Hadoop in modernen BI-Infrastrukturen

Dr. Stefan Igel
inovex GmbH

Zur Person

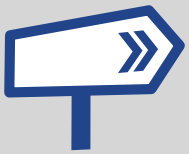
Dr. Stefan Igel

- Seit 01/2005 als Projektleiter und Systemarchitekt bei inovex
- Seit 08/2009 als Business Engineer bei 1&1 Web.Intelligence
  - Erstkontakt mit Hadoop 09/2009
  - Seit 2011 Technische Projekt-leitung für Hadoop-Projekte
- Seit 01/2013 Teamleiter Big Data Solutions bei inovex
- stefan.igel@inovex.de
- www.xing.com/profile/Stefan_Igel

Wir nutzen Technologien, um unsere Kunden glücklich zu machen.
Und uns selbst.

Zum Unternehmen Inovex Gmbh

- Inhabergeführtes IT-Projekthaus
- 120 Mitarbeiter
- Pforzheim / Karlsruhe / München / Köln
- Technologischer Fokus
- www.inovex.de



1	2	3	4	5
SHAPE Consulting	BUILD Application Development	RUN Systems Engineering	TRACK Business Intelligence	TEACH Academy

Zum Thema

- BI meets BIG DATA at 1&1
- Jede Menge Blech: Die Hardware
- Was darf es denn sein: Das Hadoop Ecosystem
- Speichern, Mappen, Reduzieren
- Spielregeln im Cluster
- Der fleißige Handwerker
- Essenz

Agenda

- **BI meets BIG DATA at 1&1**
- Jede Menge Blech: Die Hardware
- Was darf es denn sein: Das Hadoop Ecosystem
- Speichern, Mappen, Reduzieren
- Spielregeln im Cluster
- Der fleißige Handwerker
- Essenz

Das Umfeld: Beispiel 1

WEB.DE

- 1995 gestartet als Internet-Katalog
- Eines der größten deutschen Internet-Portale mit 16,44 Mio. Nutzern/Monat*
- WEB.DE FreeMail – beliebtester deutscher Mail-Dienst mit zahlreichen Sicherheitsfeatures

WEB.DE

*Unique User / AGOF internet facts 2011-12

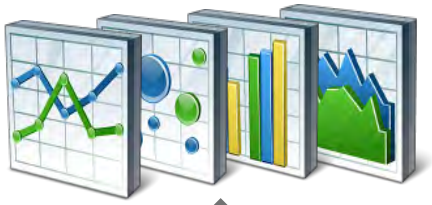
Das Umfeld: Beispiel 2

GMX

- FreeMail Pionier mit leistungsstarken Mail- und Messaging-Lösungen
- 12,75 Mio. Nutzer/ Monat*
- Umfangreiche Portalangebote, Dienste und Online-Services in Deutschland, Österreich und der Schweiz
- Internationale Mail-Angebote in sechs Ländern (A, CH, F, E, UK, USA), auch unter mail.com

*Unique User / AGOF internet facts 2011-12

Die Aufgabe | Anpassung der BI Plattform | an steigendes Datenvolumen



Die spaltenorientierte Datenbank stieß an ihre Grenzen:

Web. Intelligence BI Platform

- Verarbeitungsgeschwindigkeit nicht mehr ausreichend
- Aufrüstung teuer
- Begrenzte Ressourcen

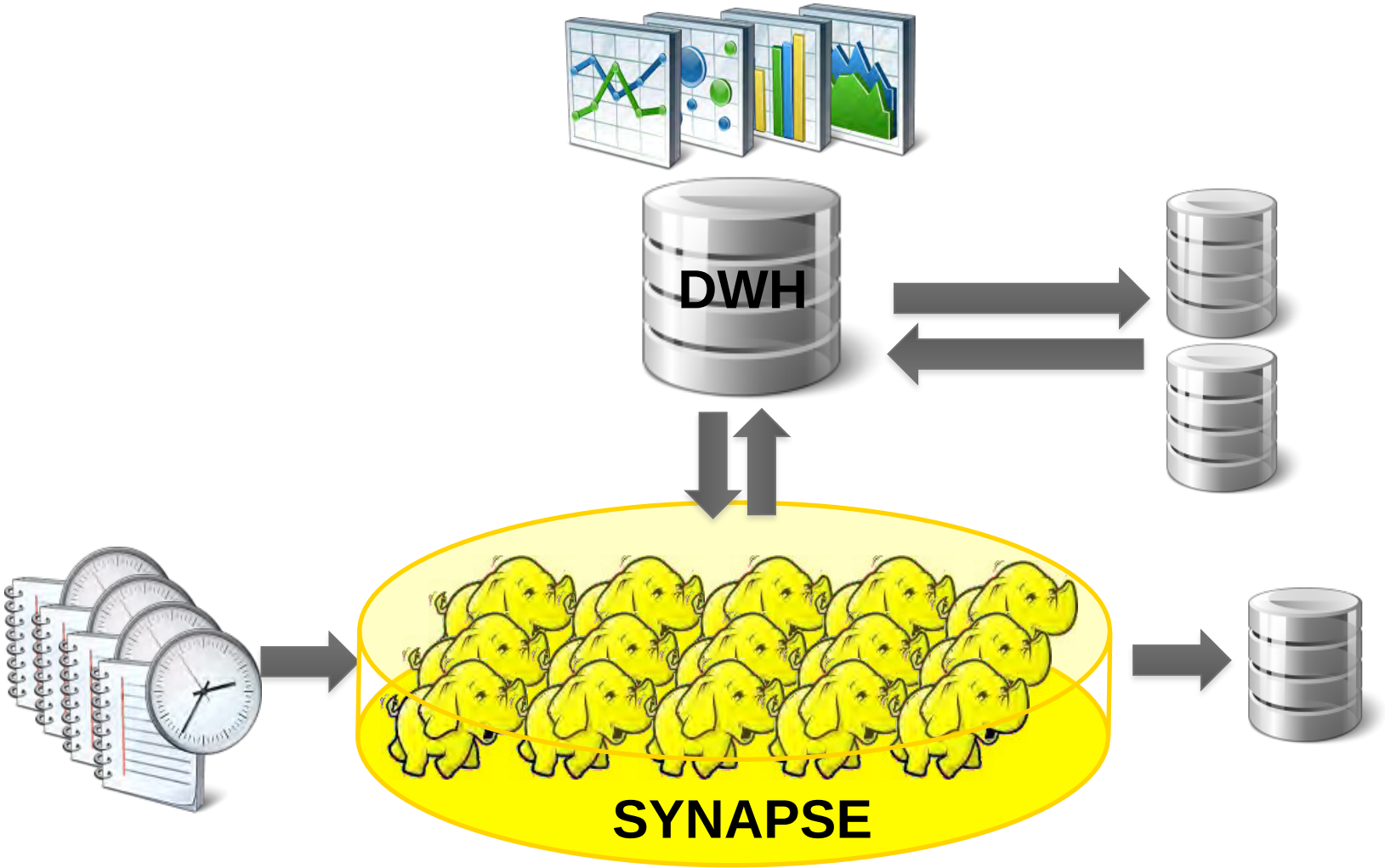
WEB.DE

1&1

GMX

Web-Analytics: 240 Files/d
200 GB/d * 90d = 18 TB
Media-Analytics: 15.000 Files/d
2.000 GB/d * 30d = 60 TB

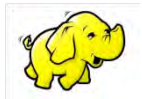
Die Lösung | SYNAPSE* | Kostengünstig und skalierbar

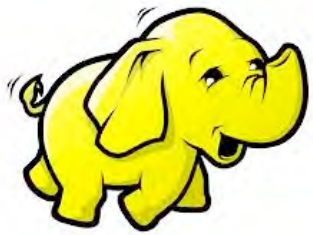


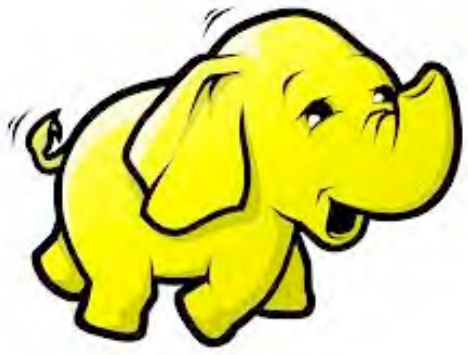
*** SYN**ergetic **A**nalytical **P**rocessing and **S**torage **E**ngine

Big Data Projekte bei 1&1 Web.Intelligence

2006	2009	2011	2012	2013
Web-Analytics mit Spalten-orientiertem Datenbank-System	Erste Evaluierung von Hadoop	**Januar:** Hadoop-Projekt „Web-Analytics“ beginnt **August:** Hadoop-Cluster mit 12 Knoten einsatzbereit **Oktober:** Hadoop verarbeitet im produktiven Einsatz ca. 200.000.000 Datensätze/ Tag	**Mai:** Cluster-Ausbau auf 20 Knoten **Juni:** Hadoop-Projekt „Media-Analytics“ geht live: 1.000.000.000 Datensätze/Tag **Dezember:** Weitere Daten-lieferstrecken	**März:** Neue Kennzahlen und Algorithmen für „Web-Analytics“ **Mai:** AdHoc-Analysen auf Rohdaten **Juli:** Real-Time Distributed Event Processing (Streaming)







