BIG DATA != HADOOP

Moving the Cheese for the Industry

Agenda

1 Obeservation about Hadoop adoption in industry

2 Big Data has moved on

3 Business case on finding the cheese
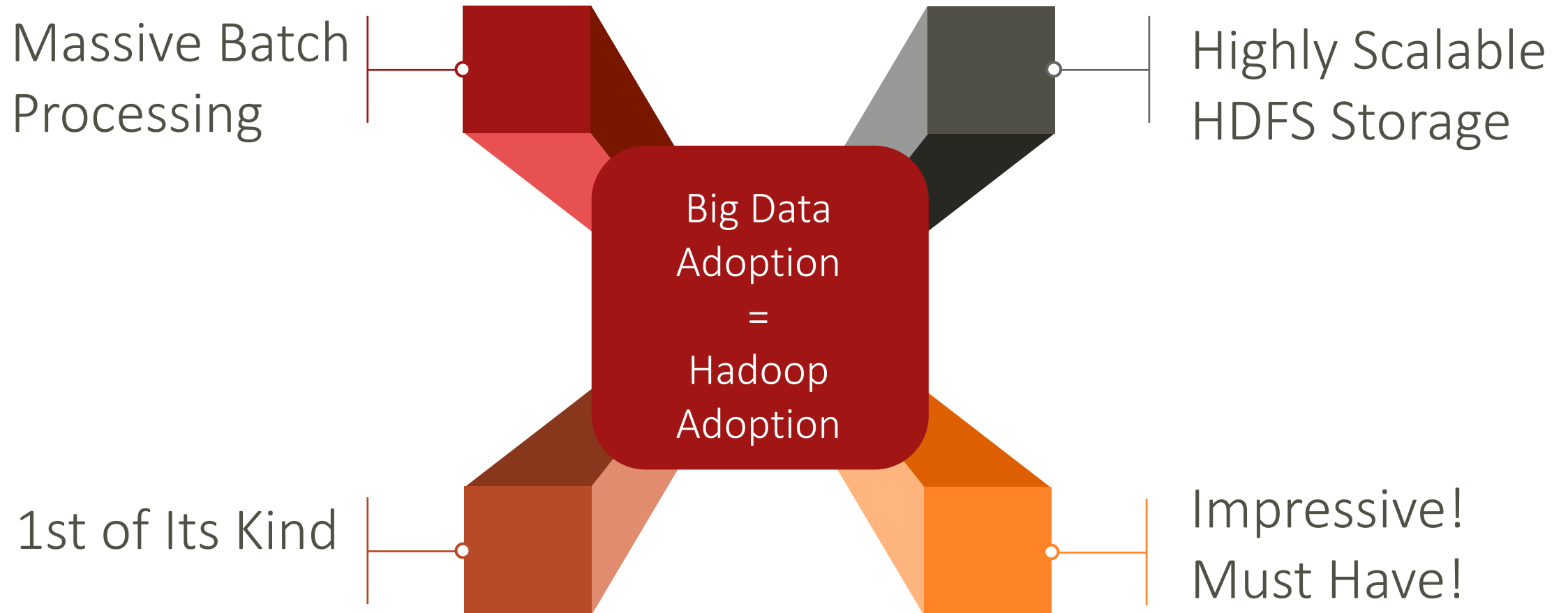
4 Conclusions

Andreas Tönne

- CTO of dibuco
- Special interests in
  - Cloud and Big Data Architecture
  - Business requirements for the new world
  - Programming language design

