Zürcher Hochschule
für Angewandte Wissenschaften
School of
Engineering
InIT Institut für angewandte
Informationstechnologie

Big Data Anwendungen – Chancen und Risiken

Dr. Kurt Stockinger
Studienleiter Data Science, Dozent für Informatik
Zürcher Hochschule für Angewandte Wissenschaften

Big Data Workshop „Squeezing more out of Data“
Regensdorf, 11. Juni 2015

Inhalt

- ZHAW Datalab
- Moore‘s Law, Big Data, Data Warehousing
- Big Data Erfahrung aus Lehre:
  - MapReduce
  - Pig
  - Hive
- Big Data Erfahrung aus Forschung:
  - Cloudera Impala

Datalab = Data Science @ ZHAW

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

The ZHAW Data Science Laboratory

datalab

dlab.zhaw.ch

- Eines der ersten Data Science Labs in Europa
- Zusammenschluss von Informatikern, Statistikern, Mathematikern und Physikern zur Lösung von Data Science Problemen in Forschung, Lehre und Praxis:
  - Institut für Angewandte Informationstechnologie
  - Institut für Datenanalyse und Prozessdesign
  - Zentrum für Sozialrecht
  - Institut für Angewandte Mathematik und Physik (neu)
  - Institut für Angewandte Simulation (neu)

Erstes DAS Data Science in

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

DAS = Diploma of Advanced Studies

Besteht aus 3 CAS (Teilzeit, ein Nachmittag pro Woche):

- CAS Information Engineering:
  - Information Retrieval & Text Analysis
  - Data Warehousing & Big Data
- CAS Data Analysis:
  - Statistical Inference, Regression, Time Series Analysis
  - Descriptive Statistics, Clustering, Classification
- Data Science Applications:
  - Machine Learning
  - Visualization of Big / High-Dimensional Data
  - Data Protection Laws & Data Security

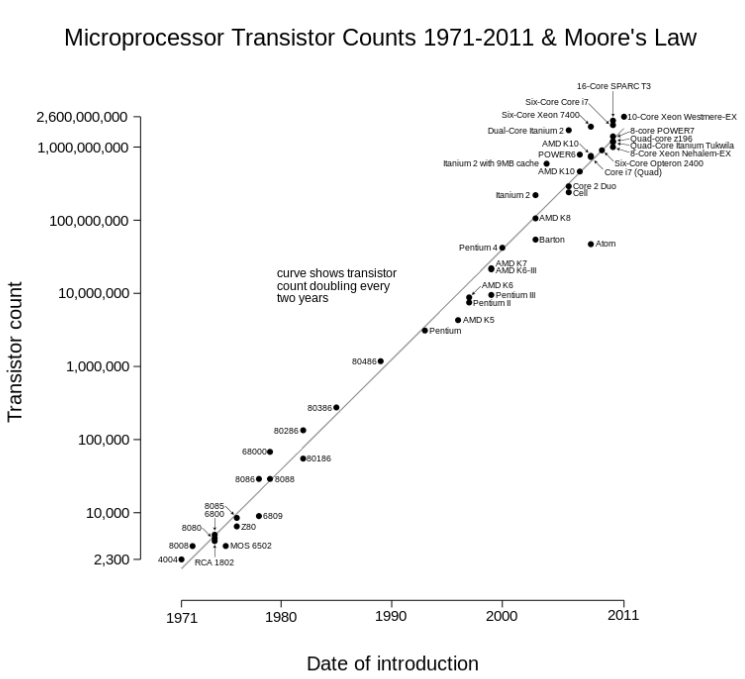
- Mehr Information:
  - http://www.weiterbildung.zhaw.ch/de/programm/das-data-science.html

Inhalt

- ZHAW Datalab
- Moore‘s Law, Big Data, Data Warehousing
- Big Data Erfahrung aus Lehre:
  - MapReduce
  - Pig
  - Hive
- Big Data Erfahrung aus Forschung:
  - Cloudera Impala

Moore‘s Law – Ursprung

- Anzahl der Transistoren in einem Chip verdoppelt sich alle 18 Monate (ursprünglich: alle 24 Monate)
- Exponentielles Wachstum




SECTIONS HOME SEARCH

The New York Times

OPINIONATOR | PRIVATE LIVES
Shopping for Antiques, Finding My Mother

EDITORIAL
A Tiny Crack in the Russian Ice

CHARLES M. BLOW
The President, Fox News and the Poor

GAIL COLLINS
Wow, Jeb Bush

The Opinion Pages | OP-ED COLUMNIST

Moore’s Law Turns 50

MAY 13, 2015

Thomas L. Friedman

Email
Share
Tweet

SAN FRANCISCO — On April 19, 1965, just over 50 years ago, Gordon Moore, then the head of research for Fairchild Semiconductor and later one of the co-founders of Intel, was asked by Electronics Magazine to submit an article predicting what was going to happen to integrated circuits, the heart of computing, in the next 10 years. Studying the trend he’d seen in the previous few years, Moore predicted that every year we’d double the number of transistors that could fit on a single chip of silicon so you’d get twice as much computing power for only slightly more money. When that came true, in 1975, he modified his prediction to a doubling roughly every two years. “Moore’s Law” has essentially held up ever since — and, despite the skeptics, keeps chugging along, making it probably the most remarkable example ever of sustained exponential growth of a technology.

Querlle: http://www.nytimes.com/2015/05/13/opinion/thomas-friedman-moores-law-turns-50.html?smid=nytcore-iphone-share&smprod=nytcore-iphone&_r=0

