



Business Intelligence, Big Data und Search

Drei Seiten einer Medaille?

Patrick Thoma

Offenburg, 7. März 2013

1. Business Intelligence – Es wird Zeit für „Intelligence“.
2. Big Data – größer, schneller, weiter. Was ist daran neu?
3. Search – ein alternativer Zugang zu BI und Big Data?
4. Fazit – Drei Seiten einer Medaille?



Die Rolle von Open Source

1. **Business Intelligence – Es wird Zeit für „Intelligence“.**
2. Big Data – größer, schneller, weiter. Was ist daran neu?
3. Search – ein alternativer Zugang zu BI und Big Data?
4. Fazit – Drei Seiten einer Medaille?

## Trends in Business Intelligence

Das BARC-Institut, führender Analyst für BI-Technologie untersucht die weitere Marktentwicklung und Trends

von Carsten Bange

27 SEPTEMBER

Der Softwaremarkt in Deutschland ist im Euro um. Das für Business-Intelligence die weitere Marktentwicklung kommenden Trends

Business Intelligence Information zu mehr Unternehmen Geschäftsprozesse Organisation Softwaremarkt hat, sondern auch organisationalen Organisation e

Aktuelle Trends auf einen Mega-Trend (Consumerization und Zugangsw

COMMENTARY

## 7 Top Business Intelligence Trends For 2013



Cindi Howson  
Founder  
See more

Short list of BI hot collaboration and

3  
Comments

Cindi Howson | Jan

Many people seem to be moving away from business intelligence, whether it's still business intelligence companies boost revenue making better, faster

Whatever you want

Forrester Blogs > Information Technology > Application Development & Delivery Professionals > Boris Evelson

## TOP 10 BI PREDICTIONS FOR 2013 AND BEYOND

Posted by Boris Evelson on December 12, 2012

265 Recommendations

It's that time of year for new predictions: We did a survey to update them. The predictions most of our Forrester analysts confirm or disprove new ones (stay tuned)

#1 (From 2012) traditionally been the truth. These will be an approach to add value react and adapt start embracing enterprise-grade



Suche

Login » Neu registrieren » zum CIO-Netzwerk

- Nachrichten
- Strategien
- Knowledge Center
- Karriere
- Partners

Sie sind hier: Homepage Knowledge Center BI

### Einsatzbereiche, Trends, Markt in Deutschland Die Zukunft von Business Intelligence

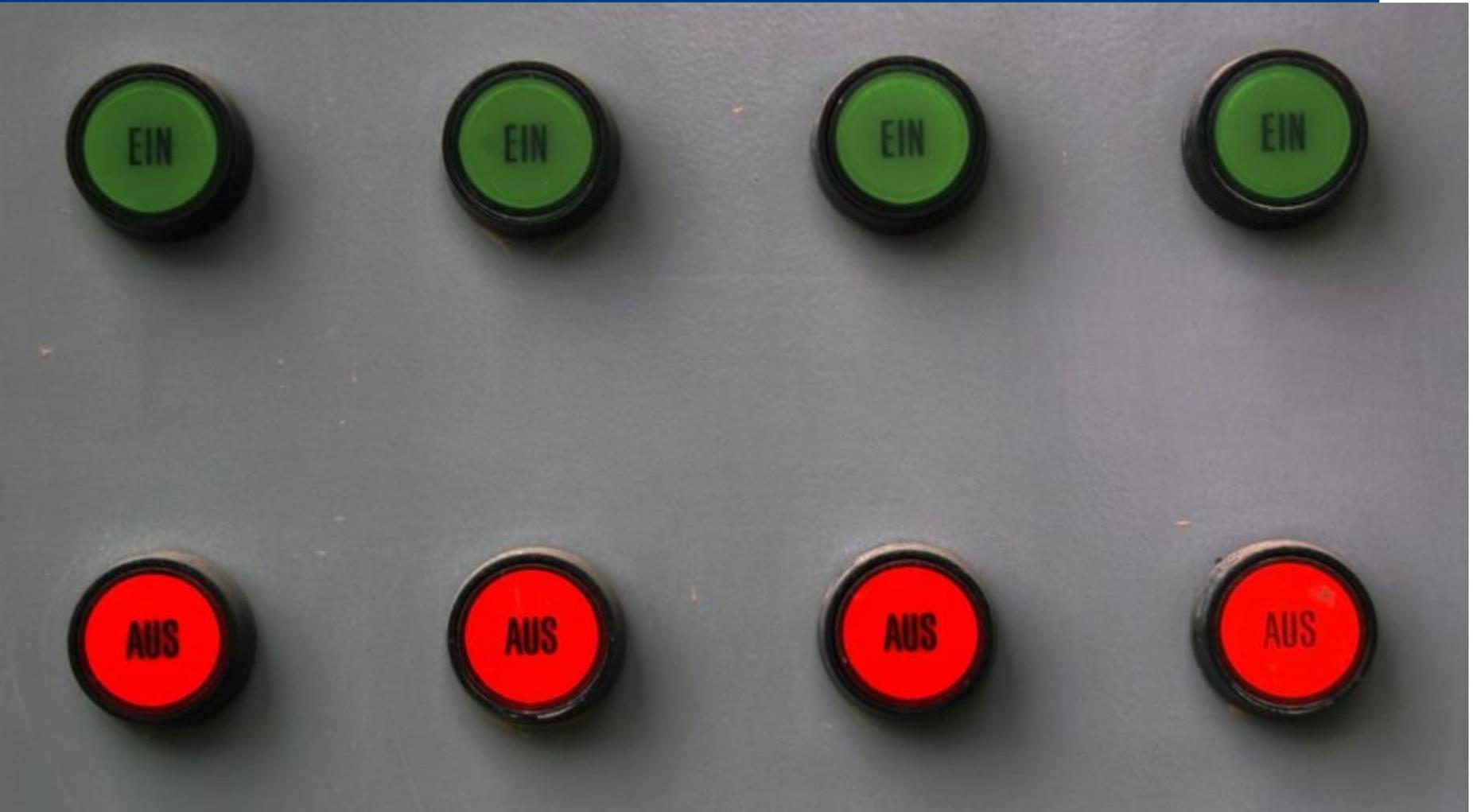
12.09.2012, von Andreas Schaffry

Drucken | Versand | PDF

XING +1 Gefällt mir Twittern

BI-Spezialanbieter machen Rekordumsätze. Mobile BI wird künftig der Mega-Trend sein, sagt eine Lünendonk-Studie.

# Self Service BI nimmt zu: Dann ist Usability gefragt!



- ▶ Für Open Source Software oft eine Herausforderung



statt



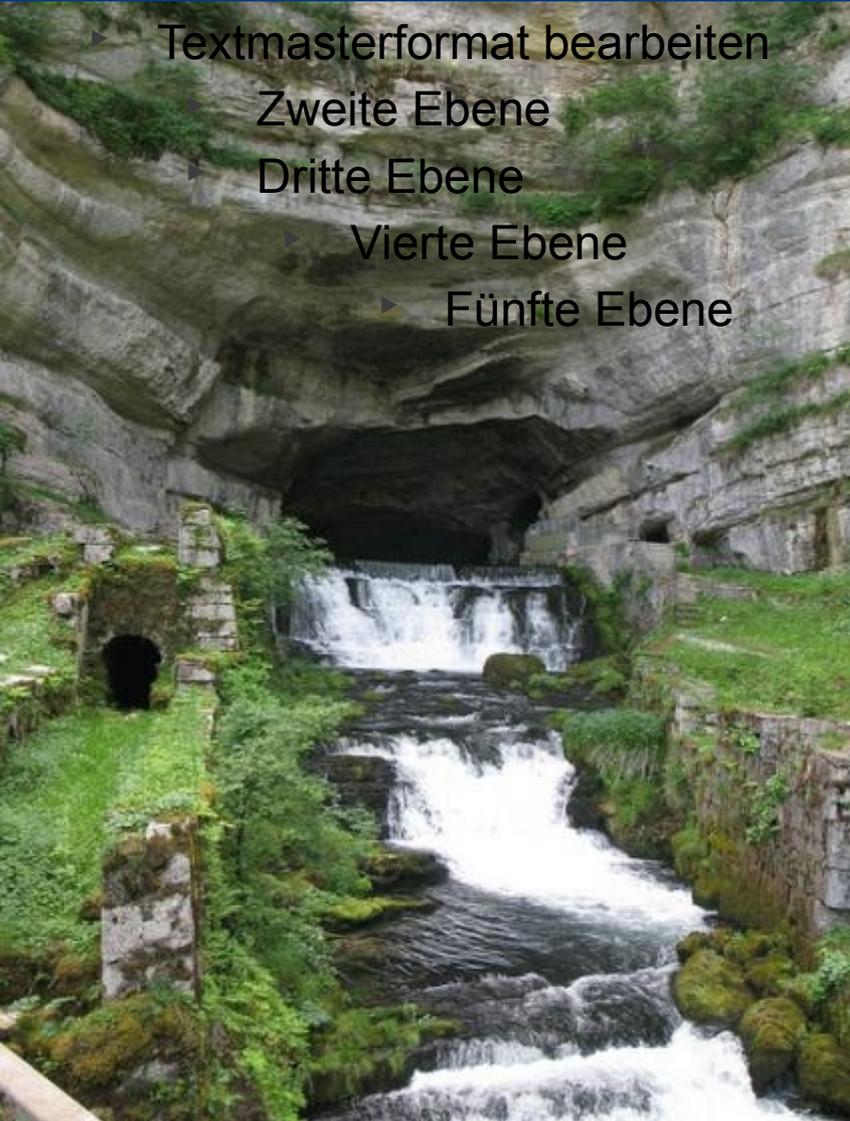
Thomas Nico Neuter Photography  
www.tnmp.de 2013 ©

- ▶ Chance für Open Source Software durch
  - ▶ niedrigere technische Einstiegshürden
  - ▶ geringes finanzielles Risiko aufgrund der Lizenmodelle
  - ▶ gute Integration durch Offenheit



- ▶ Leistungsfähige Open Source Tools vorhanden (R, Weka, Mahout, Knime,...)
- ▶ Im Unternehmenseinsatz oft noch mit Akzeptanzproblemen
- ▶ mögliche Ursachen: Usability, mangelndes Know-how
- ▶ aber große Chancen durch Kostenvorteil und steigende Datenvolumen!

