2.– 5. September 2013
in Nürnberg

Herbstcampus

Wissenstransfer
par excellence

Why you should care about Big Data

You are not Facebook or Google?

Kai Wähner

Talend

@KaiWaehner

www.kai-waehner.de

Kai Wähner

Consulting
Developing
Coaching
Speaking
Writing

Main Tasks

Requirements Engineering

Enterprise Architecture Management

Business Process Management

Architecture and Development of Applications

Service-oriented Architecture

Integration of Legacy Applications

Cloud Computing

Big Data

Contact

Email: kontakt@kai-waehner.de

Blog: www.kai-waehner.de/blog

Twitter: @KaiWaehner

Social Networks: Xing, LinkedIn

Key messages

You have to care about big data to be competitive in the future!

Start your big data projects business driven!

Big data from a technical perspective is no (longer) rocket science!

Agenda

- Big data paradigm shift
- Use cases for SMEs
- Challenges of big data

Why?

- Technology perspective
- Getting started
- Live demo

How?

Agenda

- Big data paradigm shift
- Use cases for SMEs
- Challenges of big data
- Technology perspective
- Getting started
- Live demo

Why should you care about big data?

"If you can't measure it, you can't manage it."

William Edwards Deming
(1900 –1993)
American statistician, professor, author, lecturer and consultant

Why should you care about big data?

- ➔ „Silence the HiPPOs“ (highest-paid person‘s opinion)
- ➔ Being able to interpret unimaginable large data stream, the gut feeling is no longer justified!

Harvard Business manager
DAS WISSEN DER BESTEN

ERIK BRYNJOLFSSON AND ANDREW MCAFEE
Erik Brynjolfsson is the Schussel Family Professor at MIT's Sloan School of Management and director of its Center for Digital Business. Andrew McAfee is principal research scientist at the Center. They are the coauthors of Race Against the Machine (Digital Frontier Press, 2012).

Why should you care about big data?

Friday Data Stories: Big Data driven decision making

JUNE 29, 2012 BY INFORBIX 2 COMMENTS

Like Tweet 4 +1 2 Share 3 Follow 61

Everyone wants to make good decisions. And as everyone knows, to make good decision you need data. Not just any data, but the right data. In the good old days, when data was easy and uncomplicated to access, the hard part in making a good decision was making the decision. Today, that's changed. The hard part in making a good decision is accessing data. That's because data has become complex and difficult to access: it's everywhere and there's loads of both structured and unstructured data everywhere in a company. And more of it is generated every day.

"Accessing data is now [again] the critical path in making good decisions!"

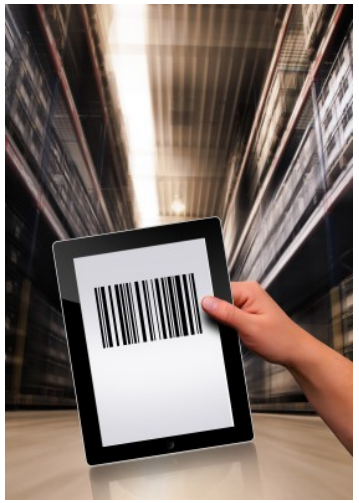
http://www.inforbix.com/friday-data-stories-big-data-driven-decision-making/

Where does big data come from?

Changing Expectations

Changing Scale

Sensors

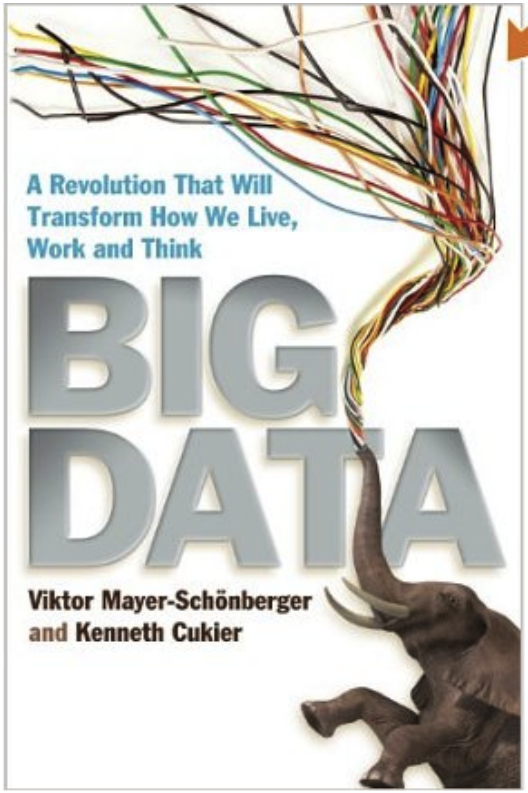


Changing Interactions

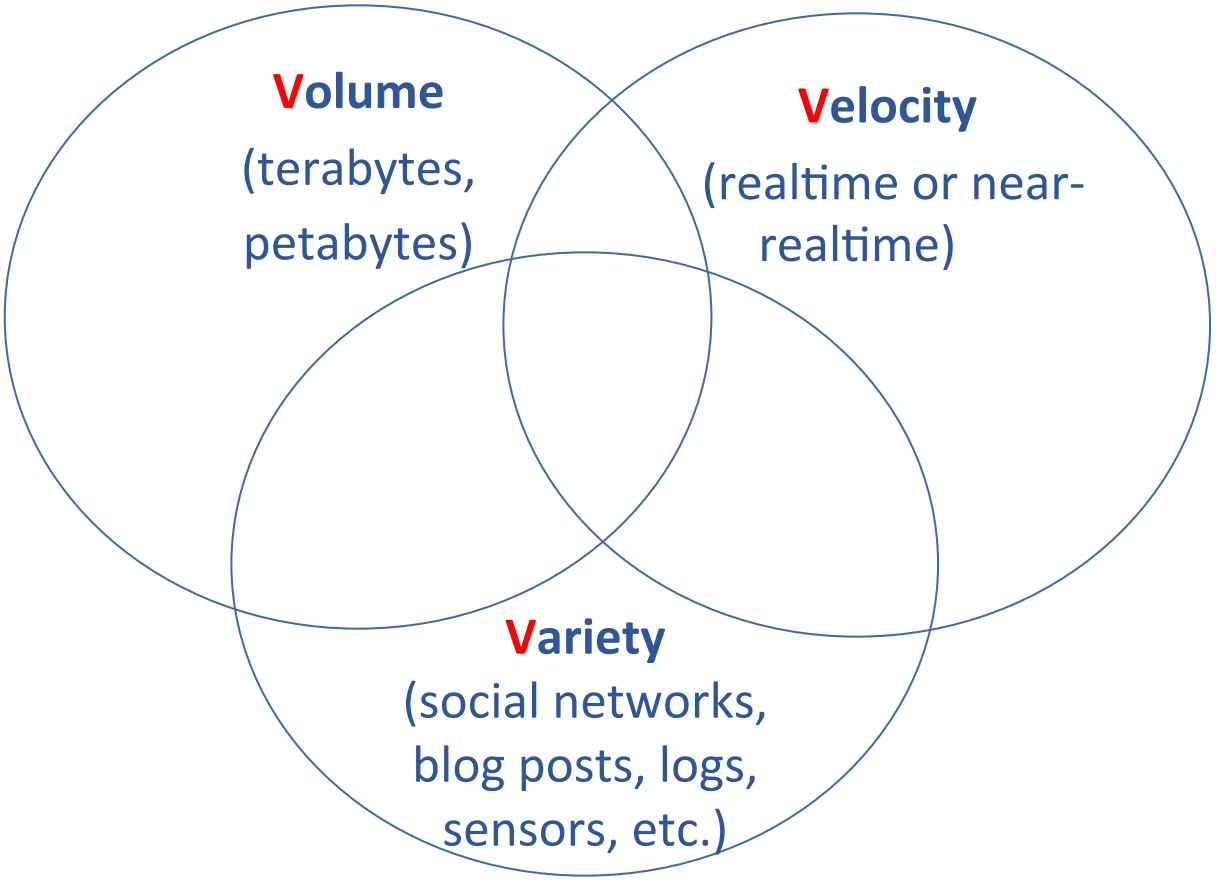
Cloud

Three shifts in the way we analyze information

- **Messiness:** Using ALL data, not just samples
  - Also bad data (e.g. Word spell checker, Google auto-complete and „did you mean...“ recommendation
- **Correlations:** Instead of causalities
  - May not tell us WHY something is happening, but THAT it is happening
  - In many situations, this is good enough
  - What drug substance cures cancer? When should I buy an airplane ticket?
- **Datafication**: Store, process, combine, reuse, enhance all data!
  - Digitalisation (Amazon Kindle → Read) vs. Datafication (Google Books → Read, Search, Process, ...)
  - Words becomes data: Google books: not just read, but also search, analyse, etc.
  - Locations becomes data: GPS: not just navigation, but also insurance costs, economic routes, etc.