Moore‘s Law auf VW Käfer angewandt

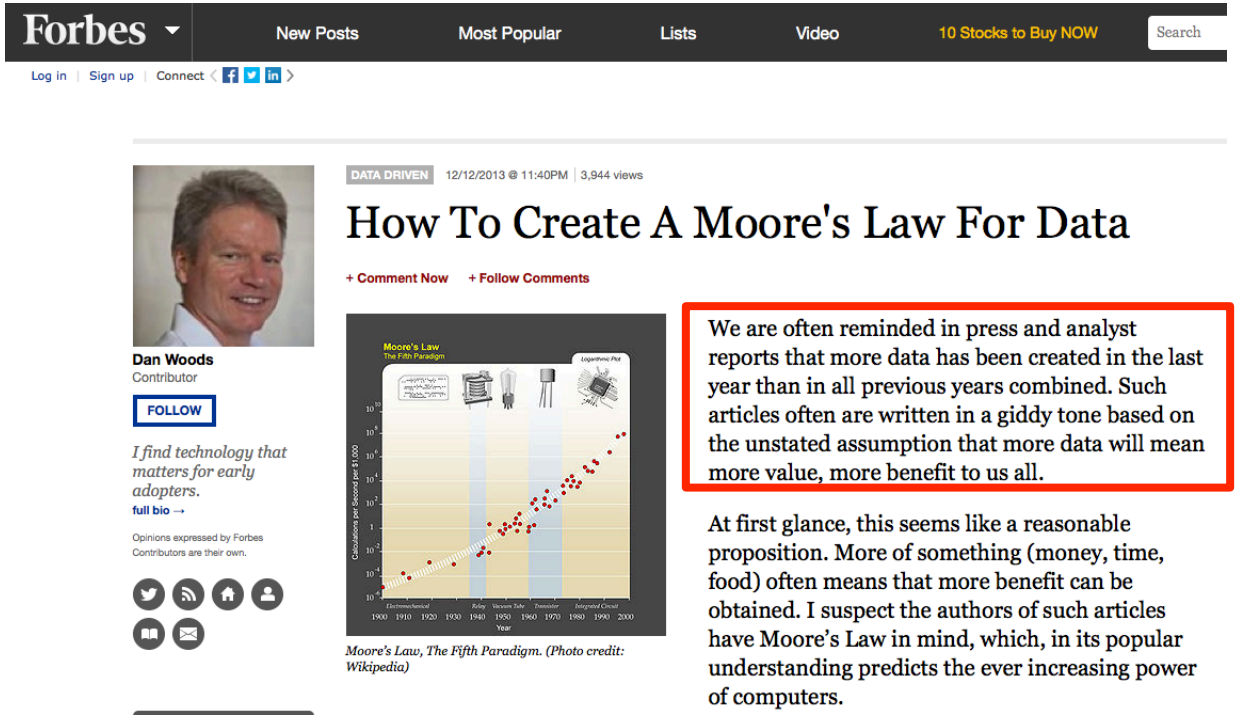
- Gedankenexperiment
  - Wie hätte sich ein VW Käfer seit 1971 gemässt Moore‘s Law entwickelt?
  - Geschwindigkeit?
  - Benzinverbrauch?

Moore‘s Law auf VW Käfer angewandt

- Gedankenexperiment
  - Geschwindigkeit: ~ 480‘000 km/h
  - Benzinverbrauch: ~ 0.00011875 Liter pro 100 km

Moore‘s Law - Daten

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie



Forbes | New Posts | Most Popular | Lists | Video | 10 Stocks to Buy NOW | Search

Log in | Sign up | Connect

Dan Woods
Contributor

FOLLOW

I find technology that matters for early adopters.

full bio →

Opinions expressed by Forbes Contributors are their own.

DATA DRIVEN 12/12/2013 @ 11:40PM | 3,944 views

How To Create A Moore's Law For Data

+ Comment Now + Follow Comments

Moore's Law, The Fifth Paradigm. (Photo credit: Wikipedia)

We are often reminded in press and analyst reports that more data has been created in the last year than in all previous years combined. Such articles often are written in a giddy tone based on the unstated assumption that more data will mean more value, more benefit to us all.

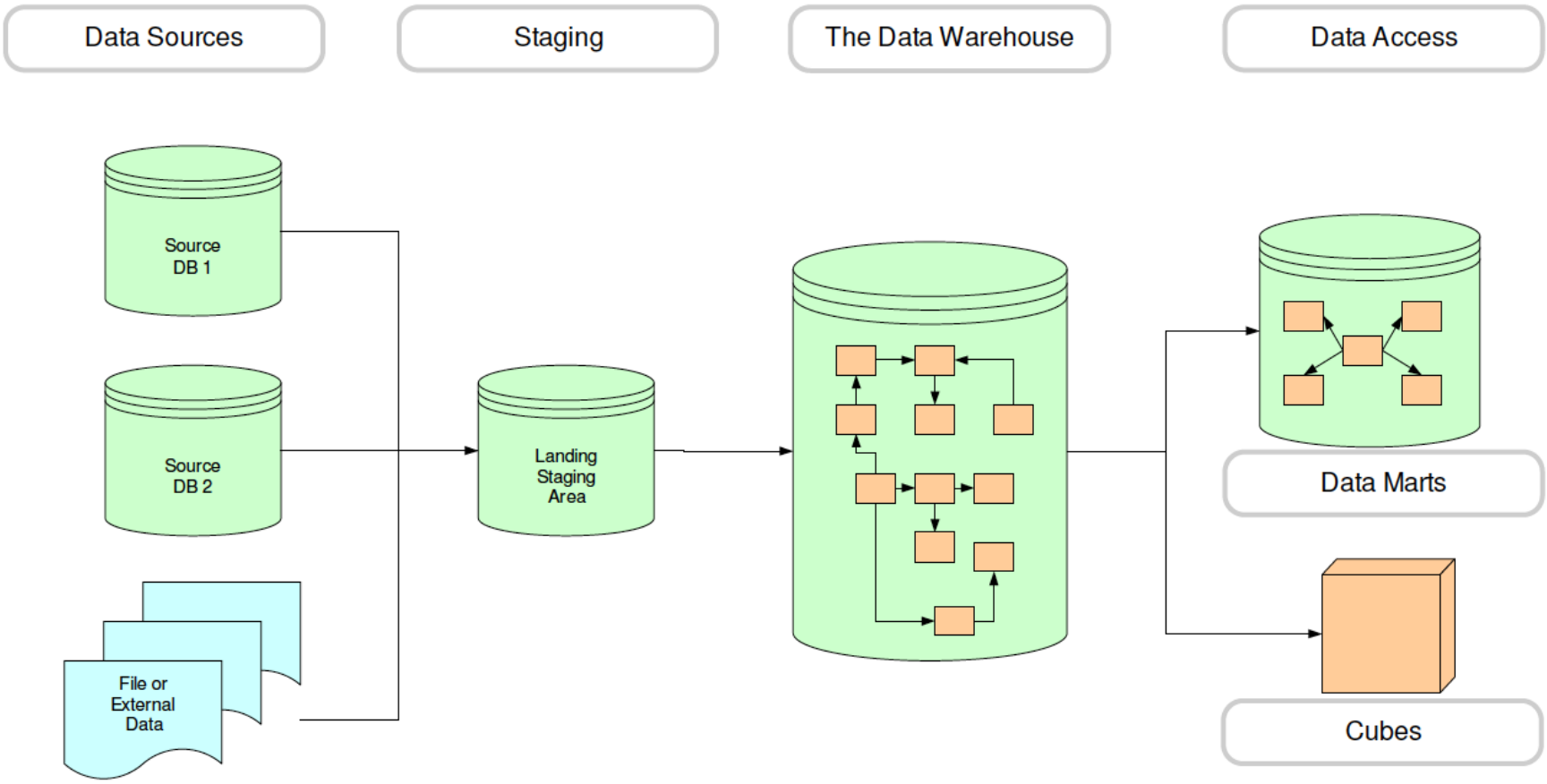
At first glance, this seems like a reasonable proposition. More of something (money, time, food) often means that more benefit can be obtained. I suspect the authors of such articles have Moore's Law in mind, which, in its popular understanding predicts the ever increasing power of computers.

