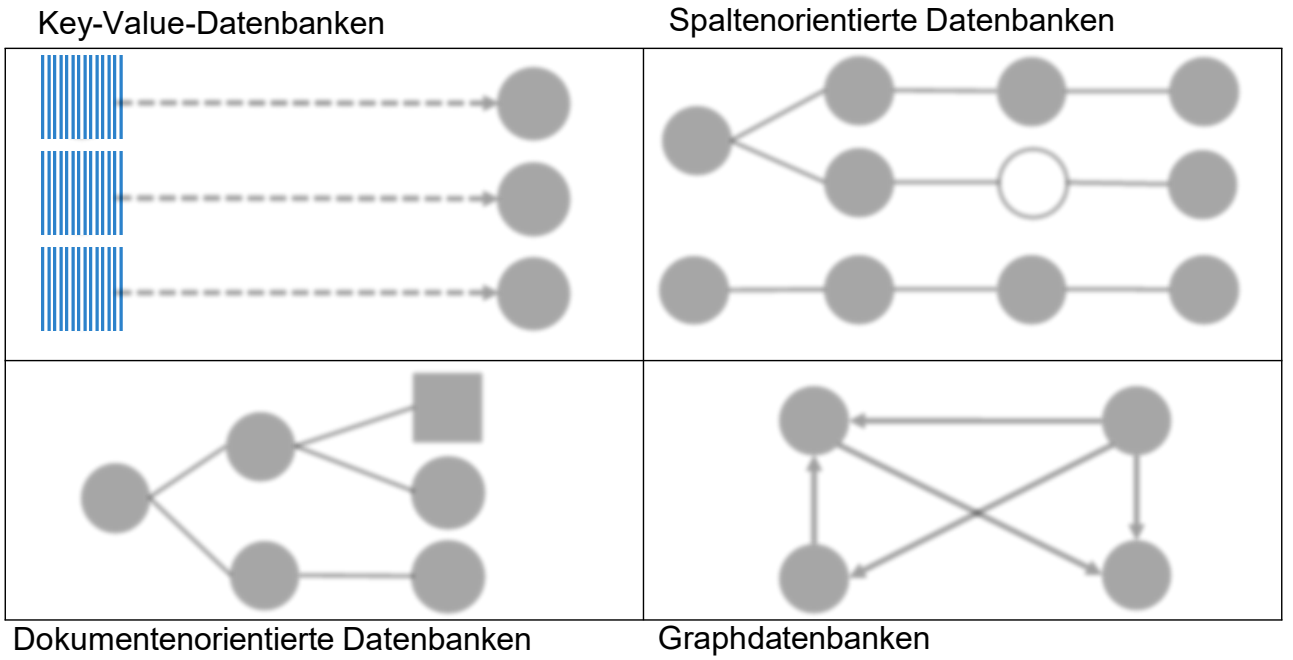
Data Warehouse vs. Data Lake

Data Warehouse	Data Lake
Kontrolle	Flexibilität
Persistieren bereits transformierter Daten	Daten werden unverändert persistiert und beim Auslesen transformiert
Strukturiert	Unstrukturiert
Hoher Kostenaufwand bei großer Datenmenge	Konzipiert für günstige Hardware (horizontale Skalierbarkeit)
Eingeschränkte Flexibilität der Konfiguration	Konfiguration ist flexibel

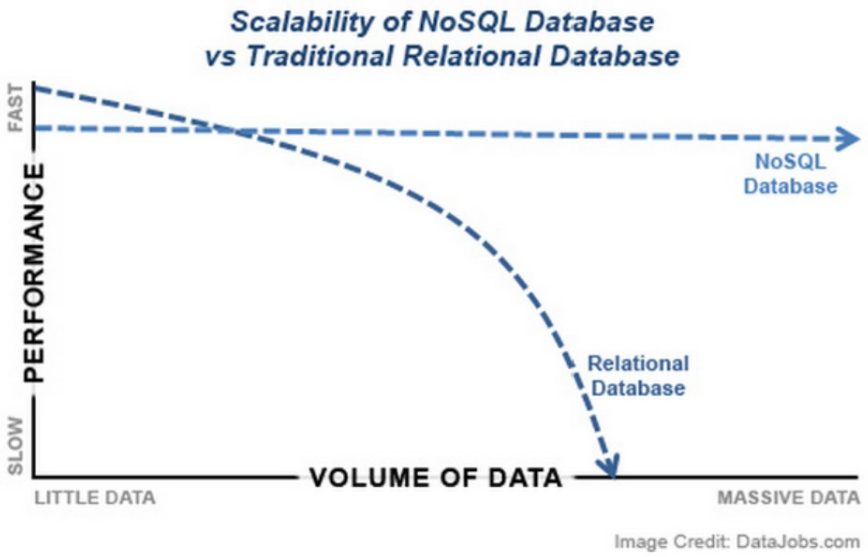
SQL vs NoSQL

SQL		NoSQL
Relational	**Schema**	Keines/Flexibel
Strukturiert	**Daten**	Unstrukturiert Semi-strukturiert Strukturiert
Vertikal	**Skalierbarkeit**	Horizontal
Fixes Schema	**Flexibilität**	Flexibles Schema
SQL	**Abfrage**	Nicht standardisiert, Parallel (Map-Reduce)
ACID	**Modell**	Eventual Consistency/CAP Theorem

Hauptkategorien von NoSQL Datenbanksystemen



Scalability: Scale Out vs. Scale Up



Stapelverarbeitung

- Daten ersistiert
- Keine Echtzeitanforderung
  - Hohe Latenz
- Beispiel:
  Kundenanalyse

Datenstromverarbeitung

- Ereignisse in Echtzeit
  - Zeitreihen
- Echtzeitanforderung
  - Geringe Latenz
- Beispiel:
  Sensordaten einer Produktionsanlage

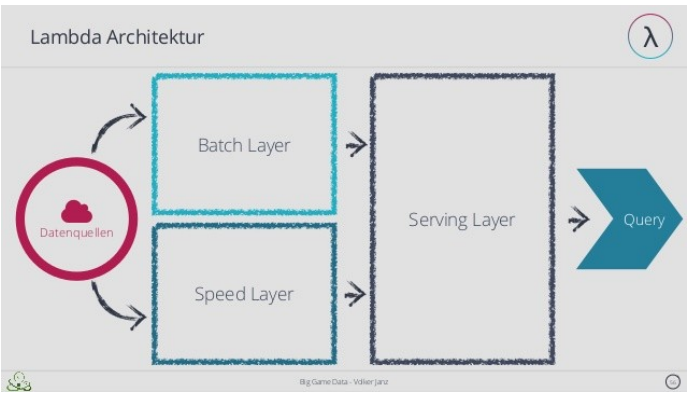
http://www.bfe-inf.org/sites/default/files/data-curve-stream-for-earthcube-01_0.png