What is big data? The Vs of big data



Value

"You are not Facebook or Google? Why you should still care about Big Data" by Kai Wähner

Big data tasks to solve - before analysis

Big Data Integration

- Land data in a Big Data cluster
- Implement or generate parallel processes

Big Data Manipulation

- Simplify manipulation, such as sort and filter
- Computational expensive functions

Big Data Quality & Governance

- Identify linkages and duplicates, validate big data
- Match component, execute basic quality features

Big Data Project Management

- Place frameworks around big data projects
- Common Repository, scheduling, monitoring

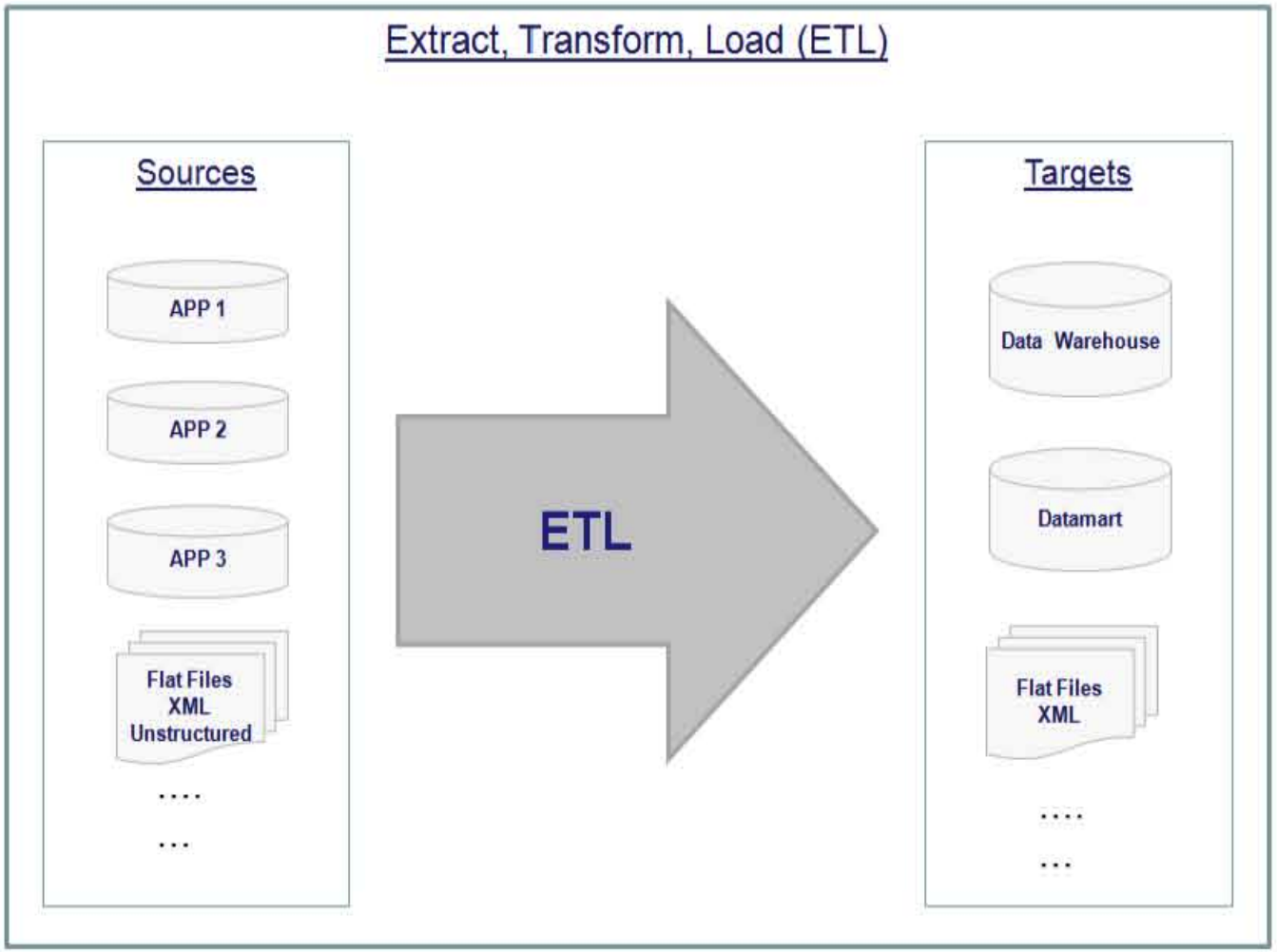
Agenda

- Big data paradigm shift
- Use cases for SMEs
- Challenges of big data
- Technology perspective
- Getting started
- Live demo

Use cases for SMEs

- ➜ Replacing ETL Jobs
- ➜ Forecast
- ➜ Logistics
- ➜ Fraud Protection
- ➜ Storage

Replacing ETL jobs



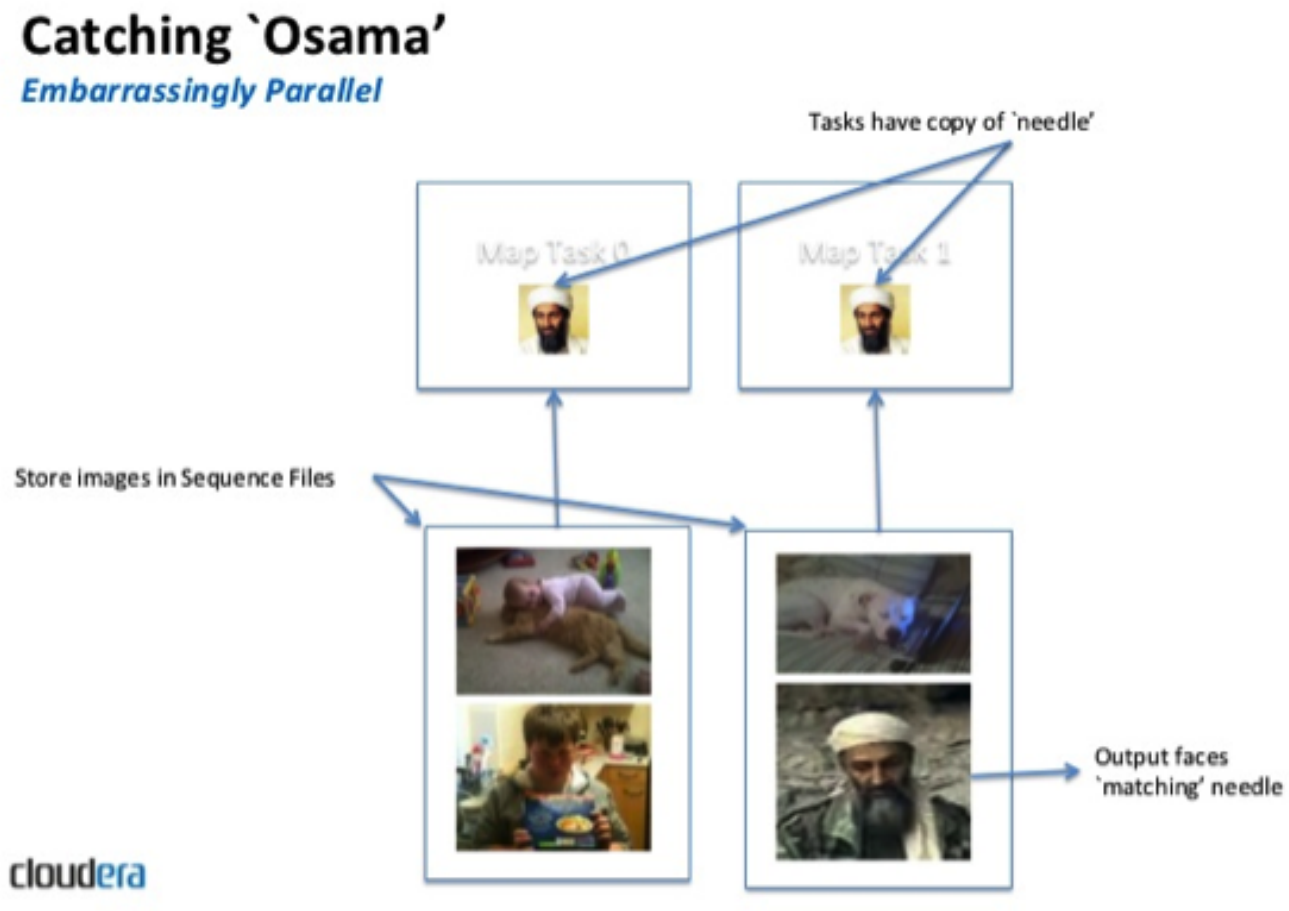
Replacing ETL jobs: Text files

"The advantage of their new system is that they can now look at their data [from their log processing system] in anyway they want:

- Nightly MapReduce jobs collect statistics about their mail system such as spam counts by domain, bytes transferred and number of logins.
- When they wanted to find out which part of the world their customers logged in from, a quick [ad hoc] MapReduce job was created and they had the answer within a few hours. Not really possible in your typical ETL system."

http://highscalability.com/how-rackspace-now-uses-mapreduce-and-hadoop-query-terabytes-data

Replacing ETL jobs: Binary files



http://www.slideshare.net/brocknoland/common-and-unique-use-cases-for-apache-hadoop

Forecast

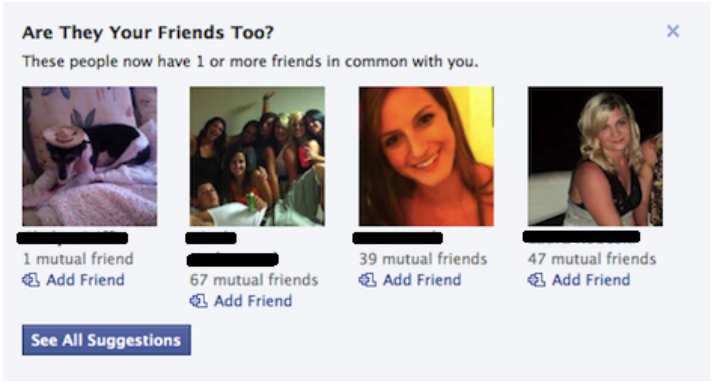
"**Forecasting** is the process of making statements about events whose actual outcomes (typically) have not yet been observed."

[Wikipedia]

Forecast: Recommendation engine

amazon

facebook



→ Is there no use case for your business (shop, community, items, etc.) ???

talend integration at any scale

Forecast: Risk management

T-Mobile

Deduce Customer Defections

T-Mobile USA has integrated Big Data across multiple IT systems to combine customer transaction and interactions data in order to better predict customer defections. By leveraging social media data (Big Data) along with transaction data from CRM and Billing systems, T-Mobile USA has been able to "cut customer defections in half in a single quarter".

http://hkotadia.com/archives/5021

Logistics

"Logistics is the management of the flow of resources between the point of origin and the point of destination in order to meet some requirements."

