

# Big Data

Hype oder Chance?

# Heute reden alle über Big Data

Handelsblatt  
Die digitale Revolution der Wirtschaft  
Wer hebt das Datengold?  
Motorenhersteller, Supermärkte, Gebrauchtwagenhändler: Wie es Unternehmen inzwischen gelingt, aus der Informationsflut im Netz großen Profit zu schlagen. Es ahnet einem neuen Goldrausch – nur geht es heute um Daten.



Süddeutsche.de Digital

2. Januar 2013 08:38 "Big Data"

## Wenn Daten sprechen

Besteht zwischen Bewegungsprofil und Schuhgröße womöglich ein Zusammenhang? Kann man durch die Auswertung von Tweets drohende Verarmung und Epidemien prognostizieren? Unter dem Stichwort "Big Data" denken IT-Firmen darüber nach, wie sie mehr Nutzen aus ihren unendlichen Datenbeständen ziehen können. Noch hat keiner eine Lösung - aber der Druck ist groß.

## Big-Data-Technologie Wissen für Entsch

Analytics: Big Data in der Praxis

Wir innovative Unternehmen über Datenbestände effektiv nutzen

The New York Times Sunday Review | The Opinion Pages

## The Age of Big Data

GOOD with numbers: Fascinated by data? The sound you hear is last summer, as a freshly minted Yale M.B.A. to join the technology consultants. They help businesses make sense of an explosion of data — Web traffic and software and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Mr. Zhou, whose job as a data analyst suits her skills.

To exploit the data flood, America will need many more like her. A report last year by the McKinsey Global Institute, the research arm of the consulting firm, projected that the United States needs 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, whether retrained or hired.

The impact of data abundance extends well beyond business. Justin Grimm, for example, is one of the new breed of political scientists. A 28-year-old assistant professor at Stanford, he combined math with political science in his undergraduate and graduate studies, seeing "an opportunity because the discipline is becoming increasingly data-intensive." His research involves the computer-automated analysis of blog postings.

McKinsey Global Institute

June 2011

## Big data: The for innovation and product

## Harness the Power of Big Data

### The IBM Big Data Platform

- Learn all about IBM's enterprise-class end-to-end Big Data platform
- Boost your Big Data I/O, details on in-motion and at-rest analytics, data asset discovery, integration, and governance
- Get details surrounding the most common Big Data use cases that are transforming organizations today
- Learn how to make down-stream Big Data use cases that deploy...and less risky
- Gain confidence in your Big Data projects with an end-to-end tour of accelerators, tool sets, and samples to get you going...FAST!

PAUL ZIKOPOULOS  
THOMAS DEUTSCH  
DIRK DEROGS  
DAVID CORRIGAN  
KRISHNAN PARASURAMAN  
JAMES GILES

# Big Data

Most "big data" research shows that high-quality data and analytics can increase ROI after investment. But no research has shown the technologies that improve existing data has...

# 10%

**PRODUCTIVITY INCREASE**  
A 10% increase in U.S. data translates to an additional **\$2.01 billion** in total revenue

Let's look at the impact of big data on various industries:

<b>PRODUCTIVITY INCREASE</b>	49%
<b>RETAIL</b>	39%
<b>CONSULTING</b>	21%
<b>AIR TRANSPORTATION</b>	20%
<b>CONSTRUCTION</b>	20%
<b>FOOD PRODUCTION</b>	20%
<b>STEEL</b>	19%
<b>AUTOMOBILE</b>	18%
<b>INDUSTRIAL INSTRUMENTATION</b>	18%
<b>PUBLISHING</b>	17%
<b>TELECOMMUNICATIONS</b>	17%

*Median survey respondent: Fortune 1000 usability increases sales per employee by 10% could increase its sales per employee figure*

## Big data—capabilities

**\$300 billion** potential annual value to US I double the total annual health

**€250** potential annual administrative

**\$600 billion** potential annual consumer spending using personal location data

**60%**

**140,000–190,000** more deep analytical talent per

## Big data—a growing torrent

**\$600** to buy a disk drive that can store all of the world's music

**5 billion** mobile phones in use in 2010

**30 billion** pieces of content shared on Facebook every month

**40%** projected growth in global data generated per year vs. **5%** growth in global IT spending

**235** terabytes data collected by the US Library of Congress by April 2011

**15 out of 17** sectors in the United States have more data stored per company than the US Library of Congress

210 x 297 mm

**Anwendungen**

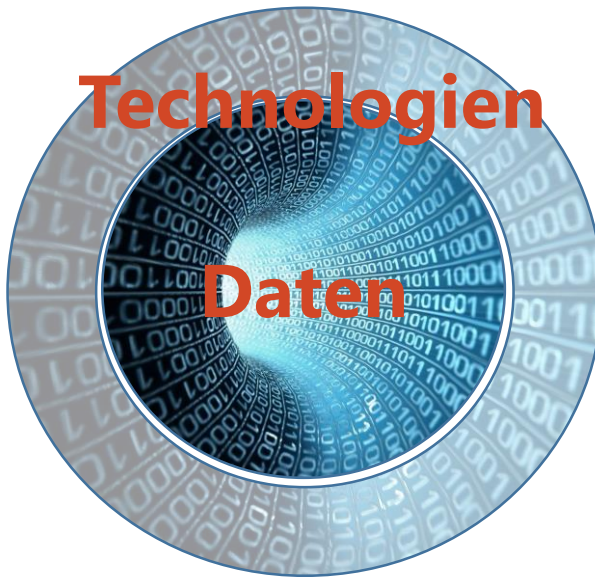
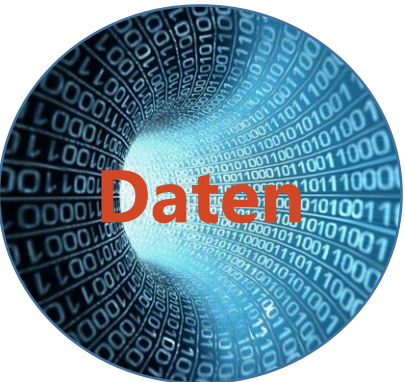
**Technologien**

**Daten**

**Technologien**

**Daten**

**Daten**



# Agenda

## Big Data

- Was ist Big Data?
- Unter welchen Aspekten kann Big Data betrachtet werden?
  - Daten
  - Technologien
  - Anwendungen – Fallbeispiele
- Welchen Mehrwert verspricht es?

## Hadoop

- Ganzheitliche Betrachtung
- Welche Vorteile hat die Verwendung von Hadoop?
- Das Filesystem
- MapReduce

## Fazit

# Was ist Big Data ?

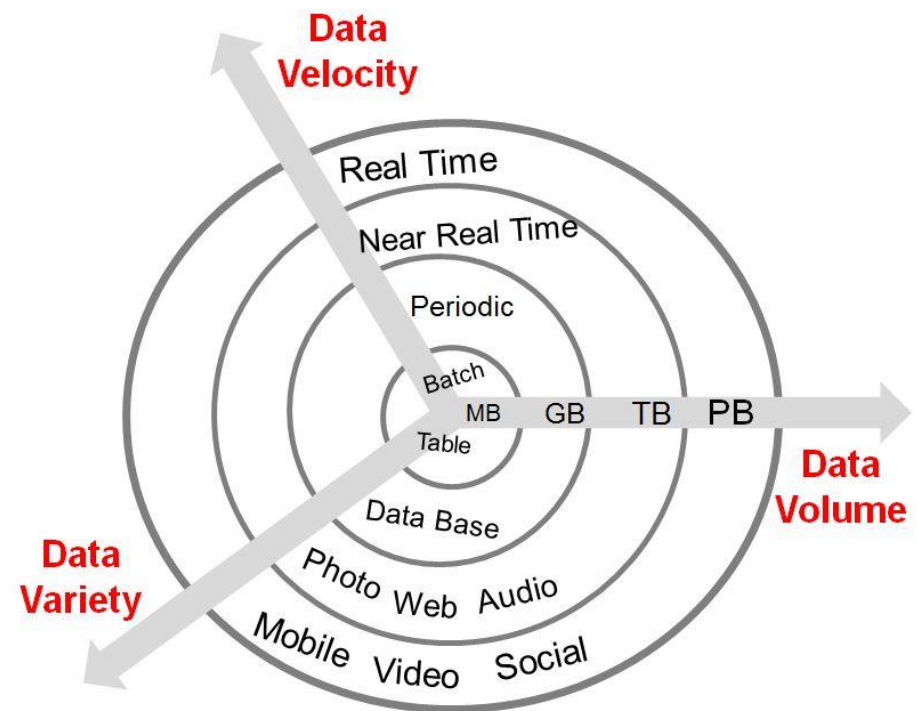
## Definitionsversuch:

- „Big data is data that exceeds the processing capacity of conventional database systems.“ Quelle: <http://strata.oreilly.com/2012/01/what-is-big-data.html>
- „“Big Data“ ist [...] die Informations- und Technologiegrundlage für die Empfehlungssysteme bzw. Entscheidungsunterstützungssysteme.“ Quelle: Baron, Pavlo. *Big Data für IT- Entscheider - Riesige Datenmengen und moderne Technologien gewinnbringend nutzen*. Carl Hanser Verlag, 2013
- „Big Data bezeichnet die Analyse großer Datenmengen aus vielfältigen Quellen in hoher Geschwindigkeit mit dem Ziel, wirtschaftlichen Nutzen zu erzeugen.“ Quelle: Bitcom. *Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte*. Leitfaden. 2012
- Bei „Big Data“ handelt sich um den Prozess aus Daten Informationen zu gewinnen und diese in Form von Wissen anzuwenden, oder anders ausgedrückt der Prozess der Informationsgewinnung und Informationsanwendung aus unterschiedlichsten Daten.