But a closer look at the world of data shows that there is no Moore’s Law in effect.
[..] More data means more costs for storage, for governance and having too much unorganized data may make it more difficult to find what you need. In other words more data can mean less value.

Quelle: http://www.forbes.com/sites/danwoods/2013/12/12/how-to-create-a-moores-law-for-data/

Klassisches Big Data im Unternehmen: Data Warehouse

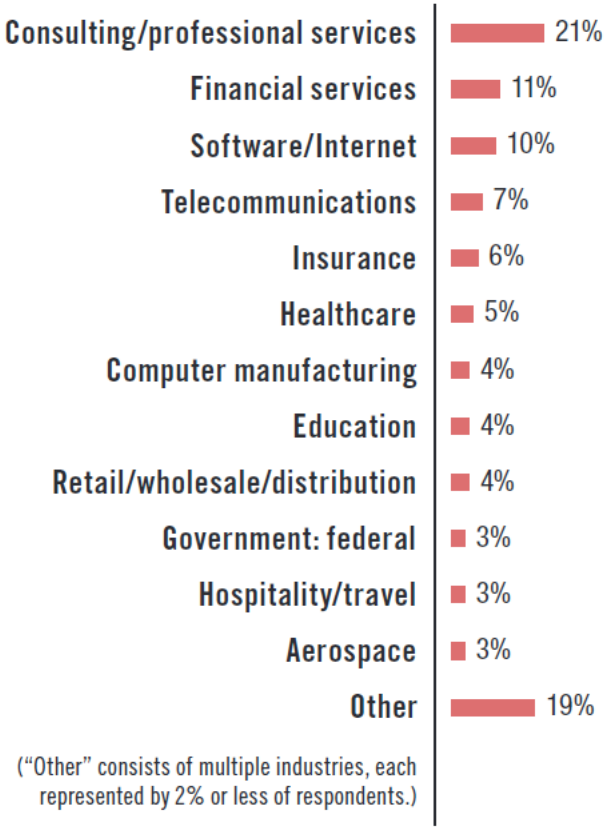
Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie



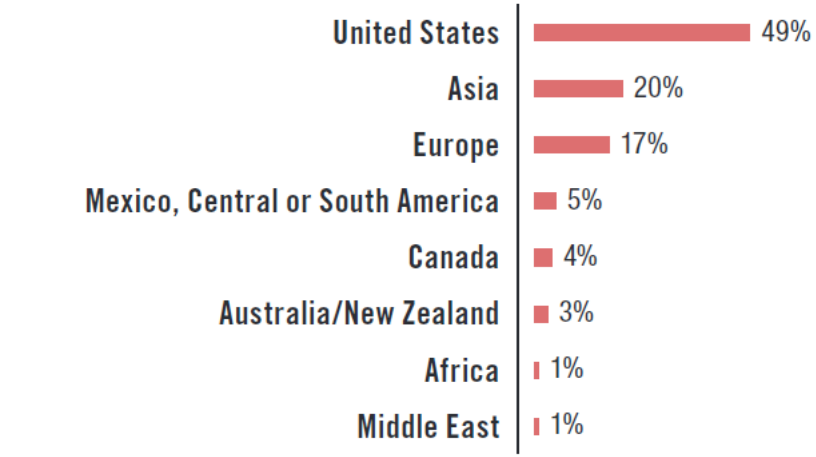
Studie vom TDWI:
The Data Warehouse Institute

Rückmeldungen von ca. 300 Firmen weltweit

Industry

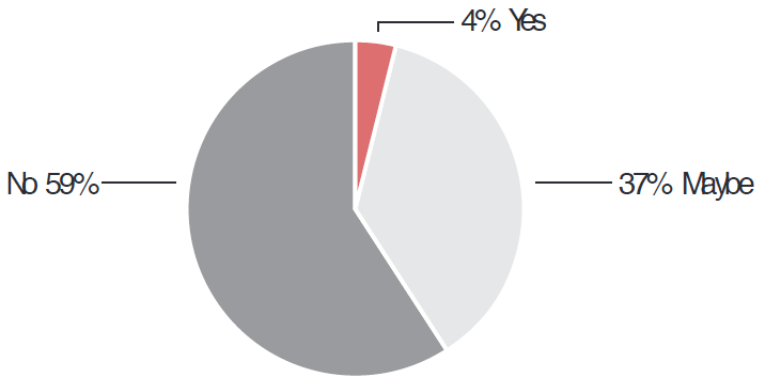


Geography

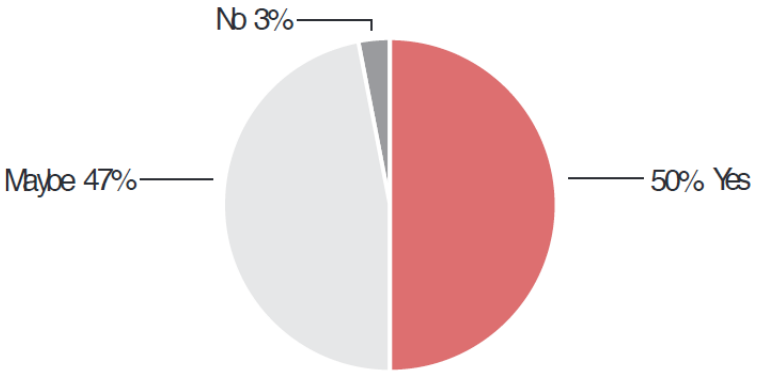


HDFS und DWH

Can the Hadoop Distributed File System (HDFS) replace your enterprise data warehouse (EDW)?