Observation: View of Big Data

Massive Batch Processing

Highly Scalable HDFS Storage

Big Data Adoption = Hadoop Adoption
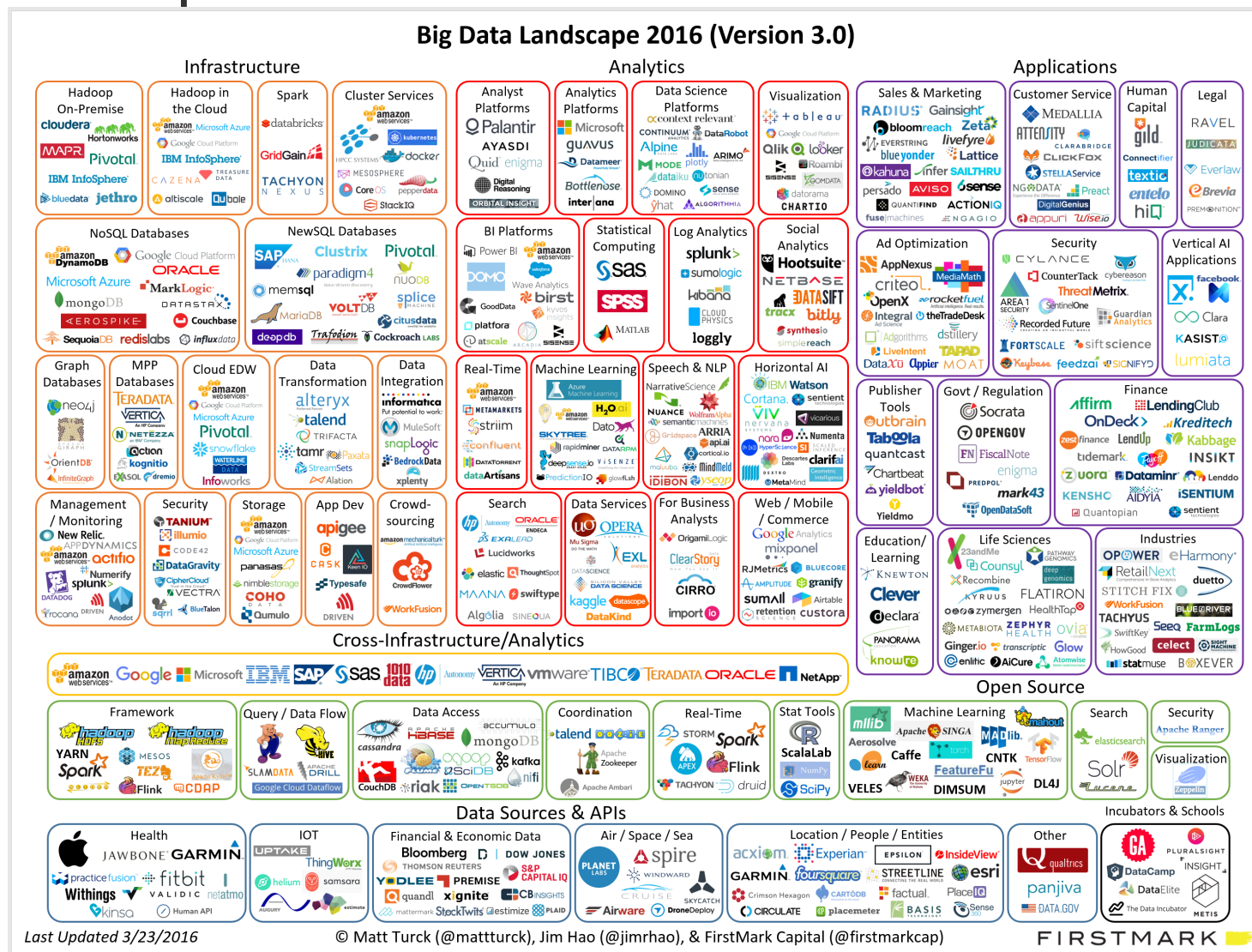
1st of Its Kind

Impressive! Must Have!

Observation: Hadoop is the Golden Hammer

Companies that have invested in Hadoop

- look for Hadoop solutions for every Big Data Problem

- think it is obvious to store Big Data in HDFS
- think it is obvious to use MapReduce

Find Hadoop!

Big Data Landscape 2016 (Version 3.0)