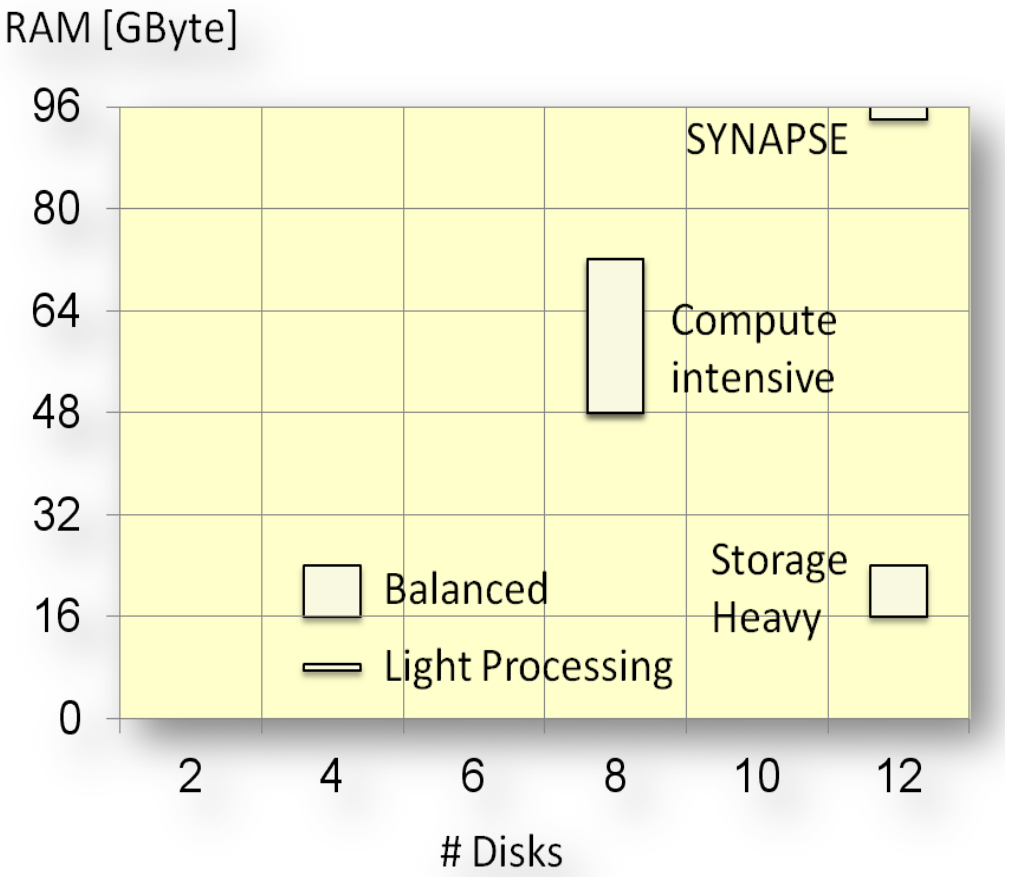
Agenda

- BI meets BIG DATA at 1&1
- **Jede Menge Blech: Die Hardware**
- Was darf es denn sein: Das Hadoop Ecosystem
- Speichern, Mappen, Reduzieren
- Spielregeln im Cluster
- Der fleißige Handwerker
- Essenz

Hardware Sizing

Hadoop Data Nodes

- Empfehlung Cloudera Support*
  - 2 Quad Core CPUs
  - Disk Drive Size 1 – 2 TByte
- Auswahl-Kriterien
  - Kosten-Effizienz
  - Stromverbrauch
  - Platz im Rechenzentrum
  - Standard-HW-Warenkorb
  - Vorbereitete Appliances



* http://www.cloudera.com/blog/2010/03/cloudera s-support-team-shares-some-basic-hardware-recommendations

Sizing der SYNAPSE

BI-Szenarien sind „Storage Heavy“

- Sizing ist bestimmt durch erforderliche Speicherkapazität: **480 TB**
  - 78 TB Nutzdaten (Web- + Media-Analytics)
  - Faktor 2 für Zwischenergebnisse und Temp-Files des Task Tracker
  - Replikationsfaktor 3 zur Datensicherung
- 480TB / 2TB / 12 = **20 Rechner** x 8 Cores = **160 CPUs!**
- **96 GB RAM**
  - DC Limit
  - Reserve vermeidet OOMEs

Unsere Definition von „Commodity Hardware“

4x 1GB Network

12 x 2TB as JBOD
direct attached

8 Cores @2,4GHz
96 GB RAM

20 x Worker-Node
(Datanode, Tasktracker)

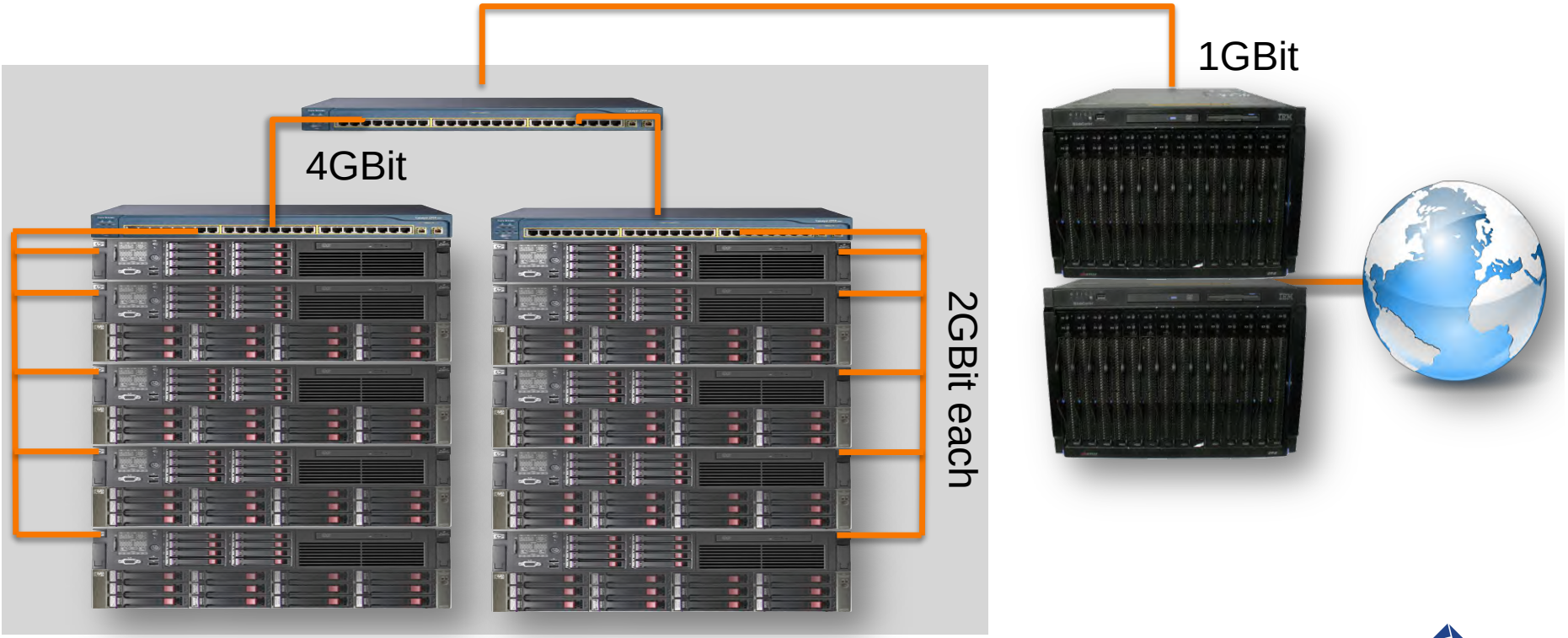
50GB local Storage

4 Cores @2,4GHz
12 GB RAM

2 x Namenode Server
(Namenode, Sec.-Namenode, Jobtracker)

Netzwerk

- Die Knoten eines Hadoop-Clusters sollten möglichst „nahe“ beieinander stehen
- und auch bei den Datenquellen
- dafür muss Platz im Rechenzentrum verfügbar sein
- **Je besser die Netzwerk-Anbindung des Clusters desto besser seine Performance**



Wie kommen die Files…

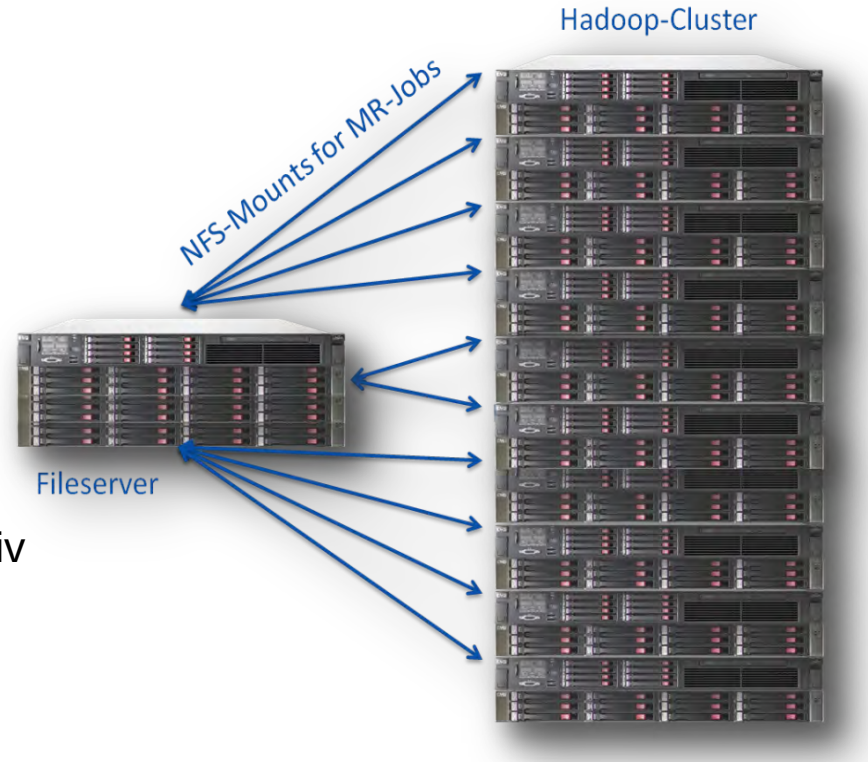
…ins Cluster?

Problemstellung

- Paralleler Zugriff der Tasktracker über zentralen Fileserver via NFS-Mounts (Import Rohdaten, Export von Aggregaten)
- Konkurrierende Zugriffe reduzieren den Durchsatz überproportional

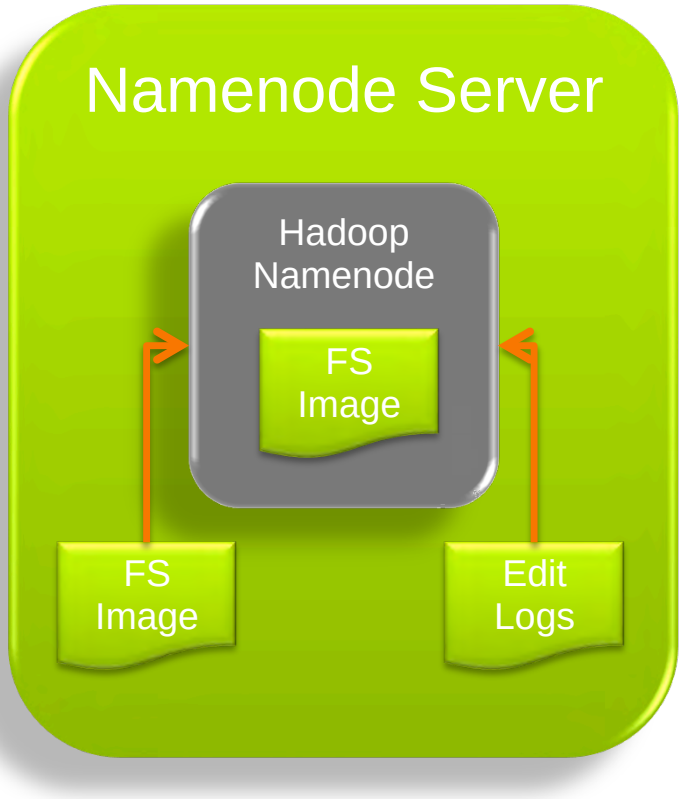
Lösung

- Reduktion der Zugriffslast
  - Mit einem Prozess im/exportieren (hadoop –fs copyFrom/ToLocal)
  - z. B. via Fair-Scheduler Pools (fair-scheduler.xml: maxMaps and maxReduces)
- Externe Systeme verwenden, die auf massiv parallele Verarbeitung ausgelegt sind:
  - z. B. Flume oder Storm für Import