# Daten



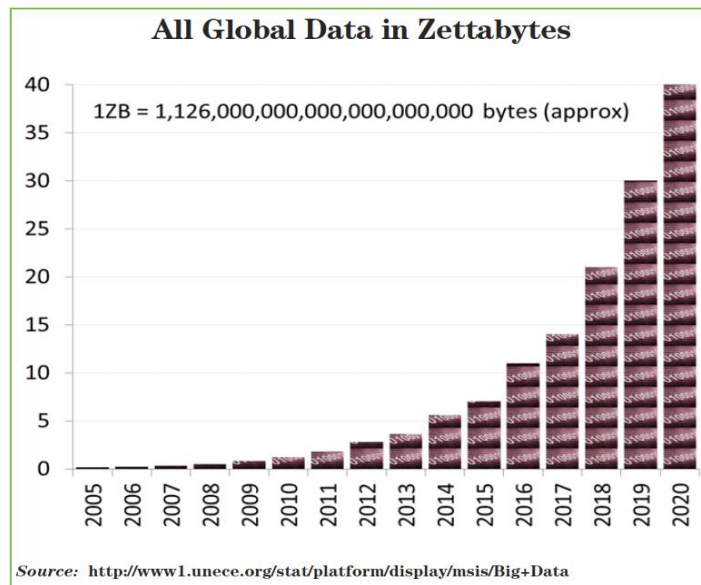
*Daten in Bezug auf "Big Data" lassen sich anhand von drei Aspekten, Problemen, Schmerzen oder auch Herausforderungen beschreiben, die im Fachjargon als die „drei Vs“ bezeichnet werden: Datenmenge oder Volume, Geschwindigkeit oder Velocity und Vielfalt oder Variety. Wer das Phänomen „Big Data“ erfassen möchte, sollte alle drei betrachten.*



# Volumen



- Die Menge an Daten, die erstellt, vervielfältigt und konsumiert werden, wird 2020 bei etwa 35 - 40 Zettabytes liegen.
- Das Datenvolumen verdoppelt sich alle 2 Jahre.



Das Datenwachstum geht im Wesentlichen auf vier Megatrends der IT zurück:

- Boom von Smartphones und Tablet Computern (App-Economy)
- Soziale Netzwerke
- Sensoren, etwa in intelligenten Maschinen
- Cloud- Services (IaaS, PaaS, SaaS)



Die Geschwindigkeit mit der sich Daten bewegen und in einem Unternehmen eintreffen hat sich ähnlich rapide erhöht wie das Datenvolumen.

Riesige Datenmengen müssen immer schneller ausgewertet werden, nicht selten in Echtzeit.

## Die Herausforderungen:

- Analysen großer Datenmengen mit Antworten im Sekundenbereich
- Datenverarbeitung in Echtzeit
- Datengenerierung und Übertragung in hoher Geschwindigkeit

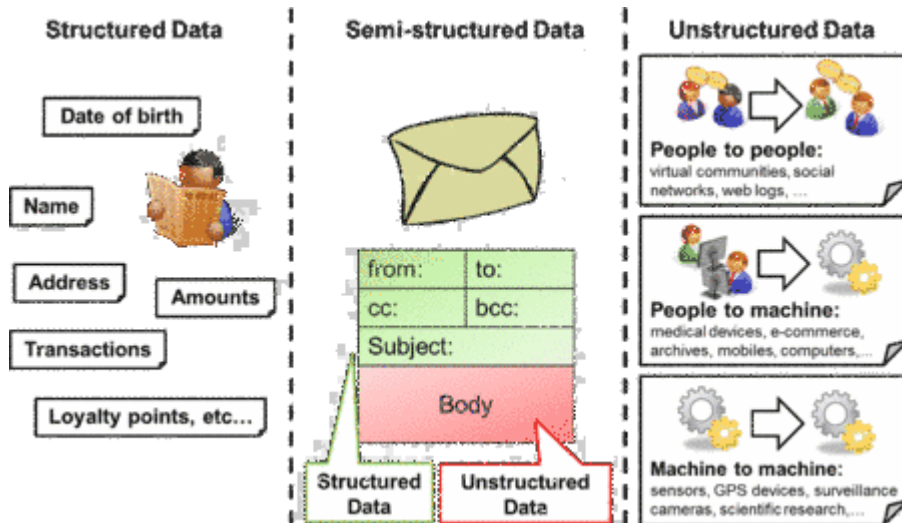
# Variety



Sinnvollen Nutzung qualitativ unterschiedlich strukturierter Daten, die einem schnellen Wandel unterliegen und in großem Umfang anfallen

Man unterscheidet dabei grob Drei Kategorien:

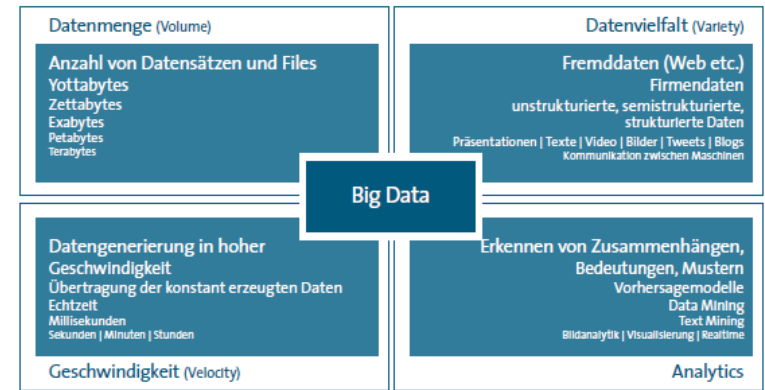
- strukturierte Daten: z.B. Datensatz in einer relationalen Datenbank
- semi- strukturierte Daten: z.B. Email, bestehen aus Anschrift, Sender und Absender (strukturiert) und Text (unstrukturiert)
- unstrukturierte Daten: Ein Text kann eine beliebige Struktur aufweisen, Videos usw.



# Die Erweiterungen der drei V's



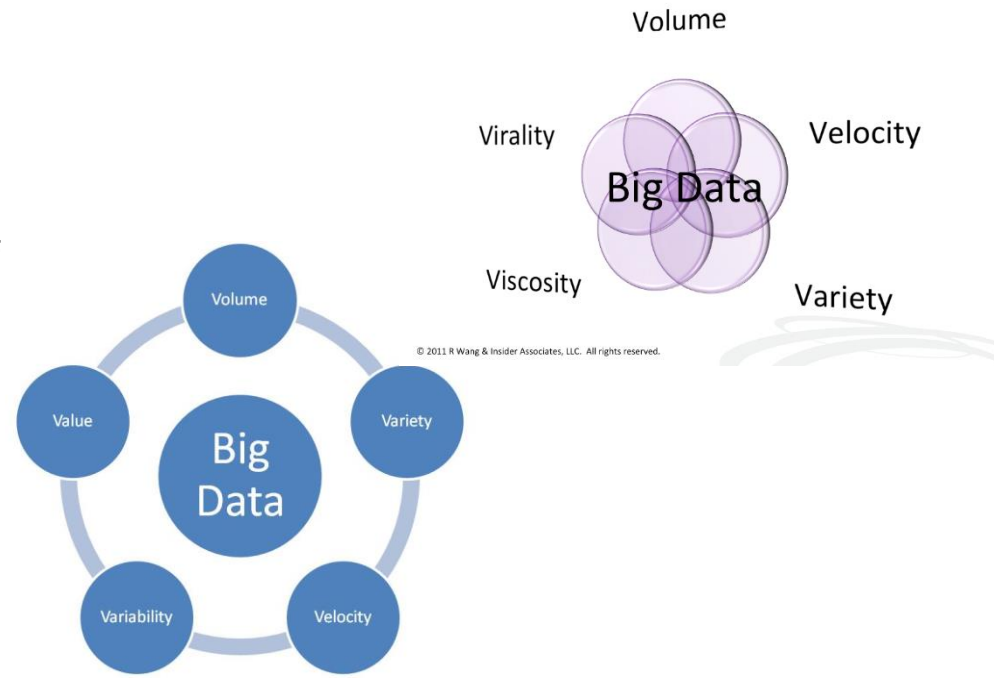
Es häufen sich die Meinungen, dass die Drei Aspekte nicht ausreichen um „Big Data“ in vollen Umfang zu erfassen. Dies hat zu unterschiedlichen Erweiterungen geführt.



