News von der Hadoop World 2011

Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Hadoop World 2011

- 1.400 Teilnehmer
- 580 Unternehmen
- 27 Länder
- 120 Server/Cluster
- Größter Cluster > 20PB

Trakken {technologies:GmbH}

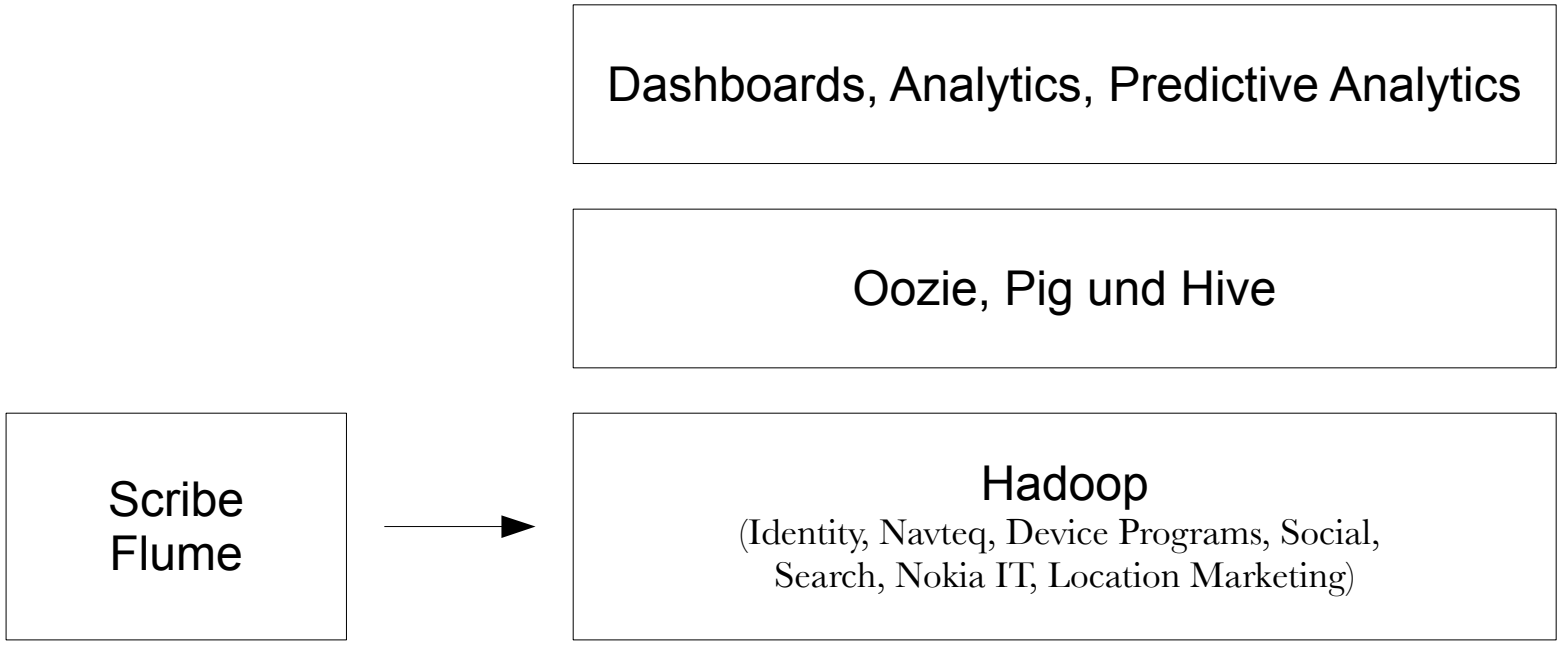
Quelle: http://www.cloudera.com/resources/Hadoop+World/

Nokia Fakten

- 2 TB Data/Tag
- 350 M Messages/Tag via Scribe, Flume
- 3000 MR Jobs/Tag
- 10 TB/Tag Daten die verarbeitet werden
- Hadoop (Identity, Navteq, Device Programs, Social, Search, Nokia IT, Location Marketing)