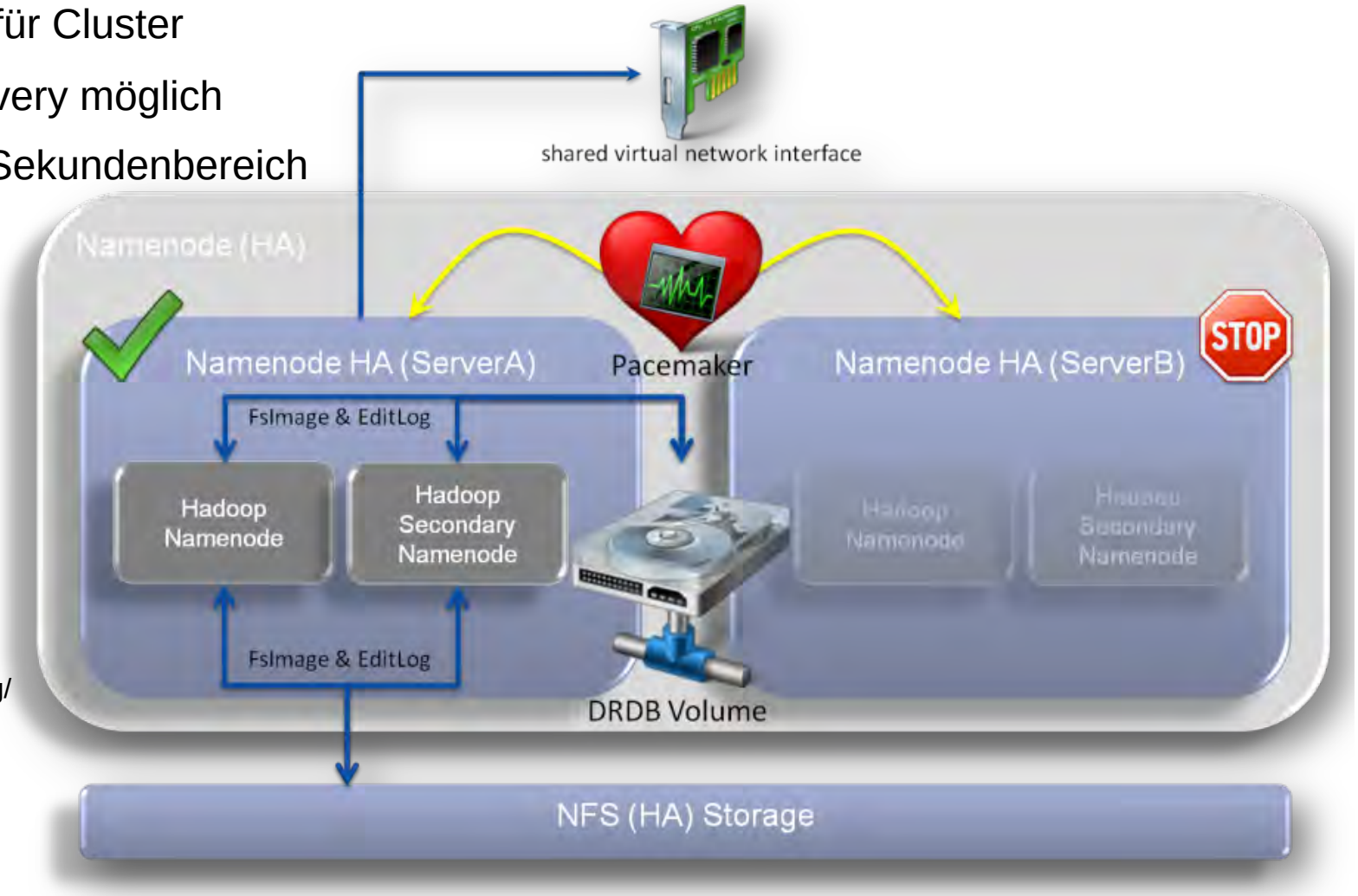
Namenode

- Namenode = HDFS-Metadaten-Verwaltung (Dateien, Verzeichnisse,…)
- Es gibt (in Hadoop 1) nur genau EINE Namenode im Cluster
- Fällt diese aus oder wird korrupt, gibt es „keine Daten“ mehr im Cluster!
- Die Secondary Namenode dient lediglich dem Verdichten der persistierten Änderungslogs
- Die Namenode ist (in Hadoop 1) ein SPOF und muss entsprechend abgesichert werden!



Namenode HA … selbst gestrickt

- Redundanz für Namenode
- transparent für Cluster
- Crash Recovery möglich
- Failover im Sekundenbereich



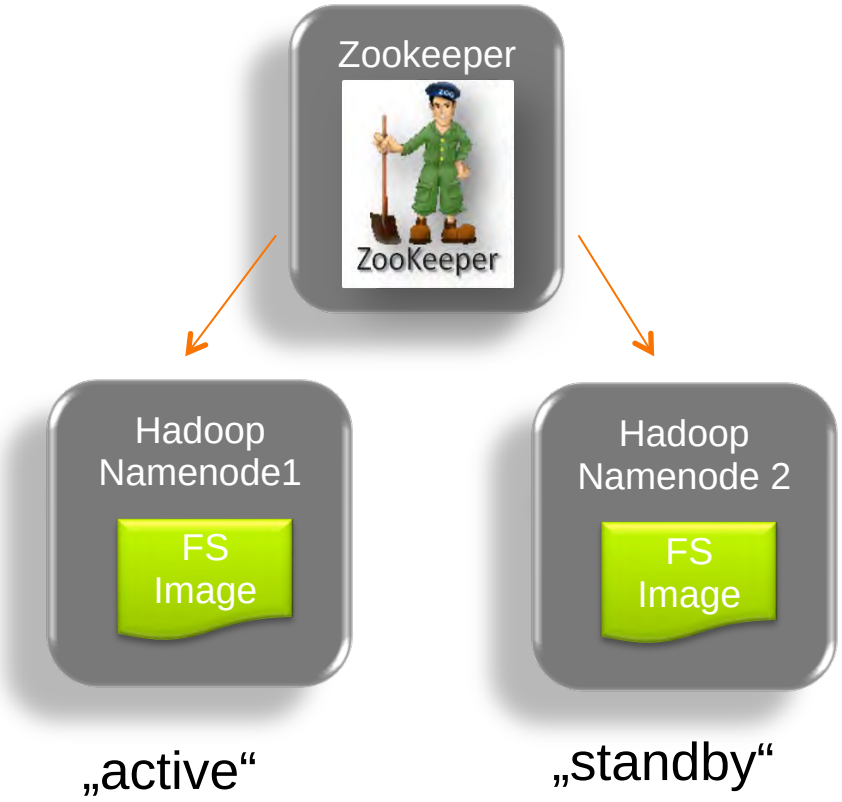
Pacemaker

http://www.clusterlabs.org/

DRBD

http://www.drbd.org/

- „Namenode HA“ in Hadoop 2 ootb
  - „Federated Mode“ : 2 Namenodes verwalten 1 Cluster
- Automatisches Failover koordiniert durch Zookeeper-Infrastruktur
  - 1 Namenode „active“
  - 1 Namenode „standy“
- **Hadoop 2 hat noch Beta-Status**



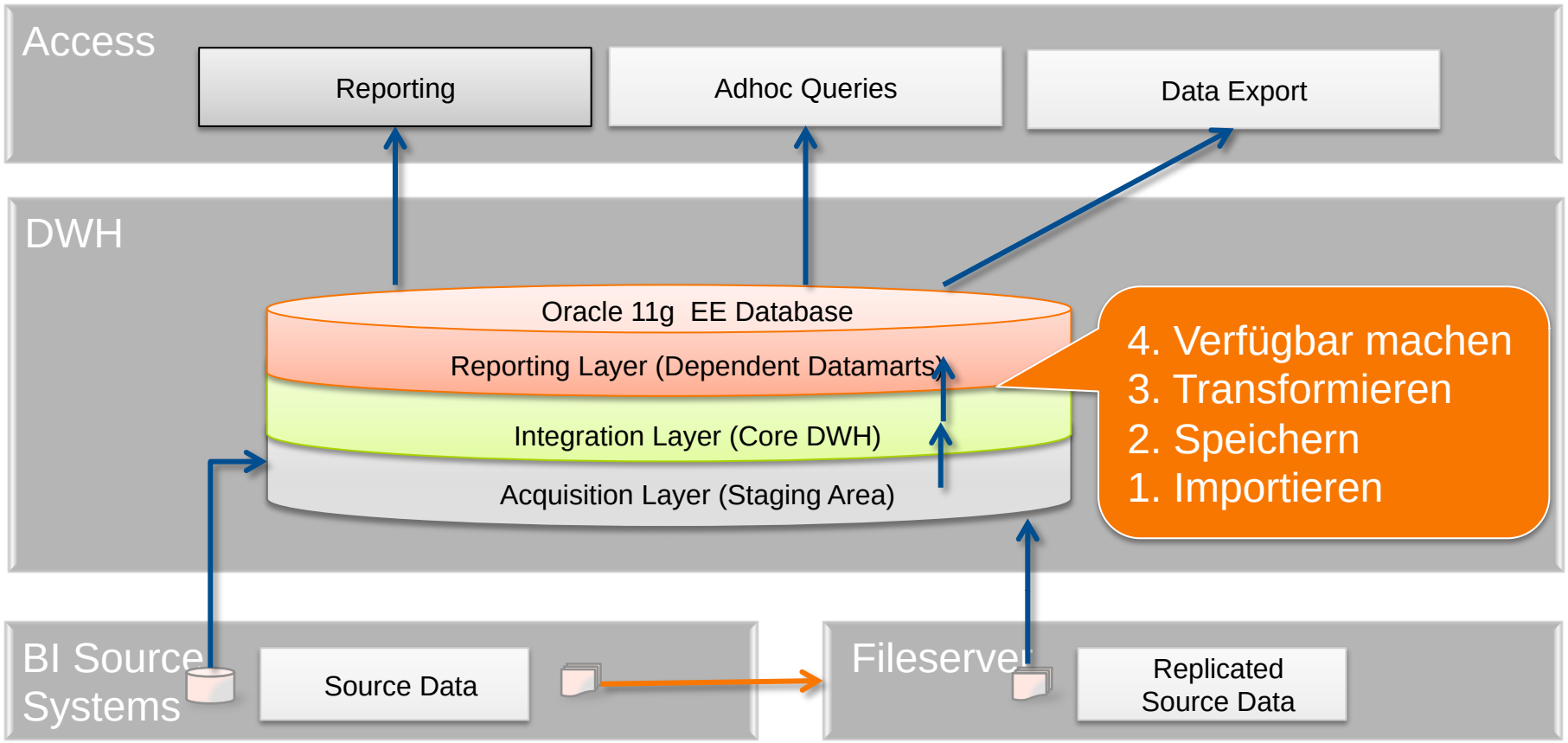
Agenda

- BI meets BIG DATA
- Jede Menge Blech: Die Hardware
- **Was darf es denn sein: Das Hadoop Ecosystem**
- Speichern, Mappen, Reduzieren
- Spielregeln im Cluster
- Der fleißige Handwerker
- Essenz

Wir machen unsere Werkzeugkiste auf | und schauen mal, | was wir darin finden ...