Wikipedia

Logistics: Time planning

- **Use case:** Guessing Aiport Flight arrival times
- **Problem:** Every minute of a delay costs a lot of money (crew, parking, delays of other machines, etc.)
- **Solution:** Computation of flight arrival with public and private data (weather forecast, flight plans, own radar stations, statistics of past flights

Andrew Mcafee, Erik BRYNJOLFSSON, Harvard Business Manager November 2012

Logistics: Flexible pricing

- With revenue of almost USD 30 billion and a network of 800 locations, Macy's is considered the largest store operator in the USA
- Daily price check analysis of its 10,000 articles in less than two hours
- Whenever a neighboring competitor anywhere between New York and Los Angeles goes for aggressive price reductions, Macy's follows its example
- If there is no market competitor, the prices remain unchanged

http://www.t-systems.com/about-t-systems/examples-of-successes-companies-analyze-big-data-in-record-time-l-t-systems/1029702

Fraud detection

pwc

"Fraud is a million dollar business and it is increasing every year. The PwC global economic crime survey suggests that close to 30% of companies worldwide have reported being victims of fraud in the past year."

http://www.pwc.com/gx/en/economic-crime-survey

Fraud detection: Anti-counterfeiting

BRIGHTPLANET CASE STUDY: PROTECTION AND ANTI-COUNTERFEITING PROBLEM

PROBLEM:

A Fortune 100 company in a high-margin industry was hemorrhaging potential profits to overseas counterfeiters. These counterfeiters advertised brand name products at a fraction of the retail price on trade boards, fly-by-night websites, e-commerce site, message boards, and social media.

The company's traditional strategy included hiring an external brand protection firm; however, this solution wasn't scalable to the wide scope of the Internet, where legitimate profits were unknowingly being siphoned off by fraudulent websites.

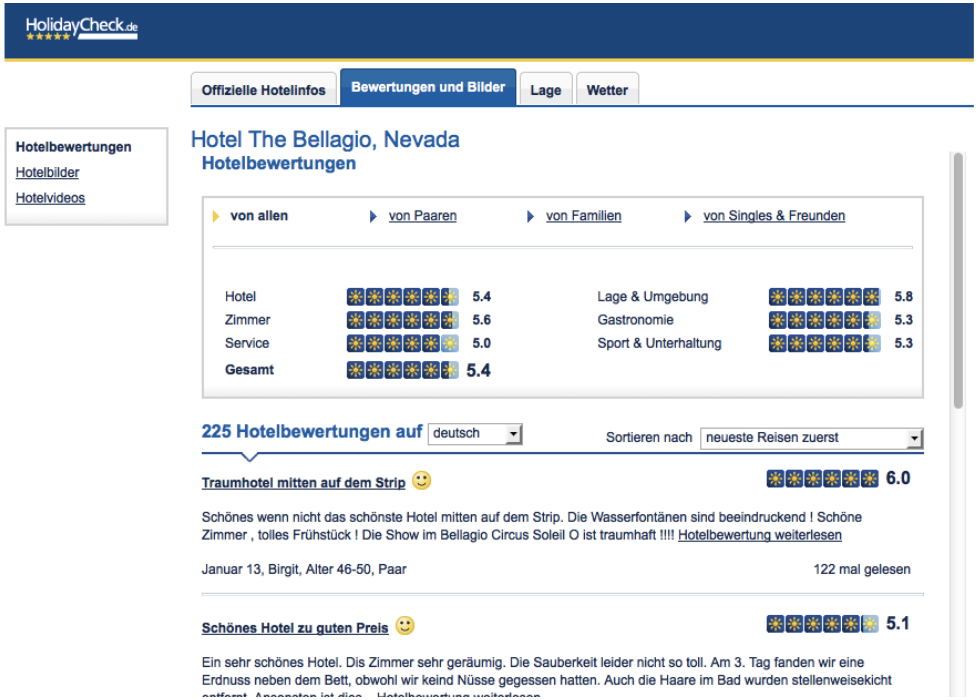
Solution: A scalable process to automatically monitor the internet for any mention of the company's brand name products. Websites, message boards, trade boards, and social media are automatically monitored.

http://www.brightplanet.com/2013/01/pharmaceutical-fraud-and-counterfeiting-finding-a-deep-web-solution/

Fraud detection: Fraud mining

HolidayCheck.de

Content manipulation at a traveling portal – Which customer reviews are trustworthy?

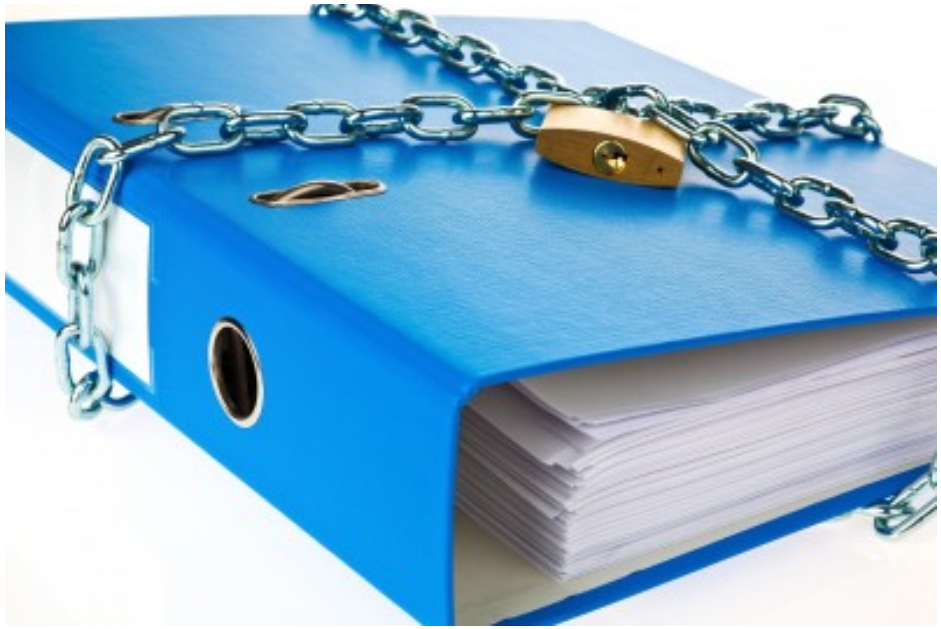


Jean-Paul Schmetz, Harvard Business Manager, November 2012

Storage

"being stored somewhere until needed"

http://www.macmillandictionary.com



Storage: Compliance

Global Parcel Service

- A lot of data must be stored „forever“
- Numbers increase exponentially
- Goal: As cheap as possible
- Problem: (Fast) queries must still be possible
- Solution: Commodity servers and „Hadoop querying“

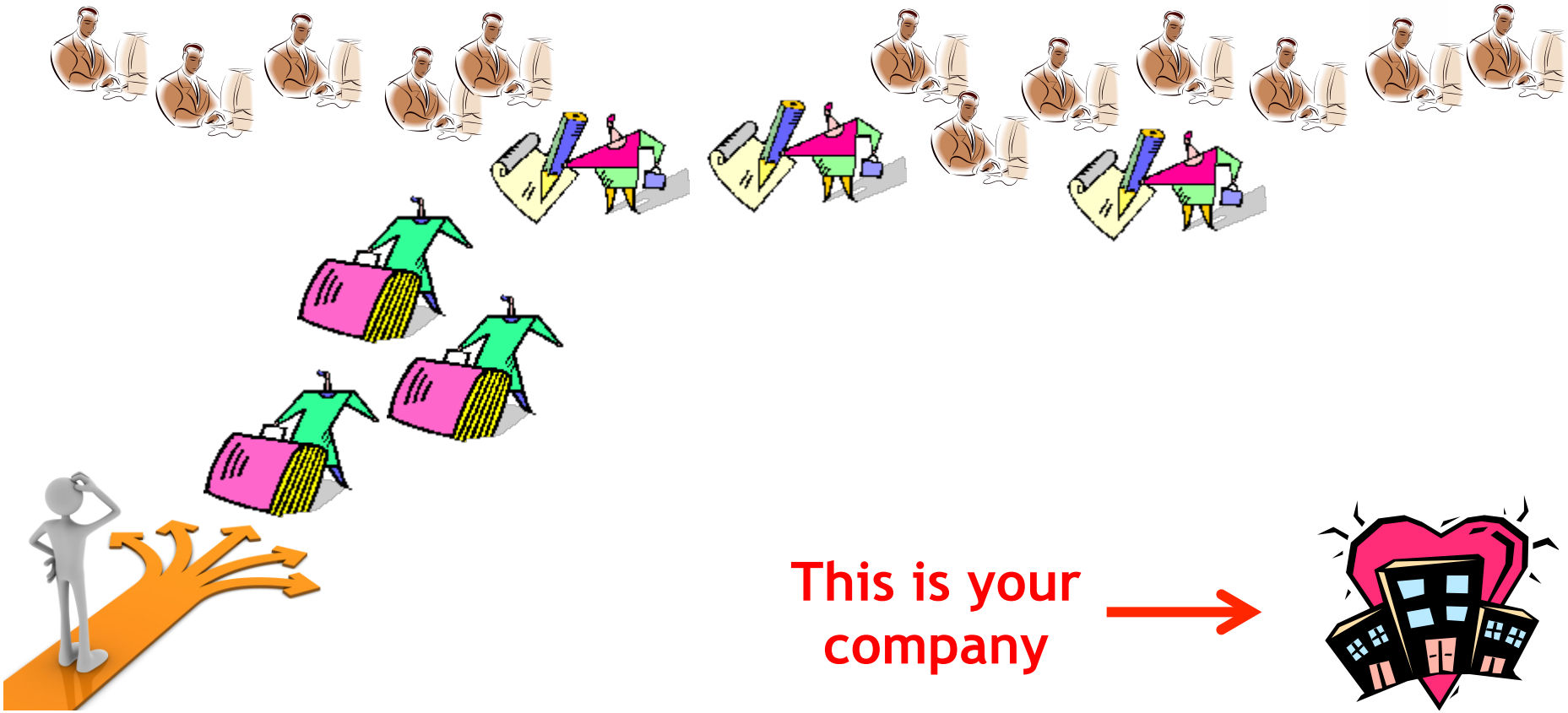
http://archive.org/stream/BigDataImPraxiseinsatz-SzenarienBeispieleEffekte/Big_Data_BITKOM-Leitfaden_Sept.2012#page/n0/mode/2up

Agenda

- Big data paradigm shift
- Use cases for SMEs
- Challenges of big data
- Technology perspective
- Getting started
- Live demo

Limited big data experts

facebook twitter Google cloudera Hortonworks MAPR TECHNOLOGIES



This is your company

Big Data Geek

Big data tool selection (technical perspective)

Looking for ‚your' required big data product?

Support your data from scratch?