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Nokia Analytics



Quelle: http://www.cloudera.com/resources/Hadoop+World/

Ebay – Fakten

- 2 Milliarden Pageviews/Tag
- 250 Mio SearchEngine Queries
- 75 Milliarden DB Calls
- 200 Mio Product Items
- 9 PB Data

- Titel wird als Default für die Suche verwendet
- Wenig Suchfunktionalität

Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Ebay – Projekt Cassini

- Hbase speichert Product Items
- Transfer in Hadoop
- Inverted Index für Items
- Indexing in Hadoop
- SearchNodes für den Index

- Start 2012
- Neues Rechenzentrum

Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

LinkedIn – Transform Raw Data to Rich Features

- Empfehlungen: Jobs, Artikel, Unternehmen, Personen, Match auf Skills
- 50% Empfehlungen basieren auf Hadoop

- Clean Data: Job Titel in Skills umwandeln
- Clean Data: Unternehmensvariationen (IBM 8000)

Quelle: http://www.cloudera.com/resources/Hadoop+World/

LinkedIn – Transform Raw Data to Rich Features 2

- **Blending Recommendations**: Was sehen sich Nutzer mit ähnlichen Profilen an, kann man diese Cross-Empfehlen
- **Job Seeker Score**: Wahrscheinlichkeit des nächsten Karriereschritts
- **A/B Testing**: Ausliefern einer Idee an kleine Nutzergruppe

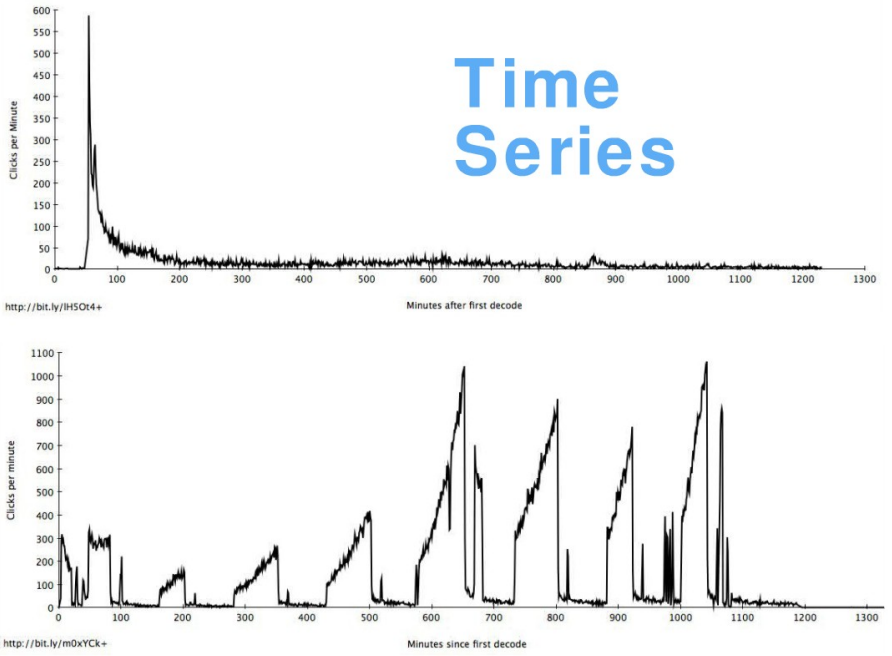
- Technik: Hadoop, Mahout, Lucene, Zoie, Voldemort, Kafka

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Bit.ly – Measuring Organic Traffic

- Predict clicks remaining
- Predictions in realtime

- organic vs. in-organic
- spot sharing events
- spot abnormal patterns (spam, inorganic, ...)