Lambda Architektur

Kombiniert Vorteile von

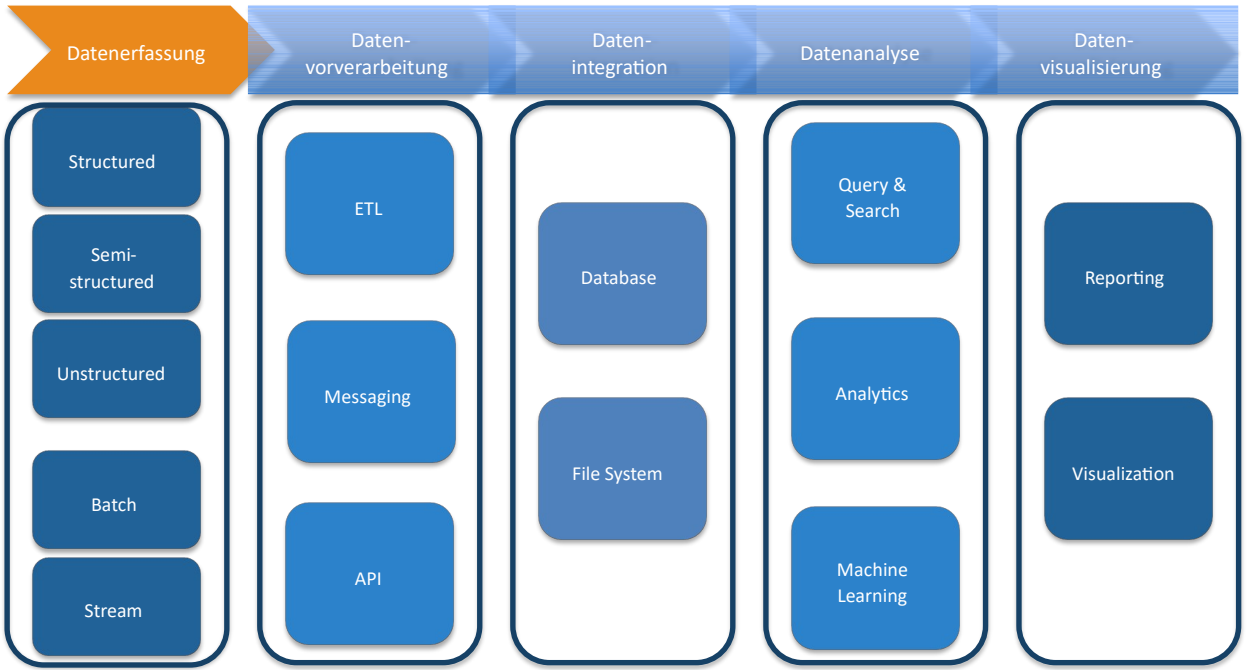
Batch & Stream Processing

- Balance zwischen
  - Latenz
  - Durchsatz
  - Fehlertoleranz



https://de.slideshare.net/VolkerJanz/big-game-data-event-tracking-mit-storm-kestrel-und-der-lambda-architektur

Big Data Analytics



Anwendungsgebiete von Big Data

1. Konsumentenverhalten verstehen und vorhersagen

- Ziel: 360° Sicht auf Kunden --> Zielgerichtete Werbung schalten
- Kunden und deren Verhalten besser verstehen und vorhersagen
- Ziel: ein möglichst vollständiges Profil der Kunden
- Risiko: Persönlichkeitsschutz

Beispiel: Target

- Jeder Target Kunde bekommt eine ID
  - Alle Einkäufe werden zum Kunden verlinkt
- Einkaufsmuster von Kunden werden analysiert
- Produkte identifiziert, die Schwangere vermehrt vor der Geburt kaufen
- Bier - Windel Korrelation
- Ziel: Gezieltes senden von Werbung, passend zum Kunden.

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#72a83ze06668

NETFLIX

- Big Data Analyse:
  - Welche Inhalte wollen Kunden sehen?
  - Wer sieht welche Inhalte und wie präsent sind diese in Sozialen Medien?

- Ziel: Mit hoher Gewissheit Kundenwünsche vorhersagen

https://blog.kissmetrics.com/how-netflix-uses-analytics/

2. Geschäftsprozesse verstehen und optimieren

- Warenlager anpassen, basierend auf:
  - Daten aus Sozialen Medien
  - Web search trends
  - Wettervorhersagen

Beispiel: Sendungsverfolgung

- "Delivery Route Optimisation"
  - Analyse von Echtzeit-Verkehrsdaten, Wetterdaten, GPS Position, …
  - Ziel: Lieferungen möglichst schnell zustellen

3. Quantified Self

- Big Data für den persönlichen Nutzen
- Daten werden von tragbaren Geräten erzeugt
  - Smartwatch / Wearables
  - Kalorienverbrauch, Aktivität, Schlafmuster

- Most online dating sites apply big data tools and algorithms to find us the most appropriate matches

4. Gesundheitsversorgung

- Gesammelte Daten von Smartphones und Wearables bereitstellen

- Ziel: (Medizinische) Daten nutzen um
  - Epidemien vorherzusagen
  - Krankheiten zu heilen
  - Lebensqualität erhöhen

Beispiel: Google

- Von 2008 – 2014: "Google Flu Trends"
- Datenerhebung:
  - Abgleich von Suchbegriffen mit realen Krankheitsdaten
  - Identifizieren von Suchbegriffen, die mit Grippe korrelieren
- Ziel: Vorhersagen von Epidemien
- Problem: Vorhersagen zu oft unpräzise

5. Performance-Optimierung

- Idee: Maschienen und Geräte "smarter" bzw. autonom machen
- Beispiel: Einsatz von Kameras, GPS und Sensoren für das selbstfahrende Auto

https://ecomento.de/wp-content/uploads/2014/02/Rinspeed-Tesla-Model-S-Cockpit-Lenkrad-Innen-740x425.jpg

