Endocode AG

Big Data und Open Source

Thomas Fricke

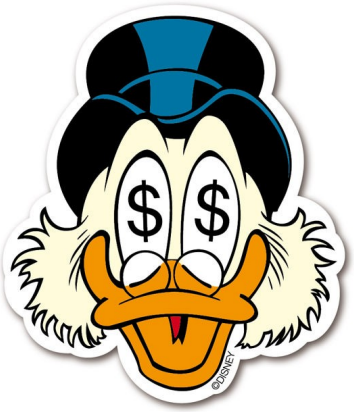
Partner Endocode AG
Regular changes from Dev to OPS and back
Big Data Architect

Überblick Big Data

Apache Hadoop Ökosystem

- Distributionen
  - Cloudera, HortonWorks, MapR
  - Teradata, IBM BigInsights
- Hosted
  - Amazon
  - Google
  - Azure

Markterwartung



1 Billion €

aufbauend auf Open Source

“Information is the Oil of the 21st century”

Was war noch mal Deepwater Horizon?

Apache Hadoop Ecosystem

Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

Sqoop
Data Exchange

Flume
Log Collector

Zookeeper
Coordination

Oozie
Workflow

Pig
Scripting

Mahout
Machine Learning

R Connectors
Statistics

Hive
SQL Query

Hbase
Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System

Programmierung

Philosophie:

- Code zu den Daten
- Map Reduce
- Verteiltes Filesystem: HDFS
  - 128 MB Blöcke
  - 3 oder mehr Kopien
  - Nodes
  - Namenodes
- Hochverfügbar bei Design

Konseqenz

- Hohe Netzbandbreite
- Redundanz aller Komponenten
- Faktor 15 Performanzeinbußen gegen Im Memory

Neue Berufsbilder

Alles Big ...

- Big Data Admin
- Big Data Developer
- Big Data Engineer
- Data Analyst
- Data Scientist
- Neural Network Scientist?

Statististik ist geil!

Programmierung

Alte Bekannte und new Kids on the Block

- Java
- Scala
- Python
- R
- F#
- …

Auswirkungen

M$

- kauft Revolution Analytics
  - Programmiersprache R
  - Standard für Statistische Datenanalyse
  - tritt R Consortium bei der Linux Foundation bei (zusammen mit Oracle)
- für Azure
  - Azure Cloud Switch
  - HDFS belegt jede Bandbreite

Steuerung

Daten werden nicht zum Selbstzweck gesammelt, sondern zu Steuerung!

- Kaufverhalten in Webshops, Apps
- Platzierung von Werbung
- Börsendaten
- Kraftwerkssteuerung
- Produktionsprozesse
- Verkehrslenkung
- Verbrechensvorhersage

Warnung

Algorithmen enthalten alle unsere Werte

und unsere Urteile,

… und Vorurteile