Last Updated 3/23/2016 — © Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)
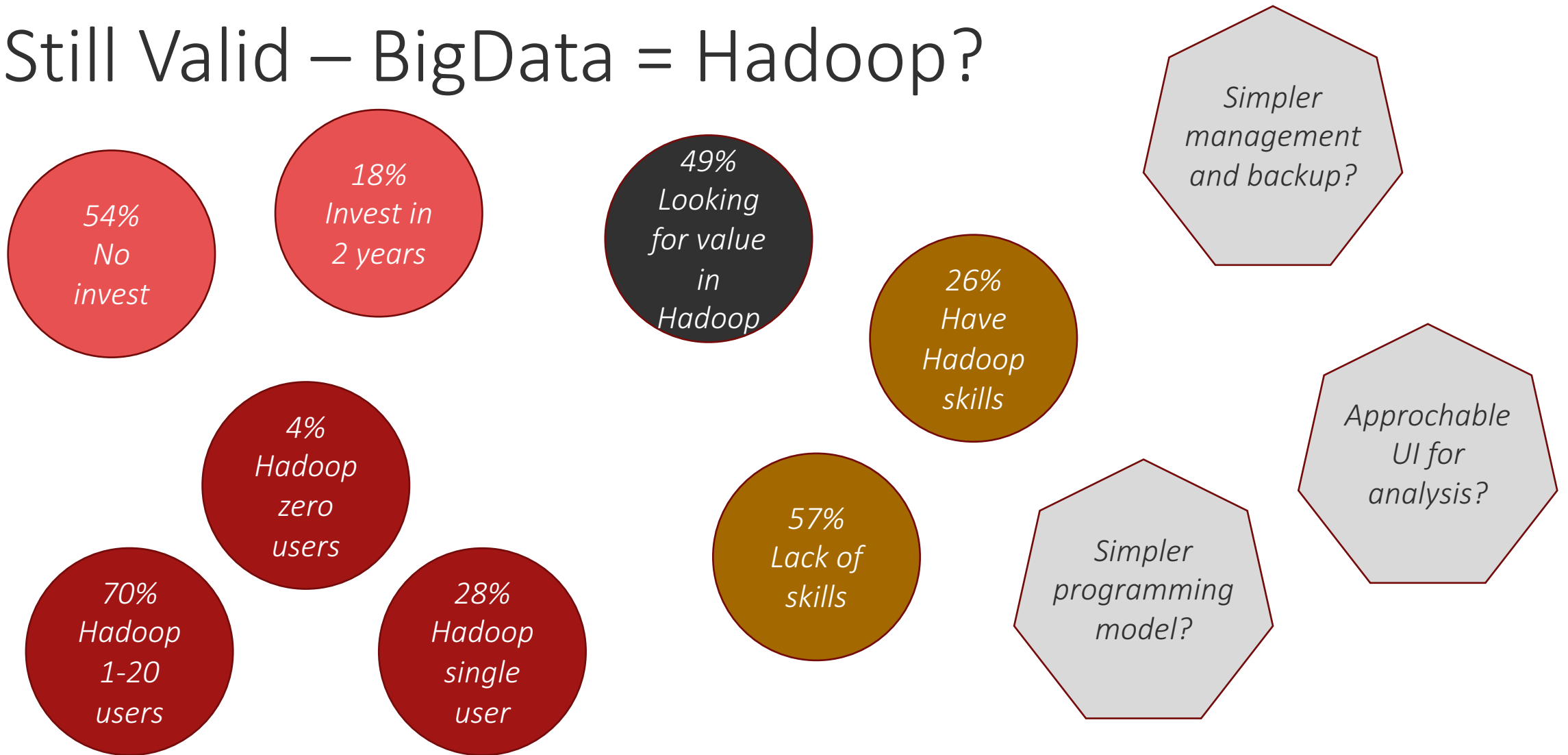
What Makes the Industry Cling on to Hadoop?

"Hadoop adoption is driven primarily by technology executives, especially those in the C-suite, including the CFO or COO." (Gartner)

Large investment! (Often) Small return

• Time and Personnel – Missing elsewhere

• Skills – Requirements and algorithms for MapReduce?

• IT-Operations – How to operate this alien?

Who dares to sack this investment driven by the C**?

Still Valid – BigData = Hadoop?

54% No invest

18% Invest in 2 years

49% Looking for value in Hadoop

26% Have Hadoop skills

Simpler management and backup?

4% Hadoop zero users

70% Hadoop 1-20 users

28% Hadoop single user

57% Lack of skills

Simpler programming model?

Approchable UI for analysis?

Data: Gartner Hadoop Adoption Study 2015

There are Good Reasons to Say – No!