Distributed Storage
Beispiel: Hadoop

Technologietreiber: Cloud-Systeme

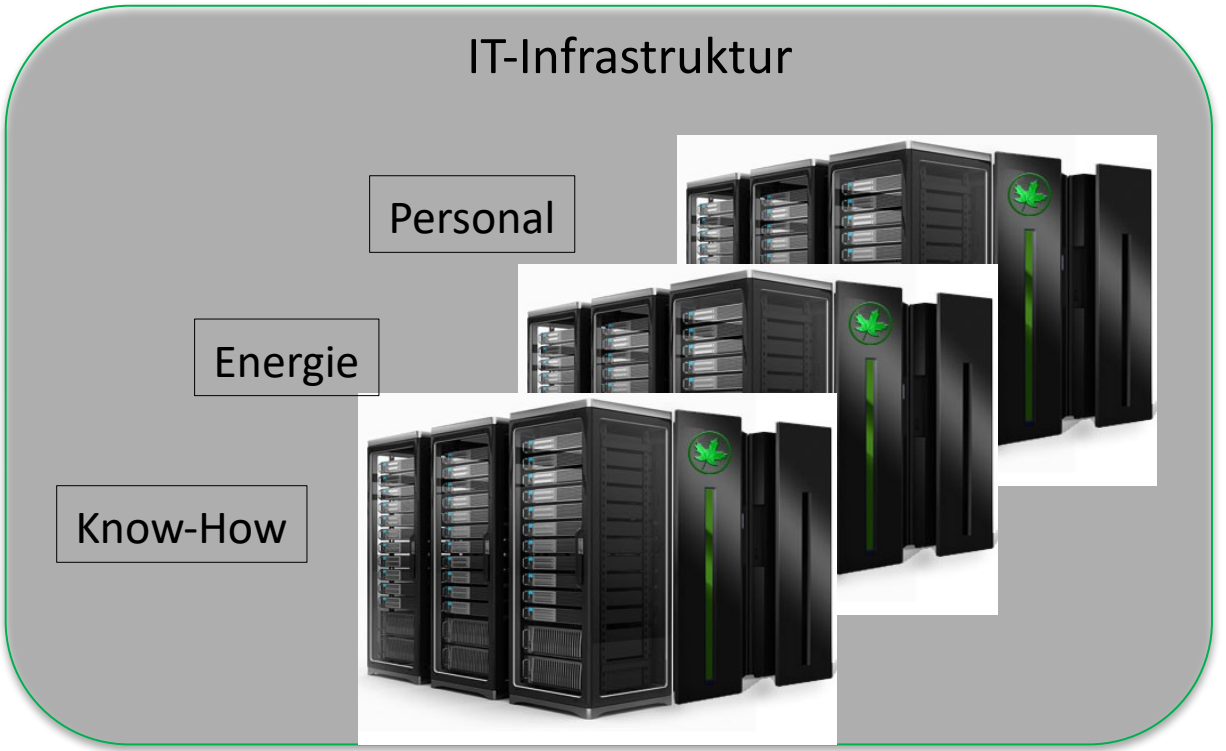
ERP

CRM

BA/BI

Office

Entwicklunginfrastr.



Grundidee Cloud

IT-Infrastruktur

ERP

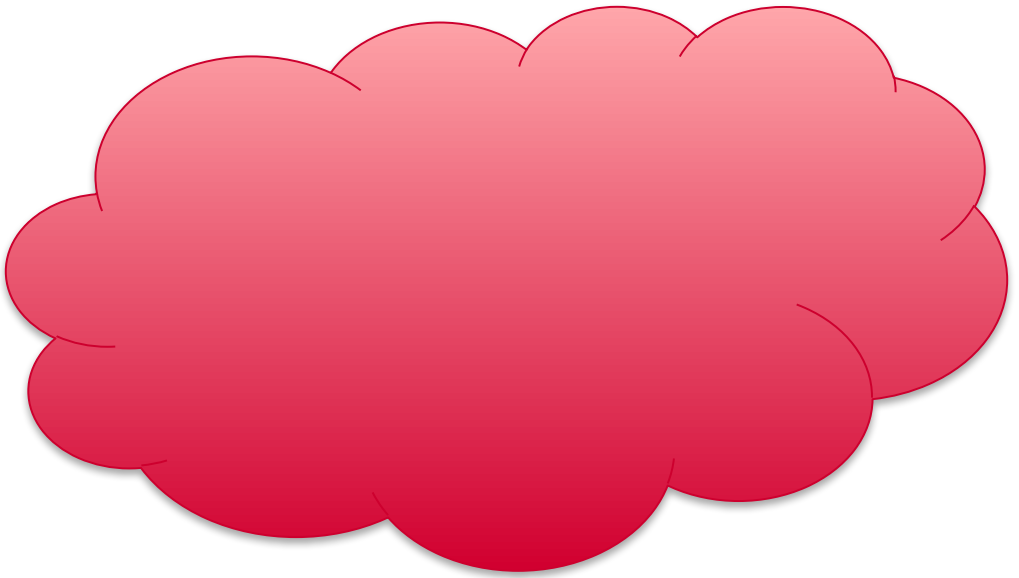
CRM

BA/BI

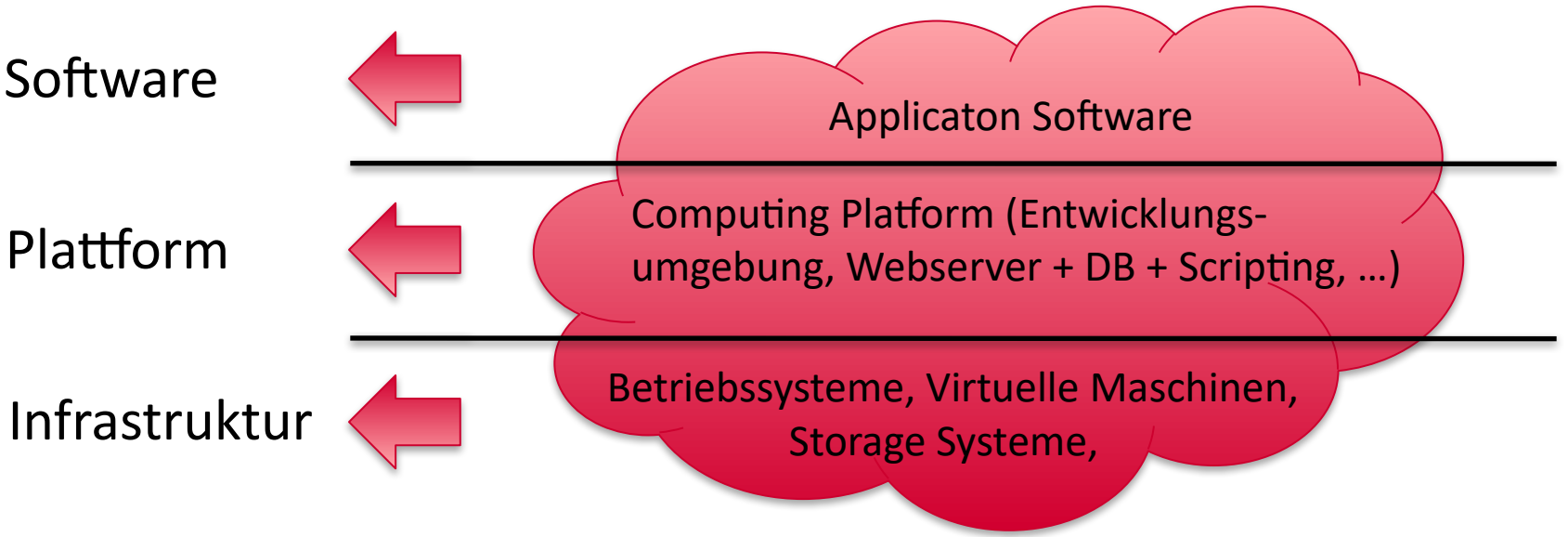
Office

Entwicklunginfrastr.