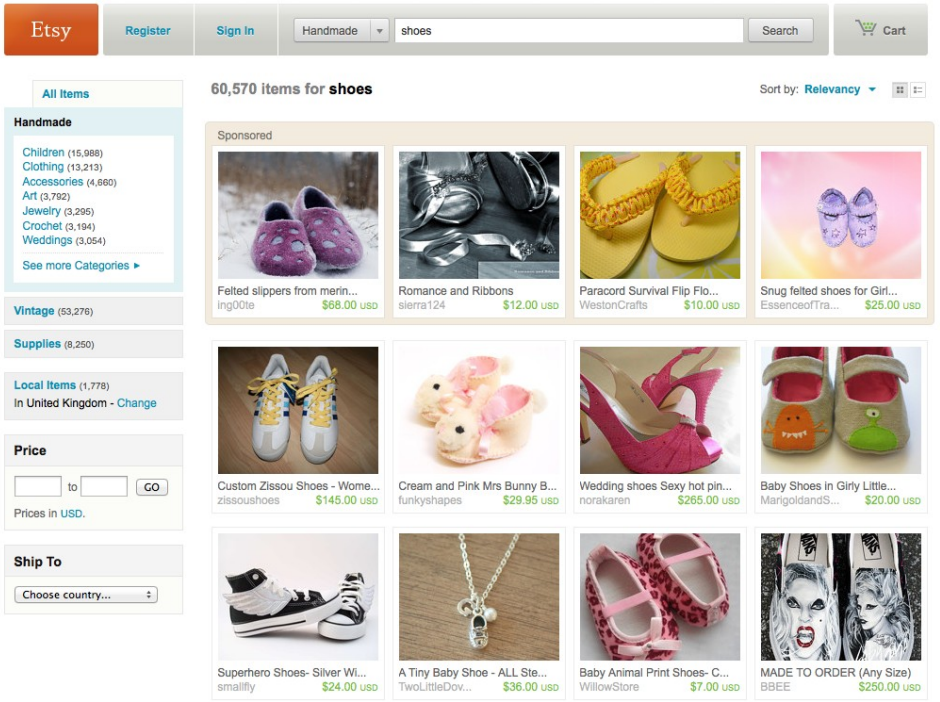


Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Etsy.com – Data Mining for Product Search Ranking

- Suchanfragen analysieren
- Nutzverhalten bewerten
- Produktattribute
- Produktbeschreibungen