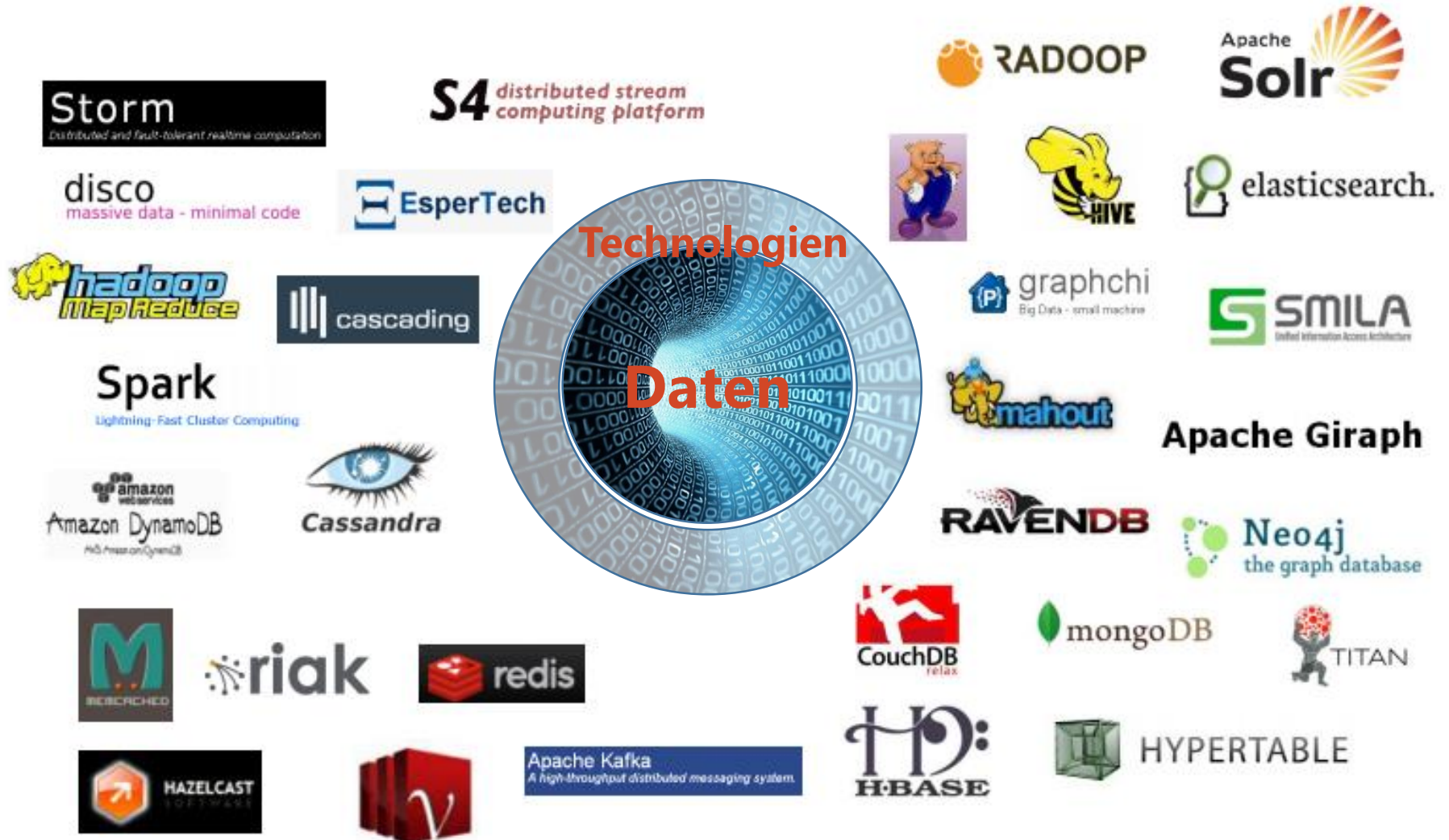
Quelle: Bitcom. Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Leitfaden. 2012

“The "big" in "big data" is a function of the volume, variety, and velocity of the information that constitutes it. If you read a dozen articles on big data, there's a good chance the 3 Vs -- volume, variety, and velocity -- will be cited in at least half of them.

This strikes many industry veterans as wrong-headed because if big data is understood solely on the basis of these trends, it isn't clear that it's at all hype-worthy. It isn't clear, in other words, that what we mean by "big data" comprises a distinct departure from the data management (DM) status quo.” Quelle: Stephen Swoyer. Big Data -- Why the 3Vs Just Don't Make Sense. 2012



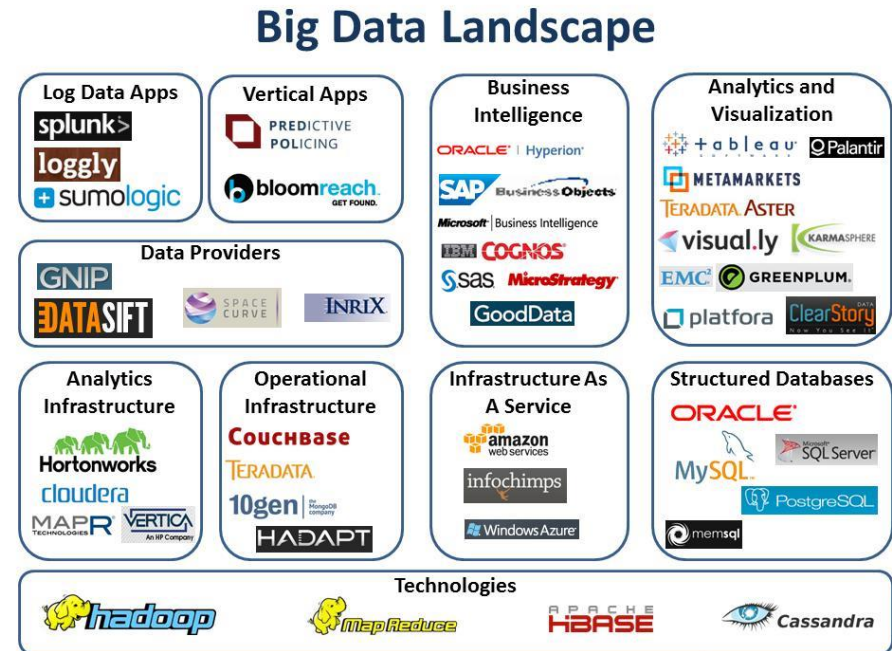
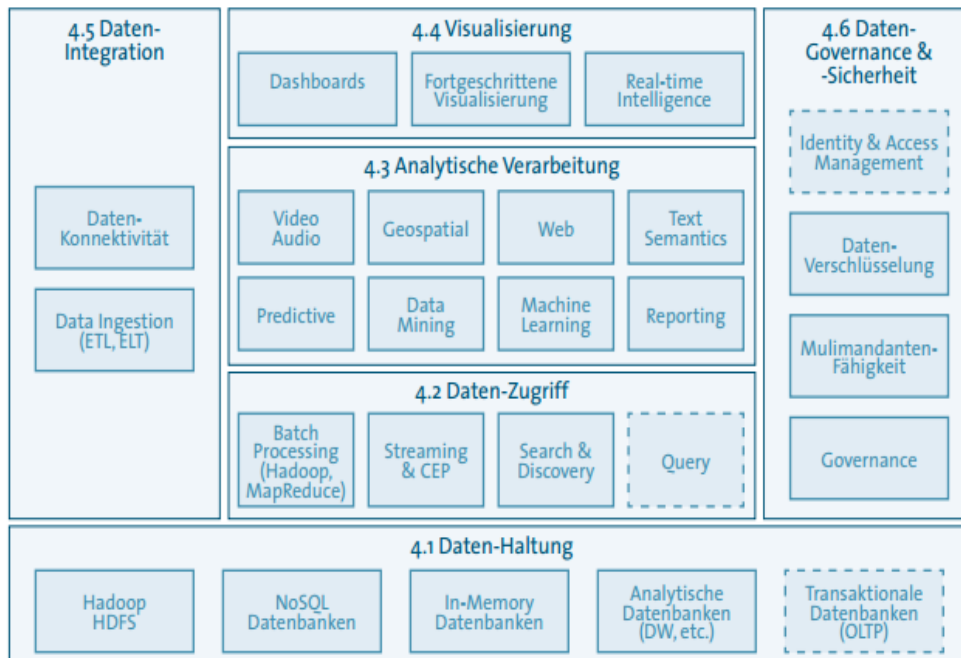
# Open-Source Technologien im Kontext von Big Data



# Technologien



Big Data basiert nicht auf einer singulären Technologie, sondern ist vielmehr das Resultat des Zusammenwirkens einer ganzen Reihe von Innovationen in verschiedenen Gebieten. Dies bedeutet das man polyglotte Architekturen bei einer Gesamtlösungen vorfindet.





# Technologien - NoSQL



Ein wichtiger Technologiebaustein stellen die NoSQL Datenbank-Infrastrukturen dar, da diese sehr gut an die hohen Anforderungen von „Big Data“ angepasst sind.

Def.:  
NoSQL ist eine Bewegung, welche im Jahre 2009 ihren Anfang nahm. Die Bewegung hat sich zur Aufgabe gemacht die Problemstellung des modernen Datenmanagement ohne relationalen Datenbanken zu lösen. So bedarf es bei einer NoSQL-Datenbanken meist keiner festen Tabellen- Schemata um Daten zu speichern.

Eigenschaften von NoSQL-Datenbanken (Diese können je nach Implementierung und Data Store abweichen):

- Skalieren horizontal
- Verteiltes System
- Speicherung von großen Datenmengen
- Open- Source
- Nicht relational
- Commodity Hardware
- Eventually Consistent/ CAP / BASE (nicht wie bei den relationalen Datenbanken ACID)

# Technologien - NoSQL



NoSQL Datenbank unterscheidet man grundsätzlich zwischen mehreren Arten:

- Key/Value Datenbanken: z.B. Riak
- Spaltenorientierte Datenbanken: z.B. Cassandra
- Document Stores: z.B. MongoDB
- Graphendatenbank: z.B. Neo4j
- In- Memory- Datenbanken: z.B. Redis

# Anwendungen - Fallbeispiele



## **Big-Data-Beispiel 1: Verkehrssteuerung in Stockholm**

Ziel: Staus verringern, Fahrzeit verkürzen

Stockholm hat hierfür ein intelligentes Verkehrsmanagement auf Basis von Big-Data-Technologie eingeführt. Die Echtzeit-Analyse bezieht unter anderem über 250.000 GPS-Daten pro Sekunde sowie Daten von Sensor- und Videosystemen, etwa vom Mautsystem, in die Analyse ein. Auch Stau- und Unfallmeldungen werden berücksichtigt

Ergebnis: individuellen Fahrzeiten um bis zu 50 Prozent verringert, Die Emissionen wurden um 20 Prozent reduziert, Der Verkehr ging um 20 Prozent zurück

## **Big-Data-Beispiel 2: Krebstherapie**

Ziel: Verbesserung der individualisierten Krebstherapie

Bisher ist dies ein extrem zeitaufwendiger Prozess. Zum einen liegt das an der Fülle unterschiedlicher Daten, zum anderen an ihrer Verschiedenartigkeit. Alleine durch die Genomsequenzierung fallen pro Patient rund 2 Terrabyte an Daten an. Der Einsatz von Big-Data-Werkzeugen ermöglicht die Analyse dieser vielfältigen und großen Datenmengen in kürzester Zeit, um eine individuell optimierte Therapie vorzuschlagen.