Good luck! ☺

Big data tool selection (business perspective)

- Wanna buy a big data solution for your industry?
- Maybe a competitor has a big data solution which adds business value?
- The competitor will never publish it (rat-race)!

Data quality

Big Data + Poor Data Quality = Big Problems

How to solve these big data challenges?

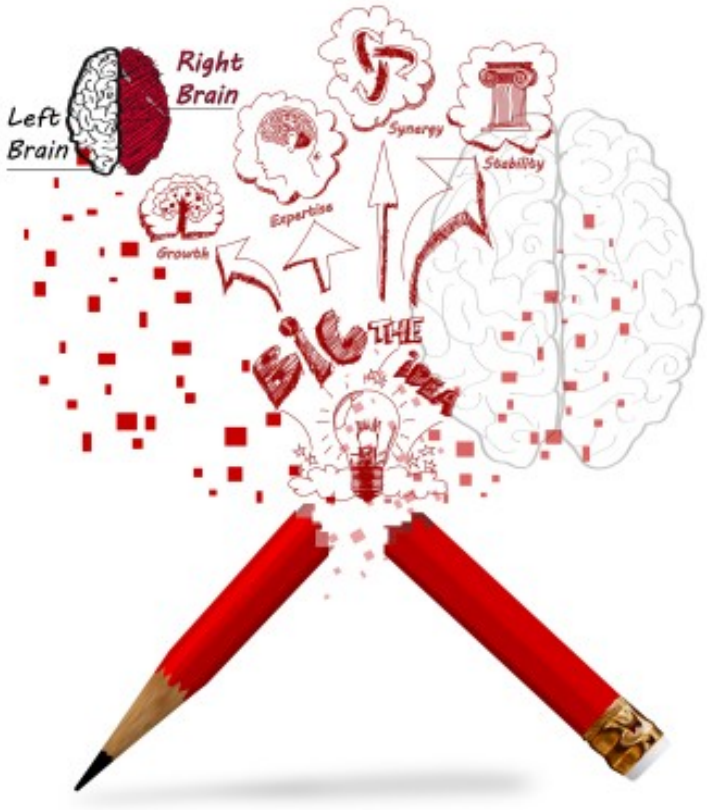
Be no expert! Be simple!

→ "[Often] simple models and big data trump more-elaborate [and complex] analytics approaches"

→ "Often someone coming from outside an industry can spot a better way to use big data than an insider"

Erik Brynjolfsson / Lynn Wu
http://alfredopassos.tumblr.com/post/32461599327/big-data-the-management-revolution-by-andrew-mcafee

Be creative!



→ Look at use cases of others
(SMU, but also large companies)

→ How can you do something similar with your data?

→ You have different data sources?
Use it! Combine it! Play with it!

What is your Big Data process?

1) Do not begin with the data, think about business opportunities
2) Choose the right data (combine different data sources)
3) Use easy tooling

http://hbr.org/2012/10/making-advanced-analytics-work-for-you

Agenda

- Big data paradigm shift
- Use cases for SMEs
- Challenges of Big data
- Technology perspective
- Getting started
- Live demo

Technology perspective

How to process big data?

How to process big data?

Parallel ETL Tools Are Dead

AUGUST 29, 2012

Post by Dave Nahmias

Global Director of Data Integration Sales Engineering

They just don't know it yet.

The critical flaw in parallel ETL tools is the fact that the data is almost never local to the processing nodes. This means that every time a large job is run, the data has to first be read from the source, split N ways and then delivered to the individual nodes. Worse, if the partition key of the source doesn't match the partition key of the target, data has to be constantly exchanged among the nodes. In essence, parallel ETL treats the network as if it were a physical I/O subsystem. The network, which is always the slowest part of the process, becomes the weakest link in the performance chain.

http://blog.syncsort.com/2012/08/parallel-etl-tools-are-dead

How to process big data?

python™

ERLANG

RabbitMQ
Open Source Enterprise Messaging

riak

disco
massive data - minimal code

Slides: http://www.slideshare.net/pavlobaron/100-big-data-0-hadoop-0-java

Video: http://www.infoq.com/presentations/Big-Data-Hadoop-Java

How to process big data?

The defacto standard for big data processing

IBM

EMC²

amazon

SAP

talend
integration at any scale

ORACLE

INFORMATICA

sas

How to process big data?

Microsoft makes its move with Hadoop on Azure and Windows Server

Microsoft is working with Hadoop core committers from Hortonworks to bring the …

by **Sean Gallagher** - Oct 12 2011, 11:41am PDT

14

Microsoft

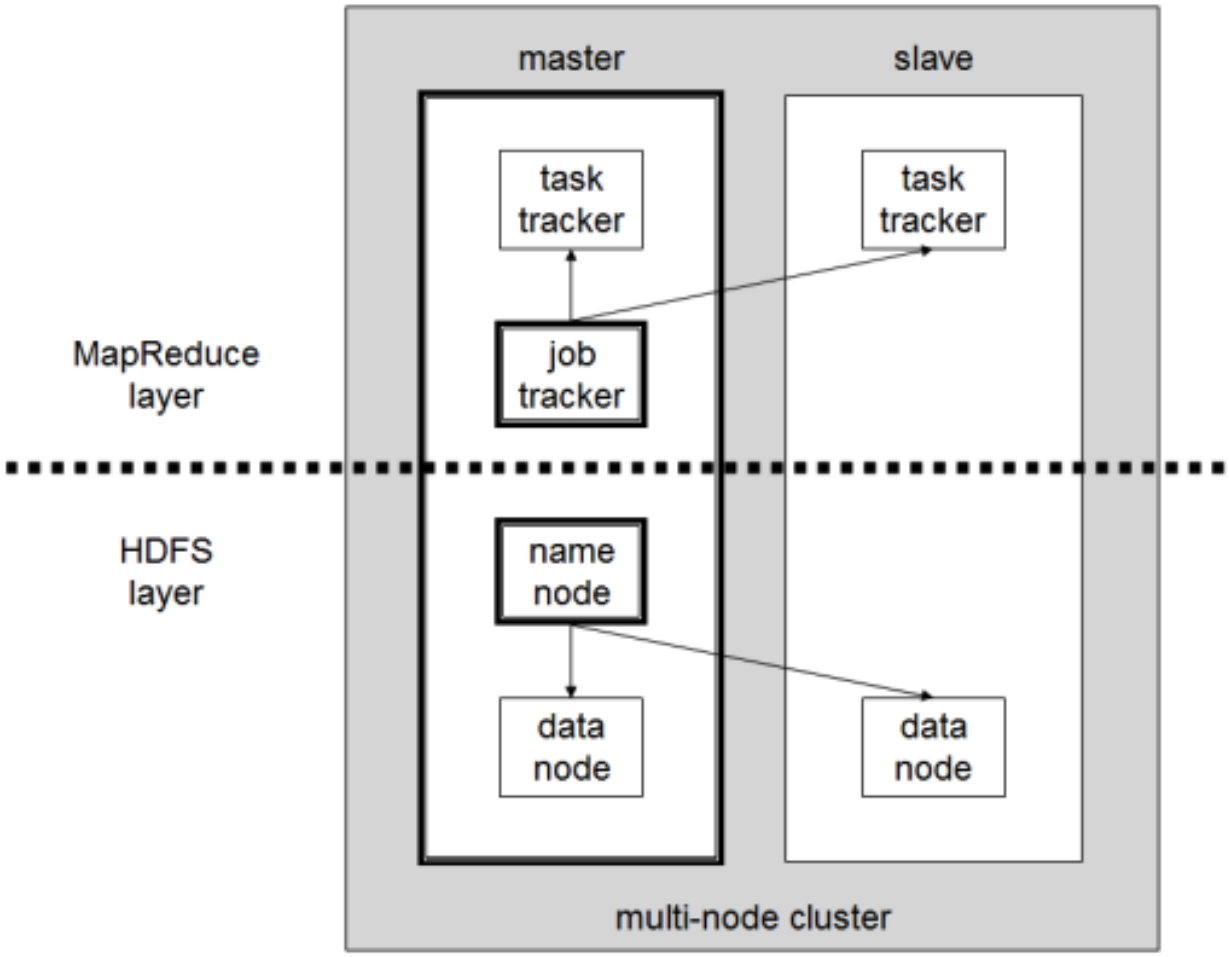
"A big part of [the company's strategy] includes wiring SQL Server 2012 (formerly known by the codename "Denali") to the Hadoop distributed computing platform, and bringing Hadoop to Windows Server and Azure"

Even Microsoft (the .NET house) relies on Hadoop since 2011

What is Hadoop?

Apache Hadoop, an open-source software library, is a framework that allows for the distributed processing of large data sets across clusters of commodity hardware using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

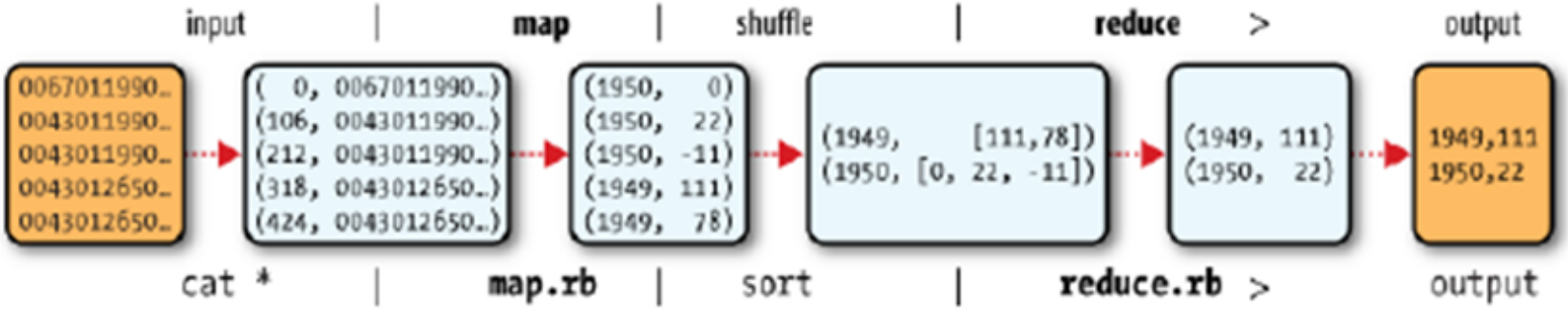
Hadoop architecture



Map (Shuffle) Reduce

Simple example

- Input: (very large) text files with lists of strings, such as:
  „318, 00430126509999919490324120040500001N9+01111+99999999999...“
- We are interested just in some content: year and temperate (marked in red)
- The Map Reduce function has to compute the maximum temperature for every year



Example from the book “Hadoop: The Definitive Guide, 3rd Edition”

Map (Shuffle) Reduce

MapReduce is a distributed data processing model and execution environment that runs on large clusters of commodity machines. MapReduce is a programming model for data processing. The model is simple, yet not too simple to express useful programs in.