# Und was wird aus der „Single Source of Truth“?

- 
- ▶ Textmasterformat bearbeiten
    - ▶ Zweite Ebene
    - ▶ Dritte Ebene
      - ▶ Vierte Ebene
      - ▶ Fünfte Ebene

- ▶ weiterhin notwendig, aber nicht mehr hinreichend
  - ▶ Deshalb aber auch nicht weniger wichtig!
- ▶ Zunehmender Commodity-Charakter erzeugt Kostendruck
- ▶ Chance für Open Source Lösungen, z. B.
  - ▶ Datenbanken
  - ▶ ETL

Also alles ganz einfach?

Nicht ganz...

1. Business Intelligence – Es wird Zeit für „Intelligence“.
2. **Big Data – größer, schneller, weiter. Was ist daran neu?**
3. Search – ein alternativer Zugang zu BI und Big Data?
4. Fazit – Drei Seiten einer Medaille?





# Skalierung am Beispiel eines Transportunternehmens



# Skalierung am Beispiel eines Transportunternehmens



# Skalierung am Beispiel eines Transportunternehmens



# Skalierung am Beispiel eines Transportunternehmens

\$5.000.000 Anschaffung

Spezial-Know-how für  
Betrieb und Wartung



Hoher Schaden bei Ausfall

\$46.000 pro Reifen

# Skalierung am Beispiel eines Transportunternehmens

Horizontal statt vertikal

40.000€ Anschaffung pro Fahrzeug

skaliert in beide Richtungen



modernisierbar

keine Spezialkenntnisse erforderlich

Ausfall einzelner Fahrzeuge kompensierbar



*Die spaltenorientierte Datenbank stieß an ihre Grenzen:*

Web.  
Intelligence  
BI Plattform

- Verarbeitungsgeschwindigkeit nicht mehr ausreichend
- Aufrüstung teuer
- Begrenzte Ressourcen