Infrastruktur extern



Cloud Service Models




hadoop
HDFS

HDFS (Hadoop File System): Idee

1990

1,4 GB

4,4 MB/s

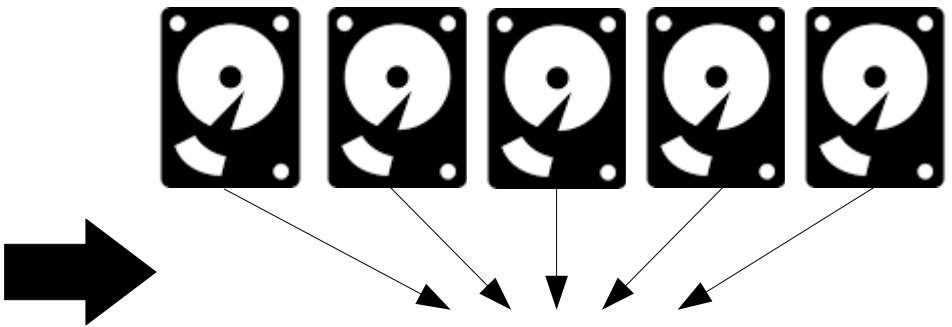
5 Min/Disk

2015

1 TB

100 MB/s

2,5 Std/Disk

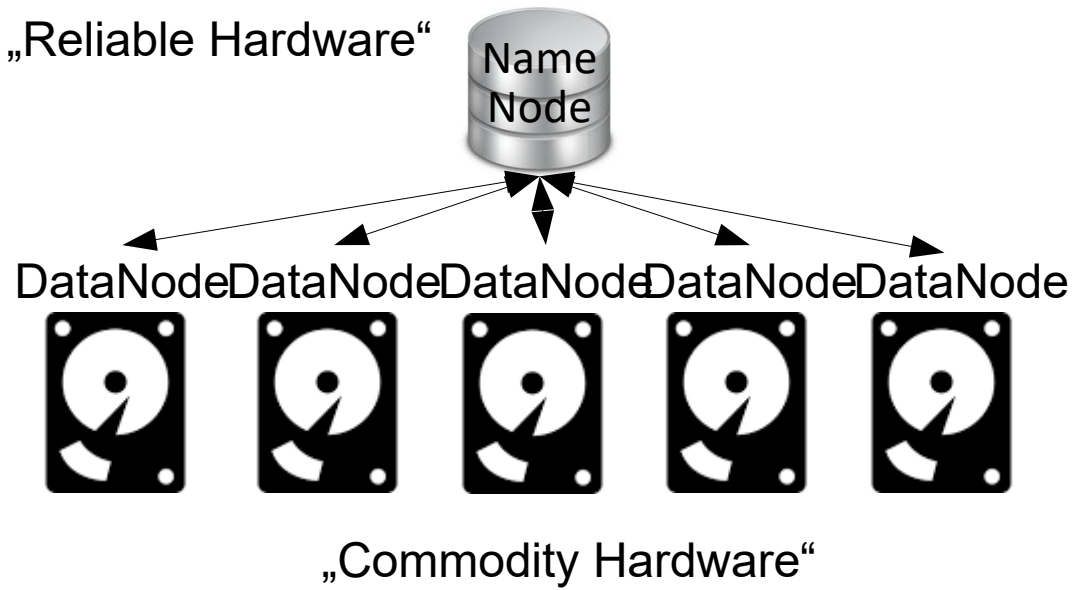


Lösung: Simultanes Lesen von vielen Disks

~ 30 Disks -> 1 TB in 5 Min

Problem: Ausfallgefahr bei hoher Diskzahl

HDFS - Grobe Architektur



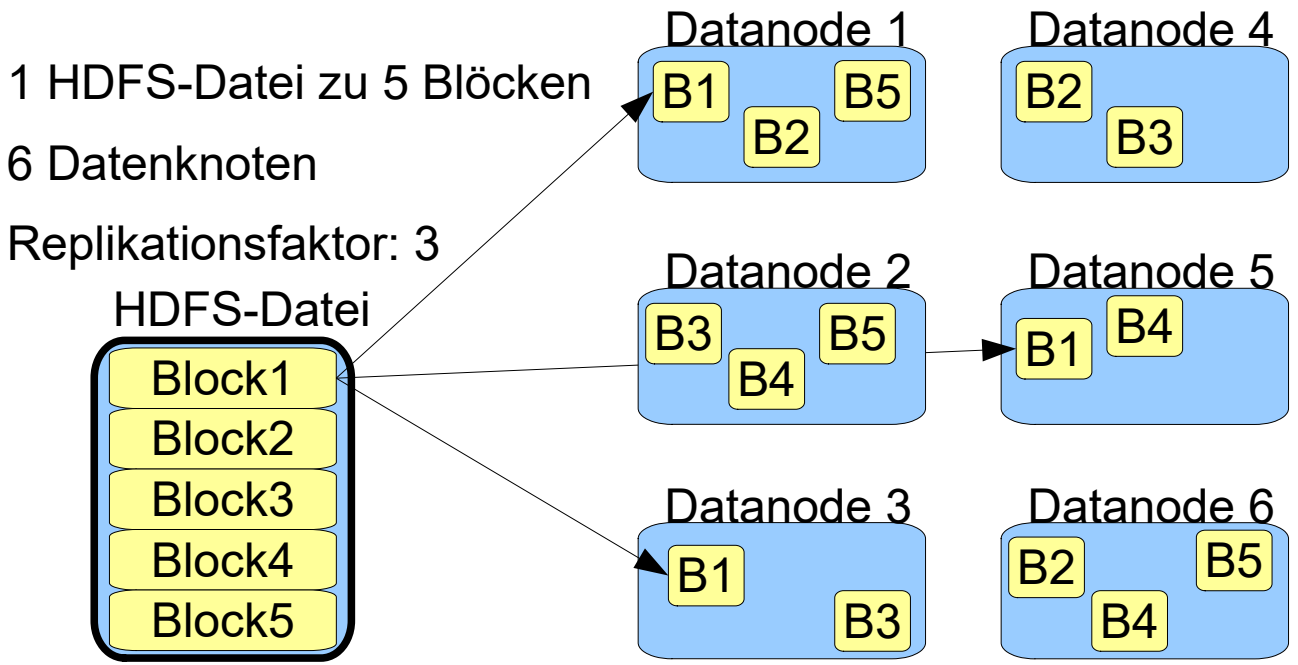
NameNode:

- Regelt Zugang zu DataNode
- Überwacht und Verwaltet DataNodes
- Öffnen, schließen, umbenennen von Dateien und Verzeichnissen

DataNode:

- Führt Lese/Schreibbefehle durch

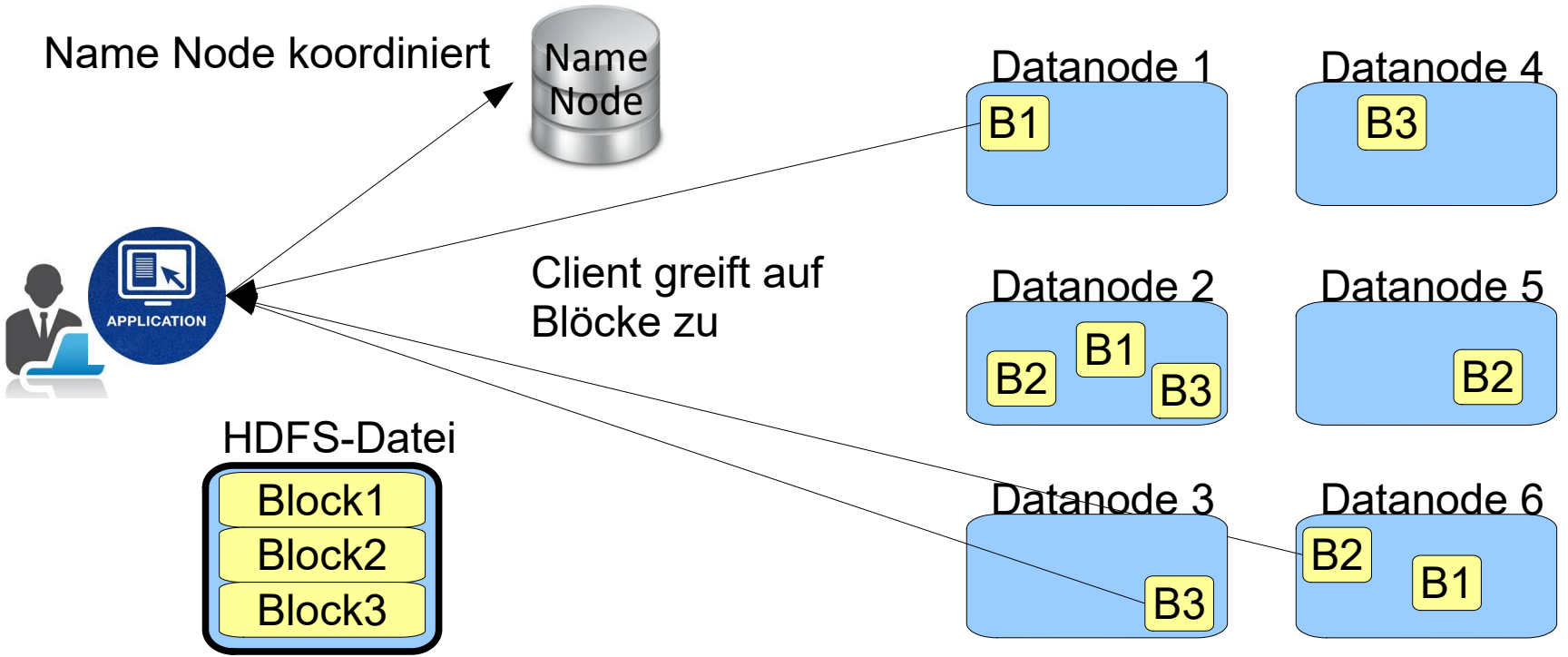
HDFS - Details: Blöcke - Sharding - Replikation



Zusammenfassung: Blöcke - Sharding - Replikation

- HDFS Dateien bestehen aus mehreren Blöcken.
- Standardgröße der Blöcke ist 128 MB.
- Die Blöcke einer HDFS-Datei werden auf unterschiedlichen Datenknoten gespeichert
- Die Blöcke werden zusätzlich auf weitere Knoten repliziert
- Standardreplikationsfaktor ist 3

HDFS Lese/Schreib-Zugriff:



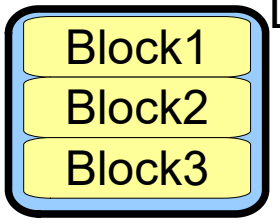
HDFS Write: Replikation

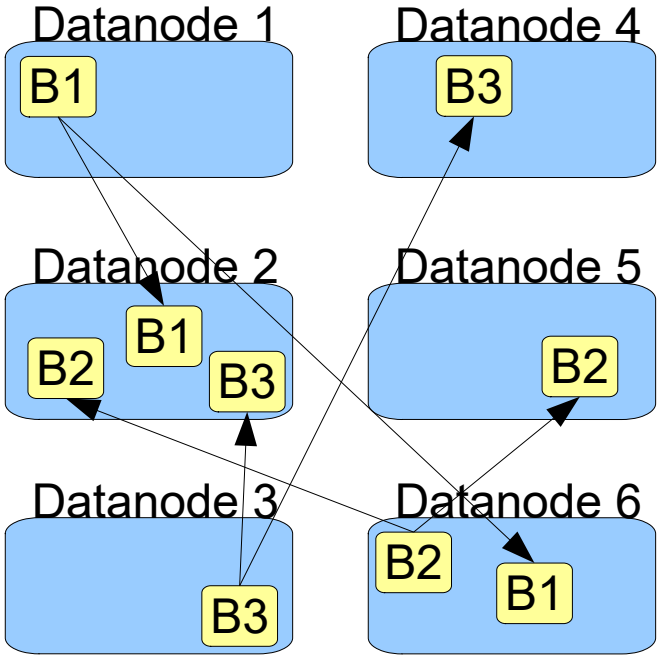
Client wartet
auf Rückmeldung

Datenknoten
replizieren

die neuen
Datenblöcke

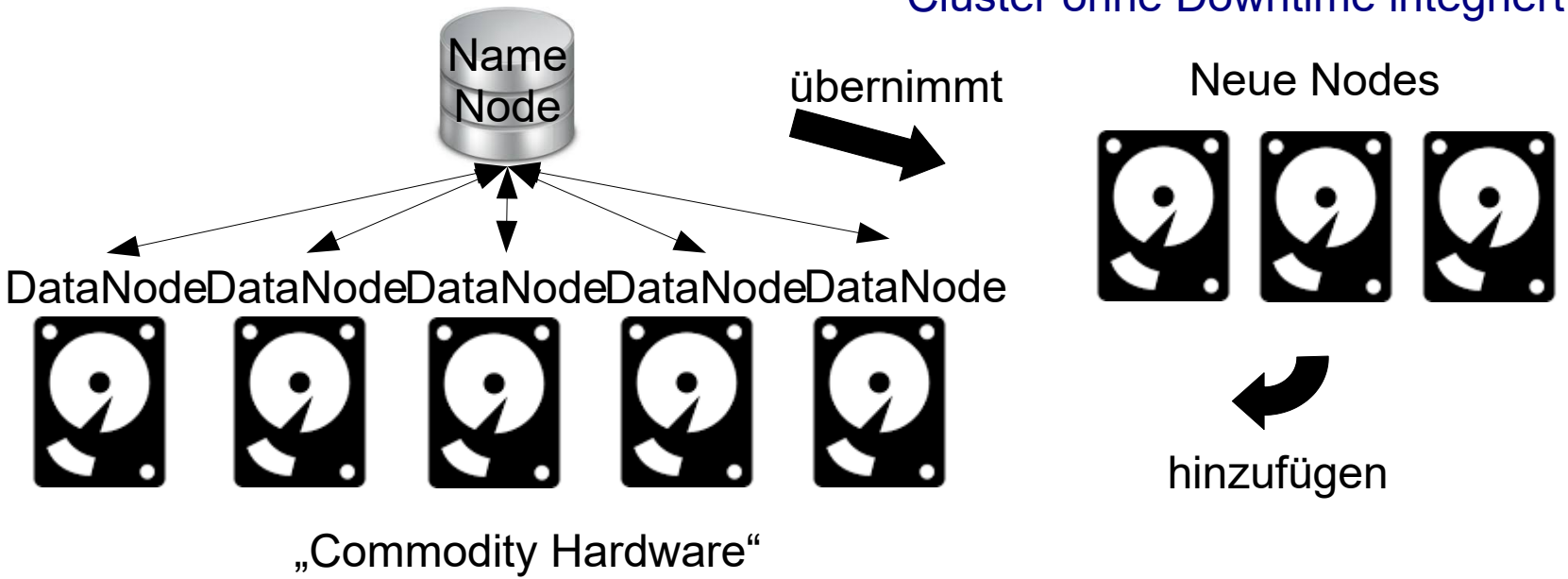
HDFS-Datei





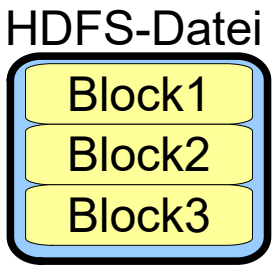
HDFS - Skalierbarkeit

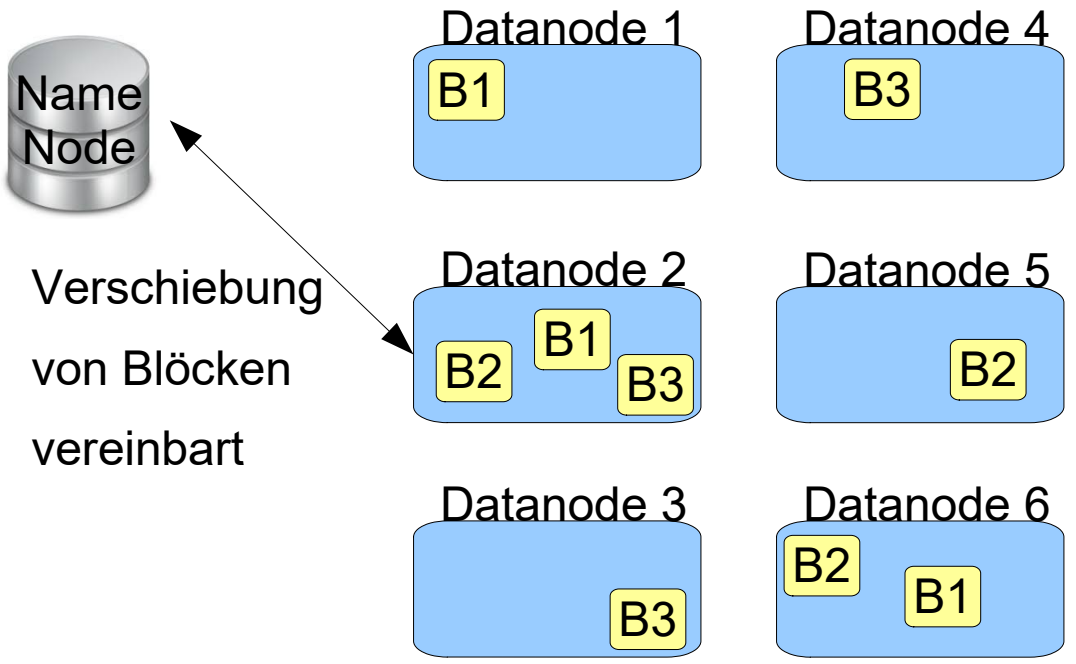
Neue Nodes können in bestehenden Cluster ohne Downtime integriert werden



HDFS Cluster Rebalancing

Datanode2 ist deutlich mehr belegt, als die anderen Datanodes.

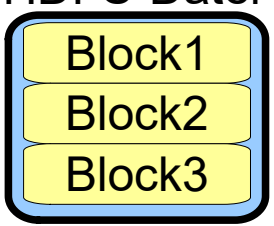




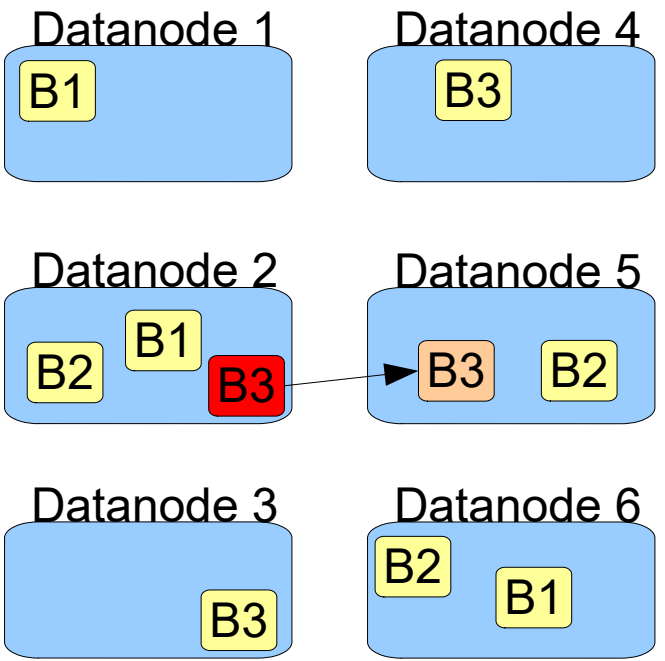
HDFS Cluster Rebalancing

Datanode2 ist deutlich mehr belegt, als die anderen Datanodes.