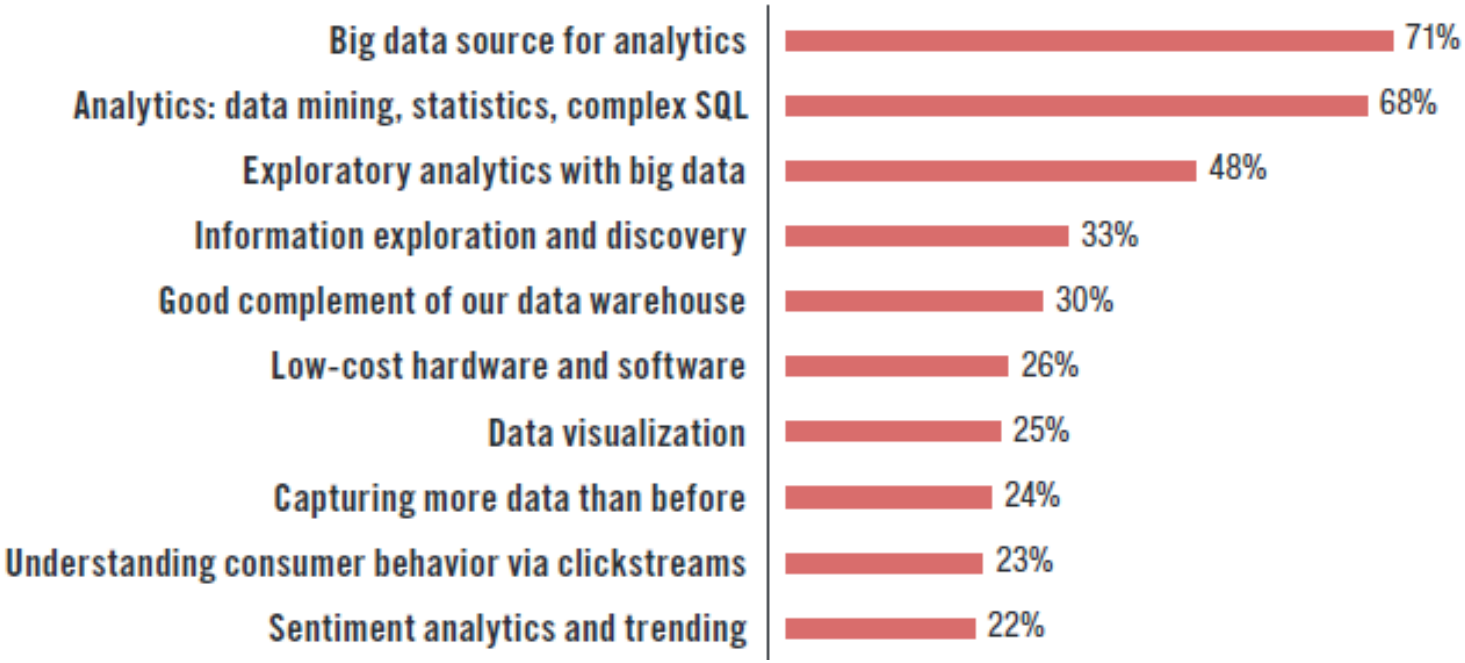


Can HDFS augment your enterprise data warehouse (EDW) or other data infrastructure?



Anwendungen von Hadoop

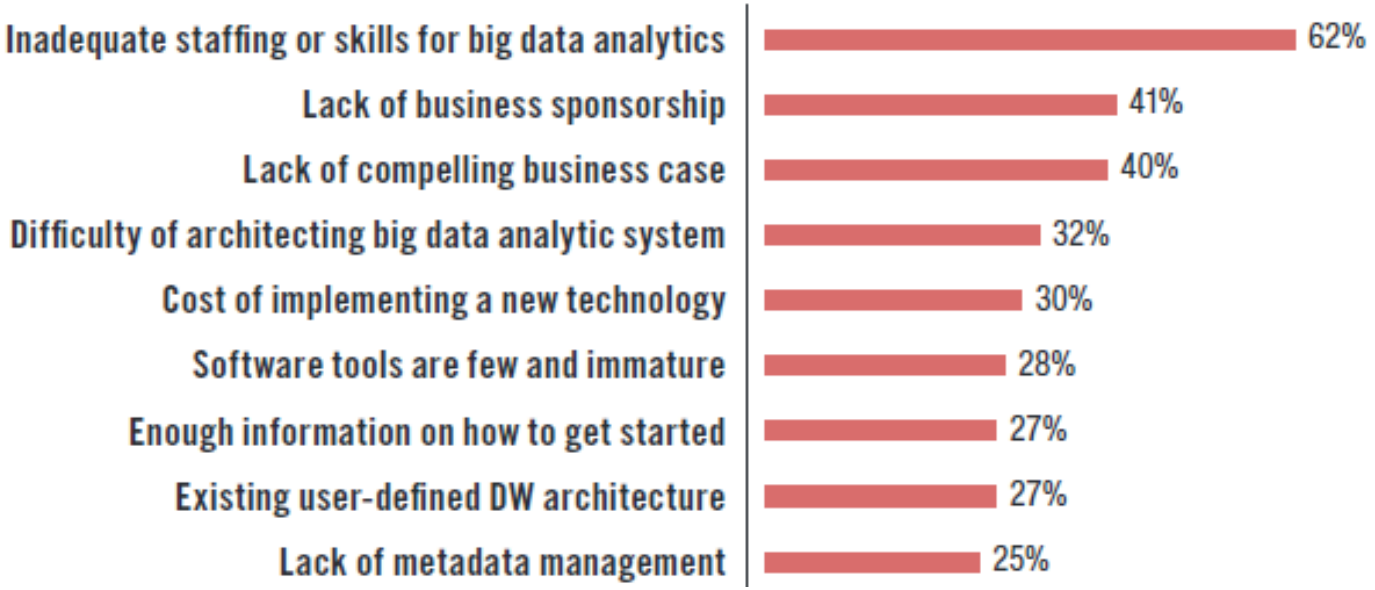
If your organization were to implement Hadoop technologies, which business processes, data, and applications would most likely benefit? Select eight or fewer.



Hindernisse für die Anwendung von Hadoop

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

What are the most likely barriers to implementing Hadoop technologies in your organization? Select eight or fewer.