Ergebnis: gesteigerte Heilungschancen

# Anwendungen - Fallbeispiele



## **Big-Data-Beispiel 3: ProSiebenSat1 Digital Web Traffic**

Ziel: Gesamte Analyse des Web Traffic in einem Data Warehouse

Die ProSiebenSat.1 Digital GmbH benötigt für die Messung, Steuerung und Optimierung ihrer zahlreichen Online-Plattformen detaillierte Datenanalysen des Web Traffic. Primär sind dabei nicht-transaktionale Online-Plattformen relevant, aber es sollen auch Transaktionsdaten der Video-Plattform MyVideo und von maxdome.de eingebunden werden. Während der IT-technischen Konzeption einer geeigneten Analyse-Plattform (Data Warehouse) stößt man sehr schnell auf zwei typische Big-Data-Problemstellungen: ein sehr stark skalierendes Datenvolumen und eine Vielzahl von heterogenen Datenquellen.

### **1. Datenvolumen 2012: 10 Terrabyte -> jährliches Wachstumsrate von 300 %**

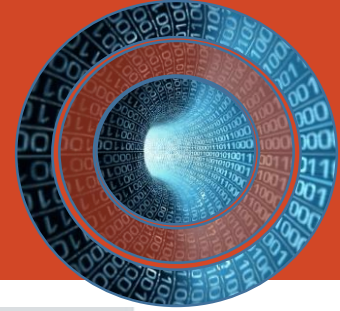
Eine vertikale Skalierung der Systeme stößt hier an ihre Grenzen

Lösung: Rohdatenhaltung in einem Hadoop- Cluster -> horizontale Skalierung

**2. Fünf Quellsysteme:** Online-Trafficdaten aus verschiedenen Messsystemen, Online-Trafficdaten aus verschiedenen Messsystemen, Markt- und Konkurrenzdaten, Transaktionsbezogene Daten, Weitere interne Daten zur Geschäftssteuerung

Lösung: Aggregation mit Mapreduce und ETL-Technologie für die Übertragung in das DWH

# Anwendungen - Fallbeispiele



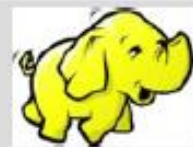
## Lösungsansatz

Hybrides System aus relationaler Datenbank und Hadoop Cluster

PostgreSQL



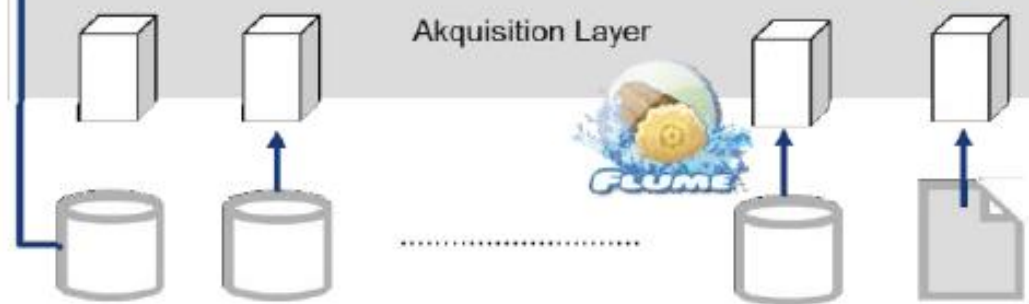
Hadoop Cluster



Integration Layer



Source Systems



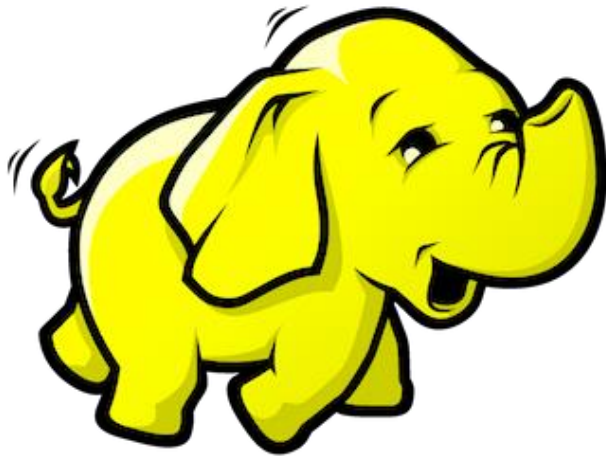
# Unternehmen die auf Big Data setzen



# Was sind die positiven Effekte von der Daten- Sintflut?

- Big Data schafft mehr **Transparenz**, was Unternehmen hilft, den Überblick zu bewahren und schneller bessere Entscheidungen zu treffen.
- Big Data erlaubt mehr Planspiele und Simulationen, da Unternehmen auf unerhört großen Datenmengen sitzen und sie zeitnah auswerten können.
- Big Data verbessert den Zugang zum einzelnen Kunden, sodass Produkte und Dienstleistungen **auf eine Person zugeschnitten werden können**.
- Big Data unterstützt Firmen dank Analysewerkzeugen, Simulationen und Prognosen bei der **Entscheidungsfindung**.
- **Big Data sorgt für die Entstehung neuer Geschäftsmodelle, Produkte und Dienstleistungen** – entweder von etablierten Unternehmen oder vollkommen neuen Firmen.

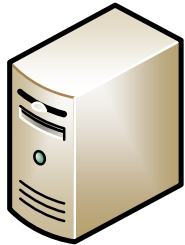
# Apache Hadoop



Bei Hadoop handelt es sich um ein Java-basiertes Open- Source- Framework für die skalierbare und verteilte Verarbeitung von großen Datenmengen.

# Wieso Hadoop ?

Aufgabe= 1 TB Daten lesen



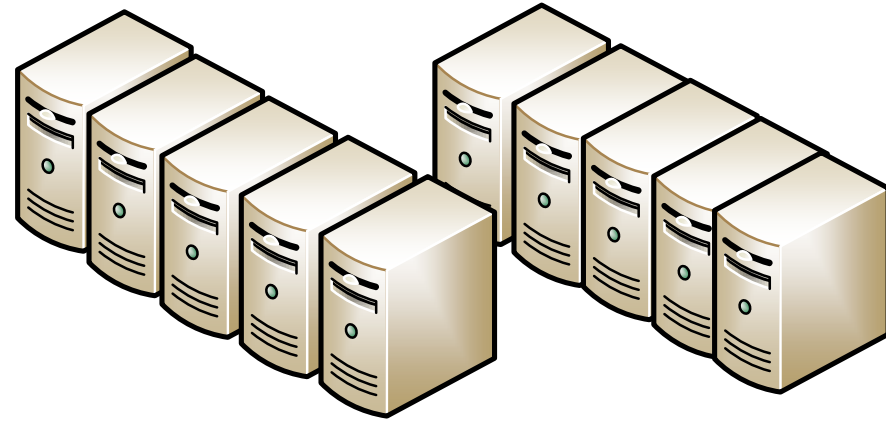
1 Knoten

- 4 I/O- Kanäle
- 100 MB/s  
Lesegeschwindigkeit

=



Ca. 45 Minuten ?



10 Knoten

- 4 I/O- Kanäle
- 100 MB/s  
Lesegeschwindigkeit

=



Ca. 4,5 Minuten ?

# Wann nutzt man Hadoop

- Billiger
  - Skaliert mit Petabytes und mehr
- Schneller
  - Parallele Datenverarbeitung
- Besser
  - Gut geeignet für bestimmte Probleme von Big Data

# Für welche Probleme von Unternehmen ist Hadoop als Lösung geeignet?

Risk Modeling

Customer Churn  
Analysis

Recommendation  
Engine

Ad Targeting

Point of Sale  
Transactional  
Analysis

Threat Analysis

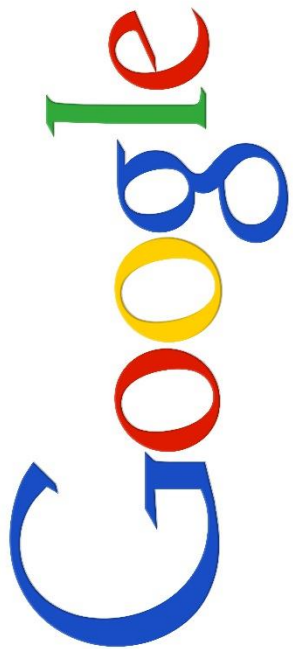
Trade Surveillance

Search Quality

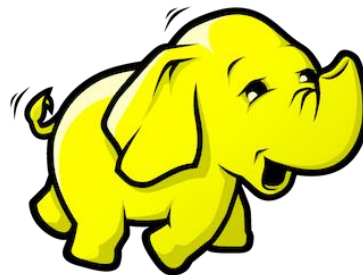
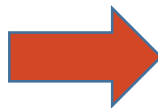
Data Sandbox

# Wer nutzt Hadoop

- Facebook
- Yahoo
- Amazon
- eBay
- American Airlines
- The New York Times
- Federal Reserve Board
- IBM
- Orbitz
- ProSiebenSat1 Digital