HDFS-Datei



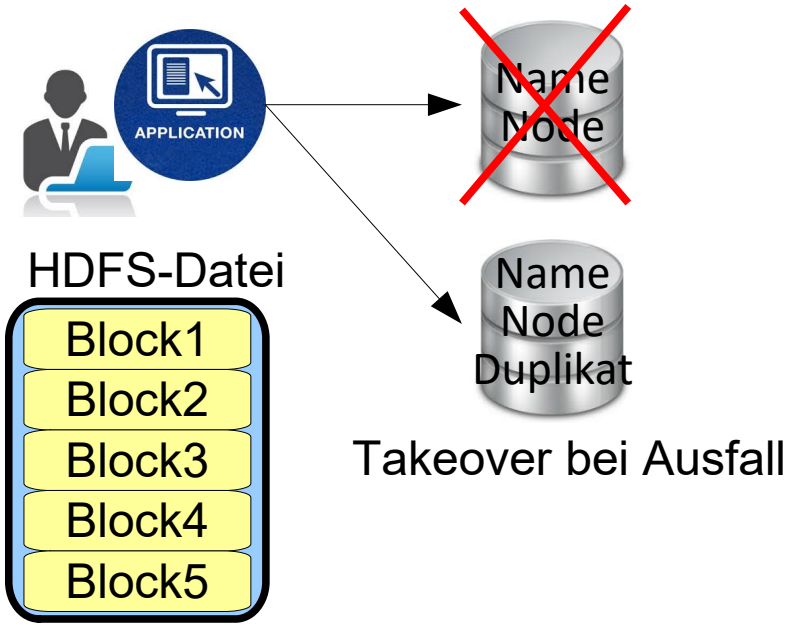
Verschiebung wird durchgeführt!

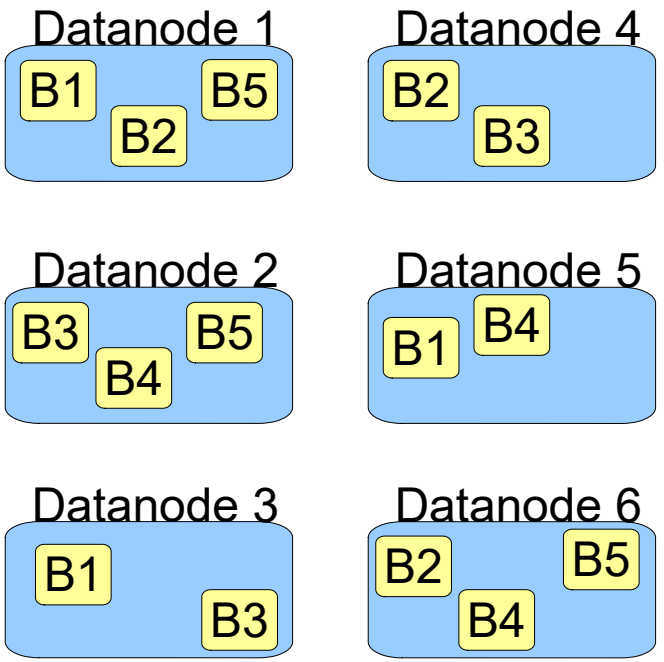


HDFS: Partition-Tolerance

Primärkontakt ist immer ein Namenode.

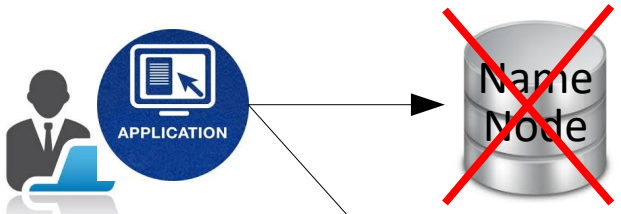
Ausfall des Namenodes: wäre Systemausfall!



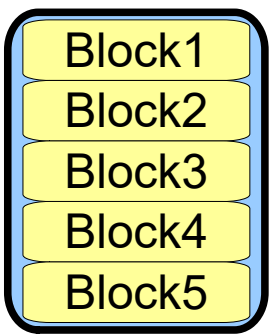


HDFS: Availability

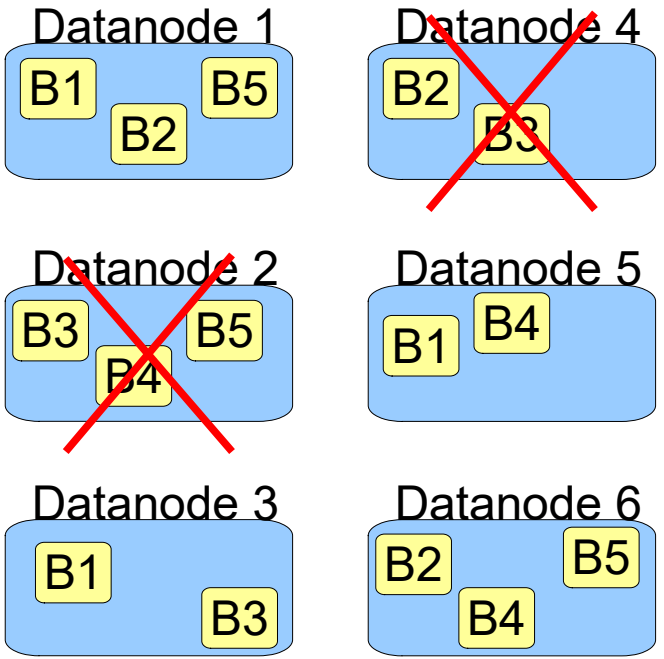
Ausfall von Datenknoten:
kann je nach Replikationsfaktor verkraftet werden



HDFS-Datei



Ausfall von Namenodes ist kritisch!