"One of the core value propositions of Hadoop is that it is a lower cost option to traditional information infrastructure," Heudecker (Gartner) …

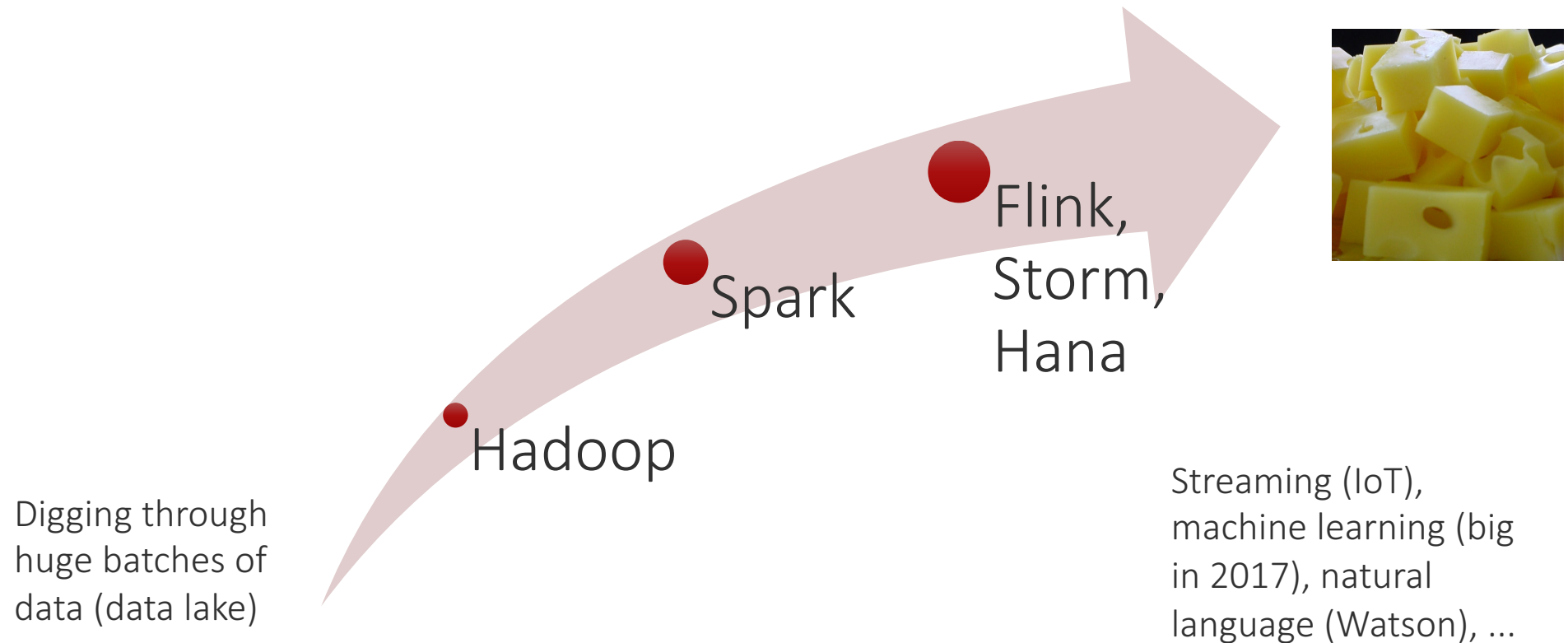"However, the low numbers of users relative to the cost of cluster hardware, as well as any software support costs,

may mean Hadoop is failing to live up to this promise."

That was 2015 – What about today?

Big Data Evolved-Where Is Your Cheese Now?

Gartner Hype Cycle 2016 – What goes missing? **Big Data**!

• Understanding of Big Data value matured

Flink,
Storm,
Hana

Spark

Hadoop

Digging through
huge batches of
data (data lake)

Streaming (IoT),
machine learning (big
in 2017), natural
language (Watson), ...