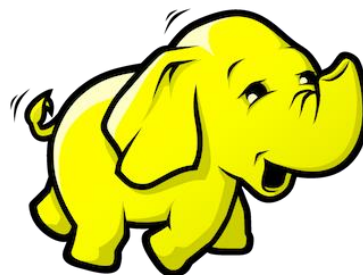
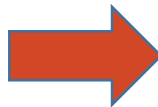
GFS



HDFS



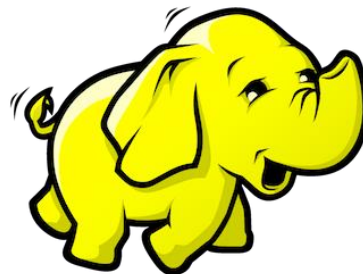
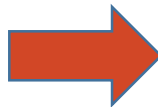
MapReduce



MapReduce

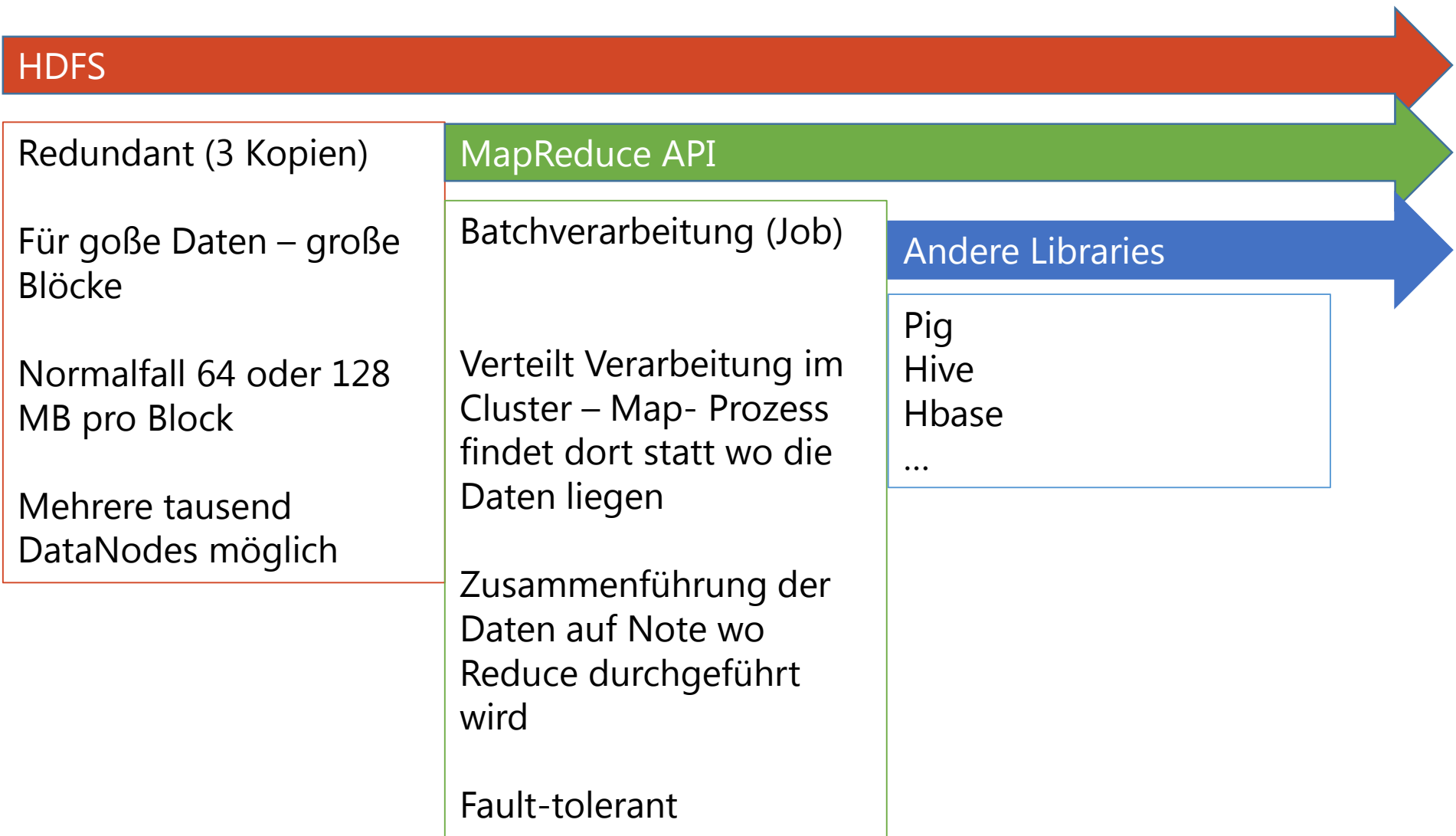


BigTable



Apache HBase

# Mehr Details zu der Architektur



# Das Filesystem unter HDFS



ext3

ext4

XFS

- wird von Yahoo! genutzt
- wird von Google genutzt
- sehr schnell

HDFS ist ein abstraktes Filesystem, das auf die oben genannten Filesysteme aufgesetzt wird.

# Architektur von Hadoop



Master

NameNode

JobTracker

/

Architektur



Slave

DataNode

TaskTracker

# Hadoop HDFS

## NameNode



Metadaten der Files:  
/mandant/VoDMax/log1.txt ↗ 1,2,3  
/mandant/websiteXY/log2.txt ↗ 4,5