A MapReduce ***job*** is a unit of work that the client wants to be performed: it consists of the input data, the MapReduce program, and configuration information. Hadoop runs the job by dividing it into ***tasks***, of which there are two types: ***map tasks*** and ***reduce tasks***.

There are two types of nodes that control the job execution process: a ***jobtracker*** and a number of ***tasktrackers***. The jobtracker coordinates all the jobs run on the system by scheduling tasks to run on tasktrackers. Tasktrackers run tasks and send progress reports to the jobtracker, which keeps a record of the overall progress of each job. If a task fails, the jobtracker can reschedule it on a different tasktracker.

Hadoop divides the input to a MapReduce job into fixed-size pieces called *input splits*, or just ***splits***. Hadoop creates one map task for each split, which runs the user defined map function for each *record* in the split.

How to process big data?

Sqoop

Apache Drill

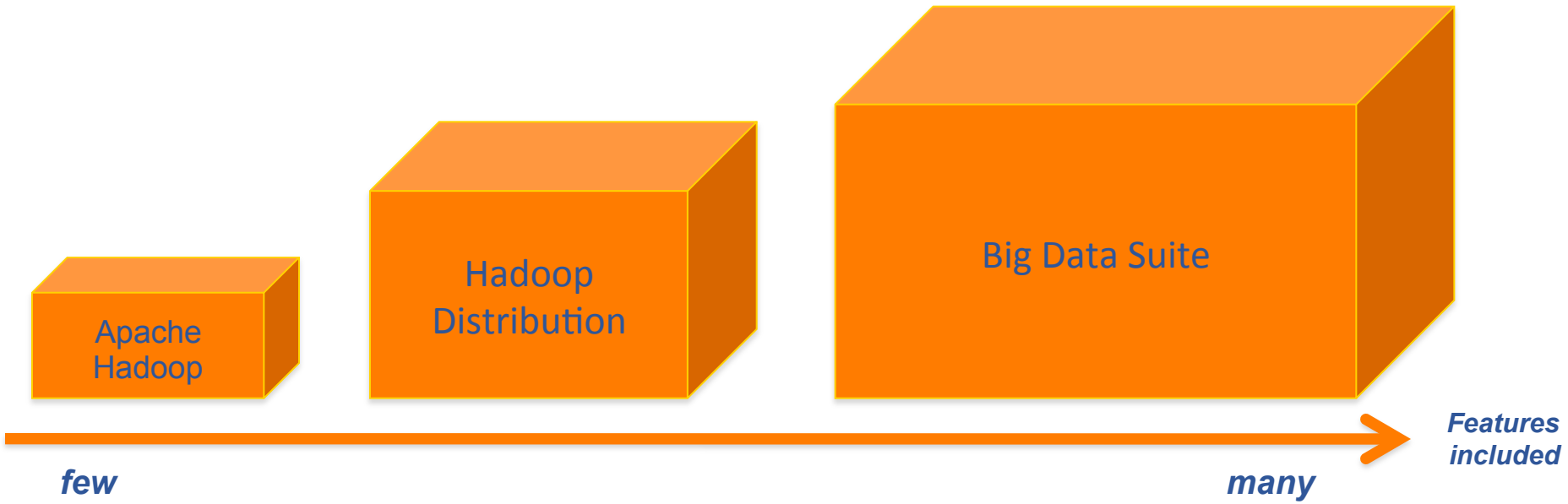
HCatalog
Table Management

cascading

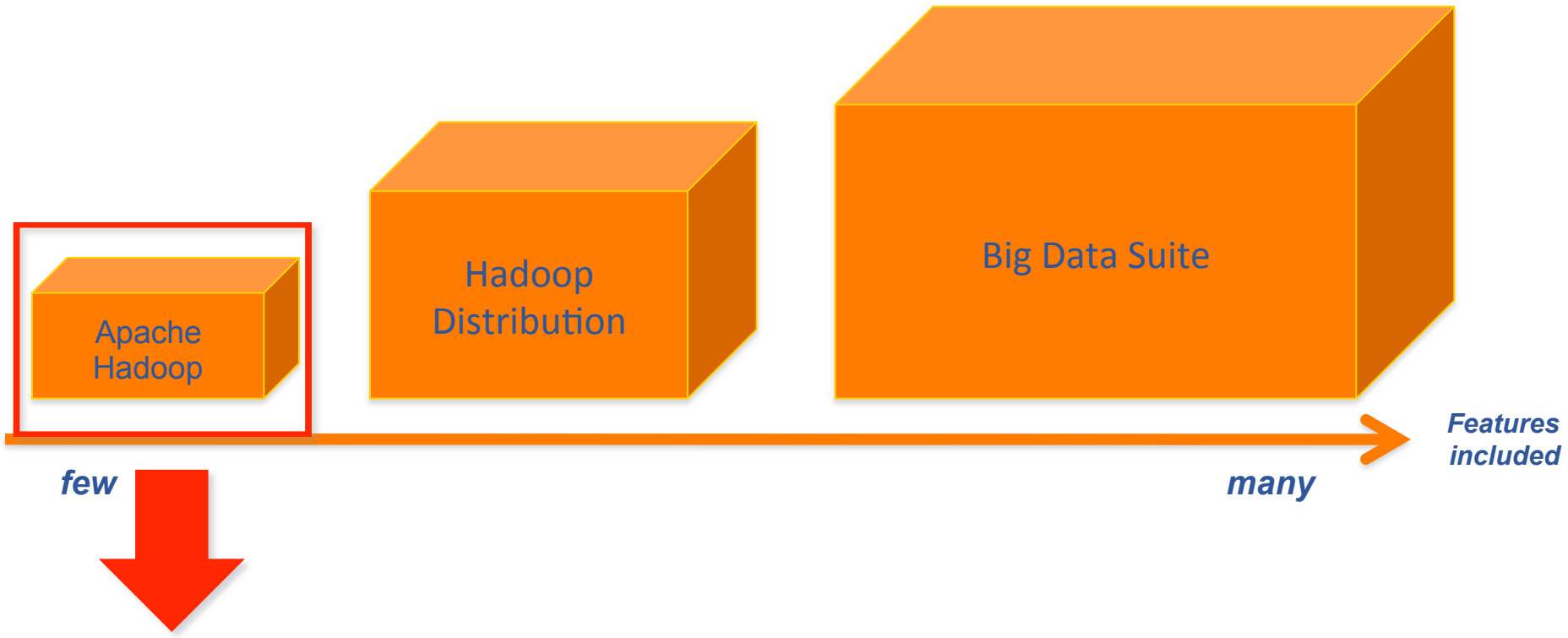
Agenda

- Big data paradigm shift
- Use cases for SMEs
- Challenges of Big data
- Technology perspective
- Getting started
- Live demo