Big Data Technologien: NoSQL-Systeme

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

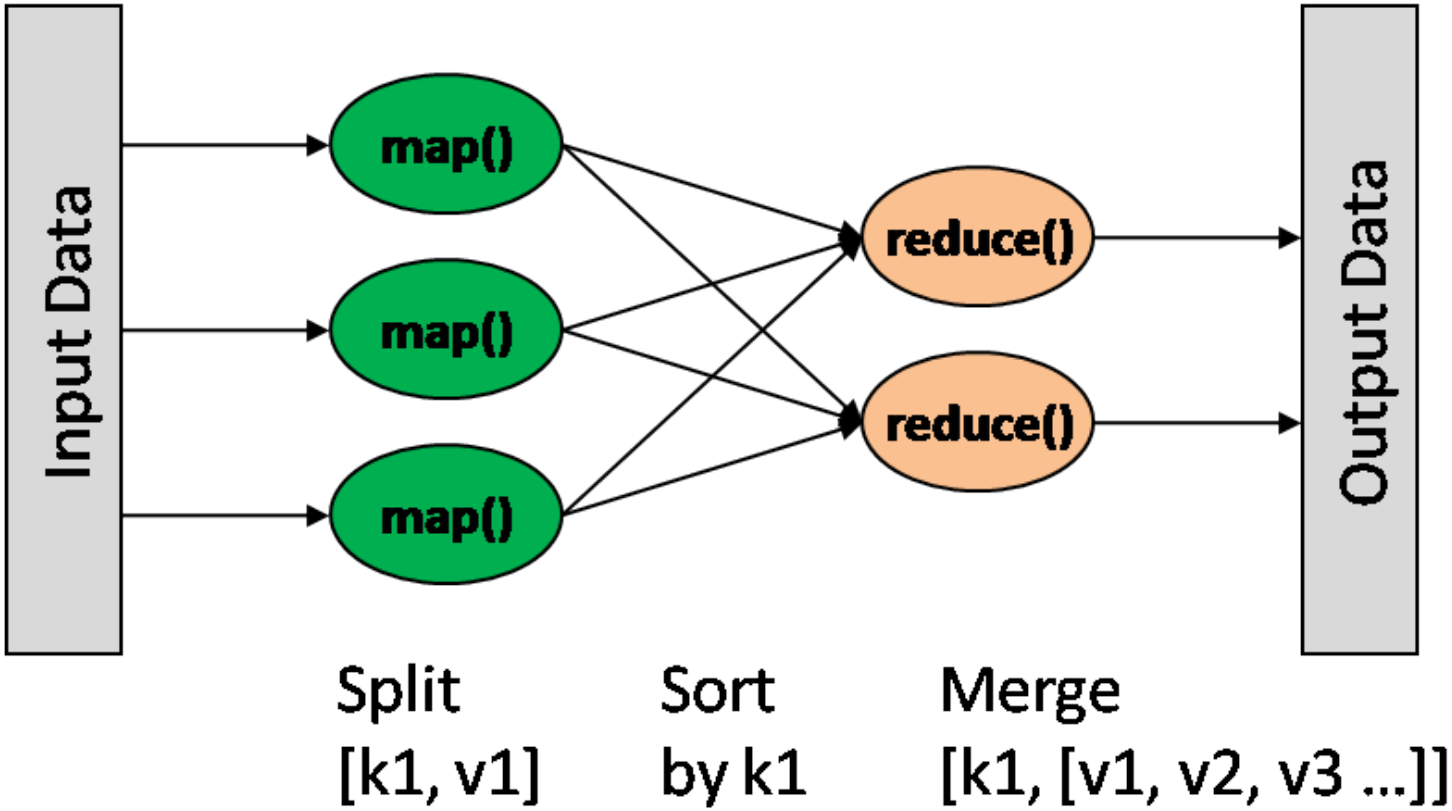
Year	System/ Paper	Scale to 1000s	Primary Index	Secondary Indexes	Transactions	Joins/ Analytics	Integrity Constraints	Views	Language/ Algebra	Data model	my label
1971	RDBMS	O	✔	✔	✔	✔	✔	✔	✔	tables	sql-like
2004	MapReduce	✔	O	O	O	✔	O	O	O	key-val	batch
2005	CouchDB	✔	✔	✔	record	MR	O	✔	O	document	nosql
2006	BigTable/Hbase	✔	✔	✔	record	compat. w/MR	/	O	O	ext. record	nosql
2007	MongoDB	✔	✔	✔	EC, record	O	O	O	O	document	nosql
2008	Pig	✔	O	O	O	✔	/	O	✔	tables	sql-like
2008	HIVE	✔	O	O	O	✔	✔	O	✔	tables	sql-like
2008	Cassandra	✔	✔	✔	EC, record	O	✔	✔	O	key-val	nosql
2009	Voldemort	✔	✔	O	EC, record	O	O	O	O	key-val	nosql
2009	Riak	✔	✔	✔	EC, record	MR	O			key-val	nosql
2010	Dremel	✔	O	O	O	/	✔	O	✔	tables	sql-like
2011	Megastore	✔	✔	✔	entity groups	O	/	O	/	tables	nosql
2011	Tenzing	✔	O	O	O	O	✔	✔	✔	tables	sql-like
2011	Spark/Shark	✔	O	O	O	✔	✔	O	✔	tables	sql-like
2013	Impala	✔	O	O	O	✔	✔	O	✔	tables	sql-like

Tabelle übernommen von Bill Howe, University of Washington (Coursera «Introduction to Data Science»), 2013.

Inhalt

- ZHAW Datalab
- Moore‘s Law, Big Data, Data Warehousing
- Big Data Erfahrung aus Lehre:
  - MapReduce
  - Pig
  - Hive
- Big Data Erfahrung aus Forschung:
  - Cloudera Impala

MapReduce - Überblick

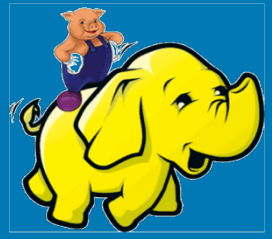


MapReduce - Erfahrung

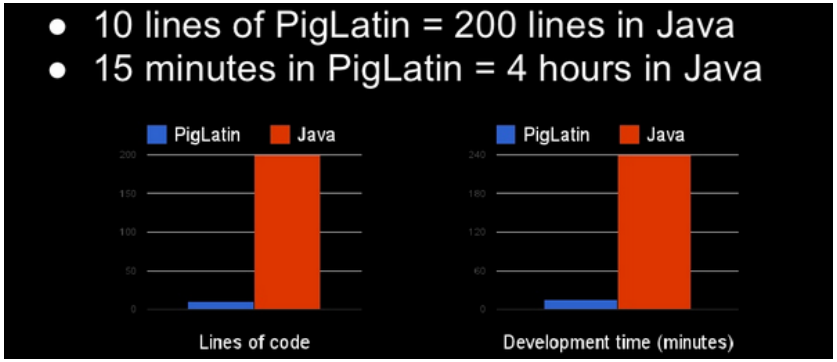
- Vorteil:
  - Mächtiger Programmieransatz, um Daten parallel zu bearbeiten
  - Programmierer muss sich um die Datenaufteilung nicht direkt kümmern, da dies vom Hadoop Distributed File System verwaltet wird
  - Sehr gute Skalierbarkeit für grosse Datenmengen