$r = 3$

hdfs-site.xml



dfs.replicaion



3	
4	2
	1

DataNode 1



	3
5	4

DataNode 2



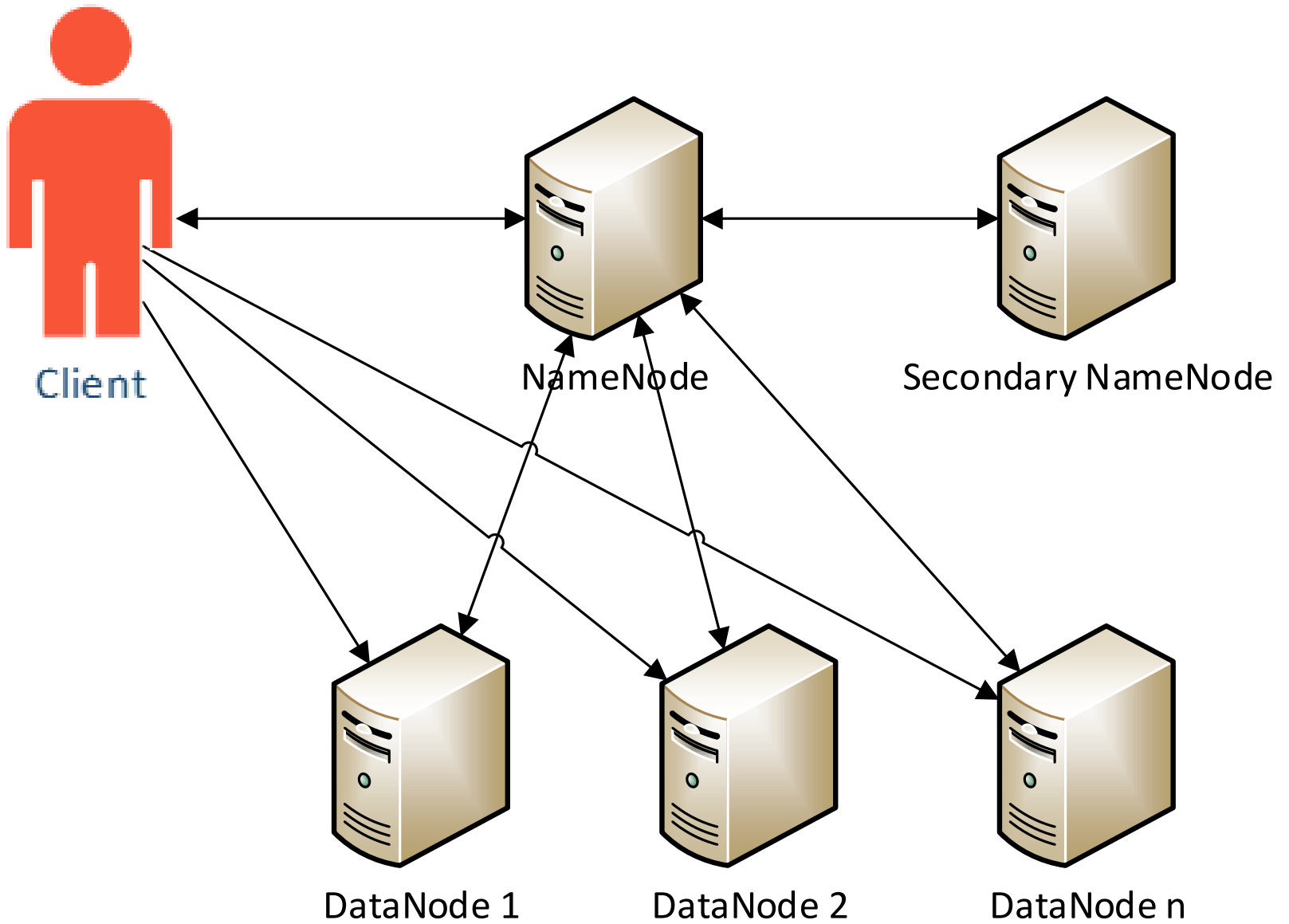
	3
1	5
	2

DataNode 3

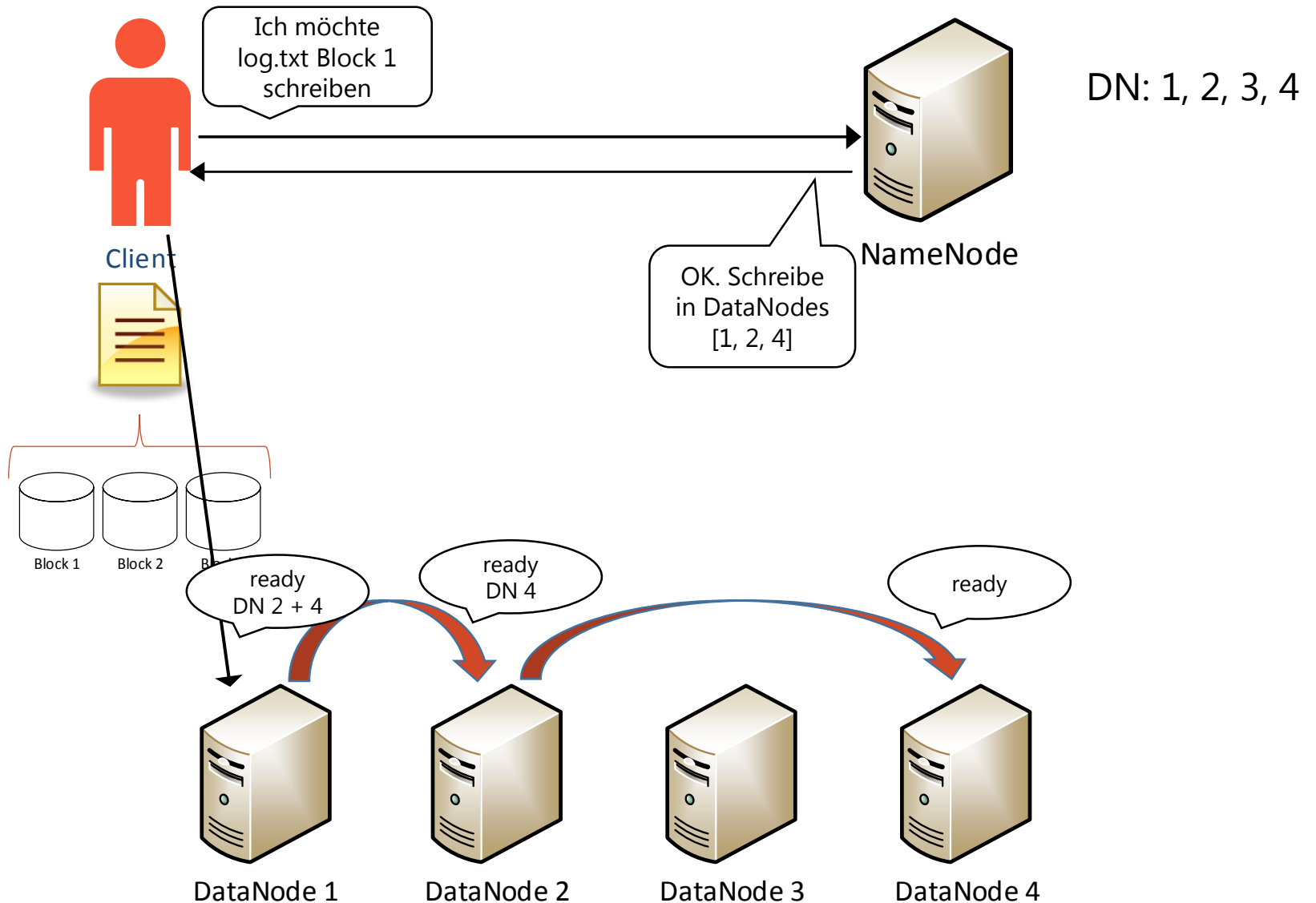


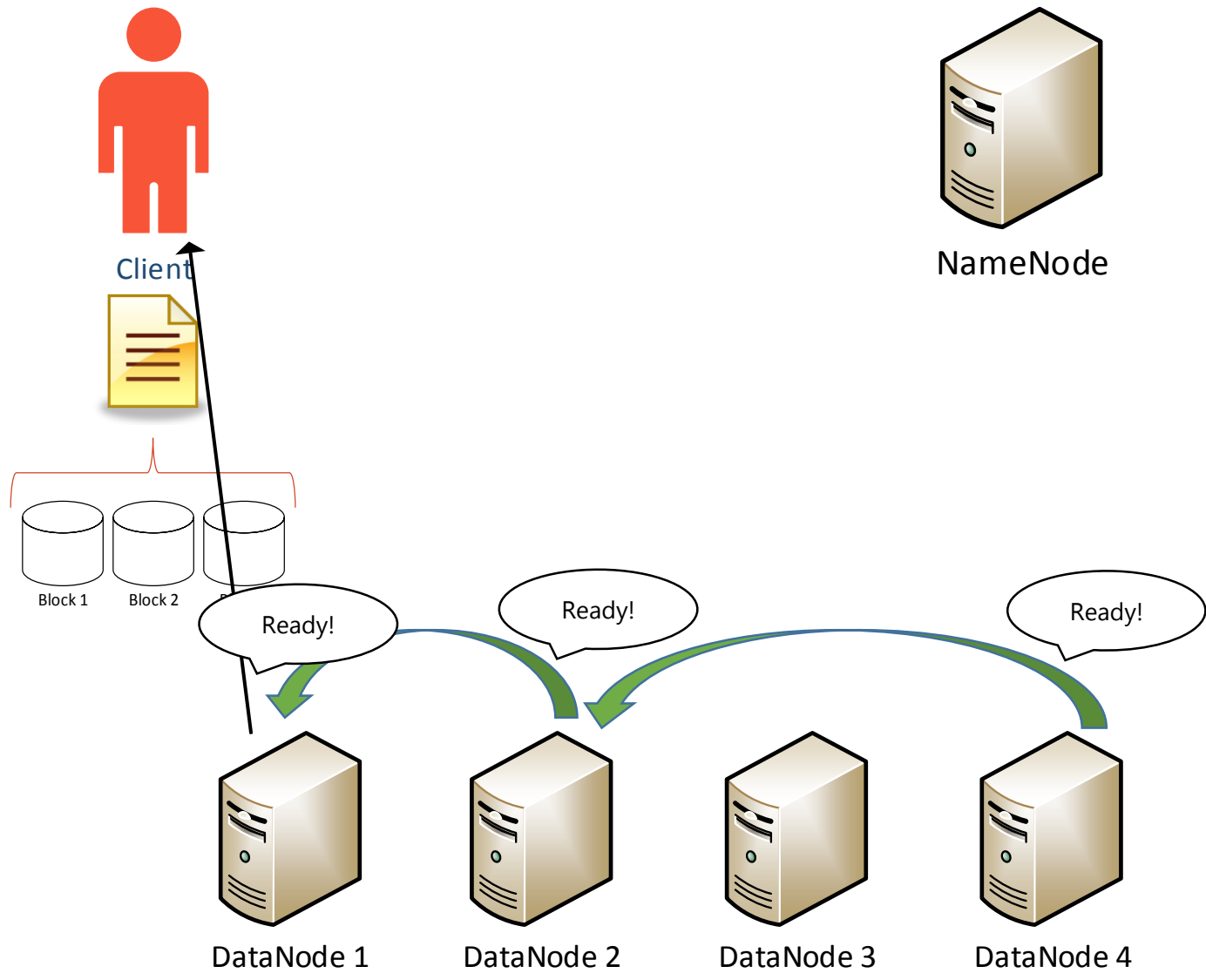
	4
5	1
	2

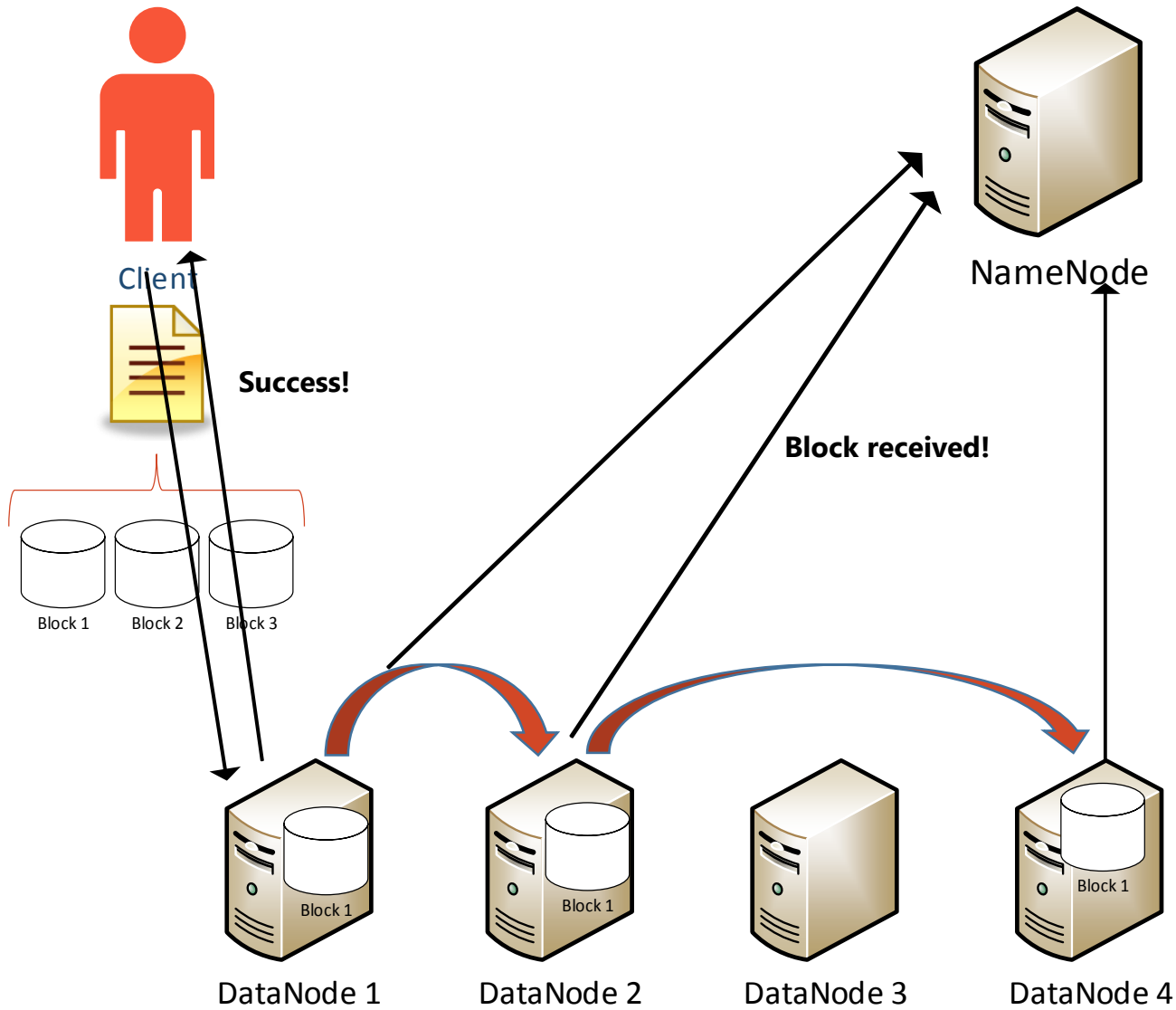
DataNode 4



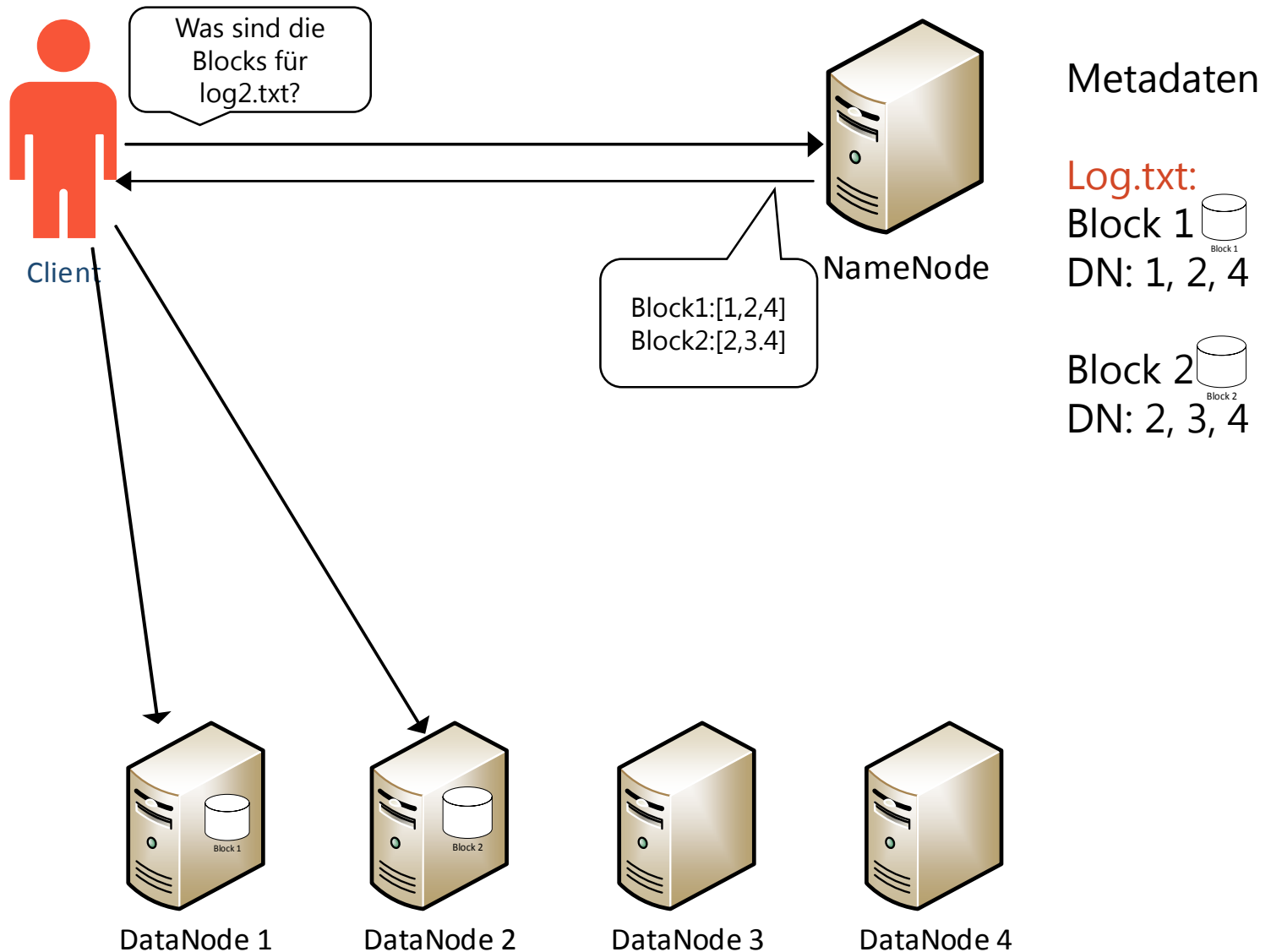
# Schreibzugriff auf HDFS







# Lesezugriff auf HDFS



# NameNode Web UI

## NameNode 'ip-10-140-6-87.ec2.internal:8020'

**Started:** Tue Feb 21 19:30:19 EST 2012  
**Version:** 1.0.0, r1224962  
**Compiled:** Sat Jan 21 03:22:22 UTC 2012 by hrt\_qa  
**Upgrades:** There are no upgrades in progress.

[Browse the filesystem](#)  
[Namenode Logs](#)

---

### Cluster Summary

1575 files and directories, 2091 blocks = 3666 total. Heap Size is 960 MB / 960 MB (100%)