Hadoop alternatives



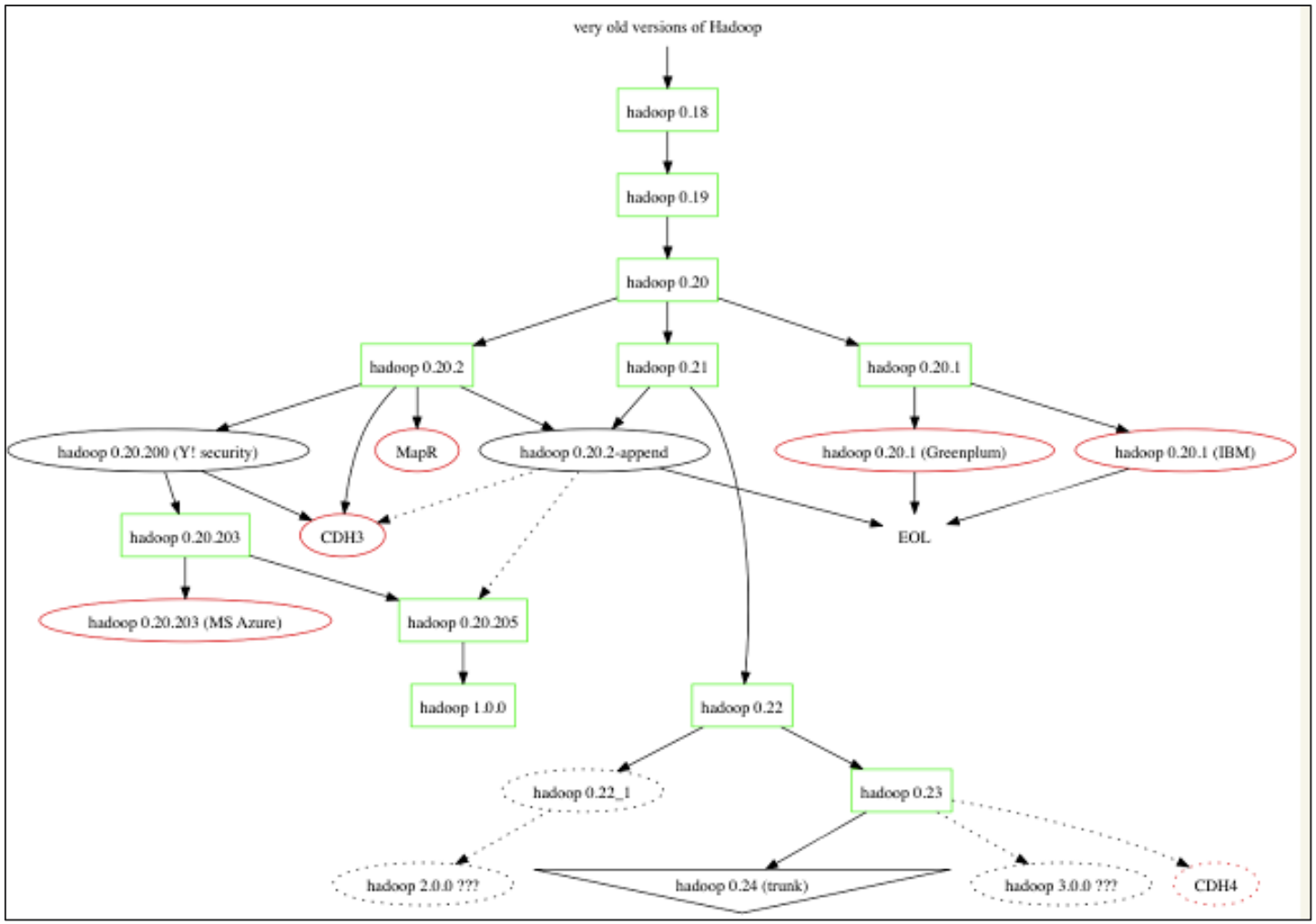
Hadoop alternatives



Apache Hadoop (MapReduce code)

```
public static class AllTranslationsReducer
extends Reducer<Text,Text,Text,Text>
{
   private Text result = new Text();
   public void reduce(Text key, Iterable<Text> values,
   Context context
   ) throws IOException, InterruptedException
   {
      String translations = "";
      for (Text val : values)
      {
         translations += "|"+val.toString();
      }
      result.set(translations);
      context.write(key, result);
   }
}
public static void main(String[] args) throws Exception
{
   Configuration conf = new Configuration();
   Job job = new Job(conf, "dictionary");
   job.setJarByClass(Dictionary.class);
   job.setMapperClass(WordMapper.class);
   job.setReducerClass(AllTranslationsReducer.class);
```

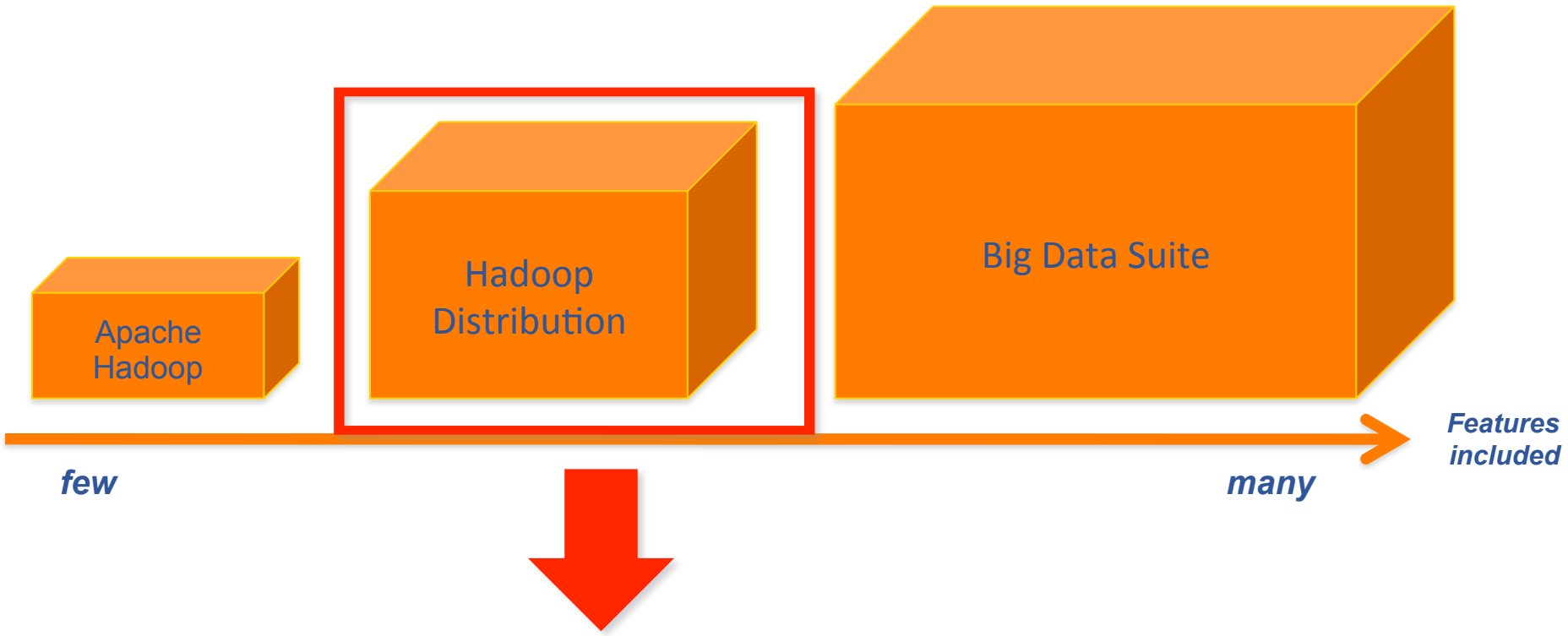
Genealogy of elephants



A little bit outdated (from February 2012), but you understand the problem ?!

https://blogs.apache.org/bigtop/entry/all_you_wanted_to_know

Hadoop alternatives



MapReduce
HDFS
Ecosystem

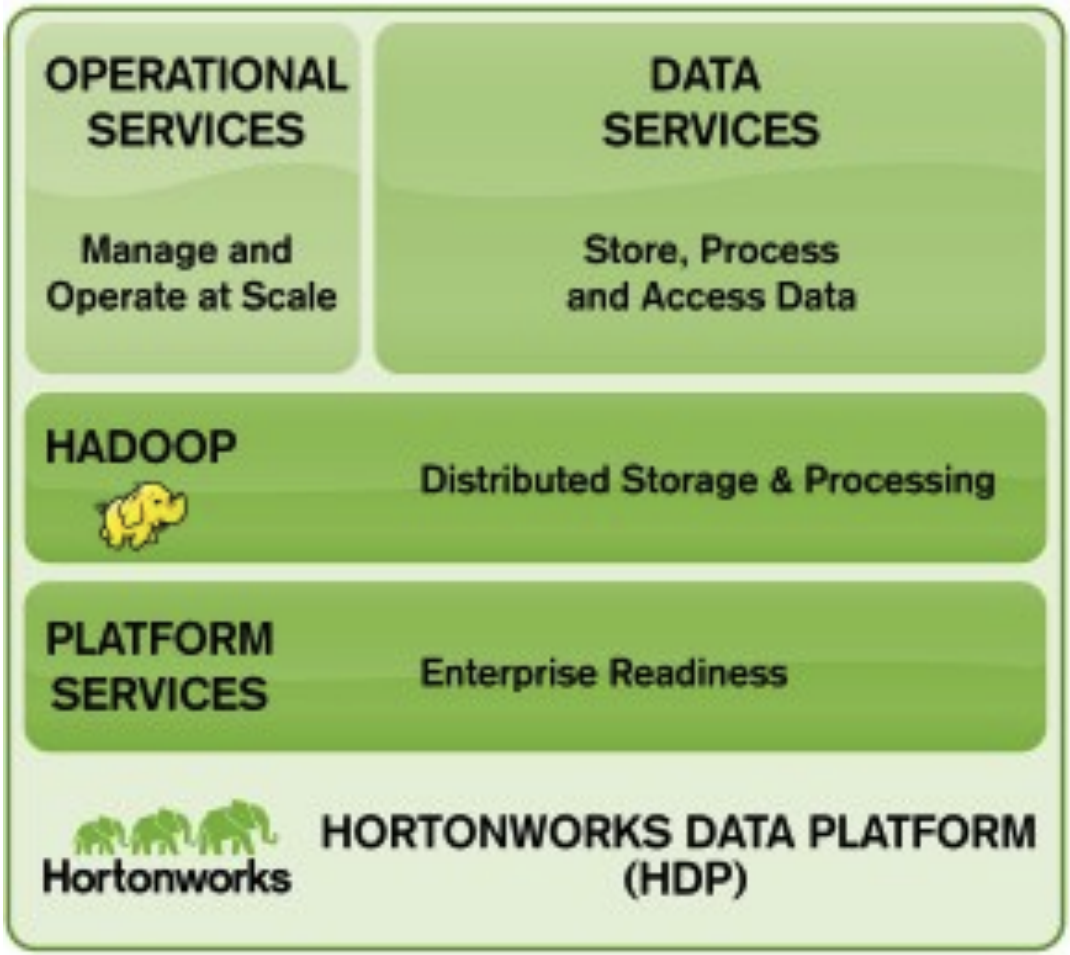
+

Packaging
Deployment-Tooling
Support

Hadoop distributions

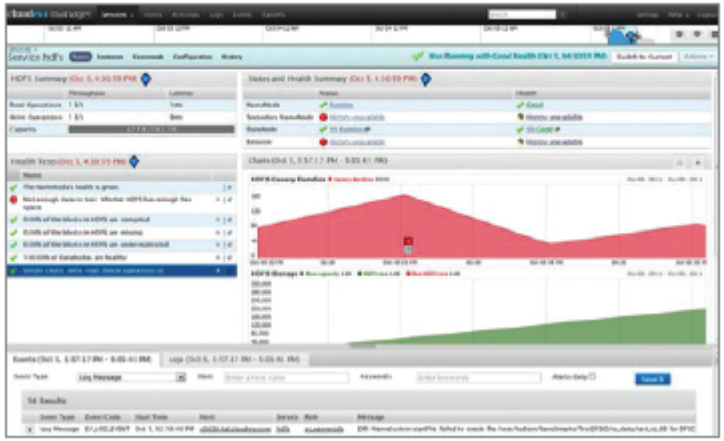
(… some more arising)