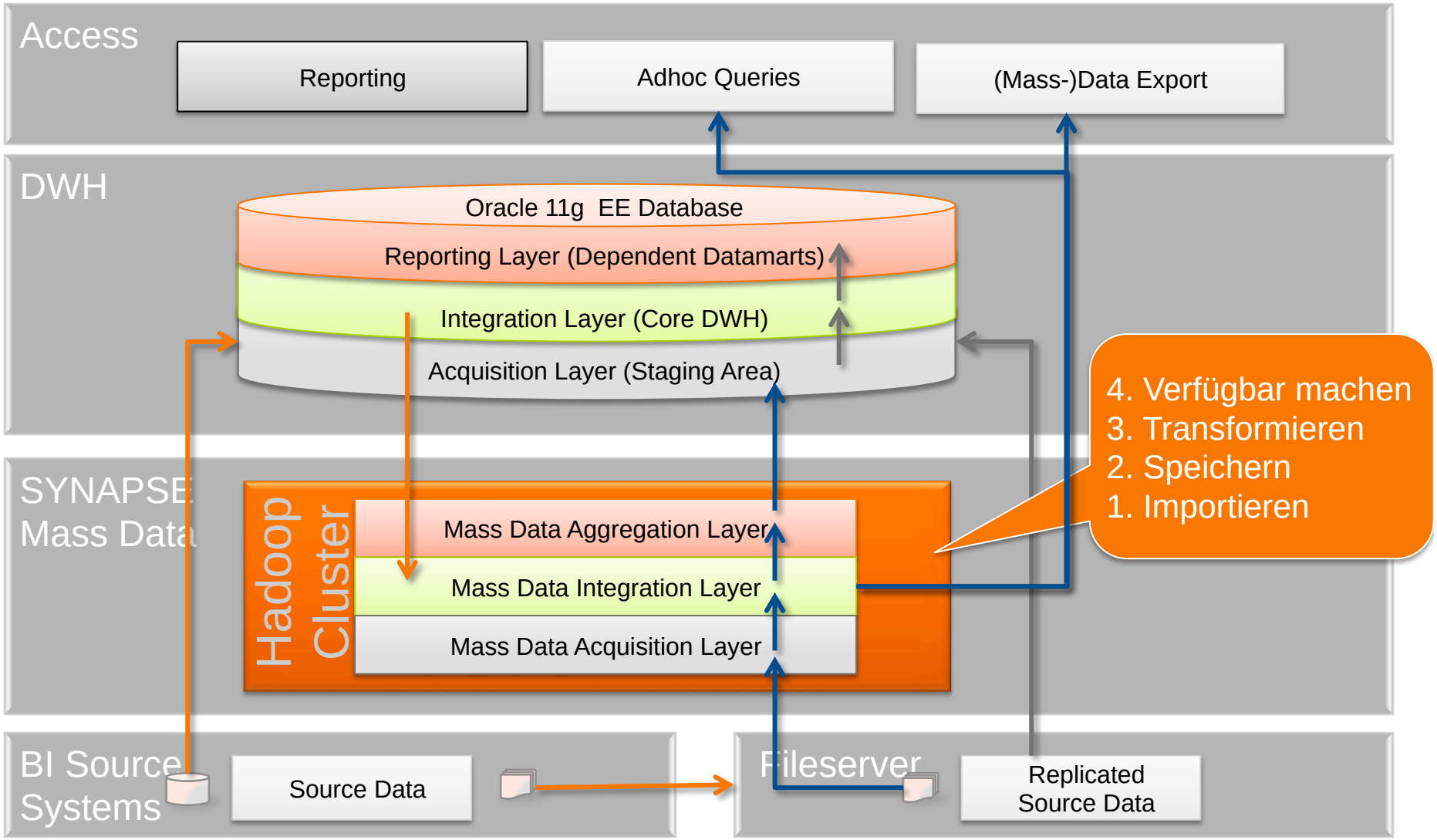
Architektur BI-Plattform

Datawarehouse



Architektur BI-Plattform

DWH + SYNAPSE



- Ähnliche Toolsets unterscheiden sich oft nur in Details
  - **Queries:** Pig, Hive, Cascading, …
  - **KV-Stores:** HBase, Cassandra, …
  - **Streaming:** Flume, Storm, …
  - **Steuerung:** Askaban, Oozie, …
  - **Serialisierung:** Avro, Thrift, Protocol Buffers, …
  - …
- Aber gerade die können manchmal entscheidend sein!
  - Manchmal landet man auch in Sackgassen!

Framework Auswahl | Stabilität

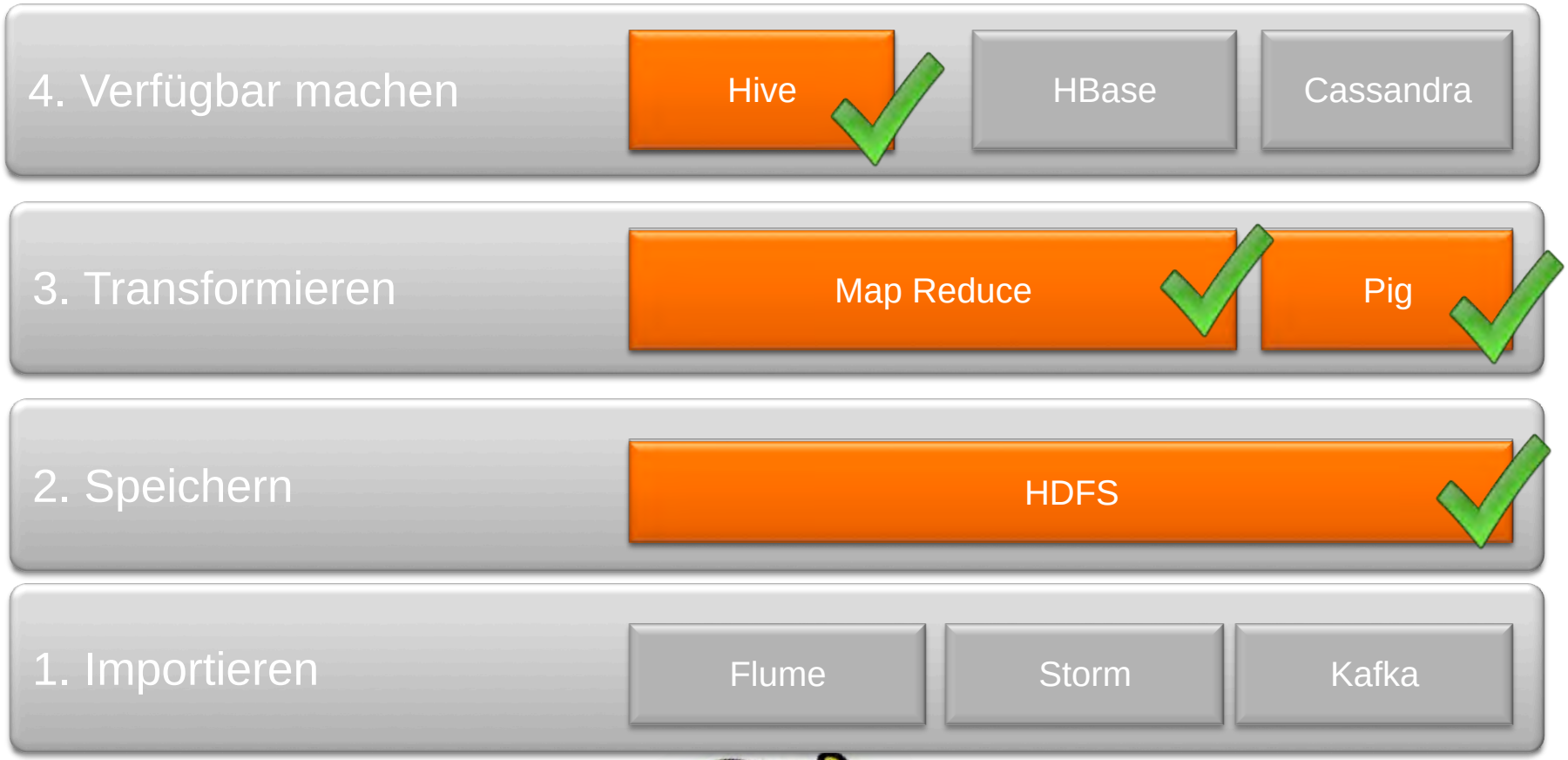
- Das Hadoop-Ecosystem entwickelt sich rasant weiter
- Hadoop- Kern ist bereits sehr stabil und für den produktiven Einsatz geeignet - aber (noch) kein Vergleich bspw. mit einer DB-Lösung
- Manche Sub- und Neben-Projekte erscheinen noch nicht ausgereift – im Handumdrehen ist man Committer eines OpenSource-Projektes ;-)
- Es kann sich lohnen, nach ein paar Monaten wieder vorbeizuschauen!
- Abgestimmte Software-Versionen der Toolsets erforderlich!