Configured Capacity	:	55.57 GB
DFS Used	:	152.38 MB
Non DFS Used	:	18.16 GB
DFS Remaining	:	37.27 GB
DFS Used%	:	0.27 %
DFS Remaining%	:	67.06 %
<a href="#">Live Nodes</a>	:	3
<a href="#">Dead Nodes</a>	:	0
<a href="#">Decommissioning Nodes</a>	:	0
Number of Under-Replicated Blocks	:	2

### NameNode Storage:

Storage Directory	Type	State
/grid/0/hdp/hdfs/name	IMAGE_AND_EDITS	Active

---

This is [Apache Hadoop](#) release 1.0.0

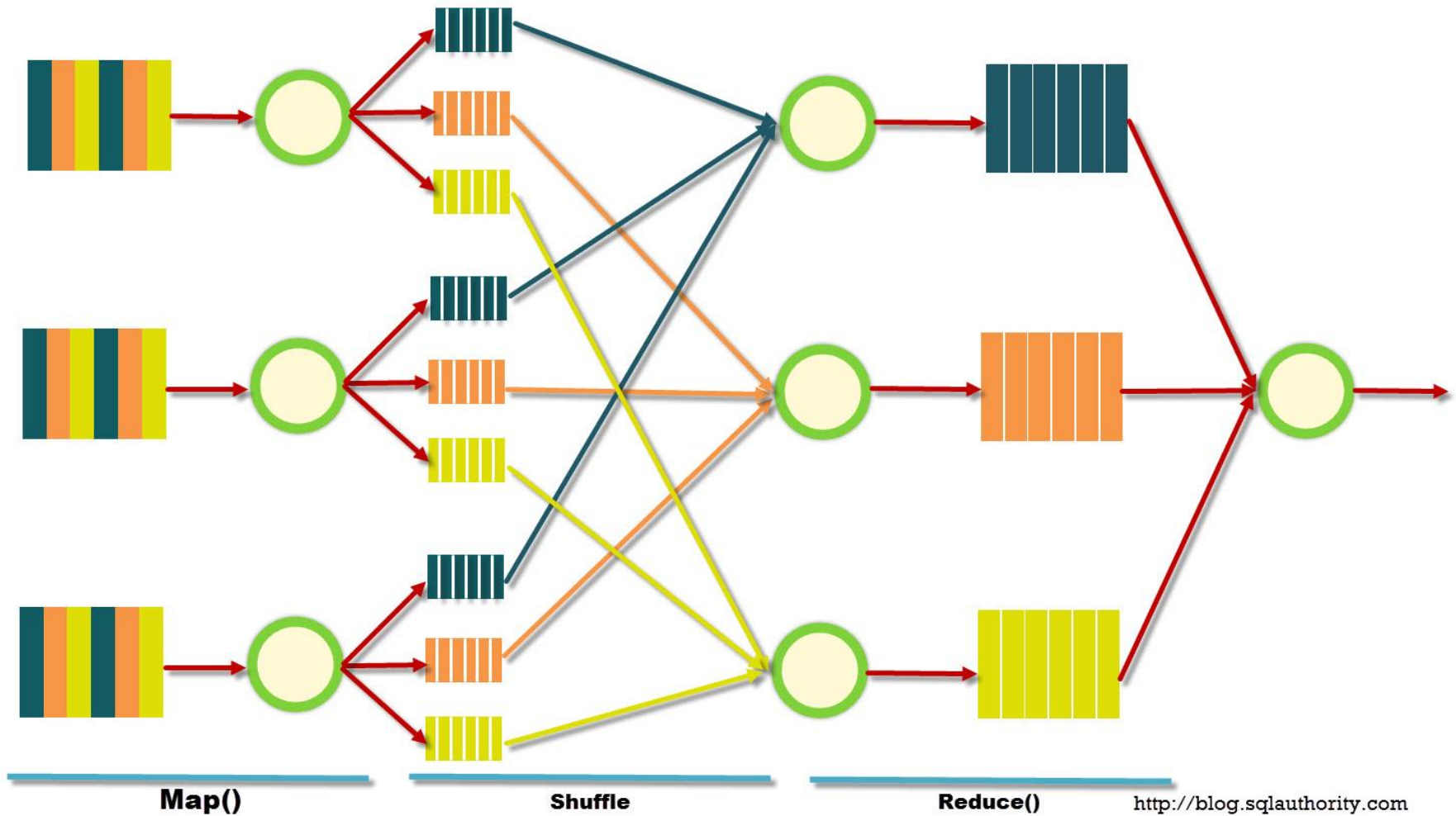
# MapReduce

Im Kern besteht der Map Reduce Ansatz aus eigenständigen Funktionen, die nacheinander, in einer sogenannten Pipeline, auf die Datensätze angewendet werden. Sowohl Map als auch Reduce stellen eigenständige benutzerdefinierte Funktionen dar.