- Nachteil:
  - Programmieren mit MapReduce ist nicht trivial
  - Java:
    - Grosser Unterschied zwischen Hadoop 1.x und 2.x für Java-API
    - Grossteil der Dokumentation ist für 1.x bzw. für 2.x nur lückenhaft
  - Python:
    - Programmierung in Python ist bedeutend einfacher
    - Python etabliert sich als neben SQL und R als eines der wichtigsten Data Science Tools

Pig - Überblick



- Pig ist eine high-level Plattform, um einfacher MapReduce Jobs auf Hadoop auszuführen:
  - MapReduce Jobs werden automatisch generiert
  - Gute Skalierbarkeit für Big Data
  - Macht Hadoop auch für „Nicht-Programmierer-Experten“ zugänglich
- Die Script-Sprache der Plattform heisst Pig Latin:
  - Ähnlich wie SQL
  - Kein Java-basiertes “low-level” MapReduce notwendig



Quelle: http://www.slideshare.net/AdamKawa/apache-pig-at-whug

Pig - Erfahrung

- Vorteil:
  - Analysen lassen sich viel schneller entwickeln als unter MapReduce
  - Prozeduraler Ansatz:
    - Mischung aus Programmiersprache und SQL
  - Lokaler Modus erlaubt schnelles, iteratives Entwickeln

- Nachteil:
  - „Gewöhnungsbedürftig“, wenn bereits SQL-Kenntnisse vorhanden
  - Debugging von Code nicht trivial, da Pig nicht konsistent bei Case-Sensitivität ist

Hive - Überblick

- Traditionelle DWHs haben Restriktionen:
  - Bestimmter Vendor, teuer, limitierte Skalierbarkeit
- Hadoop/MapReduce hat diese Restriktionen nicht, doch ist das Programmiermodell zu low-level:
  - Speziell geschriebenes Programm
  - Limitierte Wartbarkeit und Wiederverwendbarkeit
- Hive: „Data Warehouse“ basierend auf Hadoop:
  - Vereinigung zweier Welten
  - Ideal für Verarbeitung von Big Data (Batch-Prozessierung)
  - Nicht designed für:
    - Online Transaction Processing (OLTP)
    - Real-Time Queries: Overhead von HDFS

Hive - Erfahrung

- Vorteil:
  - Geringere Lernkurve, wenn SQL-Kenntnisse vorhanden
  - Kostengünstige DWH-Variante
  - Gute Integration mit Business Intelligence Tools

- Nachteil:
  - Nur ein Teil vom SQL-Standard umgesetzt (jedoch rasche Entwicklung)
  - Grossen Antwortzeiten sind Hindernis für Entwicklung:
    - Mindestantwortzeiten von 20 Sekunden pro Query
    - Joins (bei kleinen Datenmengen) können mehrere Minuten dauern

Inhalt

- ZHAW Datalab
- Moore‘s Law, Big Data, Data Warehousing
- Big Data Erfahrung aus Lehre:
  - MapReduce
  - Pig
  - Hive
- Big Data Erfahrung aus Forschung:
  - Cloudera Impala

Cloudera Impala

- Impala ist eine parallele Query Engine basierend auf HDFS
- Impala verwendet nicht MapReduce:
  - Implementiert eigene Parallelisierung ähnlich wie parallele Datenbanken
  - Kein Overhead von MapReduce
  - Ermöglicht batch-oriented und real-time Queries

- Impala vs. Hive:
  - Hive:
    - Fehlertolerante Prozessierung von Batch-Queries
    - Overhead von mindestens 20 Sekunden
  - Impala:
    - Nicht-fehlertolerante Prozessierung von Batch- und near-real Queries

Query Performance von Impala

How Impala Supports Mixed Workloads in Multi-User Environments

by **Justin Kestelyn (@kestelyn)** | **September 17, 2014** | **no comments**

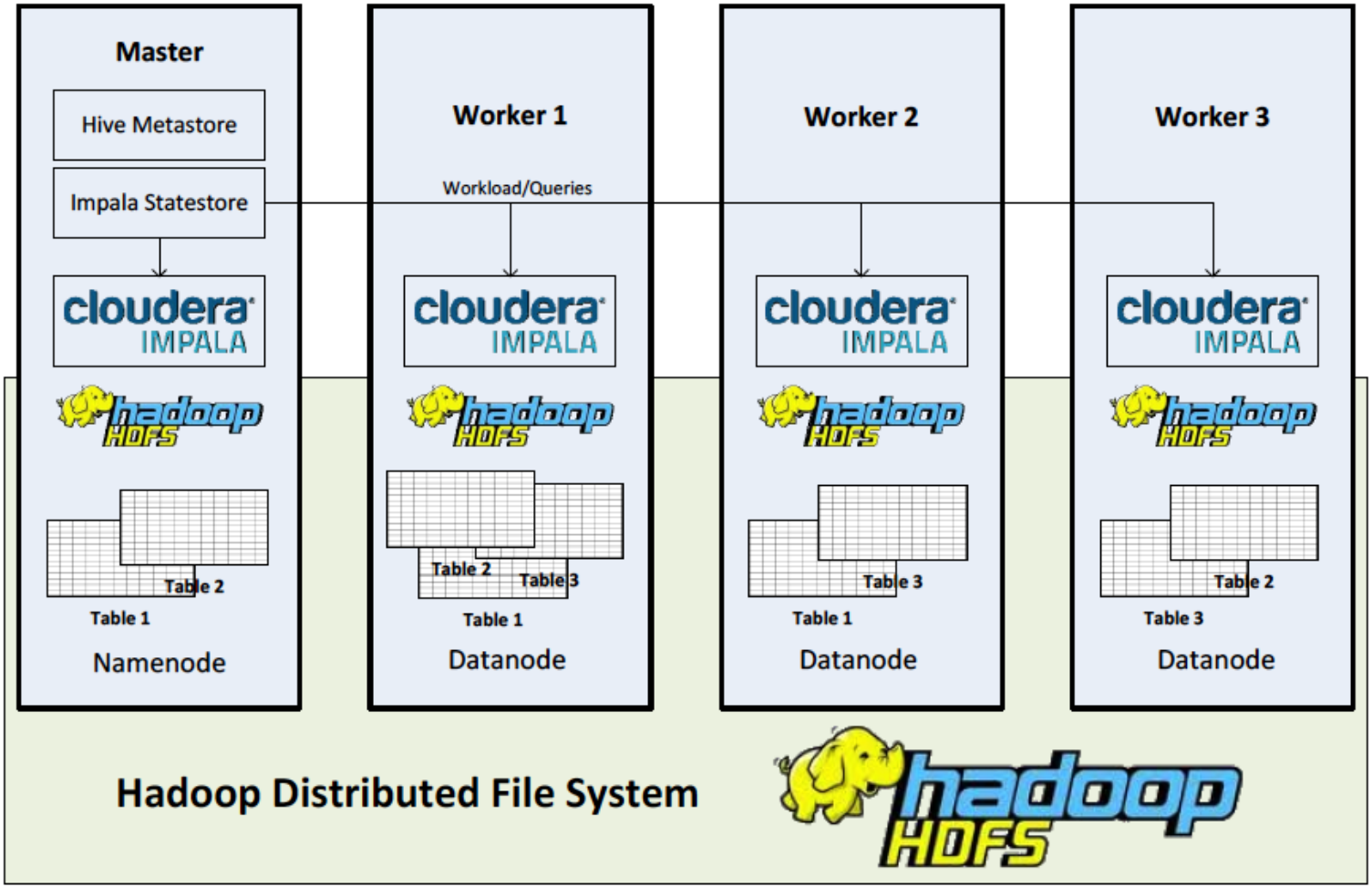
Our thanks to Melanie Imhof, Jonas Looser, Thierry Musy, and Kurt Stockinger of the Zurich University of Applied Science in Switzerland for the post below about their research into the query performance of Impala for mixed workloads.