Schön, dass es Distributionen gibt! | Hadoop-Solution-Provider ... | geben Sicherheit

- Beispiel: Cloudera
- Eigenes Hadoop-Release basierend auf den Apache Repositories
- bietet konsistente, „in sich schlüssige Release-Stände“, Patches und Bugfixes
- Möglichkeit des kommerziellen Supports

Der Technologie-Stack

weniger ist mehr

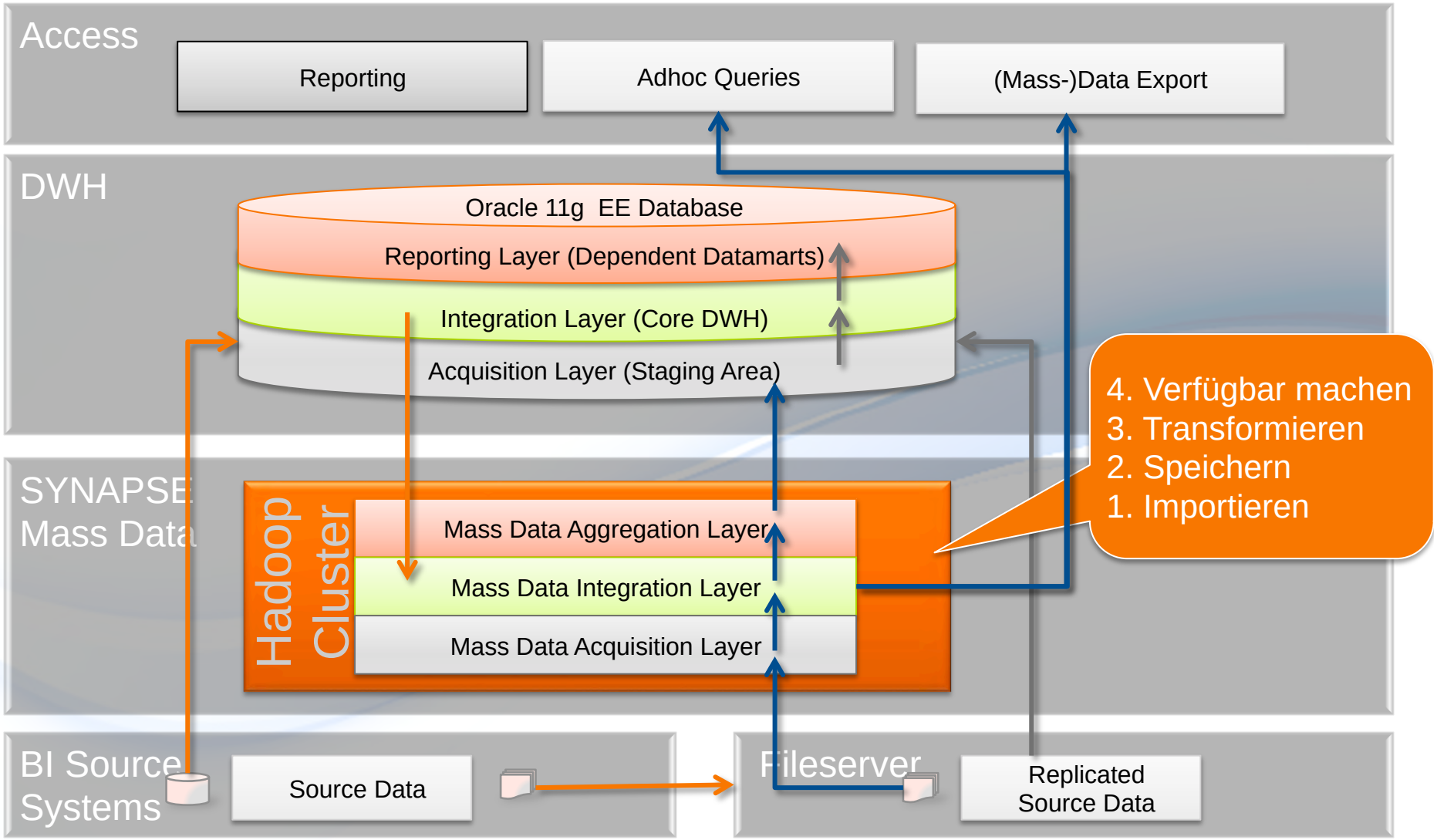


Agenda

- BI meets BIG DATA at 1&1
- Jede Menge Blech: Die Hardware
- Was darf es denn sein: Das Hadoop Ecosystem
- **Speichern, Mappen, Reduzieren**
- Spielregeln im Cluster
- Der fleißige Handwerker
- Essenz

Architektur BI-Plattform

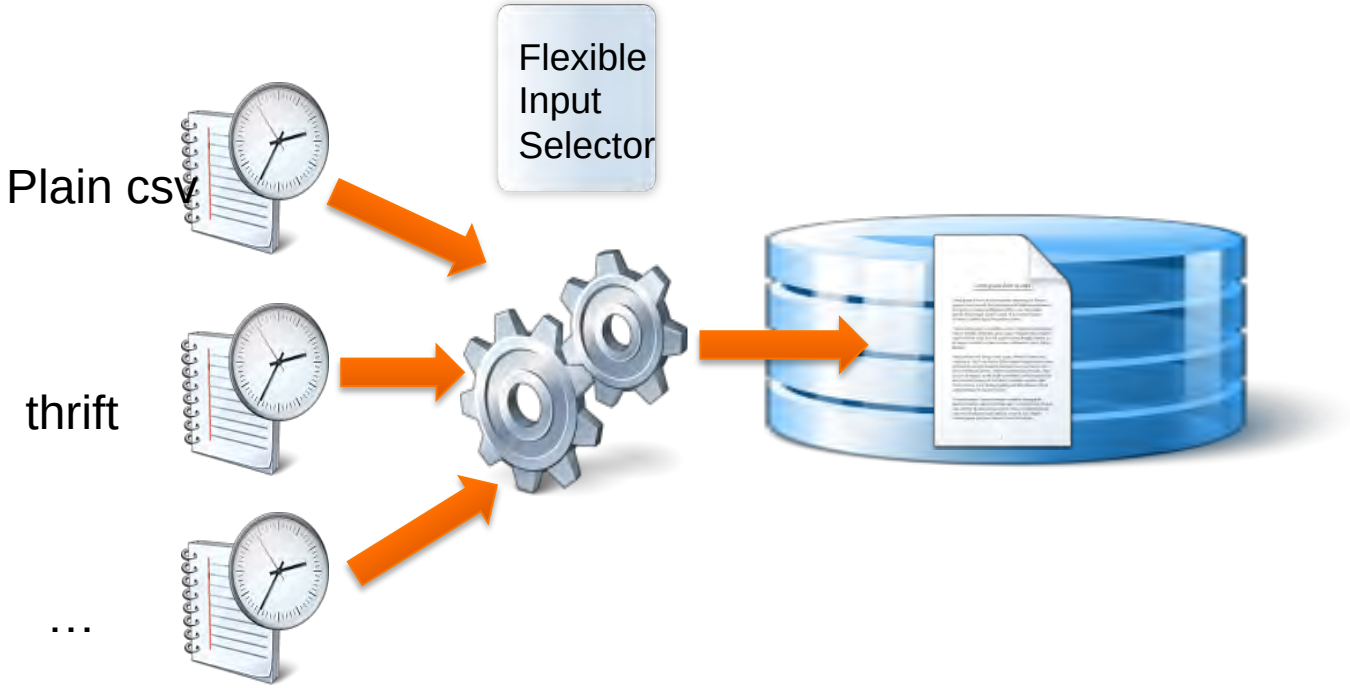
SYNAPSE



Importieren

Herausforderungen

- Umgang mit verschiedenen Datenformaten
  - Semi-strukturierte Tracking-Logfiles im CSV-Format
  - Thrift-Logfiles
- Konvertieren in internes Format
- Umgang mit “Late Arrivals”
- Sicherstellen einer genau einmaligen Verarbeitung der Datensätze



Speichern

Das Serialisierungs-Framework

- **Serialisierung**
  - Custom Hadoop-Writables
  - Thrift
  - Avro
  - Google Protocol Buffers
- **Kompression**
  - Snappy / LZO: Trade Offs Kompressionsrate vs. CPU
- **Partitionierung der Daten**
  - Struktur der Datenablage sauber konzipieren
  - Weiterverarbeitung ermöglichen

```
# avro textual
representation
{"type": "record", "name":
"Point",

 "fields": [

  {"name": "x", "type":
"int"},

  {"name": "y", "type":
"int"}

 ]
}
5 8
-3 4
2 -7
```

Transformieren

Java Map Reduce

- Low Level Programmierung von Map-Reduce-Jobs
- Die volle Mächtigkeit von MR steht zur Verfügung
- Java Know-How erforderlich
- Kenntnis über Hadoop-Internas für effiziente Implementierung erforderlich
- Design Pattern helfen bei Standard-Problemen (Sort, Join, ...)
- Bibliotheken entstehen

```
map(key , value) {
 // key = byte offset in log file
 // value = a line in the log file
 if ( value is an account access
      audit log) {
  account number = parse account
  from value
  output key = account number,
  value = 1
 }
}

reduce(key, list of values) {
 // key = account number
 // list of values {1,1,1,1.....}
 for each value
  count = count + value
 output key , count
}
```

http://www.javacodegeeks.com/2011/05/mapreduce-soft-introduction.html

Transformieren

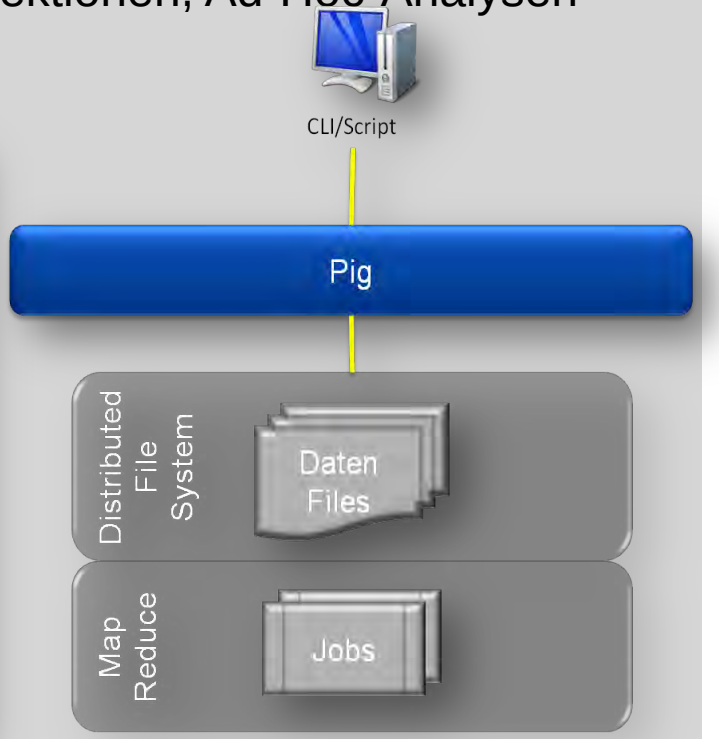
Abstraktion durch PIG

- Ursprünglich von Yahoo entwickelt
- File basiertes High Level Daten-Manipulations-Framework
- Deskriptive Sprache, die durch das Framework in Map/Reduce Jobs übersetzt und ausgeführt wird
- Gut geeignet für Aggregationen, Selektionen/Projektionen, Ad-Hoc-Analysen
- http://hadoop.apache.org/pig/