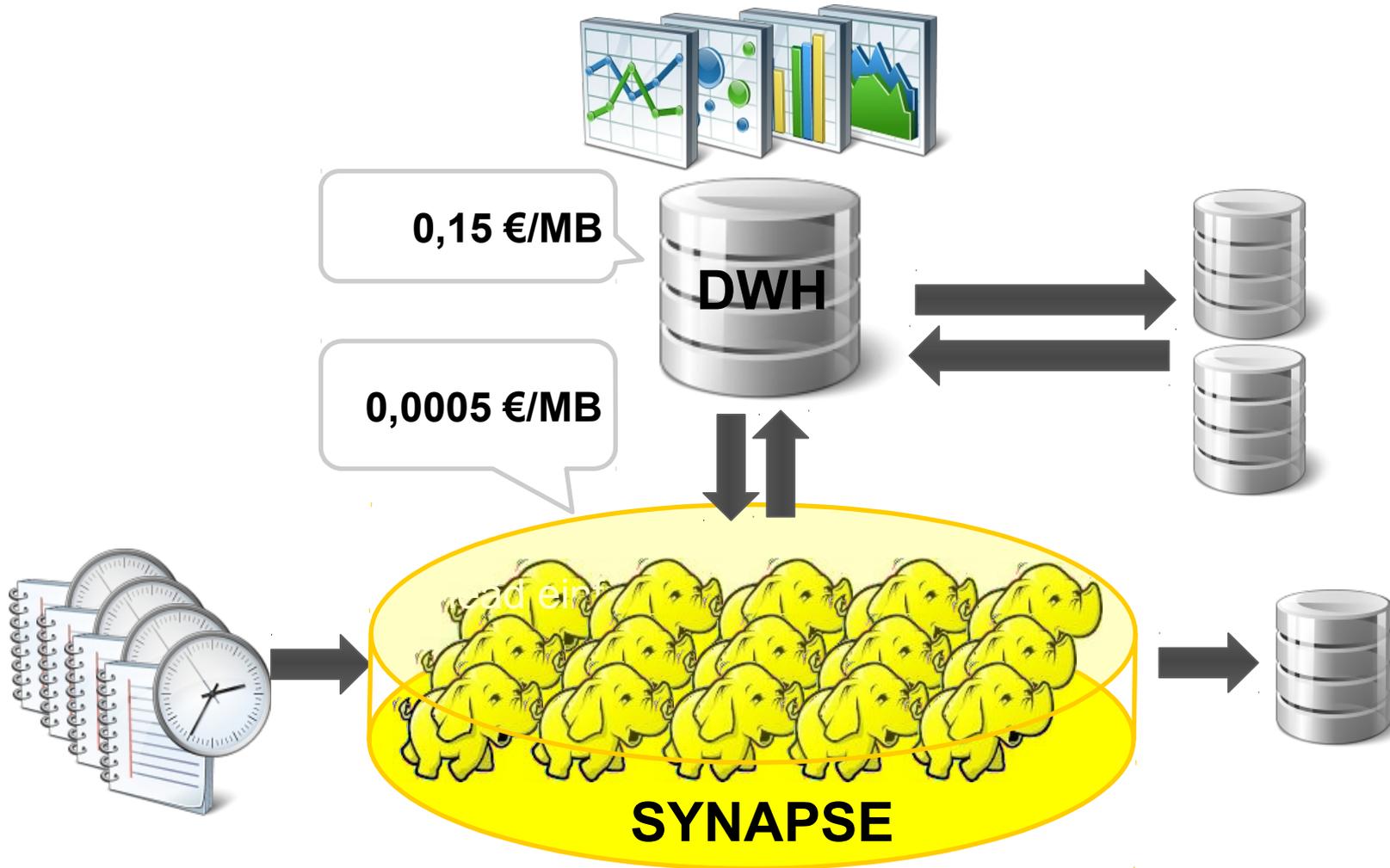
**Web-Analytics: 240 Files/d**

200 GB/d \* 90d = 18 TB

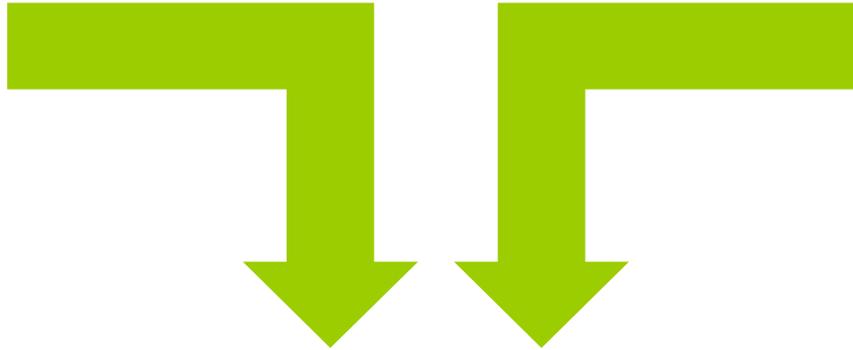
**Log-Analytics: 15.000 Files/d**

2.000 GB/d \* 30d = 60 TB

# Hadoop-Cluster bietet kostengünstige und skalierbare Speicherung und Vorverarbeitung



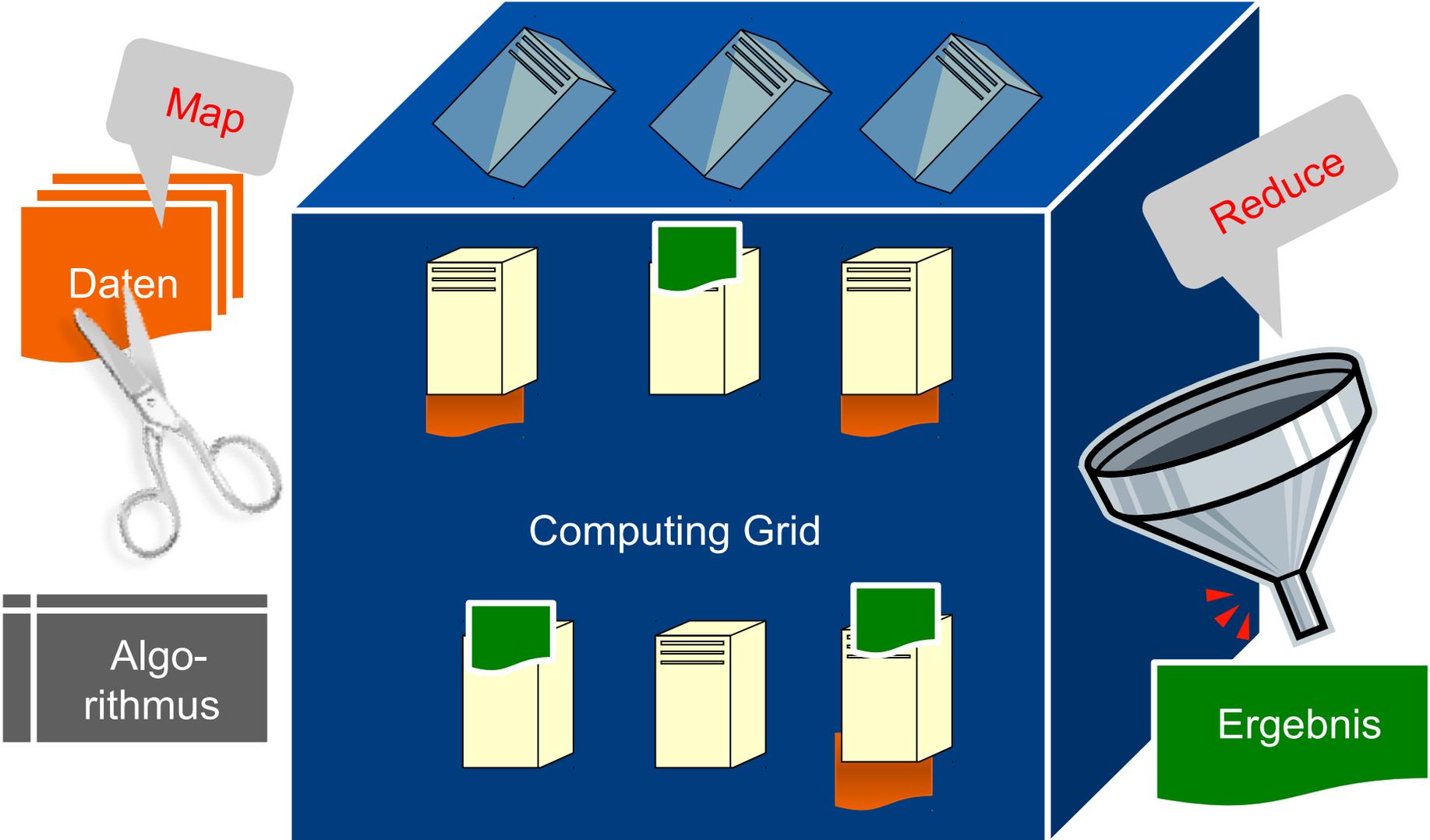
\* **SY**Nergetic **A**nalytical **P**rocessing and **S**torage **E**ngine



Ausführungsumgebung  
Programmiermodell

Commodity  
Hardware

# „Teile und beherrsche“: Das Prinzip von MapReduce



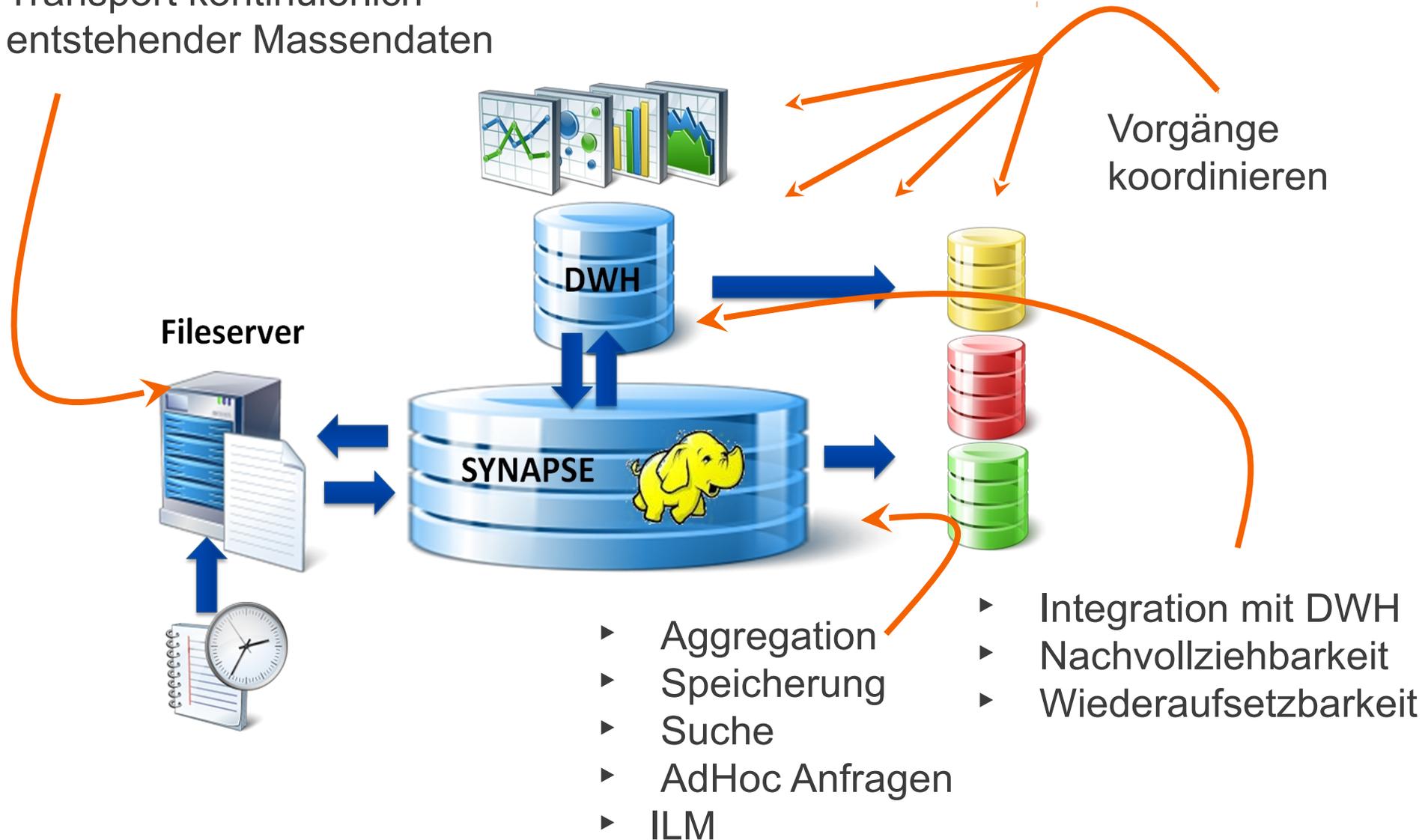
# Hadoop Ökosystem recht unübersichtlich

Aber alles da, was man braucht. Und alles open source.

- ▶ Textmasterformat bearbeiten
  - ▶ Zweite Ebene
  - ▶ Dritte Ebene
  - ▶ Vierte Ebene
    - ▶ Für



Transport kontinuierlich  
entstehender Massendaten



# Weniger ist mehr

## Der Technologie-Stack

Hive      ZooKeeper

HBase      Cassandra

Flume

Pig ✓

Avro ✓

MapReduce ✓

HDFS ✓



- ▶ **Unterstützt die Verarbeitung durch...**
  - ▶ Header-Zeile pro File mit Typ-Information (JSON-Format)
  - ▶ Versionsprüfung möglich
- ▶ **Effiziente Verarbeitung durch...**
  - ▶ Disk: Speicher-effizientes Binärformat
  - ▶ (De-)Serialisierung ok
- ▶ **Zukunftssicher da...**
  - ▶ Künftiger Hadoop-Standard
  - ▶ Aktive Weiterentwicklung
  - ▶ Sukzessive Integration in die Hadoop-Teilprojekte



```
# avro textual
representation
{"type": "record", "name":
"Point",
  "fields": [
    {"name": "x", "type":
"int"},
    {"name": "y", "type":
"int"}
  ]}
5 8
-3 4
2 -7
```

- ▶ **Unterstützt die Verarbeitung durch...**
  - ▶ Header-Zeile pro File mit Typ-Information (JSON-Format)
  - ▶ Versionsprüfung möglich
- ▶ **Effiziente Verarbeitung durch...**
  - ▶ Disk: Speicher-effizientes Binärformat
  - ▶ (De-)Serialisierung ok
- ▶ **Zukunftssicher da...**
  - ▶ Künftiger Hadoop-Standard
  - ▶ Aktive Weiterentwicklung
  - ▶ Sukzessive Integration in die Hadoop-Teilprojekte
- ▶ Variabilität: Anreichern von Datenstrukturen herausfordernd
- ▶ Unvollständige Implementierungen
- ▶ Ineffizientes Speicher-Verhalten im RAM: Schema an jedem Datensatz
- ▶ Spärliche Dokumentation
- ▶ Etliche Bugs, gefühlte Early-Adopter
- ▶ Seamless Hadoop-Integration engt Spielräume ein
- ▶ Hadoop 2 setzt auf Google Protocol Buffers!

- ▶ Entwicklung ist immer noch im Fluss
- ▶ Hadoop- Kern ist bereits sehr stabil - aber kein Vergleich bspw. mit einer kommerziellen relationalen Datenbank
- ▶ Flankierende Projekte haben gefühlt oft noch Beta-Status – im Handumdrehen ist man Committer eines OS-Projektes ;-)
- ▶ Ähnliche Toolsets unterscheiden sich meist nur in Details
  - ▶ Pig, Hive, Cascading
  - ▶ HBase und Cassandra
  - ▶ Askaban vs. OozieAber gerade die – können manchmal entscheidend sein!
- ▶ Distributionen helfen (z.B. Cloudera)
  - ▶ Eigenes Hadoop-Release basierend auf den Apache Repositories
  - ▶ bieten konsistente, „in sich schlüssige Release-Stände“ und Bugfixes
  - ▶ Möglichkeit des kommerziellen Supports

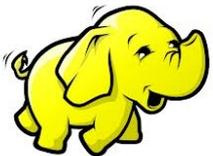
# Know-how Aufbau muss im Projekt eingeplant werden...