Die vier Phasen von Map Reduce :

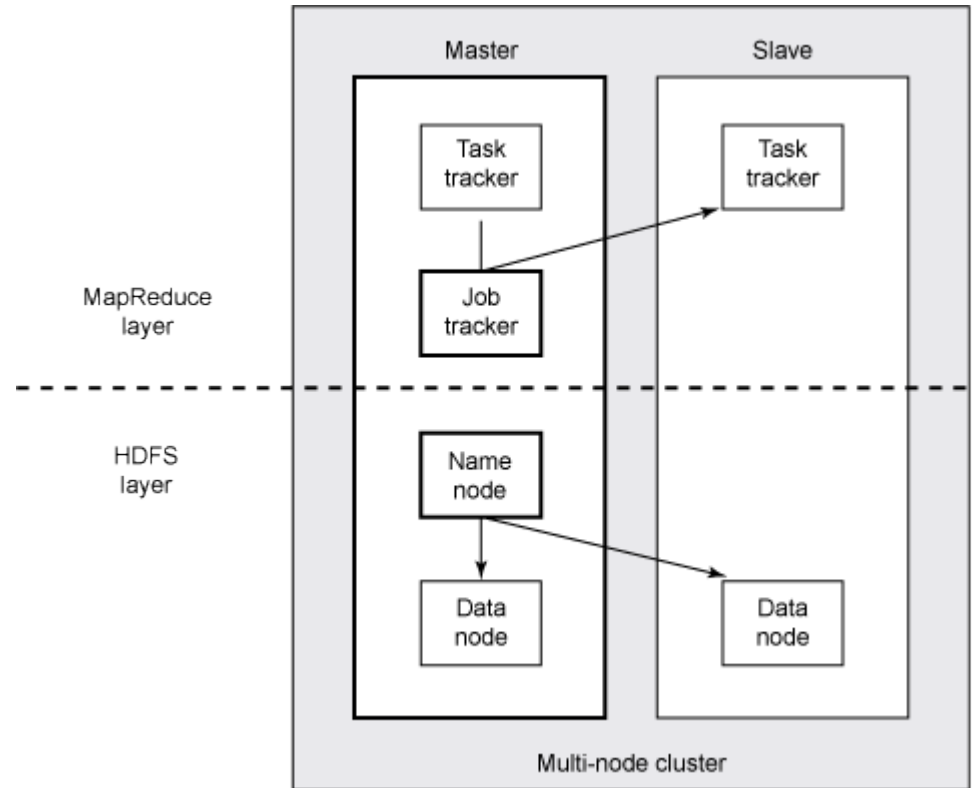
- Split: Teilt eingehende Daten in Blöcke auf und verteilt diese auf verschiedene Knoten im Cluster.
- Map: Macht aus einer Liste von Key/Value- Paaren eine andere Liste, die aus den zu verarbeitenden Key/Value- Paaren besteht, indem sie auf jedes Element der Originalliste eine benutzerdefinierte Funktion anwendet. (Die Eingabedaten lassen sich durch das Hadoop-Framework als Key/Value-Paare aufbereiten)
- Combine: Reduzierung der Daten auf dem Mapper
- Reduce: Reduziert die Liste auf weniger Werte, andere Werte oder einen einzigen Wert.

# MapReduce

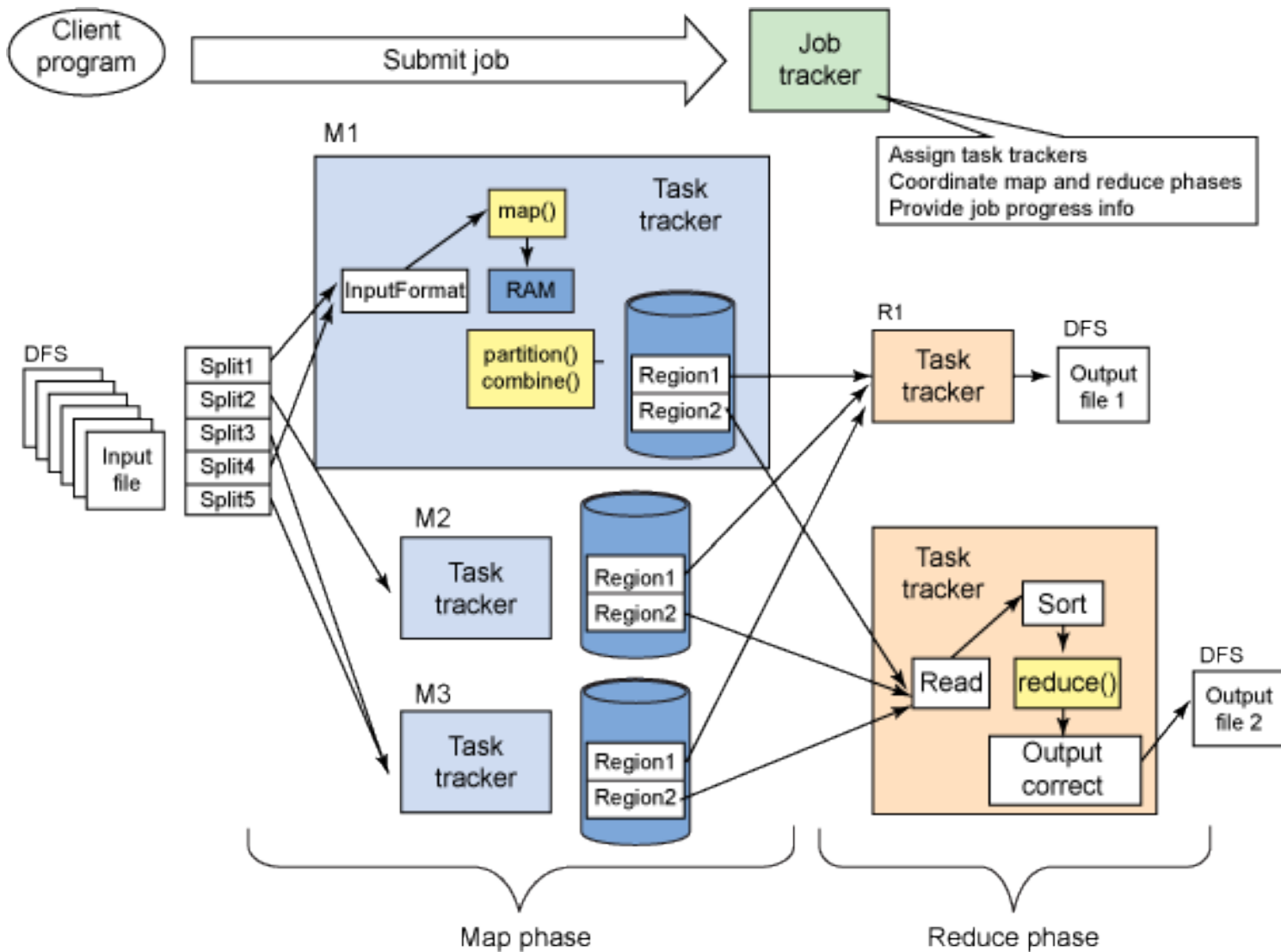


# MapReduce Framework von Hadoop

Die MapReduce- Funktion wird bei Hadoop als MapReduce- Job bezeichnet werden von den Diensten JobTracker und TaskTracker durchgeföhrt.



# MapReduce Framework von Hadoop



# Fazit

## **Big Data – Hype oder Chance?**

Betrachtung der Marktentwicklung:

- 2012 wurden ca. 4,5 Milliarden Euro in Hard- und Software sowie Services in diesem Bereich investiert. 2013 bereits ca. 6 Milliarden Euro.
- Für die nächsten Jahre wird ein Wachstum von durchschnittlich 36% prognostiziert.
- Bis 2016 soll der Umsatz mit Hard- und Software sowie Services von Big Data allein in Deutschland auf knapp 1,7 Milliarden Euro zulegen. Das entspricht einer jährlichen Steigerung um 48 Prozent und führt dazu, dass die Hälfte des europäischen Umsatzes in Deutschland gemacht wird.

## **Hadoop, die Big Data Lösung?**

Hadoop ist eine leistungsfähige Technologie, aber es ist nur ein Bestandteil der Big Data-Technologi Landschaft. Hadoop wurde für bestimmte Datentypen und Workloads entwickelt. Zum Beispiel ist es eine sehr kostengünstige Technologie für die Bereitstellung großer Mengen von Rohdaten, die dann weiter verfeinert und für die Analyse vorbereitet werden können.

Ende

Fragen?