SQL	Apache Pig
SELECT spalte1 AS alias1, ...	alias = FOREACH { gen_blk | nested_gen_blk };
SELECT ... FROM tabelle	LOAD 'data' [USING function] [AS schema];

Cross product (SELECT ... FROM tabelle1, tabelle2, ...)	alias1 = LOAD 'data1' [USING function] [AS schema];
	alias2 = LOAD 'data2' [USING function] [AS schema];
	alias3 = CROSS alias1, alias2 [, alias3 ...] [PARTITION BY partitioner] [PARALLEL n];

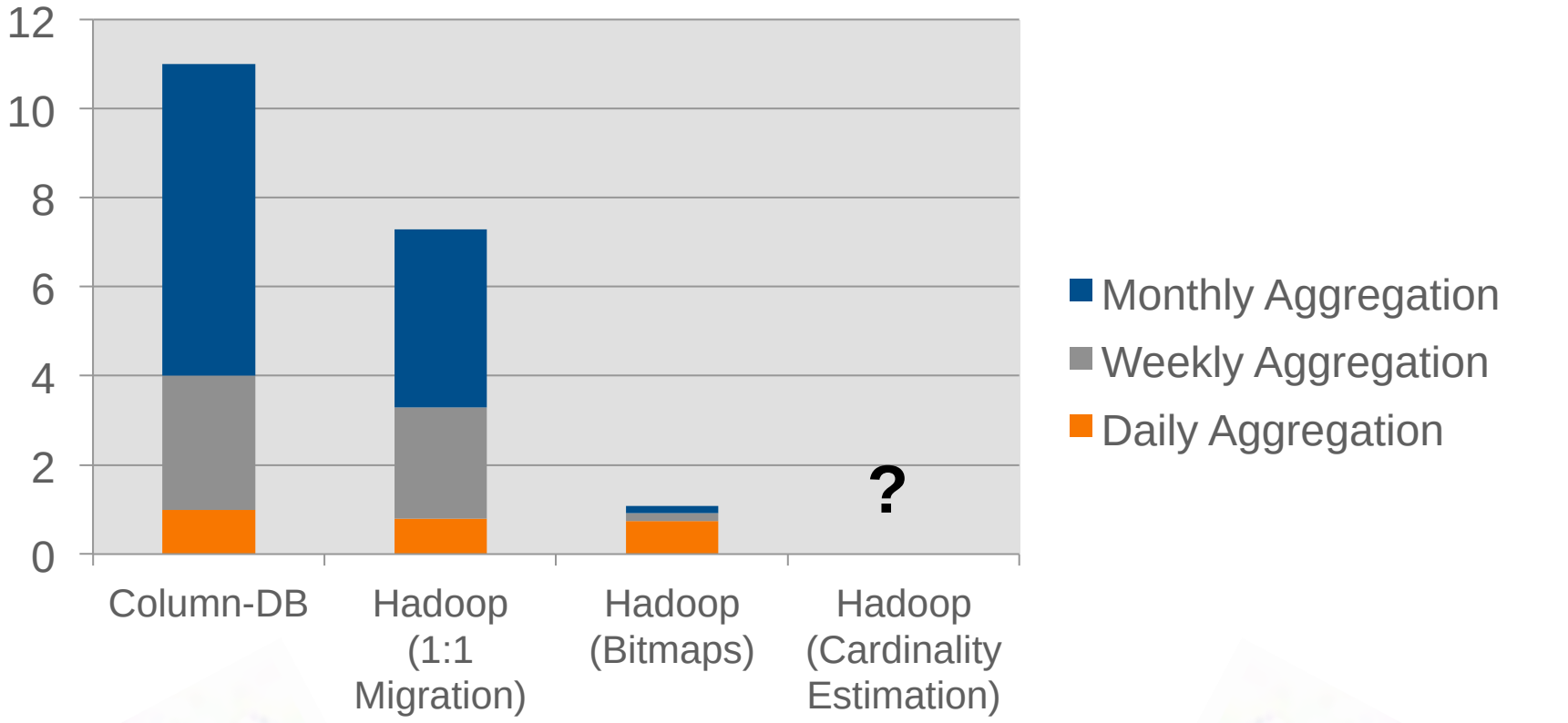
INNER JOIN	alias = JOIN alias BY {expression['('expression [, expression ...]')'} (, alias BY {expression['('expression [, expression ...]')'} ...) [USING 'replicated' | 'skewed' | 'merge'] [PARTITION BY partitioner] [PARALLEL n];
OUTER JOIN	alias = JOIN left-alias BY left-alias-column [LEFT|RIGHT|FULL] [OUTER], right-alias BY right-alias-column [USING 'replicated' | 'skewed' | 'merge'] [PARTITION BY partitioner] [PARALLEL n];
WHERE Klausel (außer Join)	alias = FILTER alias BY expression;
GROUP BY	alias = GROUP alias { ALL | BY expression} [, alias ALL | BY expression ...] [USING 'collected' | 'merge'] [PARTITION BY partitioner] [PARALLEL n];
DISTINCT	alias = DISTINCT alias [PARTITION BY partitioner] [PARALLEL n];
COUNT, SUM, AVG	COUNT, SUM, AVG
FETCH FIRST n ONLY	alias = LIMIT alias n;
ORDER BY	alias = ORDER alias BY { * [ASC|DESC] | field_alias [ASC|DESC] [, field_alias [ASC|DESC] ...] } [PARALLEL n];
-	SAMPLE alias size;
User defined Functions	alias1 = MAPREDUCE 'mr.jar' STORE alias2 INTO 'inputLocation' USING storeFunc LOAD 'outputLocation' USING loadFunc AS schema [`params, ... `];
INSERT INTO tabelle	STORE alias INTO 'directory' [USING function];
SELECT ... UNION SELECT ...	alias = UNION [ONSCHEMA] alias, alias [, alias ...];

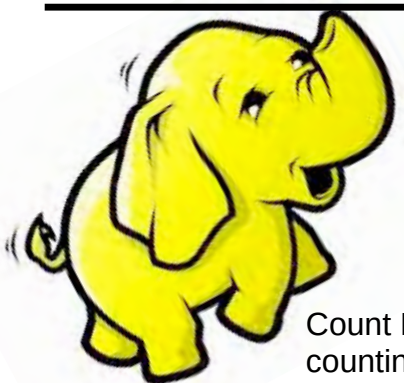


Transformieren

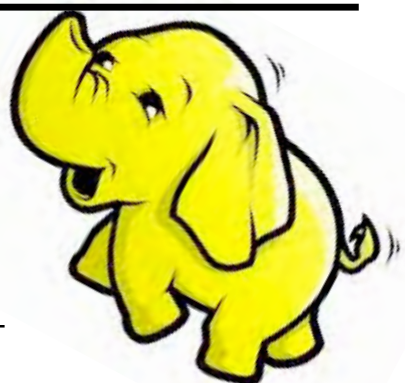
Performance-Check

Beispiel: Count Distinct





Count Distinct Approaches: http://highscalability.com/blog/2012/4/5/big-data-counting-how-to-count-a-billion-distinct-objects-us.html



Verfügbar machen ..mit Hive.

- Ad-hoc Analysen per Console oder Web-Frontend
- Anbindung existierender Reporting- und Analysetools über
  - JDBC
  - ODBC
- HiveQL als SQL-artige Abfragesprache
- Erzeugt wie PIG transparent MR-Jobs
  - Batch-Prozess
- Dateien im HDFS organisieren
  - geeignet partitionieren und indizieren
  - Hive-Metastore DB enthält Metadaten

Testen und Debuggen

- Rapid Prototyping mit Hadoop-Streaming
- Debugging im Local Mode:
  - 1 JVM, Remote-Debugging
  - Alle Logfiles in 1 Verzeichnis
- Debugging Pseudo-Distributed
  - Logfile Analyse
  - Zugriff über Browser Interface
- Debugging im Cluster: Counter
  - Gemeinsam

Automatisiertes Testen ...mit Fitnesse

- Freies OpenSource Testframework
- Kollaboratives Testen (e2e)
- Komplementär zum Unit Test
- Testdefinition im „ausführbaren Wiki“
- Verwendete Fixtures:
  - DBSlim (DB-Anfragen an Oracle, Hive, …)
  - CommandLine (Shell-Zugriff)
  - PigConsole (Zwischenergebnisse im HDFS)
  - FundiCLI (Steuerung Fundi-batch)
  - WriteCsvFile (Erzeugen von CSV-files)
- Setup-Suite zur Testvorbereitung
- Test-Suites zum testen

The fully integrated standalone wiki, and acceptance testing framework: www.fitnesse.org

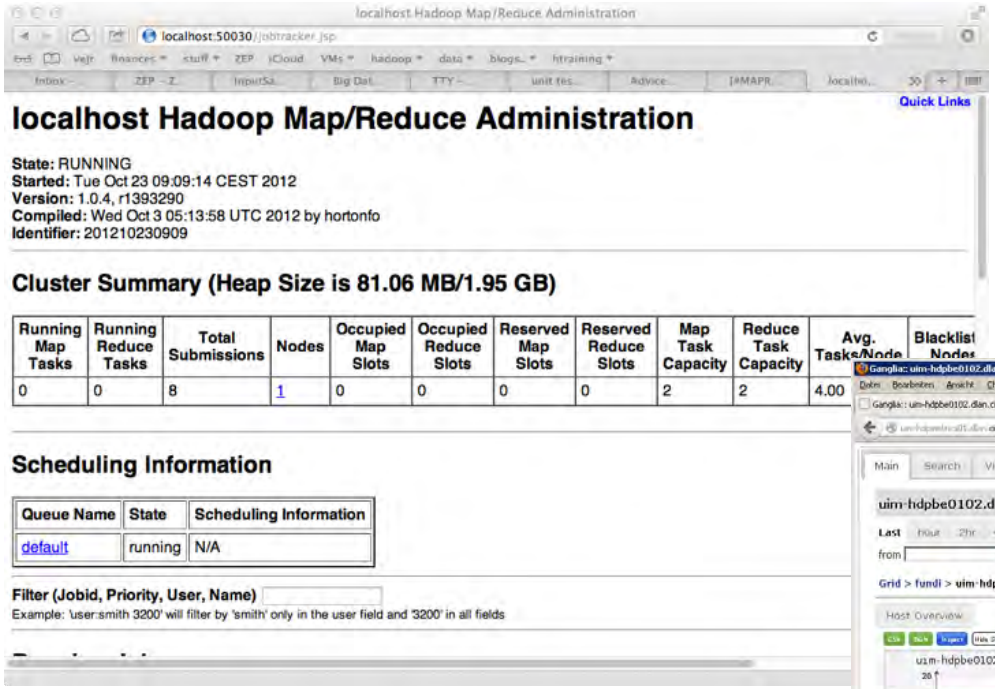
Agenda

- BI meets BIG DATA at 1&1
- Jede Menge Blech: Die Hardware
- Was darf es denn sein: Das Hadoop Ecosystem
- Speichern, Mappen, Reduzieren
- **Spielregeln im Cluster**
- Der fleißige Handwerker
- Essenz