Recently, we were approached by an industry partner to research and create a blueprint for a new Big Data, near real-time, query processing architecture that would replace its current architecture based on a popular open source database system.

Read More...

http://blog.cloudera.com/blog/2014/09/how-impala-supports-mixed-workloads-in-multi-user-environments/

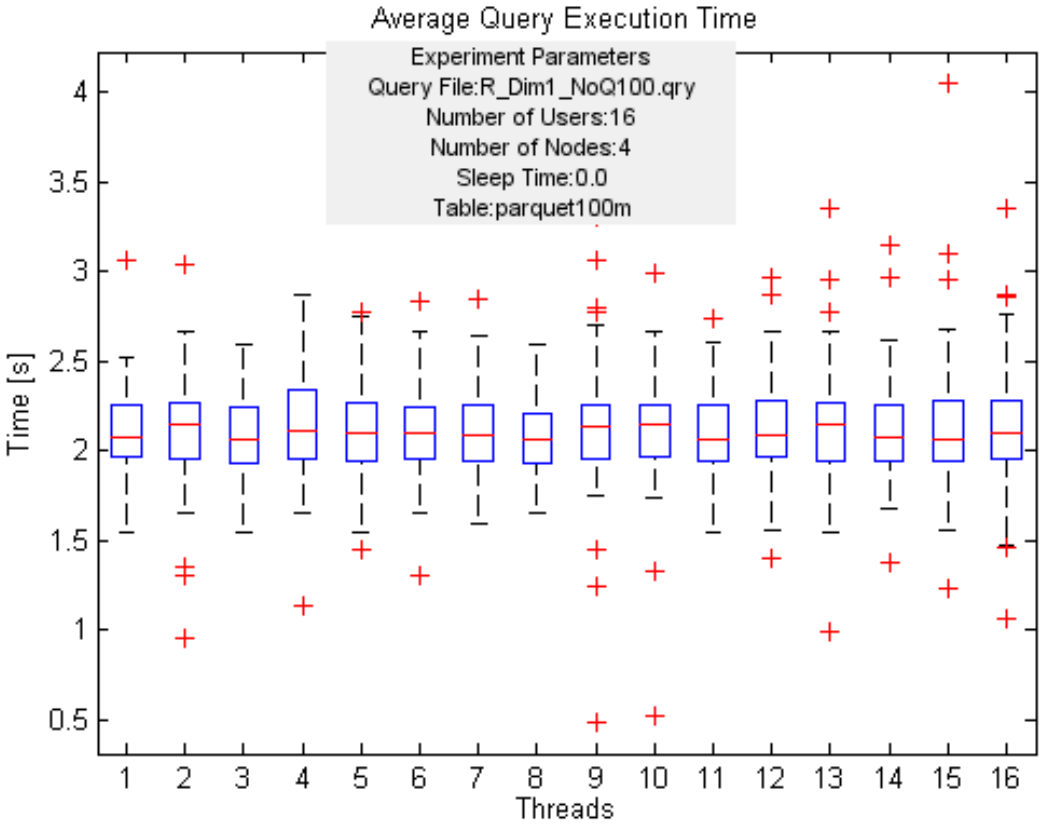
System Architektur



Query-Typen

Query Types	Description	Example: 1 Dimensional	Example: 2 Dimensional
R	*Integer and float range queries*	*SELECT count(*) FROM <tableName> WHERE a1 < 27*	*SELECT count(*) FROM <tableName> WHERE a2 > 4727 AND a3 = 19*
S	*String queries*	*SELECT count(*) FROM <tableName> WHERE s1 LIKE '%ahx%'*	*SELECT count(*) FROM <tableName> WHERE s2 LIKE '%index' AND s3 LIKE '%j8%'*
G	*Group by-queries*	*SELECT a1, count(*) FROM <tableName> GROUP BY a1*	*SELECT a2, a3 , count(*) FROM <tableName> GROUP BY a2, a3*
M	*Mixed queries including R, S and G queries*	*SELECT parse_url(s1, 'HOST') , count(*) FROM <tableName> WHERE a1 = 3 AND s2 LIKE '%86%' GROUP BY parse_url(s1, 'HOST')*	*SELECT s2 , parse_url(s1, 'HOST') , count(*) FROM <tableName> WHERE a1 < -63 AND s2 LIKE '%pcn%' GROUP BY s2, parse_url(s1, 'HOST')*