...es fehlt anfangs an allen Fronten

- **LIVE**
- Lastverhalten nur bedingt vorhersehbar  
Tuning-Möglichkeiten mannigfaltig

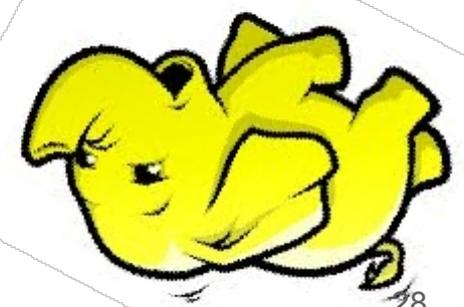
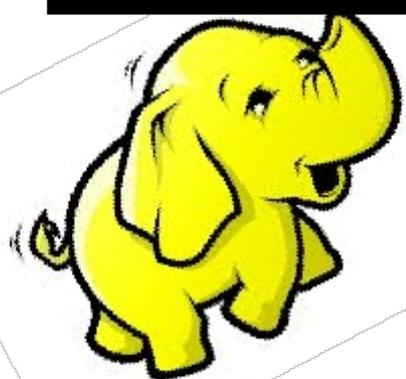
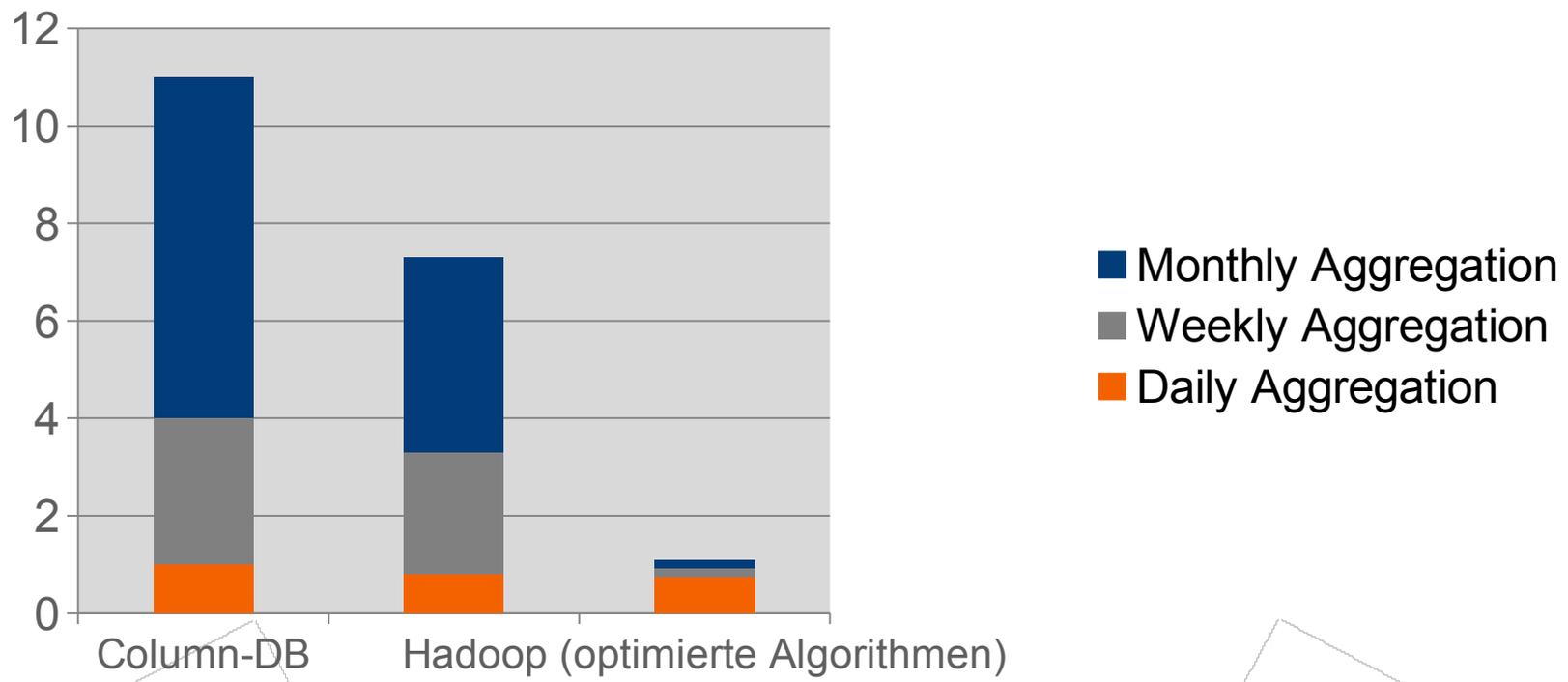
- **QS**  
Herausforderung: verteiltes Testen, echte Daten

- **DEV**  
Best Practices sind nicht vorhanden



# Performanceverbesserung um bis zu Faktor 40 erreicht

Schlüssel liegt in der Anwendung des MapReduce-Paradigmas

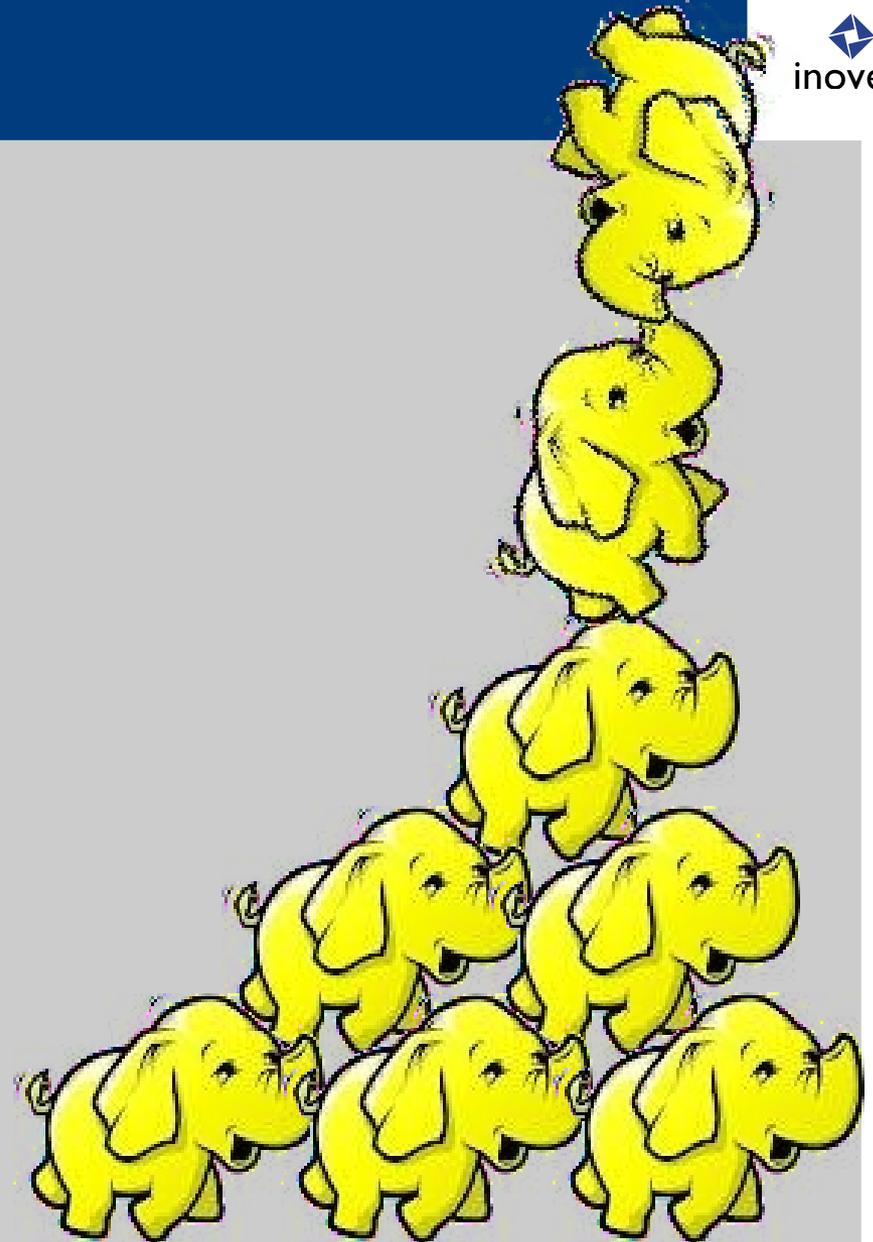


# Hadoop beeindruckt

## Das Fazit der 1&1

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

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



1. Business Intelligence – Es wird Zeit für „Intelligence“.
2. Big Data – größer, schneller, weiter. Was ist daran neu?
3. Search – ein alternativer Zugang zu BI und Big Data?
4. Fazit – Drei Seiten einer Medaille?

Google  
Deutschland

Google-Suche

Auf gut Glück!