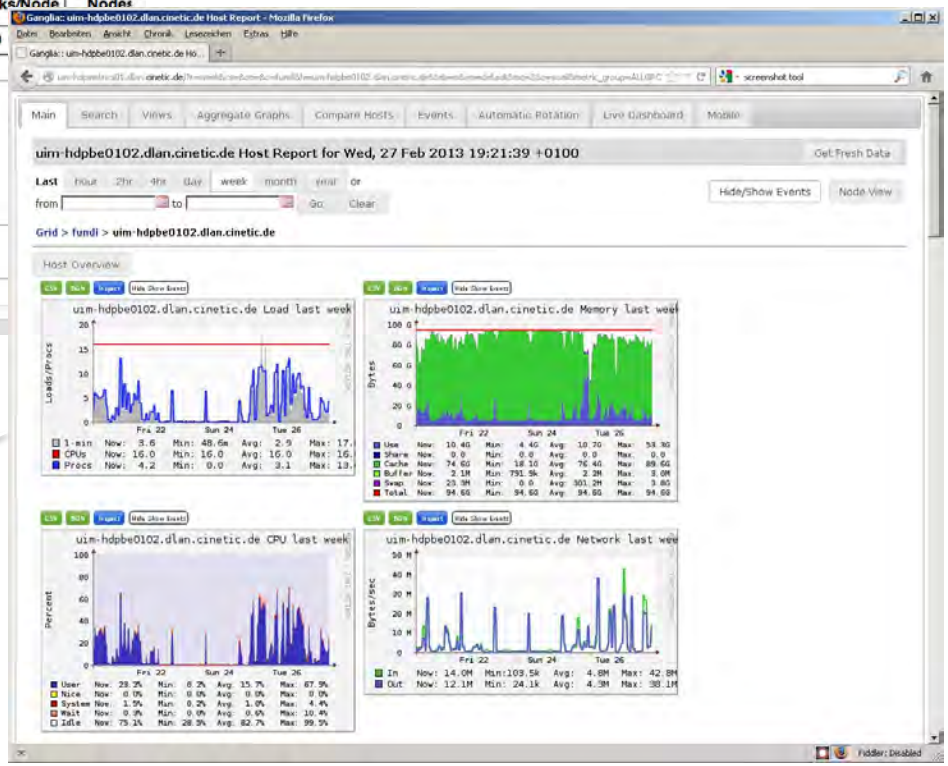
Monitoring

... mit der Admin-Console

...und Ganglia



Hadoop Jobtracker



Ganglia

Mehrparteien-Betrieb

Wer darf wann?

Hadoop Job Scheduler

- Gleichmäßige Lastverteilung über die Zeit nach Prioritäten
- Verschiedene Anwendungen können konkurrierend betrieben werden
- Ermöglicht Adhoc-Queries mit definierten Ressourcen

	Default	Capacity	Fair
Mechanismus	Vergeben von Prioritäten pro Job	Job-Queues mit festgelegten Prioritäten	Job-Queues und Pools mit Gewichten
Funktionsfähig	Ja	Ja	Ja
Clusterauslastung	Ja	Nein	Ja
Gefahr von Starvation	Ja	Nein	Nein

Mehrparteien-Betrieb

Wer darf überhaupt?

- Hadoop hat ein Zugriffsberechtigungskonzept angelehnt an POSIX (ohne sticky, setuid or setgid bits) für Files und Directories
- Hadoop hat keine eigene Benutzer-Authentifizierung, sondern übernimmt user name (whoami) und group name (bash -c groups) vom aufrufenden Client-Prozess
- Geeignetes Konzept für „Tool“-User oder Application Manager / Data Scientists

> *“… This user identity mechanism combined with the permissions model allows a cooperative community to share file system resources in an organized fashion.”*

- „Sichere Hadoop Cluster“: Authentifizierung mittels Kerberos

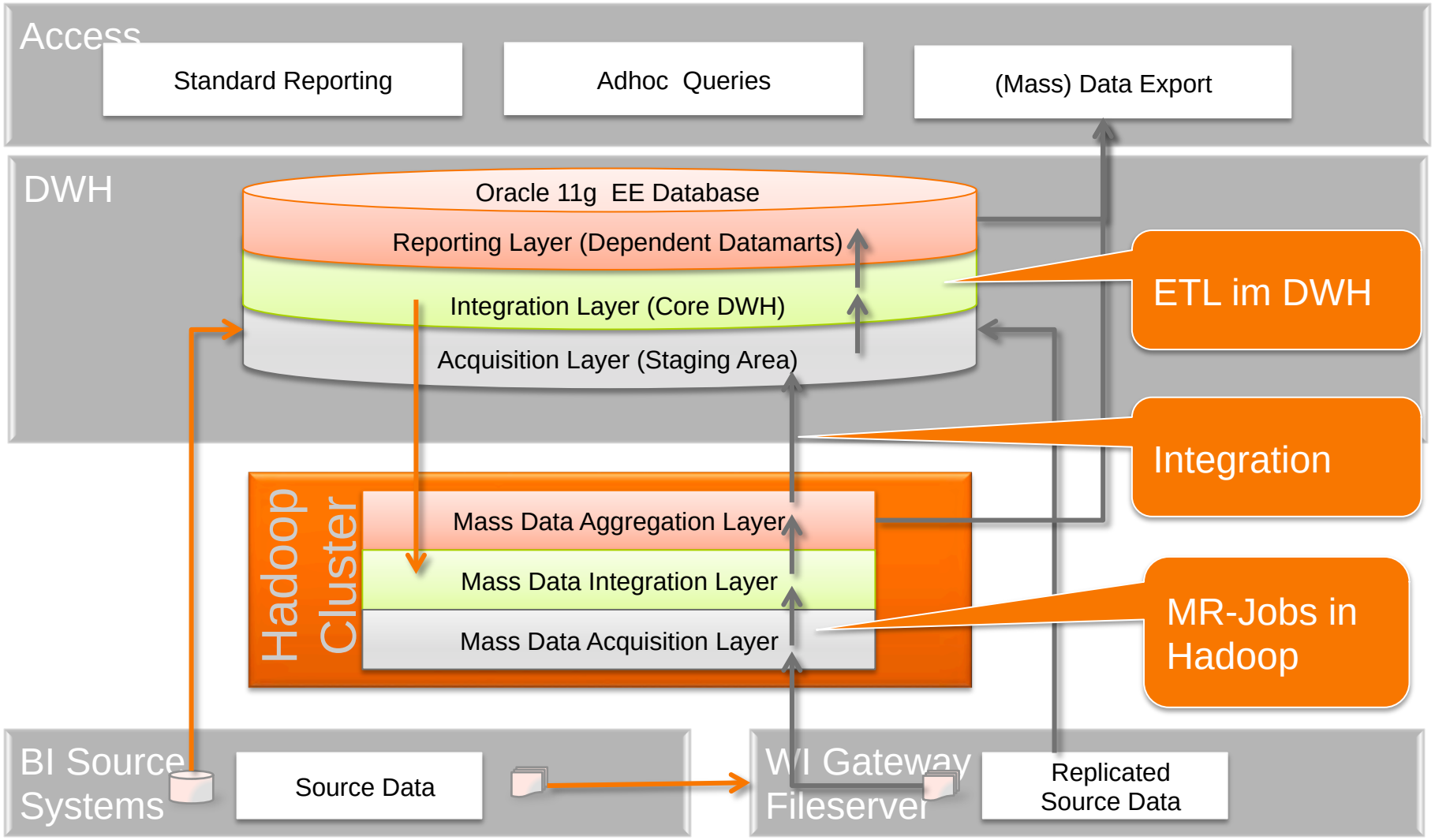
http://hadoop.apache.org/common/docs/r0.20.2/hdfs_permissions_guide.html

Agenda

- BI meets BIG DATA
- Jede Menge Blech: Die Hardware
- Was darf es denn sein: Das Hadoop Ecosystem
- Speichern, Mappen, Reduzieren
- Spielregeln im Cluster
- **Der fleißige Handwerker**
- Essenz

BI-Plattform

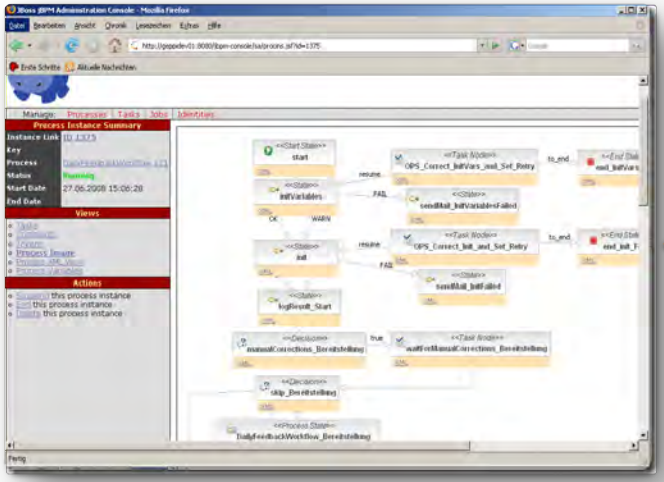
Prozess-Steuerung



- Steuerung von ETL-Prozessen im Batch-Mode
- Steuerung Datei-basierter Datenflüsse
- Verteilung der ETL-Algorithmen auf dem Cluster
- Integration in übergreifende Prozessketten
- Unterstützung des Information Lifecycle Managements (Housekeeping)
- Nachvollziehbarkeit von Ladeprozessen
- Fehler-Toleranz durch Wiederaufsetzbarkeit
- Technisches und fachliches Monitoring

Das richtige Werkzeug für die jeweilige Aufgabe

Steuerung der Verarbeitung und damit der Datenströme muss über den gesamten BI-Stack sichergestellt sein!