Big Data Evolved – Batch is only One Use Case

MapReduce is an excellent fit for embarrassingly parallel data processing

- …. real life is not embarrassingly parallel
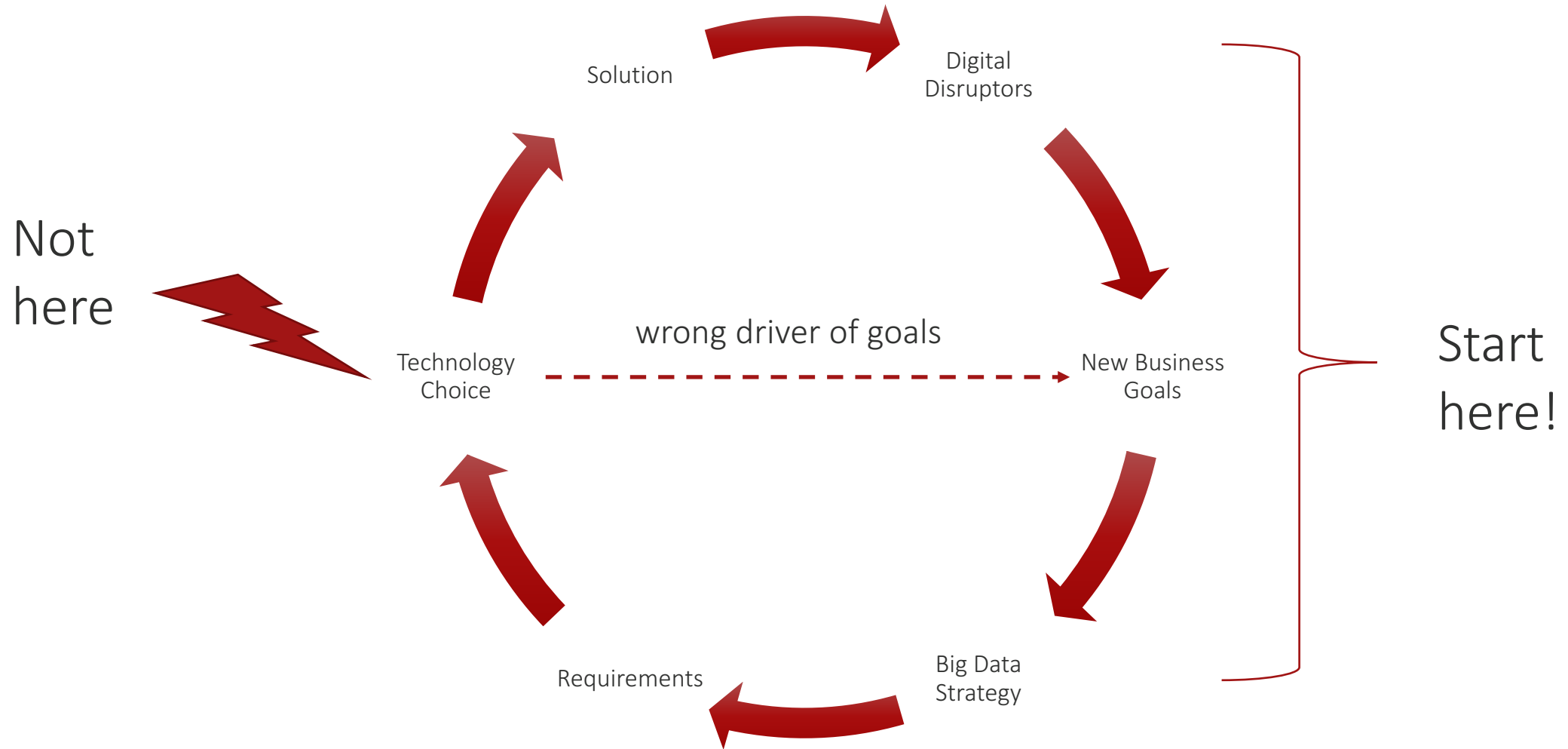- It's a stream of events, linked by history