**amazon.de** Mein Amazon | Angebote | Gutscheine | Hilfe | Impressum

Alle Kategorien ▾ Suche

Hallo! [Anmelden](#) Mein Konto ▾  Einkaufswagen ▾  ▾

Amazon.de | Gutscheine | Bestseller | Angebote | Outlet | Amazon Spar-Abo | Jetzt verkaufen

### Kategorie

#### Fremdsprachige Bücher

- Lernen & Nachschlagen
- Informationssysteme
- Informationstheorie
- Datenbanken
- Bibliotheks- & Informationswissen...

#### Kindle-Shop

- Datenbanken (englischsprachig)
- Informationstheorie (englischsprachig)

#### + Alle 3 Kategorien

#### Versandoption [\(Was ist das?\)](#)

Kostenlose Lieferung ab EUR 20 Bestellwert

### Lieblingslisten

#### Lieblingslisten

Ihre eigene Liste hier erstellen  [so geht's](#)  
Legen Sie Ihre eigene Lieblingsliste an

#### Suche Lieblingslisten

### "search based applications"

1-16 von 219 Ergebnissen

Sortieren in [Fremdsprachige Bücher](#) nach [Beste Ergebnisse](#) | [Beliebtheit](#) | [Preis: aufsteigend](#) | [Mehr](#)



**Search-Based Applications: At the Confluence of Search and Database Technologies** von Laura Wilber und Gregory Grefenstette (3. März 2010)

**EUR 8,22** Kindle Edition

Jetzt als Download verfügbar.

[Kindle-Shop: Alle 3 Artikel ansehen](#)



**Search-Based Applications: At the Confluence of Search and Database Technologies (Synthesis Lectures on Information Concepts, Retrieval, and Services)** von Gregory Grefenstette, Laura Wilber und Gary Marchionini von Morgan & Claypool Publishers (21. Dezember 2010)

**EUR 28,99** Taschenbuch

Bestellen Sie in den nächsten **1 Stunde**, um den Artikel am Dienstag, 27. November zu erhalten.

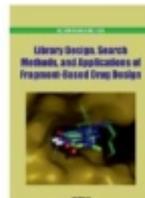
Andere Angebote - Taschenbuch

**EUR 23,03** neu (11 Angebote)

**EUR 40,16** gebraucht (1 Angebot)

Kostenlose Lieferung möglich.

[Englische Bücher: Alle 206 Artikel ansehen](#)



**Library Design, Search Methods, and Applications of Fragment-Based Drug Design (Acs Symposium Series)** von Rachele J. Bienstock von Oxford University Press Inc (20. März 2012)

**EUR 141,99** Gebundene Ausgabe

Nur noch 1 Stück auf Lager - jetzt bestellen.

Andere Angebote - Gebundene Ausgabe

**EUR 112,03** neu (14 Angebote)

**EUR 104,24** gebraucht (1 Angebot)

Kostenlose Lieferung möglich.

[Englische Bücher: Alle 206 Artikel ansehen](#)

# Enterprise Search

## Zugriff auf die im Unternehmen verfügbare Information



Anzahl Anzeige  
der Verfeinerung

Miniaturansicht

Sortierung nach jedem Feld

- Result Type
- Any Result Type
- Microsoft Excel (80)
- Microsoft Power... (62)
- Microsoft Word (47)
- Site
- Any Site
- fastdemo17.fas... (200)
- Author
- Any Author
- System Account (200)
- Alan Brewer (51)
- Eduard Dell (10)
- Spencer Low (10)
- show more v
- Modified Date
- Any Modified Date
- Past Month (11)
- Past Six Months (200)
- Past Year (200)
- Company
- Any Company

1-10 of 200 results

Language: English

Sort by: Relevance, Relevance, WordBoost, ClickRank, WoodgroveBoost, Size (Ascending), Size (Descending), Date (Newest), Date (Oldest)

**Blue Yonder Airlines - Market Analysis and Target Market Research - Recommendations**  
... Consulting Blue Yonder Airlines Market Research Recommendations Strategy ... on what market would consider an ideal airline ...  
Authors: System Account Alan Brewer Date: 10/12/2009 Size: 507KB  
<http://fastdemo17.fastsearchdemos.net/Share ... nd Target Market Research - Post Mortem.docx>  
View In Browser | Similar Results

**Blue Yonder Airlines - Market Analysis and Target Market Research - Post Mortem**  
... industry analysis to uncover opportunities and do research on what Blue Yonder's target market would consider an ideal airline experience to boost revenue. Engagement Manager: ...  
Authors: System Account Scott Bishop Date: 10/12/2009 Size: 37KB  
<http://fastdemo17.fastsearchdemos.net/Share ... nd Target Market Research - Post Mortem.docx>  
View In Browser | Similar Results

**Baldwin Museum of Science - Market Research - Sales Pitch**  
... Baldwin Museum of Science Market Research Strategy Assessment ... Contoso team conducted market research to determine how ... Proposed deliverables: Market Research documentation and ...  
Authors: System Account Alan Brewer Date: 10/12/2009 Size: 1MB  
<http://fastdemo17.fastsearchdemos.net/Share ... Science - Market Research - Sales Pitch.pptx>  
Close Preview | View In Browser | Similar Results

**Contoso Consulting**  
Baldwin Museum of Science Market Research

**Company Background**  
- Founded in 1985  
- Started with 100 employees and has grown to over 1800 worldwide  
- Revenue has grown consistently since 1995, even in times of economic downturn  
- 70% of clients are Fortune 500 or Fortune 1000 companies  
- Proven success with the following solutions:  
  - Strategy Assessments  
  - IT Implementations  
  - Organizational Redesigns  
- Read through case studies here: [www.contoso.com/casestudies](http://www.contoso.com/casestudies)

**Company Background: Service Offerings**  
- Strategy Assessments  
  - Contoso Consulting helps customers to better understand their current and opportunities in the marketplace, assess the existing strengths and weaknesses, and develop a strategic plan that meets the needs of our clients and provides a competitive edge.  
- IT Implementations  
  - Contoso Consulting addresses the technology needs of our clients and helps them develop a solution that fits the growth or process improvement management services throughout the entire lifecycle, from design to implementation and ongoing support.  
- Organizational Redesigns  
  - Contoso Consulting provides expertise and recommendations for business process re-engineering for each process in a client's company to help them increase productivity.  
Please visit other solutions for each service offering: [www.contoso.com](http://www.contoso.com)

Related Searches

- Market Research
- Market Trends Analysis
- Market Opportunity Analysis

People Matches

- Toni Poe  
Chief of Partnerships and Strategy Executive
- Christine Koch  
Managing Director Engineering
- Keith Dishmo  
Procurement Manager Operations

View more people >

Bing:

Dow Jones, Nasdaq, S&P 500, stock market data ...  
Complete financial market coverage with breaking news, analysis, stock quotes, before & after hours market data, research and earnings  
<http://money.cnn.com/data...>

Market - Wikipedia, the free encyclopedia  
A market is any one of a variety of different systems, institutions, procedures, social

Ähnliche Ergebnisse

Voransicht

- ▶ Bei klassischen Suchanwendungen steht das Auffinden von Informationen im Fokus.
- ▶ **Search Based Applications:**  
*“A software application that uses a search engine as the primary information access backbone, and whose main purpose is not classical Information Retrieval but rather performing a domain-oriented task rather than locating a document.”*  
[Gregory Grefenstette, Laura Wilber. *Search-Based Applications: At the Confluence of Search and Database Technologies*, Morgan & Claypool Publishers; 1 edition (December 21, 2010).]
- ▶ **Beispiel:**
  - ▶ Sendungsverfolgung in der Logistik
  - ▶ Kontextbezogene Werbung
  - ▶ Decision intelligence

# Navigation & dynamisches Content Rendering

## Search Based Applications



**heute.de**

Futter für das Schwarze Loch  
Mainz 8° nach Regen

1/4

- 14:10 Polizeihund fällt Kinder an - Ermittlungen gegen Beamtin
- 13:46 Schläuble: Finden Athen-Lösung
- 13:20 TÜV-Test: Jede dritte Lichterkette mit Sicherheitsmängeln
- 12:58 Verleger fordern Abschaffung der Mehrwertsteuer auf Zeitungen
- 12:47 Merkel besorgt über Entwicklung in Ägypten

Überblick  
Politik  
Wirtschaft  
Panorama  
Sport

heute 100 sec | heute 19 Uhr | heute Journal

### heute.de-Interview



**"Athen-Hilfe: EU verliert an Glaubwürdigkeit"**  
Griechenland braucht dringend frisches Geld - heute verhandeln die Euro-Finanzminister über ein neues Rettungspaket. Zsolt Darvas von der Denkfabrik Bruegel sieht die bisherige Rettungsstrategie als gescheitert an. "Sie werden auch weiterhin scheitern, wenn sie ihren Kurs nicht ändern", sagt er im Gespräch mit heute.de. [\[mehr\]](#)

Video ZDF-Korrespondent: Haushaltsloch stopfen  
Video Reiner Wein in Sachen Athen  
Infografik: Der griechische Patient

### UN-Gipfel in Doha



Klimaschutz: Altmaier bangt um Vorreiter-Rolle



SAT.1 SENDUNGEN A-Z VIDEOS TV-PROGRAMM RATGEBER NEWS /ZDF KEYLINE

### HEUTE20:15

SAT.1 FILM KEINDRHASEN

Ausgerechnet ein Puddingherd ohne Ohren bringt Til Schweiger und Nora Tschüner zusammen.