Hortonworks ecosystem



Cloudera Manager

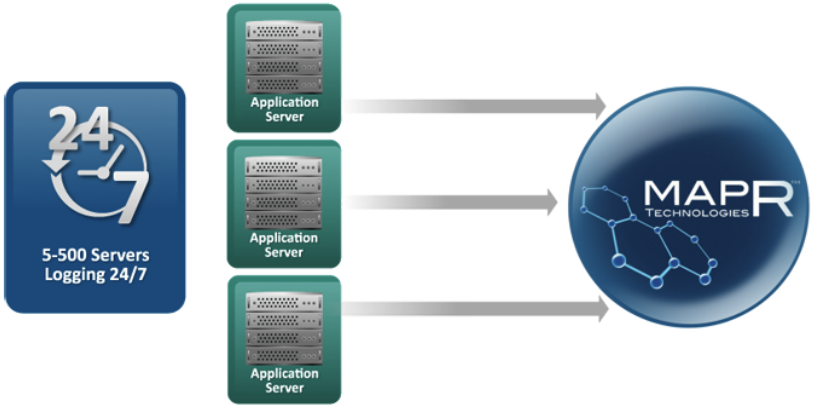
View Service Health and Performance



View Heatmaps for Status and Metrics

Monitor and Diagnose Cluster Workloads

MapR – the different alternative

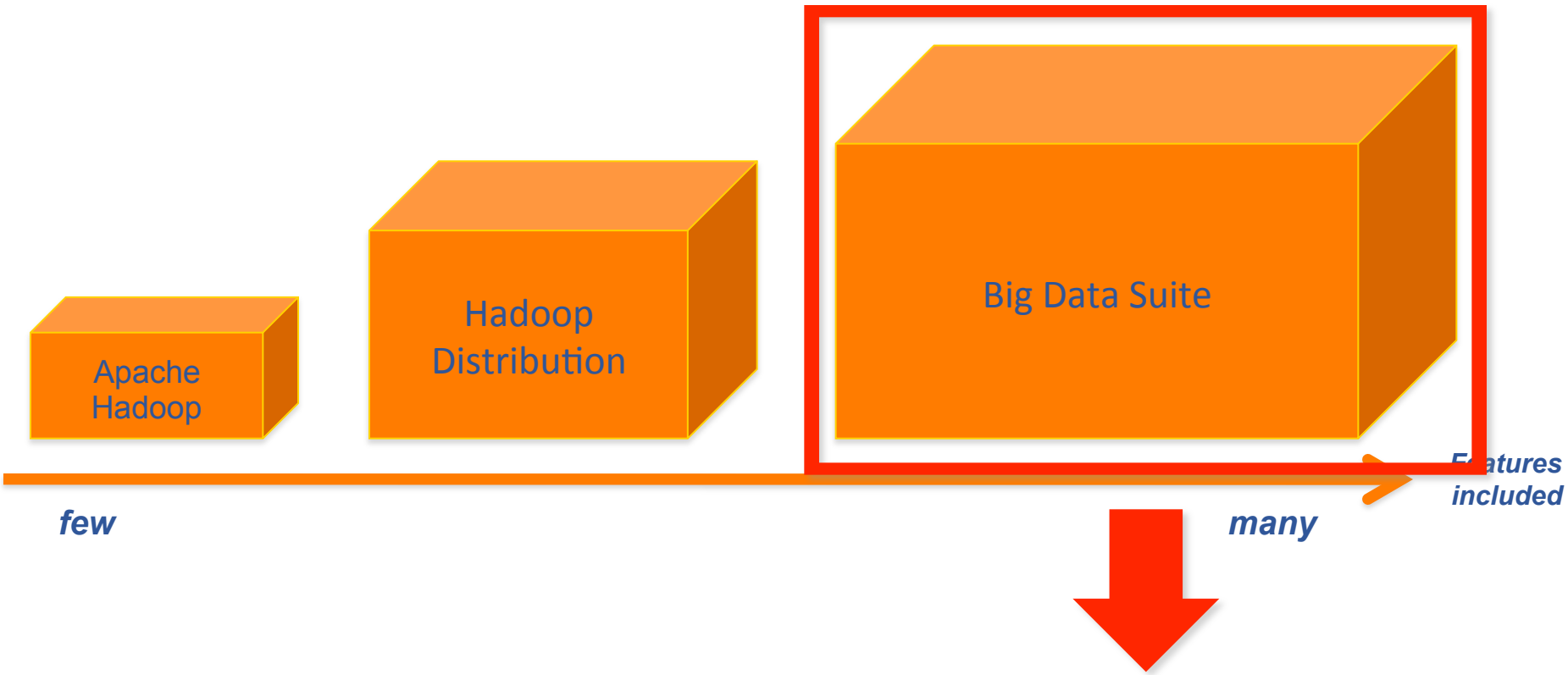


Replaced (Apache) HDFS with own implementation

MapR Direct Access NFS makes Hadoop radically easier and less expensive to use. Unlike the write-once system found in other Hadoop distributions, MapR allows files to be modified, overwritten, and enables multiple concurrent reads and writes on any file.

Users can simply browse files, automatically open associated applications with a mouse click, or drag-and-drop files and directories into and out of the cluster. Additionally, standard command-line tools and UNIX applications and utilities (such as grep, tar, sort, or tail) can be used directly on data in the cluster. With other Hadoop distributions, the user must copy the data out of the cluster in order to use standard tools.

Hadoop alternatives



MapReduce
HDFS
Ecosystem

+

Packaging
Deployment-Tooling
Support

+

Tooling / Modeling
Code Generation
Scheduling
Integration

Big data suites

IBM

talend*
*open data solutions

ORACLE®

INFORMATICA®
The Data Integration Company™

EMC²

pentaho™
open source business intelligence™

Agenda

- Big data paradigm shift
- Use cases for SMEs
- Challenges of Big data
- Technology perspective
- Getting started
- Live demo

Vision: Democratize big data

talend*
*open integration solutions

Pig

Sqoop

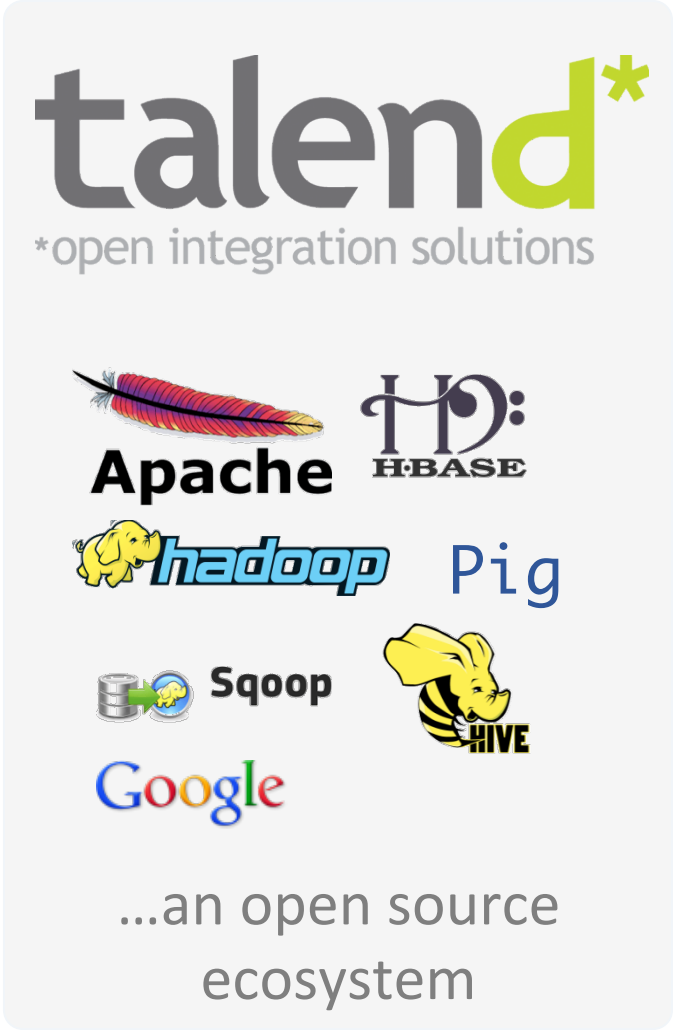

Google

…an open source ecosystem

Talend Open Studio for Big Data

- Improves efficiency of big data job design with graphic interface
- Generates Hadoop code and run transforms inside Hadoop
- Native support for HDFS, Pig, HBase, HCatalog, Sqoop and Hive
- 100% open source under an Apache License
- Standards based