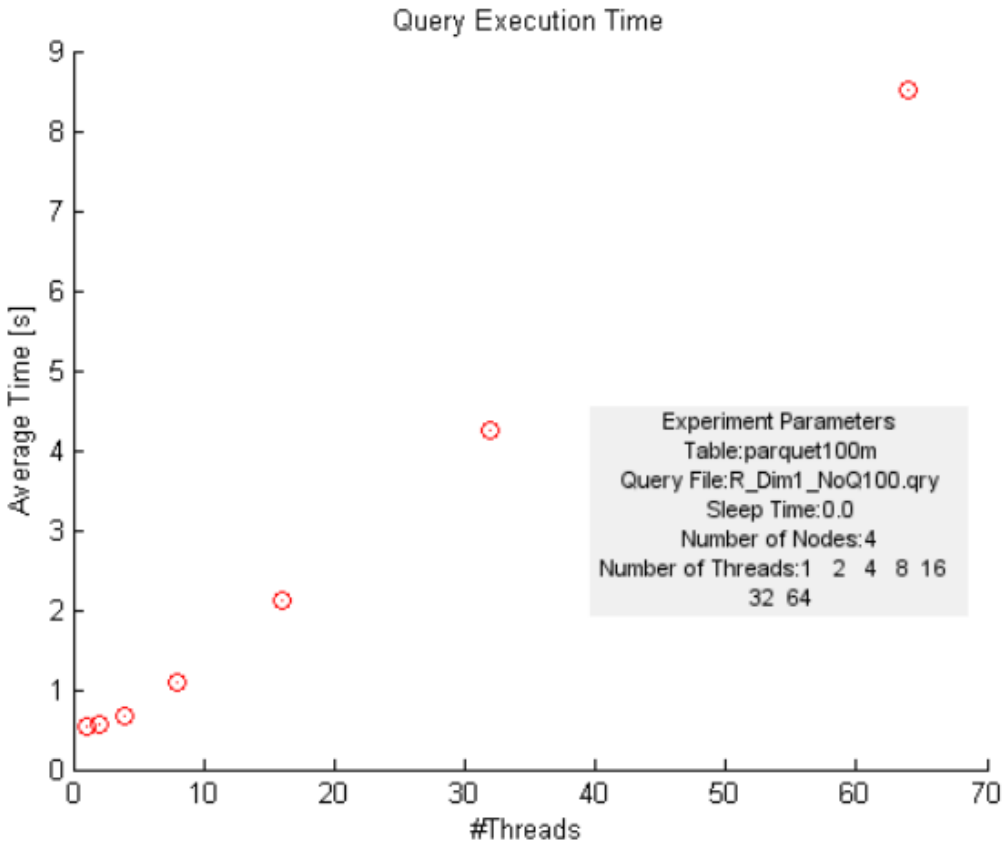
Performance Evaluierung #1



4-node-Impala cluster
~100 GB data, 16 concurrent users

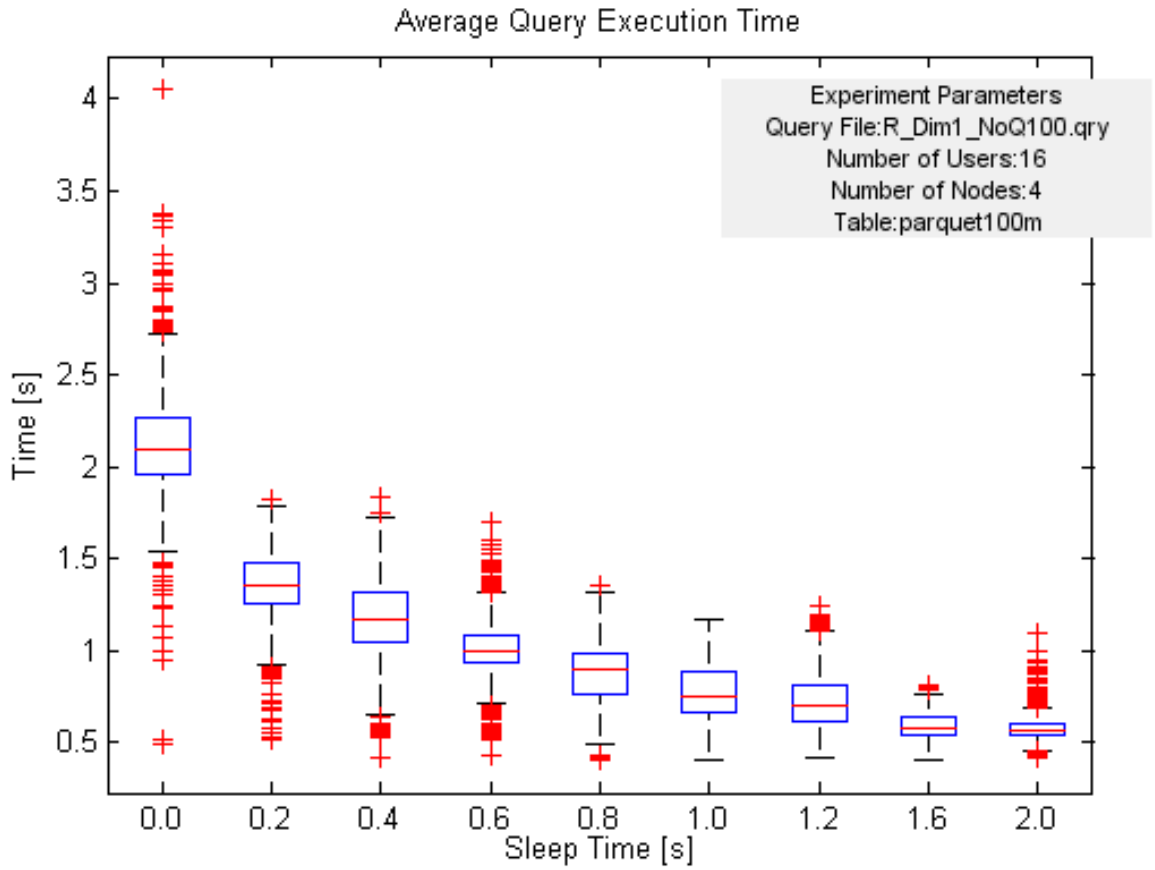
Performance Evaluierung #2

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie



4-node-Impala cluster
~100 GB data, 64 concurrent users

Performance Evaluierung #3



4-node-Impala cluster, ~100 GB data, 16 concurrent users

delay between queries

Zusammenfassung

- Verwendung von Big Data Technologien ist nicht trivial, da sich Tools sehr schnell entwickeln
- Big Data Technologie wurde ursprünglich zur Prozessierung von unstrukturierten und semi-strukturieren Daten entwickelt
- Trend in Richtung Prozessieren von strukturierten Daten:
  - Teilweise open-source Ersatz für kommerzielle Datenbanken und Data Warehouse-Lösungen
- Integration von Machine Learning und Graph-Prozessierung

- Kontakt für Forschungsprojekte bzw. Data Science Weiterbildung:
  - Kurt.Stockinger@zhaw.ch
  - http://www.weiterbildung.zhaw.ch/de/programm/das-data-science.html