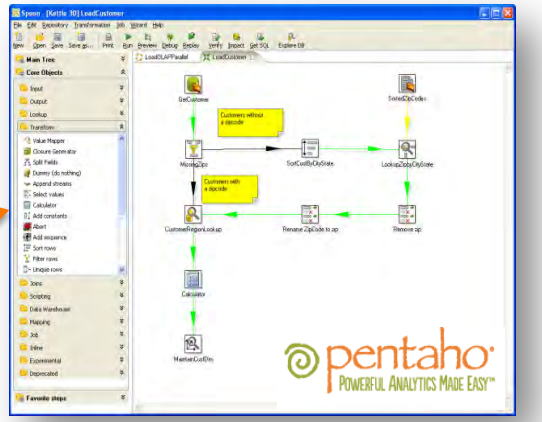
GEPPI = Workflow-Engine
Übergreifende-Steuerung

delegiert an Pentaho Kettle

delegiert an FUNDI

PDI (http://kettle.pentaho.com/)

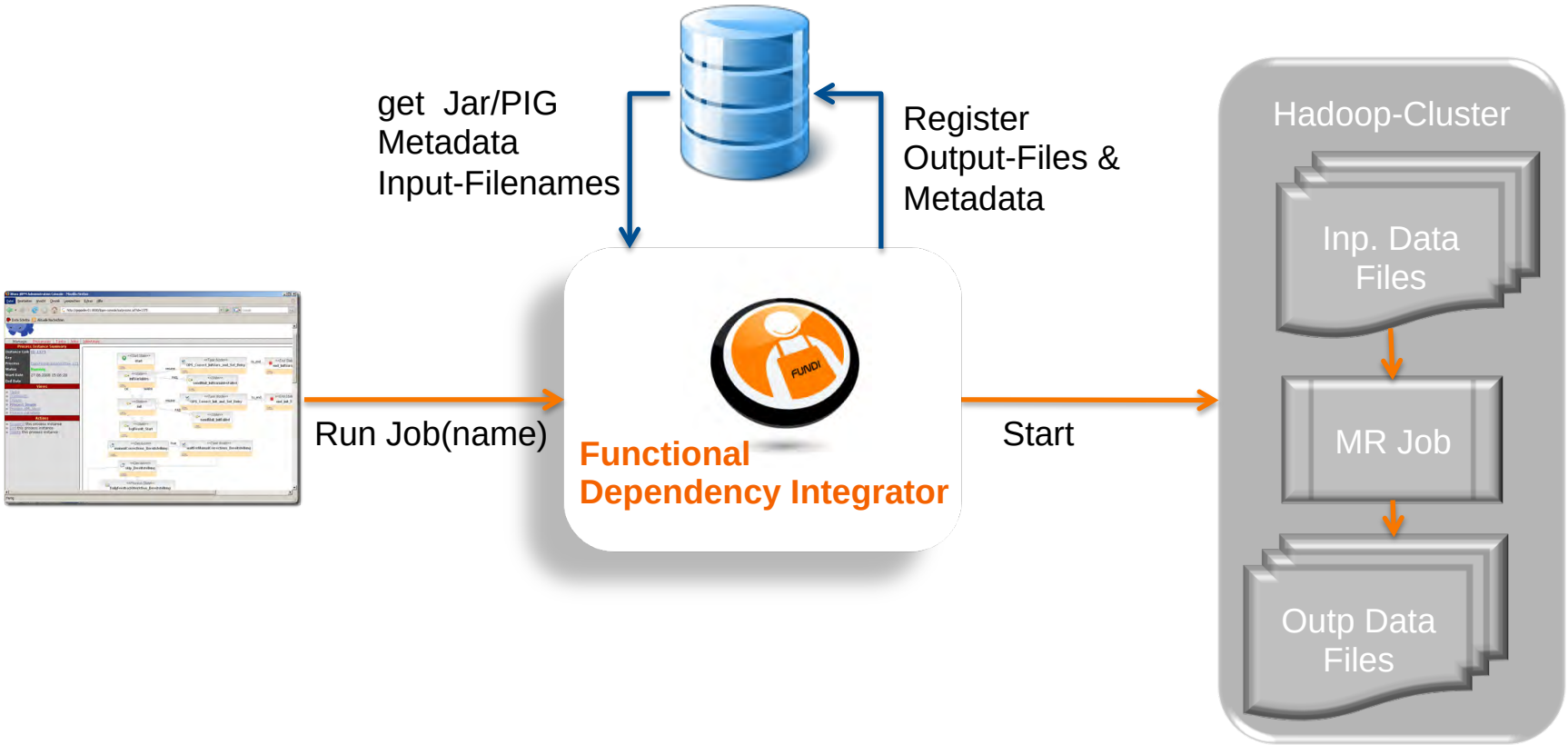
- ETL-Jobs im DWH
- HDFS-Zugriff



Functional Dependency Integrator

- Hadoop Job-Ausführung
- Data-Repository

FUNDI Job-Run



Agenda

- BI meets BIG DATA
- Jede Menge Blech: Die Hardware
- Was darf es denn sein: Das Hadoop Ecosystem
- Speichern, Mappen, Reduzieren
- Spielregeln im Cluster
- Der fleißige Handwerker
- **Essenz**

Best Practices

- **Projekt:** Identifiziere das richtige BIG DATA Problem
- **Hardware:** Etwas mehr schadet nicht: Alle Systeme müssen skalieren und benötigen Reserven, Namenode HA!
- **Frameworks:** Keep Your Ecosystem Simple, weniger kann mehr sein!
- **Implementierung:** Designe effektive und effiziente Algorithmen!
- **Spielregeln:** Sorge für geordnete Verhältnisse im Cluster!
- **Steuerung:** Es geht auch ohne Skript-Wüste und cron-Jobs!

Ohne Fleiß kein Preis ...

Know-how muss aufgebaut werden

Passende Projekte einplanen!

LIVE

Lastverhalten nur bedingt vorhersehbar

viele Tuning-Möglichkeiten, ILM,

Management der Datenflüsse, Rechte un

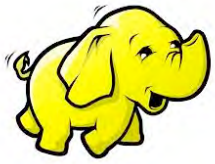
Ressourcen

QS

verteiltes Debuggen, qualitativ + quantitativ gute Test-Daten, Übertragbarkeit der Testergebnisse

DEV

Best Practices / Designpattern sind zu erarbeiten, Entwicklung von Low-level Algorithmen, „Unreife“ Frameworks, viel Prototyping



Die Belohnung: Hadoop beeindruckt!

- Hadoop und sein Ecosystem bieten hervorragende Lösungen für viele BIG DATA Probleme!

http://www.flickr.com/photos/xrm0/184379507/

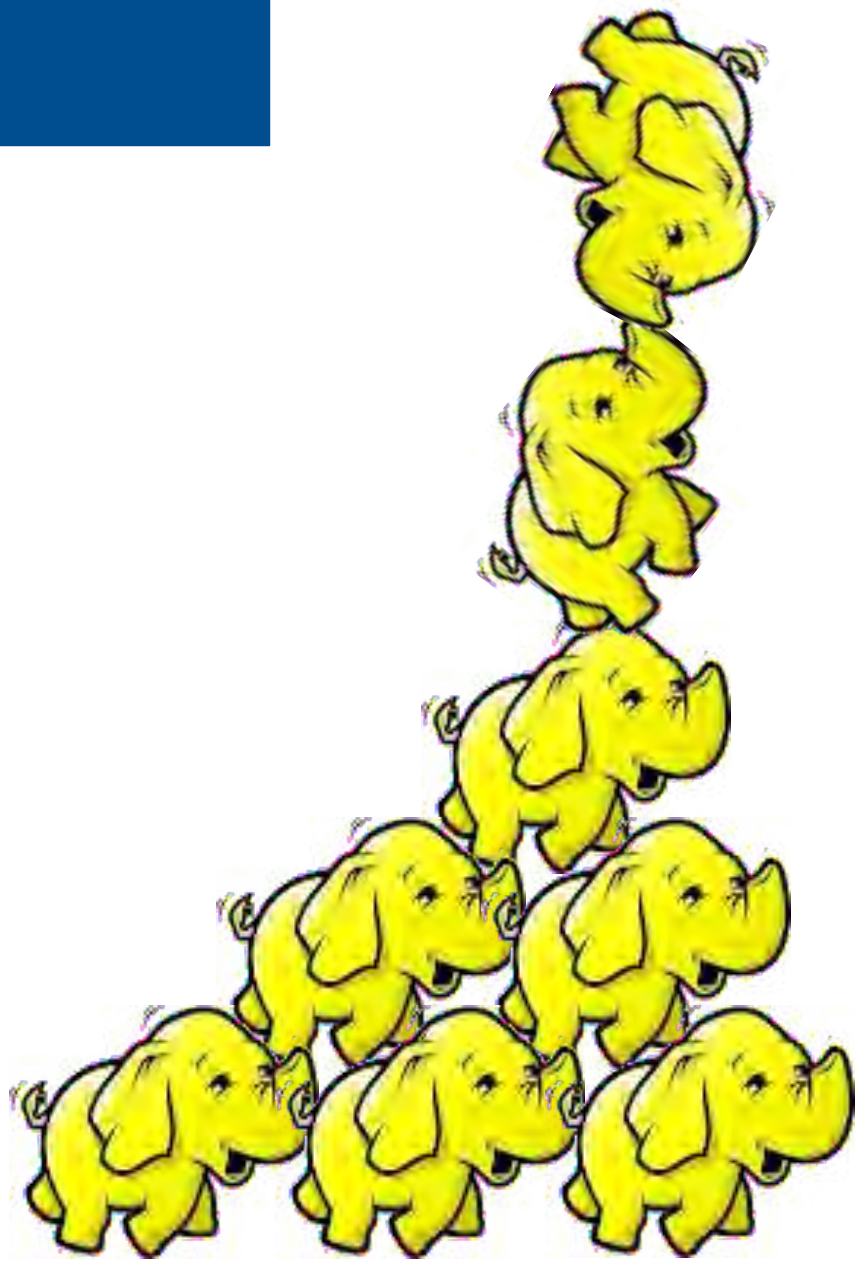
Der Nutzen

Hadoop beeindruckt

Massendatenverarbeitung bei 1&1
ist für Web- und Media-Analytics,
Logfile-Verarbeitung und
Datawarehousing
mit Hadoop messbar

- performanter,
- kostengünstiger,
- skalierbarer,
- flexibler,
- und zukunftsfähiger.

Vielen Dank!



Vielen Dank für eure Aufmerksamkeit

Wir nutzen Technologien, um unsere Kunden glücklich zu machen.
Und uns selbst.