Time is a critical factor for Big Data value gain

- Hadoop has a huge latency
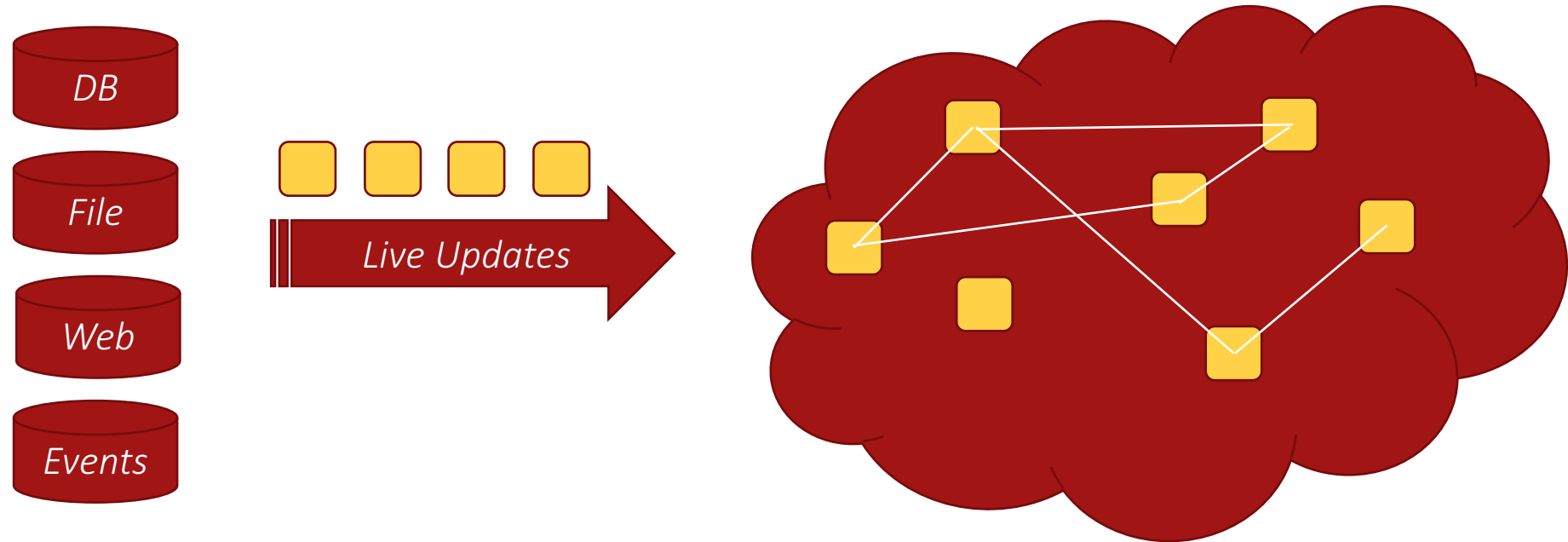- Batches are a historic extract of the real life stream of data

NB: HDFS storage can still be an excellent choice

Decision Process For Digital Transformation

Solution

Digital Disruptors

Not here

Technology Choice

wrong driver of goals

New Business Goals

Start here!

Requirements

Big Data Strategy

Business Case

dibuco

solutions for the digital age

Business Case – Getting Away From Hadoop

- Best possible linkage of data by textual contents

- Fast availability of new data
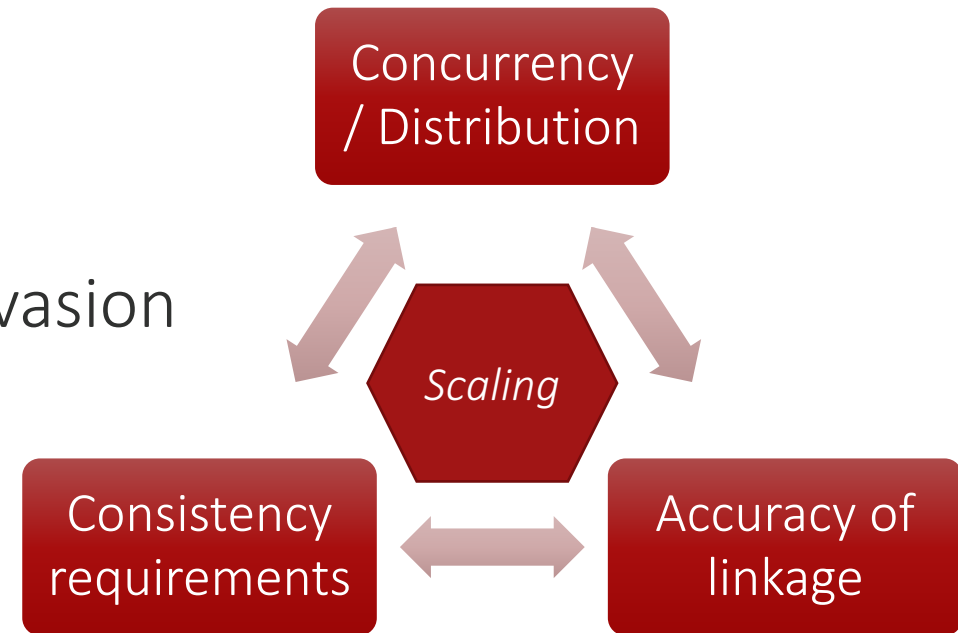
- Dealing with "language" changes