HDFS: Consistency

Alle Blöcke sind im Regelbetrieb

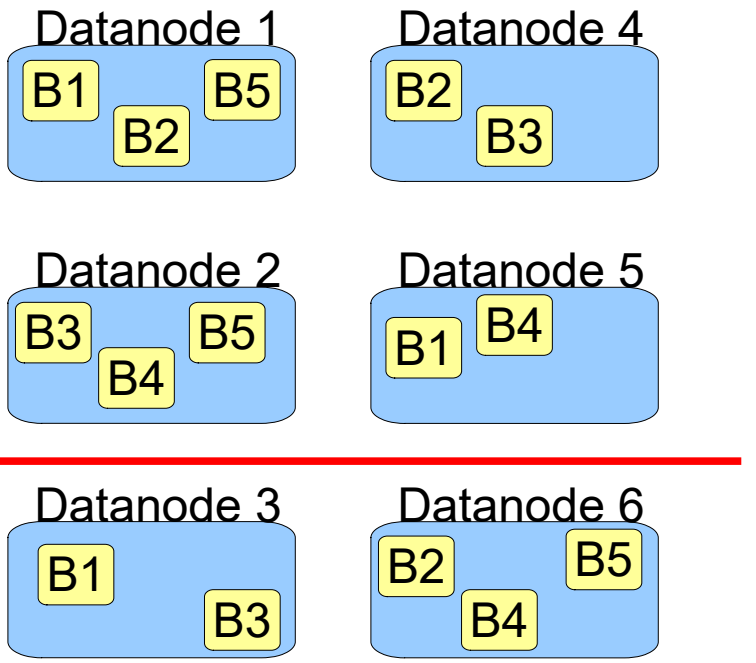
garantiert auf demselben Versionsstand

HDFS-Datei

- Block1
- Block2
- Block3
- Block4
- Block5

Getrennt Datanodes

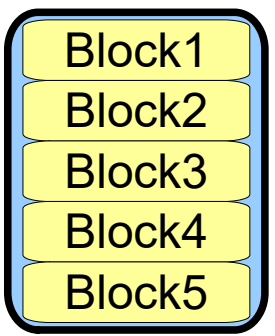
werden als tot markiert.

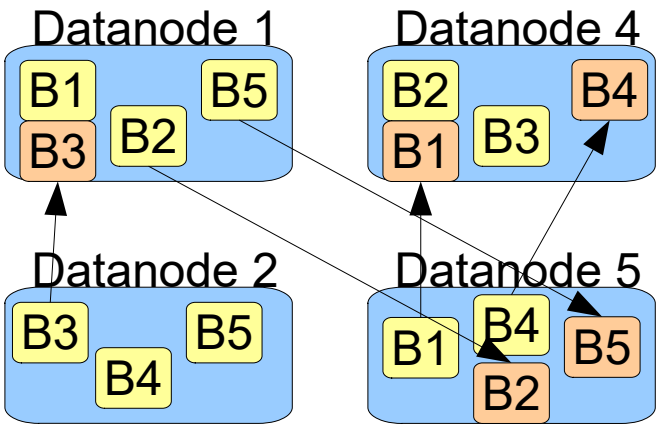


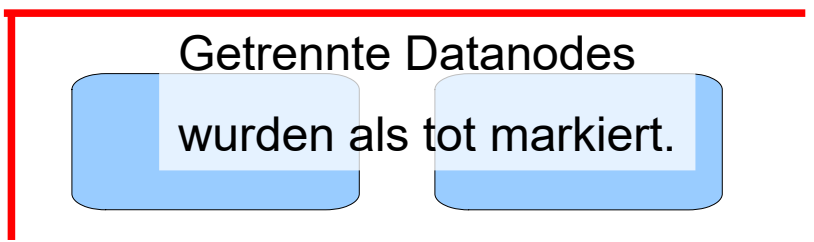
HDFS Consistency: Auto-Replikation

Durch Netzwerkpartition fehlende Blöcke werden automatisch repliziert!

HDFS-Datei

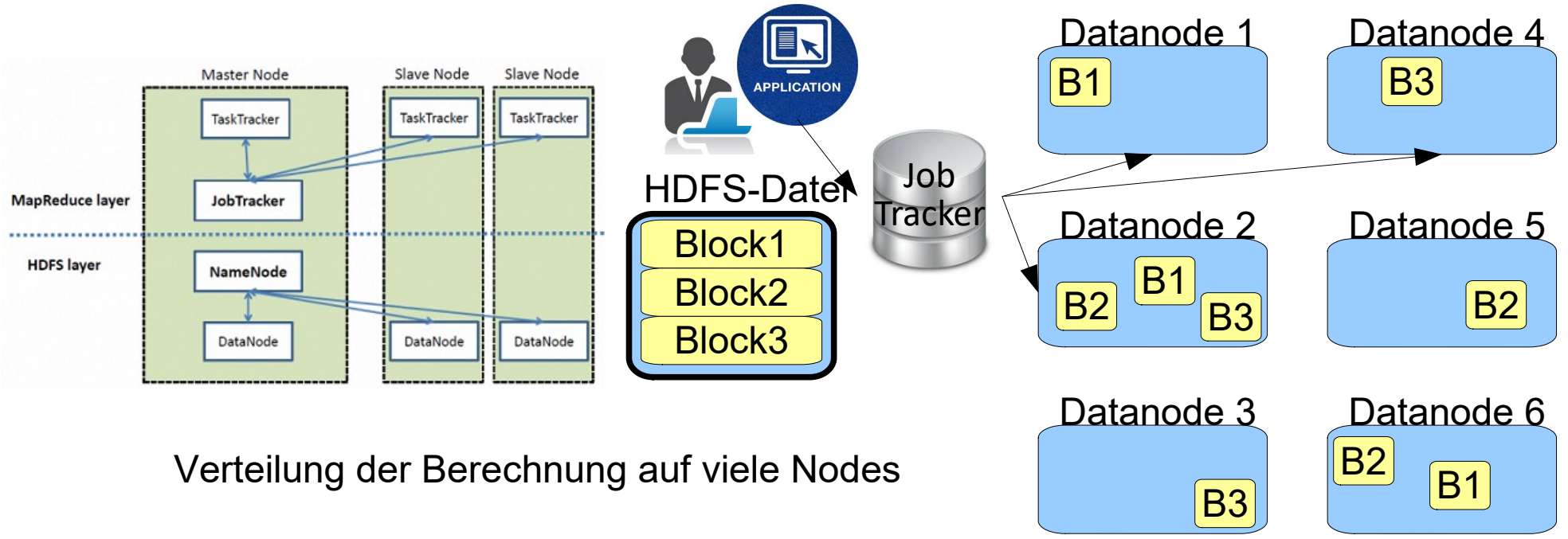






Netzwerkpartition

Hadoop/HDFS: Distributed Computing (Map-Reduce)



Verteilung der Berechnung auf viele Nodes

Fragen & Pause