• Trailer ansehen  
• Alle Film-Infos

SAT.1 FILM DIE TORE DER WELT MILLION DOLLAR SHOOTER DER COP UND DER SNOB

• TV-PROGRAMM  
• SENDUNG VERPASST

- 14:00 Richter Alexander Höp
- 15:00 Familien-Fälle
- 16:00 Familien-Fälle
- 17:00 Pures Leben - Mitten in Deutschland

• Zum TV-Programm  
• Sat.1 in HD  
• HIGHLIGHT HEUTE  
20:15 Keindhrhasen

• NEU • VOICE OF GERMANY • KNALLERFRAUEN • US-SERIE • SAT.1 FILM

Genze Folge 25.11.2012  
Common Law Gute trotz schlechter Vorbilder

Genze Folge 26.11.2012  
Der Cop und der Snob Die Flauchermädchen

Genze Folge 23.11.2012  
Knallerrfrauen Twisten, T-Papier und tanken

Genze Folge 22.11.2012  
Unforgettable Der ehrlische Tod

Sendungen

Die TORE DER WELT  
schwer verliebt  
K 11: Lieber tut weh  
ran - Vettel feiert Titel-Hatrick

SAT.1 Aktuell

Frühstücksfernsehen

Die Ehrlich Brüdern zaubern für uns im Studio Schnee. [Zum Video](#)

2:22 26.11.2012

Große Stars und größere Emotionen

Martina Hill, Alina Levshin und Ulrich Tukur stauben einen Bambi. [Zum Video](#)

2:46 26.11.2012

ran - Vettel feiert Titel-Hatrick

Nach einem packenden Rennen in Sao Paulo hat Sebastian Vettel den Titel-Hatrick vgliepdet. [Zum Video](#)

Top Videos

- 1 The Mentalist: Man nennt es [Garze Folge ansehen](#)
- 2 The Voice Kids: Jetzt bewerben! [Zum Video](#)
- 3 Knallerrfrauen: Twisten, T-Papier und tanken [Garze Folge sehen](#)
- 4 Knallerrfrauen: Gute Nacht Lied [Zum Video](#)
- 5 K 11: Liebe tut weh [Garze Folge sehen](#)

- ▶ **Was sind Search Based Business Intelligence Applications?**
  - ▶ Anwendungen zur Unterstützung von Geschäftsentscheidungen, die
  - ▶ Suchtechnologie als zentrale technische Basis verwenden
  
- ▶ **Welchen Nutzen bieten Search Based BI Applications?**
  - ▶ Einfachste Benutzeroberfläche, die von jedem Nutzer verstanden wird
  - ▶ Darstellung der Ergebnisse in visueller und interaktiver Form
  - ▶ Erlauben die Integration von und das Suchen in beliebigen Quellen,
  - ▶ insbesondere die Integration von strukturierten und unstrukturierten Daten
  - ▶ Stellen eine einheitliche Zugriffsschicht durch die Suchtechnologie bereit
  - ▶ Skalierbar in Bezug auf Datenmenge und Nutzerzahl
  - ▶ Indizierung kann Echtzeit erfolgen
  - ▶ Über Facetten und Filter sind Drill-downs zu relevanten Informationen möglich

# Beispiel: inovex search demonstrator

## Suche mit BI Features auf Wikipedia Daten

► Textmasterformat bearbeiten

Q java

Options ▾

► Zweite Ebene

► Dritte Ebene

► Vierte Ebene

► Fünfte Ebene

Standard **Bar Charts** Column Charts

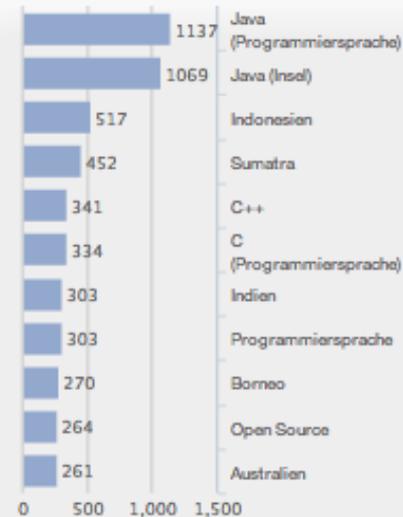
► Autor

► Kategorie

► Kontext

► Letzte Änderung

▼ Referenz



Showing 1 - 10 of 4,455 Search time: 0.061s

Score ▾ | Date ▾

**Java Card** ★★★

**Java** Card ist eine Variante der Programmiersprache **Java**, die es erlaubt, **Java** Card Applets, einem reduzierten **Java**-Standard folgende **Java** Applets, auf Chipkarten auszuführen. **Java** Card Applets werden ...

[http://de.wikipedia.org/wiki/Java\\_Card](http://de.wikipedia.org/wiki/Java_Card)  
12. Februar 2013, 01:10:27

**Java 2D** ★★★

**Java** 2D ist eine Klassenbibliothek und Programmierschnittstelle für die Umsetzung portabler zweidimensionaler Grafiken in Java. `style="background:yellow">java-media/2D/index.jsp` Mit ihr können ...

[http://de.wikipedia.org/wiki/Java\\_2D](http://de.wikipedia.org/wiki/Java_2D)  
12. Februar 2013, 01:09:41

**Java-Anwendung** ★★★

Eine **Java**-Anwendung, auch **Java**-Applikation genannt, ist ein in der Programmiersprache **Java** geschriebenes Anwendungsprogramm. Im **Java**-Umfeld unterscheidet man oft zwischen in Webbrowsern laufenden ...

<http://de.wikipedia.org/wiki/Java-Anwendung>  
08. Juli 2011, 09:40:22

# Beispiel: inovex search demonstrator

Suche mit BI Features auf Wikipedia Daten

► Textmasterformat bearbeiten

🔍 java Options ▾

► Zweite Ebene

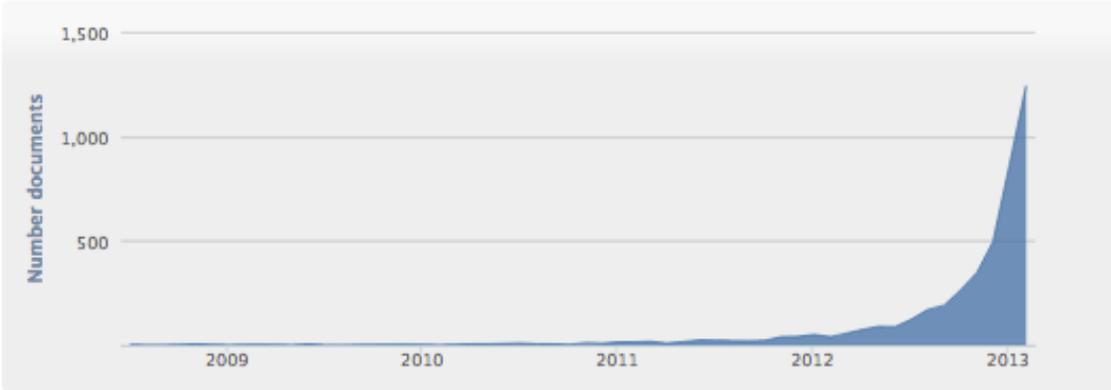
Standard Bar Charts **Column Charts**

► Dritte Ebene

► Vierte Ebene

Autor Kategorie Kontext **Letzte Änderung** Referenz

► Fünfte Ebene



Year	Number documents
2009	~10
2010	~20
2011	~50
2012	~100
2013	~1,200

Showing 1 - 10 of 4,455

Search time: 0.061s

Score ▾ | Date ▾

**Java Card**

★★★

**Java** Card ist eine Variante der Programmiersprache **Java**, die es erlaubt, **Java** Card Applets, einem reduzierten **Java**-Standard folgende **Java** Applets, auf Chipkarten auszuführen. **Java** Card Applets werden ...

[http://de.wikipedia.org/wiki/Java\\_Card](http://de.wikipedia.org/wiki/Java_Card)

12. Februar 2013, 01:10:27

**Java 2D**

★★★

# Beispiel: Einheitliche, integrierte Sicht im Kundenservice

## Search-Based Business Intelligence Applications





<http://lucene.apache.org>



<http://lucene.apache.org/solr/>

<http://www.elasticsearch.org>



▶ Vierte Ebene

▶ Fünfte Ebene

## “Lucene is an open source, pure Java API for enabling information retrieval”

- ▶ Ursprünglich entwickelt von Doug Cutting 1999 , wurde Apache TLP in 2001
- ▶ Lizensiert unter Apache License 2.0
- ▶ Reine Java Bibliothek mit Implementierungen für:
- ▶ Lucene.NET (<http://lucenenet.apache.org>)
- ▶ PyLucene (<http://lucene.apache.org/pylucene/>)
- ▶ u.a.: <http://wiki.apache.org/lucene-java/LuceneImplementations>
- ▶ Große und sehr aktive Entwickler Community
- ▶ Gut dokumentiert und supported (38 aktive Committer!)
- ▶ Aktuelles “stable release”: 4.0 (seit 12. Oktober 2012)
- ▶ Weit verbreitet und eingesetzt in kommerziellen und nicht-kommerziellen Projekten: <http://wiki.apache.org/lucene-java/PoweredBy>