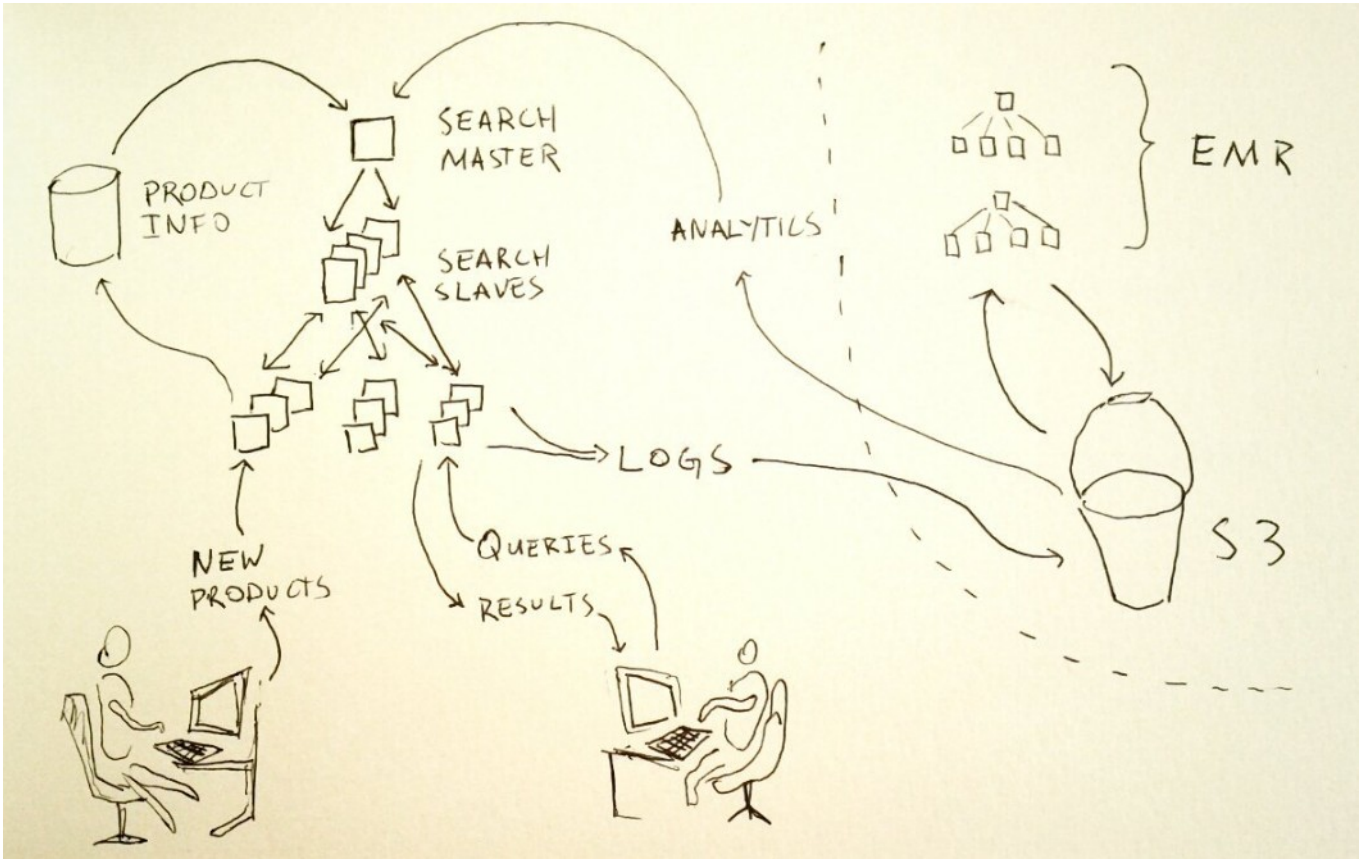
=> Zusammenhänge herstellen

Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Etsy.com – Product Search Ranking 2