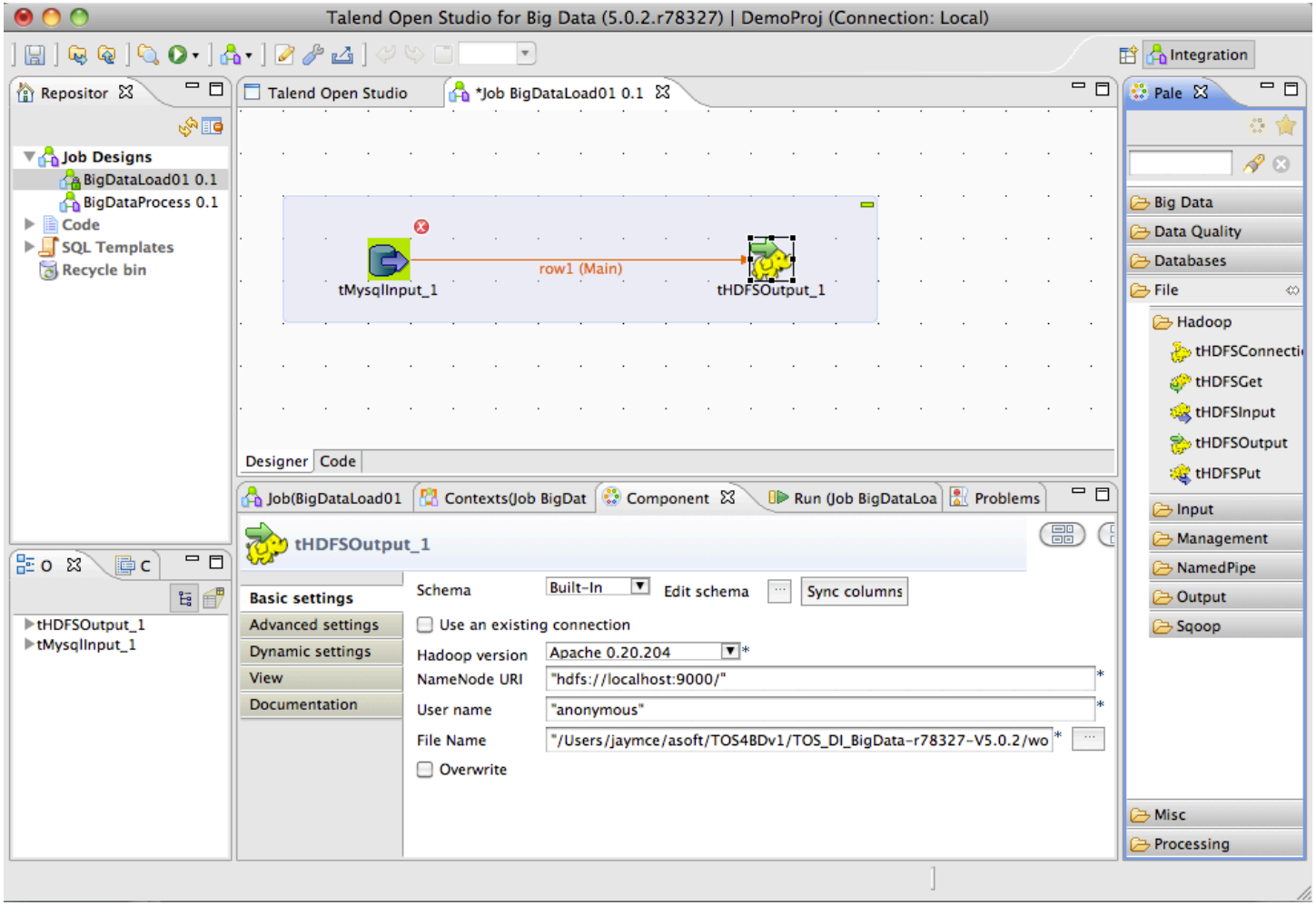
Vision: Democratize big data



Talend Platform for Big Data

- Builds on Talend Open Studio for Big Data
- Adds data quality, advanced scalability and management functions
  - MapReduce massively parallel data processing
  - Shared Repository and remote deployment
  - Data quality and profiling
  - Data cleansing
  - Reporting and dashboards
- Commercial support, warranty/IP indemnity under a subscription license

Example: Talend Open Studio for Big Data

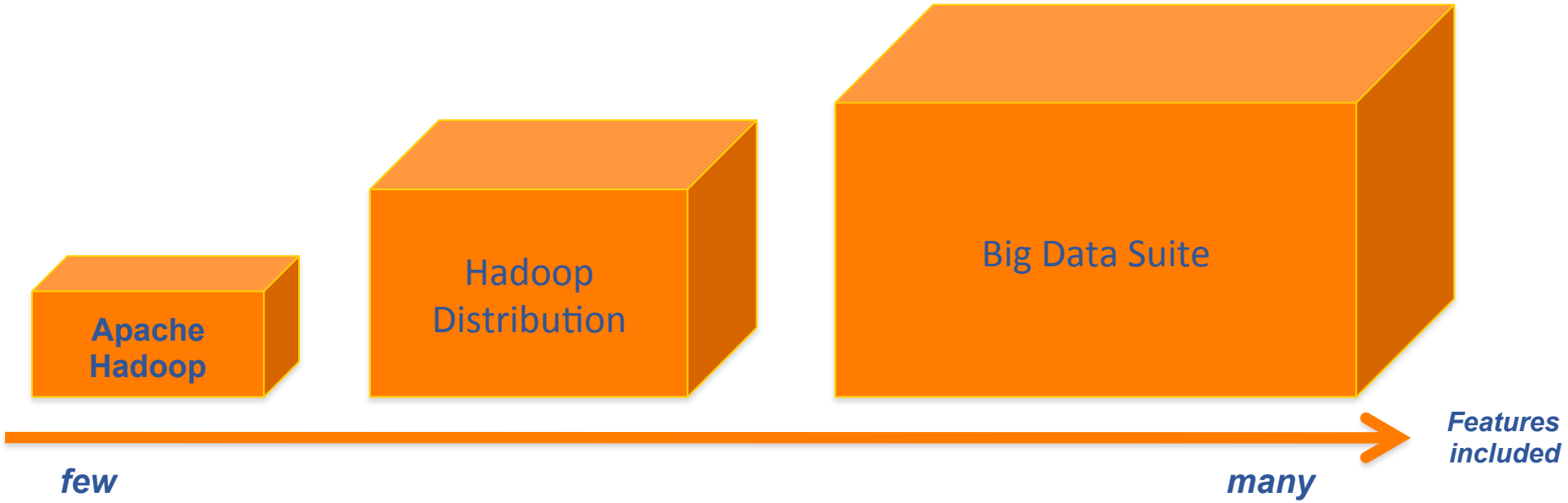


Live demo

talend*
*open data solutions

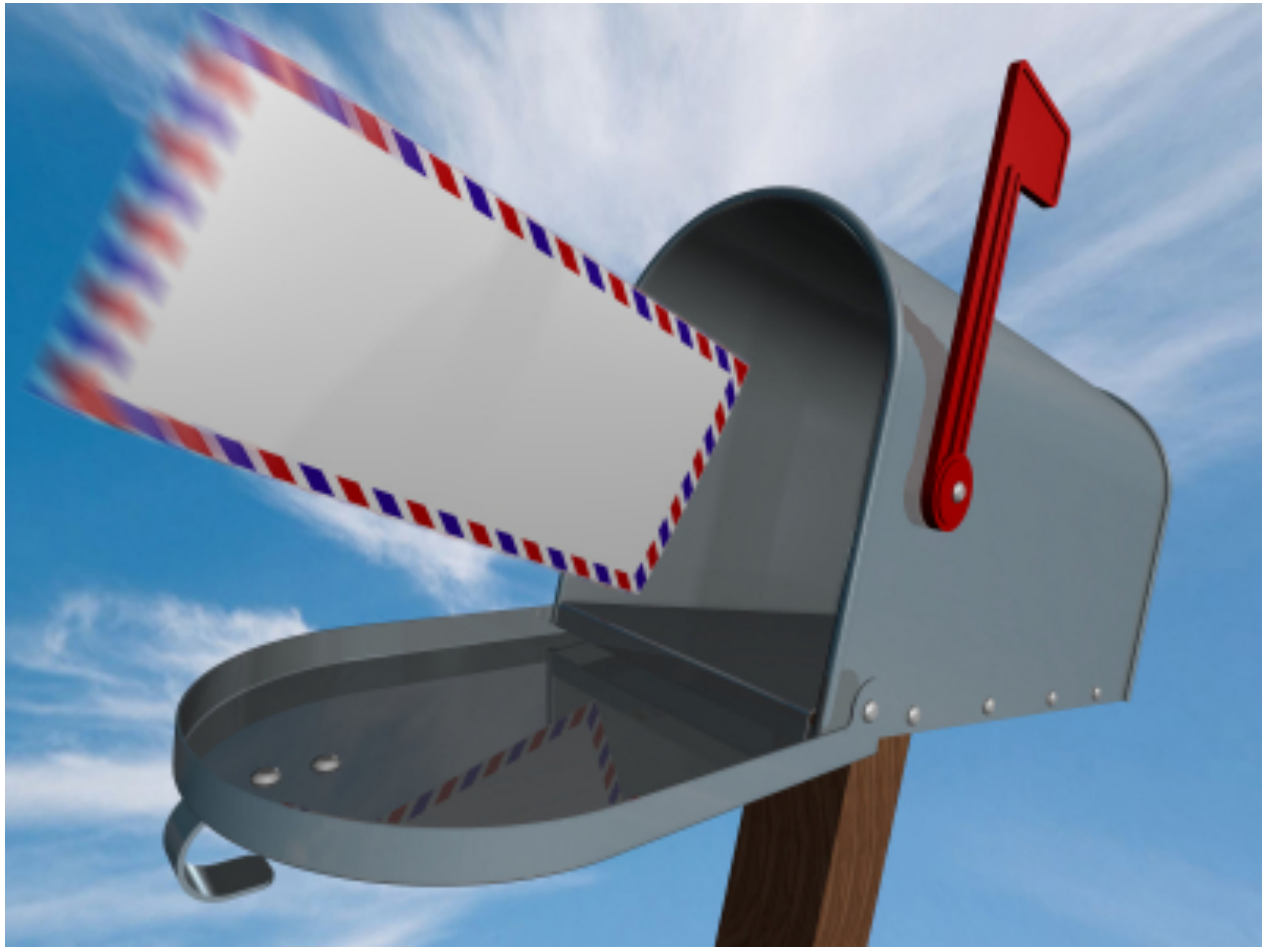
„Talend Open Studio for Big Data“ in action...

How to choose the right tooling?



- → **Simplicity** (lightweight setup, coding, deployment, usage)
- → **Prevalence** (open standards, community, extendibility)
- → **Features** (Hadoop ecosystem, connectors, monitoring)
- → **Pitfalls** (data-driven costs, no native Hadoop code, just ETL)

Did you get the key message?



Key messages

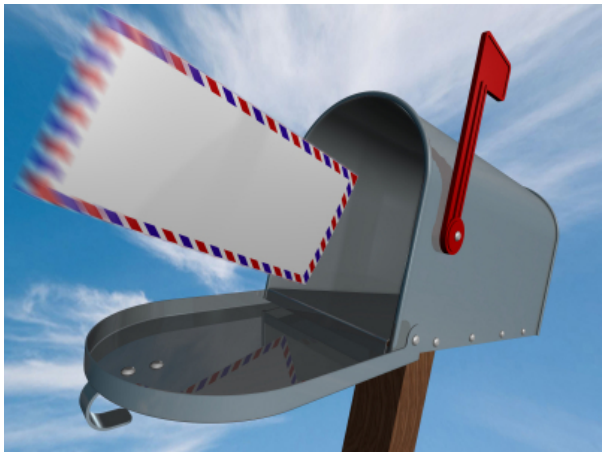
?

You have to care about big data to be competitive in the future!

Start your big data projects business driven!

Big data from a technical perspective is no (longer) rocket science!

Did you get the key message?



2.– 5. September 2013
in Nürnberg

Herbstcampus

Wissenstransfer
par excellence

Thank you!

Kai Wähner

Talend

@KaiWaehner

www.kai-waehner.de