- ▶ **Skalierbare, hoch performante Indizierung**
  - ▶ über 95GB/h auf aktueller Hardware
  - ▶ kleine Anforderungen bzgl. RAM
  - ▶ Inkrementelle Indizierung ähnlich schnell wie batchorientierte
  - ▶ Indexgröße ca. 20-30% des indizierten Texts
- ▶ **Mächtige , Treffende und Effiziente Suchalgorithmen**
  - ▶ Relevante Suche – Beste Treffer erscheinen zuerst
  - ▶ Viele mächtige Suchanfragen: Ausdrücke, Platzhalter, Ähnliche Wörter, Bereiche , etc.
  - ▶ Feldbasierte Suche (z. B. Titel, Autor, Inhalt)
  - ▶ Suche über Zeiträume
  - ▶ Sortierung nach beliebigen Feldern
  - ▶ Suche über mehrere Indizes und verschmolzene Ergebnisliste
  - ▶ Simultane Indexaktualisierung und Suche

## “Solr is a standalone enterprise search server & document store with based on Lucene”

- ▶ Initial entwickelt von Yonik Seeley bei CNET Networks in 2004
- ▶ Eingeführt als Apache Incubator in 2006, wurde TLP in 2007
- ▶ Lizensiert unter Apache License 2.0
- ▶ Seeley u.a. gründeten LucidImagination -> LucidWorks
- ▶ Große und sehr aktive Entwickler Community, gut dokumentiert und supported, enge Verbindungen zur Lucene Community
- ▶ Aktuelle “stable release”: 4.0 (seit 12. October 2012)
- ▶ Weit verbreitet und eingesetzt: <http://wiki.apache.org/solr/PublicServers>

# Solr: Die Admin-Oberfläche

Open source Suchtechnologien



Dashboard

Logging

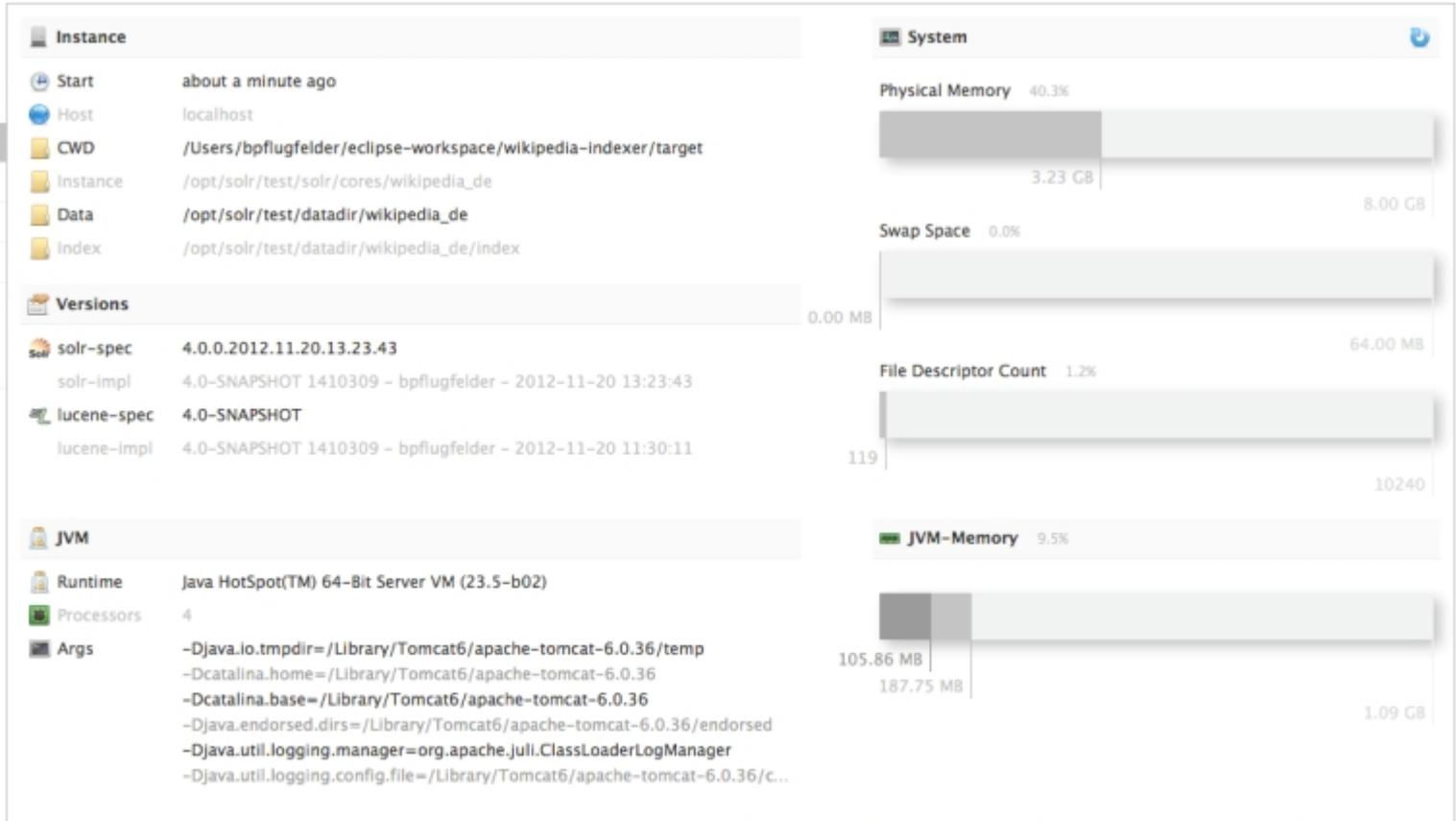
Core Admin

Java Properties

Thread Dump

wikipedia\_de

wikipedia\_en



The screenshot displays the Apache Solr Admin UI. On the left is a navigation sidebar with the Solr logo and menu items: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, wikipedia\_de, and wikipedia\_en. The main content area is divided into three sections: Instance, System, and JVM.

### Instance

Start	about a minute ago
Host	localhost
CWD	/Users/bpflugfelder/eclipse-workspace/wikipedia-indexer/target
Instance	/opt/solr/test/solr/cores/wikipedia_de
Data	/opt/solr/test/datadir/wikipedia_de
Index	/opt/solr/test/datadir/wikipedia_de/index

### Versions

0.00 MB	
<b>solr-spec</b>	4.0.0.2012.11.20.13.23.43
solr-impl	4.0-SNAPSHOT 1410309 - bpflugfelder - 2012-11-20 13:23:43
<b>lucene-spec</b>	4.0-SNAPSHOT
lucene-impl	4.0-SNAPSHOT 1410309 - bpflugfelder - 2012-11-20 11:30:11

### JVM

Runtime	Java HotSpot(TM) 64-Bit Server VM (23.5-b02)
Processors	4
Args	-Djava.io.tmpdir=/Library/Tomcat6/apache-tomcat-6.0.36/temp -Dcatalina.home=/Library/Tomcat6/apache-tomcat-6.0.36 -Dcatalina.base=/Library/Tomcat6/apache-tomcat-6.0.36 -Djava.endorsed.dirs=/Library/Tomcat6/apache-tomcat-6.0.36/endorsed -Djava.util.logging.manager=org.apache.juli.ClassLoaderLogManager -Djava.util.logging.config.file=/Library/Tomcat6/apache-tomcat-6.0.36/c...

### System

Physical Memory	40.3%	3.23 GB / 8.00 GB
Swap Space	0.0%	0.00 MB / 64.00 MB
File Descriptor Count	1.2%	119 / 10240
JVM-Memory	9.5%	105.86 MB / 187.75 MB / 1.09 GB

### + **Reife Technologie, weit verbreitet in kommerziellen Anwendungen**

- ▶ Einfaches Zusammenspiel mit Third-Party-Anwendungen
- ▶ Große Community, gute Dokumentaion, guter Support
- ▶ Auftretende Probleme wurden meist schon mal gelöst
- ▶ Hilfreiche Tools für Analyse und Monitoring

### + **Großer Funktionsumfang**

- ▶ Viele “Analzyser” und Abfragetypen
- ▶ Tuningmöglichkeiten zur Verbesserung der Relevanz
- ▶ webbasierte Administrationsoberfläche

### - **etwas “schwergewichtig”:**

- ▶ viel zu konfigurieren
- ▶ Konfiguration ist größtenteils statisch, kein Zugriff über API o.ä.
- ▶ z. T. redundante Funktionalität

## “Elasticsearch is a ‘distributed-from-scratch’ search server based on Lucene”

- ▶ Entwickelt von Shay Banon, erste öffentlich verfügbare Version in 02/2010:
- ▶ Aktuelle Version 0.19.11
- ▶ Lizenziert unter Apache License 2.0
- ▶ Kleines Kernteam von Entwicklern, starke Unterstützung von einigen Lucene Committern
- ▶ Vielsesprechende Liste von Anwendern:
  - ▶ Mozilla, StumbleUpon, Sony, Infochimps, Assistly, Klout
  - ▶ <http://www.elasticsearch.org/users/>
- ▶ Shay Banon u.a. gründeten 2012 [elasticsearch.com](http://www.elasticsearch.com) und erhielten im November 10 Mio. US\$ Venture Capital

### Architektur

- ▶ Reines Java, JSON als Datenmodell
- ▶ Suche, Indizierung und Scoring mit Lucene
- ▶ Dokumentenorientiert
- ▶ Schemalos
- ▶ HTTP & JSON API für alle Aufgaben (Suche, Indizierung, Administration)
- ▶ Verteilung ist elementarer Bestandteil