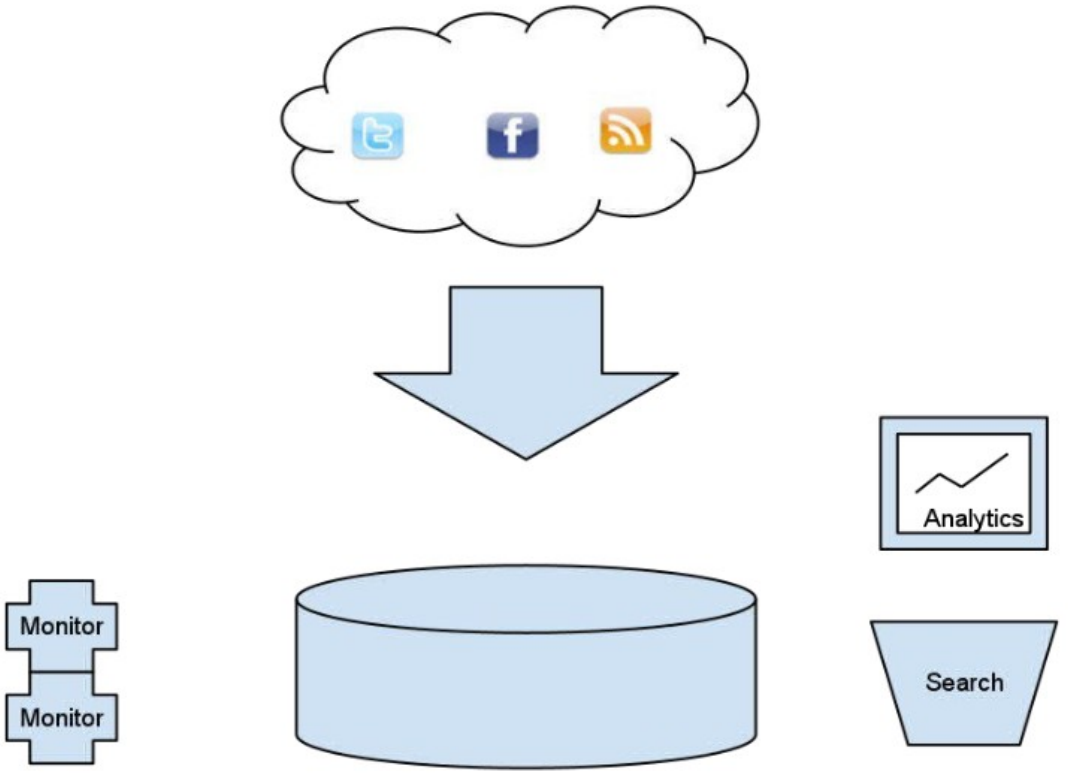
- Hadoop
- Solr
- ElasticMR
- Mahout



Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Jive – Storing and Indexing Social Media Content



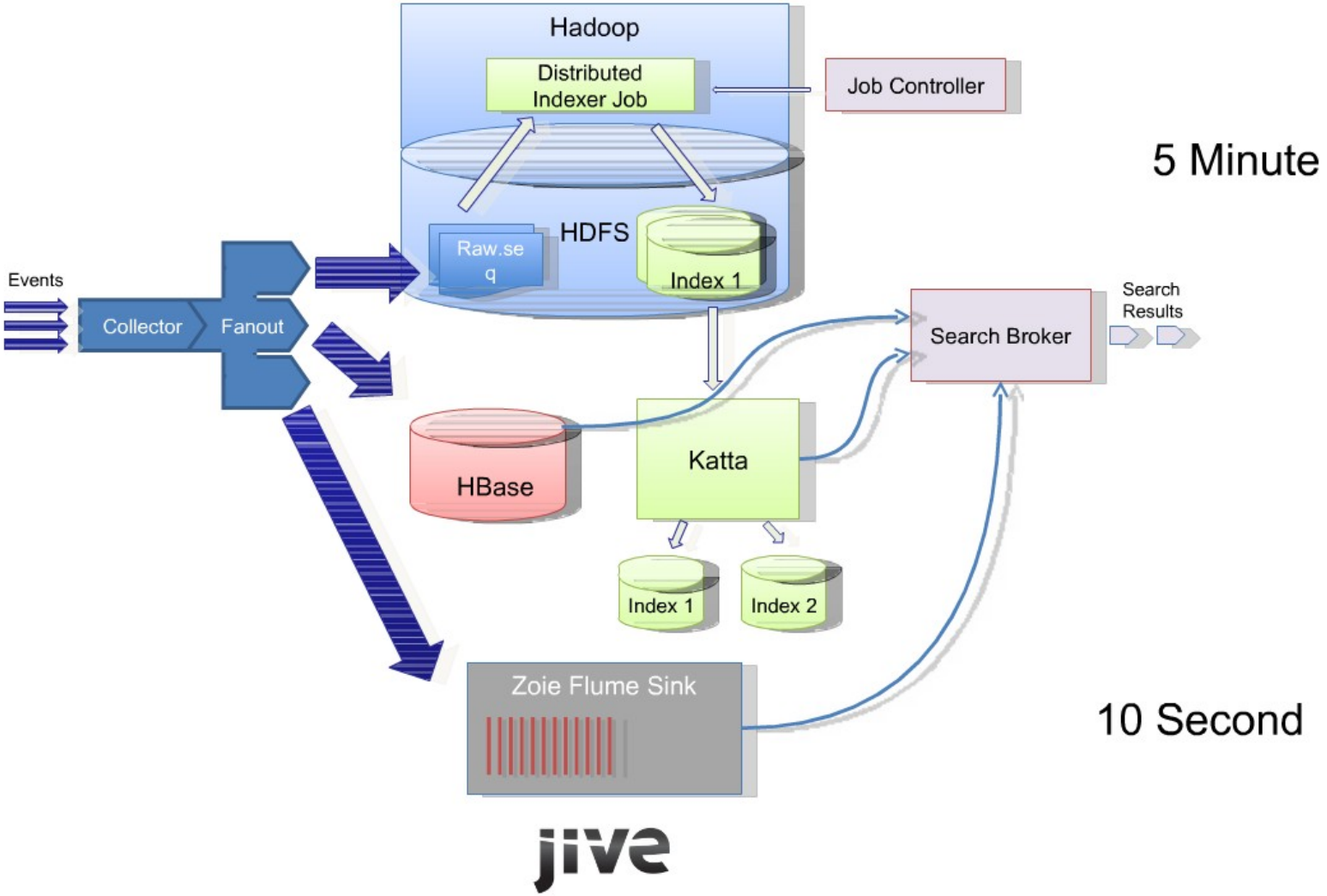
Jive Social Media Engagement stores social media for monitoring (e.g. brand sentiment), searching, and analysis