### Search Highlights

- ▶ Facetten und Filter
- ▶ Geo-Suche (“GeoShape Query”)
- ▶ Caching
- ▶ Sortierung, Highlighting
- ▶ “More Like This” Vorschläge, basierend auf Dokumenten oder Feldern
- ▶ Multi Tenancy (Unterstützung mehrerer Indizes in einer Suchanfrage)

- + **Einfache aber effektive Architektur**
- + **Einfache Anwendung, sogar in verteilten Umgebungen**
- + **Hoher Reifegrad obwohl noch junges Produkt**
- + **benutzt moderne Technologien**
- + **nur HTTP und JSON– ein Traum für Web Entwickler**
- > **prädestiniert für SBAs und Search Based BI Applications**
  
- **Noch kleine Community und kleines Team von Entwicklern**
- **Im Vergleich zu Solr:**
  - Weniger Abfrage Typen
  - Weniger Tuningmöglichkeiten
  - Weniger Analyzer
  - Fehlende Features wie Clustering, Autocomplete, Spell Checking

1. Business Intelligence – Es wird Zeit für „Intelligence“.
2. Big Data – größer, schneller, weiter. Was ist daran neu?
3. Search – ein alternativer Zugang zu BI und Big Data?
4. **Fazit – Drei Seiten einer Medaille?**

### Open Source

holt auf

#### ▶ Business Intelligence

- ▶ bleibt auch „im klassischen Sinne“ relevant
- ▶ Single Source of Truth weiterhin wichtig
- ▶ Steigenden Anforderungen bzgl. Usability und Analyse

führend

#### ▶ Big Data

- ▶ neue Datenquellen ergänzen bisher verwendete Informationer
- ▶ Menge und Art der Daten erfordern neue Verfahren und Architekturen
- ▶ Mehr Daten liefern bessere Analysen

führend

#### ▶ Search

- ▶ Für textuelle Daten besser geeignet als BI Lösungen
- ▶ Search Based BI Applications können Komplexität verbergen
- ▶ Für hohe Datenvolumen und Nutzerzahlen ausgelegt

## Kontakt

Patrick Thoma  
Head of Solution Development

inovex GmbH  
Karlsruher Str. 71  
75179 Pforzheim

Mobil: 0173/3181009

Mail: [patrick.thoma@inovex.de](mailto:patrick.thoma@inovex.de)



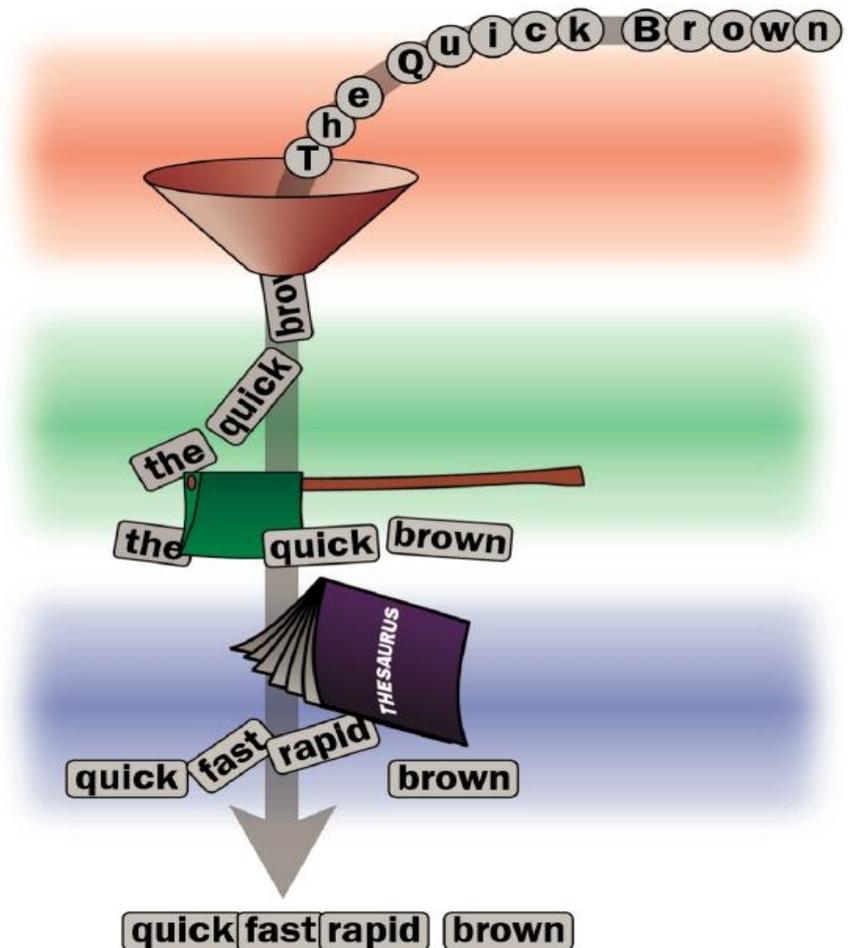
# Backup

# Analyzing / Tokenization

Search in a nutshell

## Break stream of characters into tokens / terms

- Normalization (e.g. case)
- Stopwords
- Stemming
- Lemmatizer / Decomposer
- Part of Speech Tagger
- Information Extraction



# Stopwords

## Search in a nutshell

- **function words do not bear useful information for searching**  
of, in, about, with, I, although, ...
- **Stoplist: contain stopwords, not to be used as index**
  - Prepositions
  - Articles
  - Pronouns
  - Some adverbs and adjectives
  - Some frequent words (e.g. document)
- **The removal of stopwords usually improves IR effectiveness**
- **A few “standard” stoplists are commonly used**

# Linguistic normalization

Search in a nutshell

- **Why is linguistic normalization important?**
  - Due to inflection forms different words may bear the same meaning (e.g. search, searching)
  - If search algorithms do not handle such inflection forms explicitly, queries like search & searching would result different search results
  - Goal: create a “standard” representation at indexing and query time
- **Stemming:**
  - Apply strict algorithmic normalization of inflection forms (e.g. Porter, M.F., 1980, An algorithm for suffix stripping, Program, 14(3) :130-137)
  - Strategy: removing some endings of words. Example:
  - *computer, compute, computes, computing, computed, computation* are all normalized to *comput*
  - But: going -> go, king -> k ??????????????

# Linguistic normalization (cont'd)

Search in a nutshell

- **Morphological Analyzer:**

- **Lemmatizer:** maps words to their base form

## English

going -> go (Verb)

bought -> buy (Verb)

bags -> bag (Noun)

## German

lief -> laufen (Verb)

rannte -> rennen (Verb)

Bücher -> Buch (Noun)

- **Decomposer:** decomposes words into their compounds

Kinderbuch (children's book) -> Kind (Noun) | Buch (Noun)

Versicherungsvertrag (contract of insurance) -> Versicherung (Noun) | Vertrag (Noun)

Holztisch (wooden table), Glastisch (table made of glass)

# Linguistic normalization (cont'd)

Search in a nutshell

## Suchergebnisse für "buchen"

IHRE SUCHE

DATUM FILTERN

GENAUES DATUM VOM   BIS

8044 ERGEBNISSE

Sortieren nach:

Filter:



STUDENTENANDRANG

### Willkommen in der großen Maschine Universität

... rückt, der lieber dort als in Saarbrücken studiert usw. Darum müssen die Unis die Studiengänge "über**buchen**". Zum Semesterstart merken sie dann oft, dass sie sich geirrt

haben: Das "Annahmeverhalten" war besser als geschätzt. Dann haben mehr

Studierende [\[weiter...\]](#)

18.10.2011, ZEIT ONLINE



SACHBUCH

### Liebe und solche Sachen

..., wird nie mehr in aller Unschuld in diesem angesagten Restaurant für einen Jahrestag der Liebe den teuren Fensterplatz **buchen** oder naiv die Reise zu zweit in den

Süden für eine individuelle Entscheidung halten, also blind sein gegenüber dem

Ausleben von [\[weiter...\]](#)

15.10.2011, DIE ZEIT



ZUFLUCHTSORT DEUTSCHLAND

### Das gelobte Land

... Familie schon einmal einen Kurzurlaub in Deutschland **gebucht**, auf dem Oktoberfest waren sie. Die Lebensfreude der Deutschen habe seinen Kindern gefallen, sagt er. Der Rest

habe sie eher kaltgelassen, vor allem das Essen. Er lächelt. »Aber vielleicht habe

[\[weiter...\]](#)

03.10.2011, DIE ZEIT



HANDY-FAHRSCHEIN

### Von der Deutschen Bahn verfolgt

... Endpunkt jeder Fahrt einzugeben. Das System berechnet den Fahrpreis und **bucht** ihn vom Konto des Nutzers ab.

Selbstverständlich traut die Bahn ihren Kunden nicht und will überprüfen, ob deren Angaben stimmen. Sie sammelt dazu Bewegungsdaten. Das ist

[\[weiter...\]](#)

27.09.2011, ZEIT ONLINE



SPITZENGASTRONOMIE

### Frankreichs Köche setzen Jean-François Piège auf Platz 1

... ständig **ausgebucht**. Die Köche als Leser des Branchenmagazins Le Chef konnten den ihrer Ansicht nach

besten Kollegen selbst bestimmen. Es gibt keine Liste oder Vorauswahl. Neben Piège kürten sie Alexandre Jean zum "Sommelier des Jahres" [\[weiter...\]](#)

27.09.2011, ZEIT ONLINE

<http://www.zeit.de/suche/index?q=buchen>