jive

Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Jive – Storing and Indexing Social Media Content 2

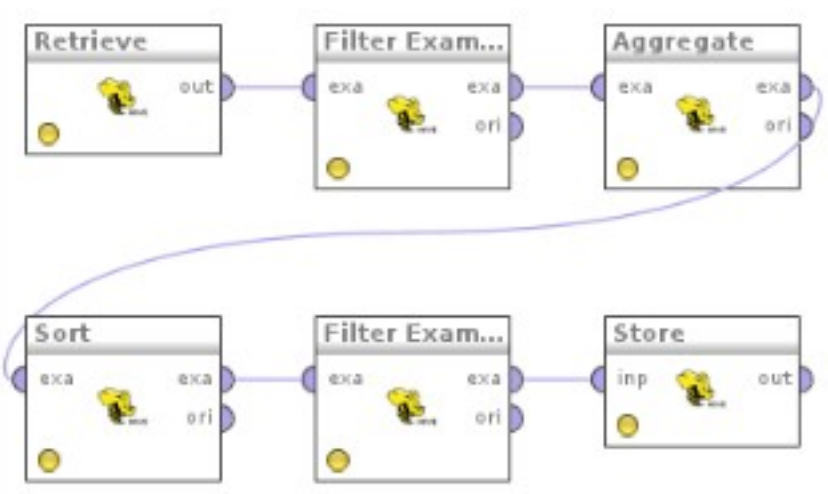


Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Radoop

- Grafische Analyse-Oberfläche für Hadoop, Hive und Mahout
- Soll Einstiegsbarriere zu Hadoop senken
- Aus Operatoren und Prozessen lässt sich ein Workflow erstellen
- OpenSource



Trakken {technologies:GmbH}

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Ecosystem

- Hadoop als Betriebssystem

Entwicklung von Erweiterungen mit Hadoop als Basis:

- Radoop
- Mahout, Lucene, Solr, EMR, Hive, ….
- Karmasphere (Netbeans, Eclipse)
- Cloudera (Service and Configuration Manager)

Quelle: http://www.cloudera.com/resources/Hadoop+World/

Zukunft

- Cloudera => VC $100 Millionen
- Accel Partners => BigData Fund von $100 Millionen
- Skalierbarkeit
- Verfügbarkeit
- Performance
- MapReduce 2
- Mahout
- Avro

Quelle: http://www.cloudera.com/resources/Hadoop+World/

R und Hadoop

- R Addon für Hadoop und HBase
- R MapReduce sehr nah an R
- Keine Grenzen im Vergleich zu Hive and Pig
- Deutlich weniger Code für einen MapReduce Job
- https://github.com/RevolutionAnalytics/RHadoop/wiki/rhbase

Quelle: http://www.cloudera.com/resources/Hadoop+World/