Business Case – Upfront Technology Decisions

- Hadoop for **everything global**
  - Maintaining identity constraints
  - Keeping global language statistics current
  - Mass import and maintenance
- Single NoSQL DB choice (Titan)
  - Graph DB matched logical data model perfectly
- Microservice/Queuing architecture for the rest

The Technology Started Biting our Goals

- Business value harmed by hard scaling problems
  - Single Data Swamp Storage
  - Storage Model inefficient
  - Oops we built a monolith!

- Problems solved by Hadoop batch invasion
- Time was running out. Literally!

Concurrency / Distribution

Scaling

Consistency requirements

Accuracy of linkage

It was time to...

Think Again!

Rethinking the Solution

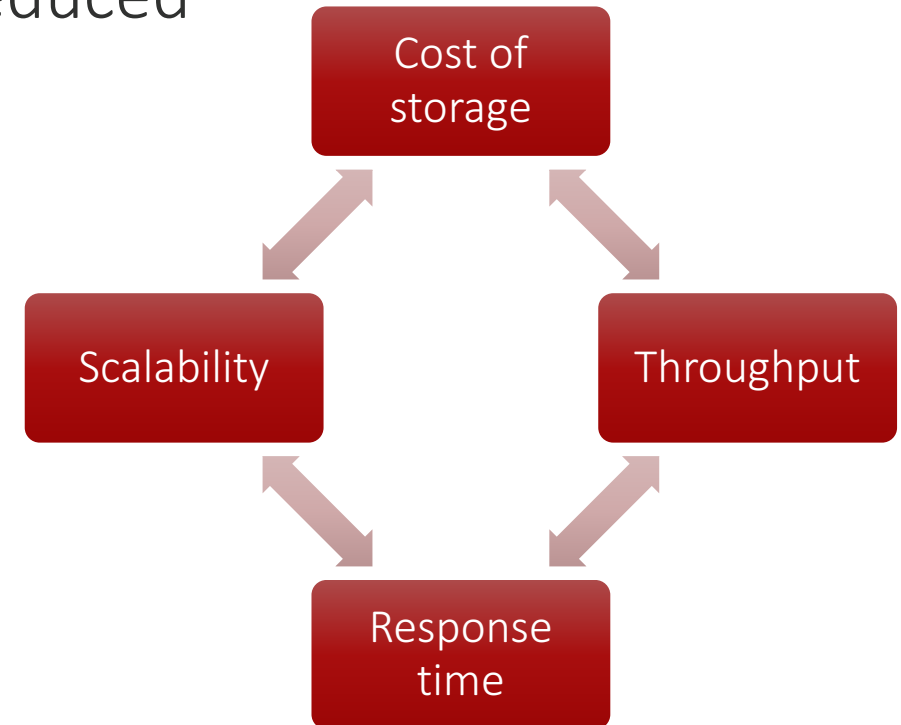|		
Data	▶	Split by service and usage needs
Data	▶	Consistency requirements reasonable?
Data	▶	Distribution of creation and usage?
Requirements	▶	Balance of requirements and scalability
Requirements	▶	Find scalable algorithmic solution or bin requirement
Business Goals	▶	What are the risks?
Legal Considerations	▶	What is allowed and accepted?
Data Governance	▶	E.g. data ownership, IAM system vs. scalability

💡 Insights 💡

- We have a streaming situation
- Time is of high importance
- Ideal consistency requirements can be reduced
- New algorithms allow to reduce the data model storage

Cost of storage

Throughput

Response time

Scalability

Solution

- True microservice architecture with polyglot persistence
- The best technology and model for each service

- Modular streaming architecture
- Multiple streaming topologies cut by service

- New streaming and big-data-optimized analysis algorithms
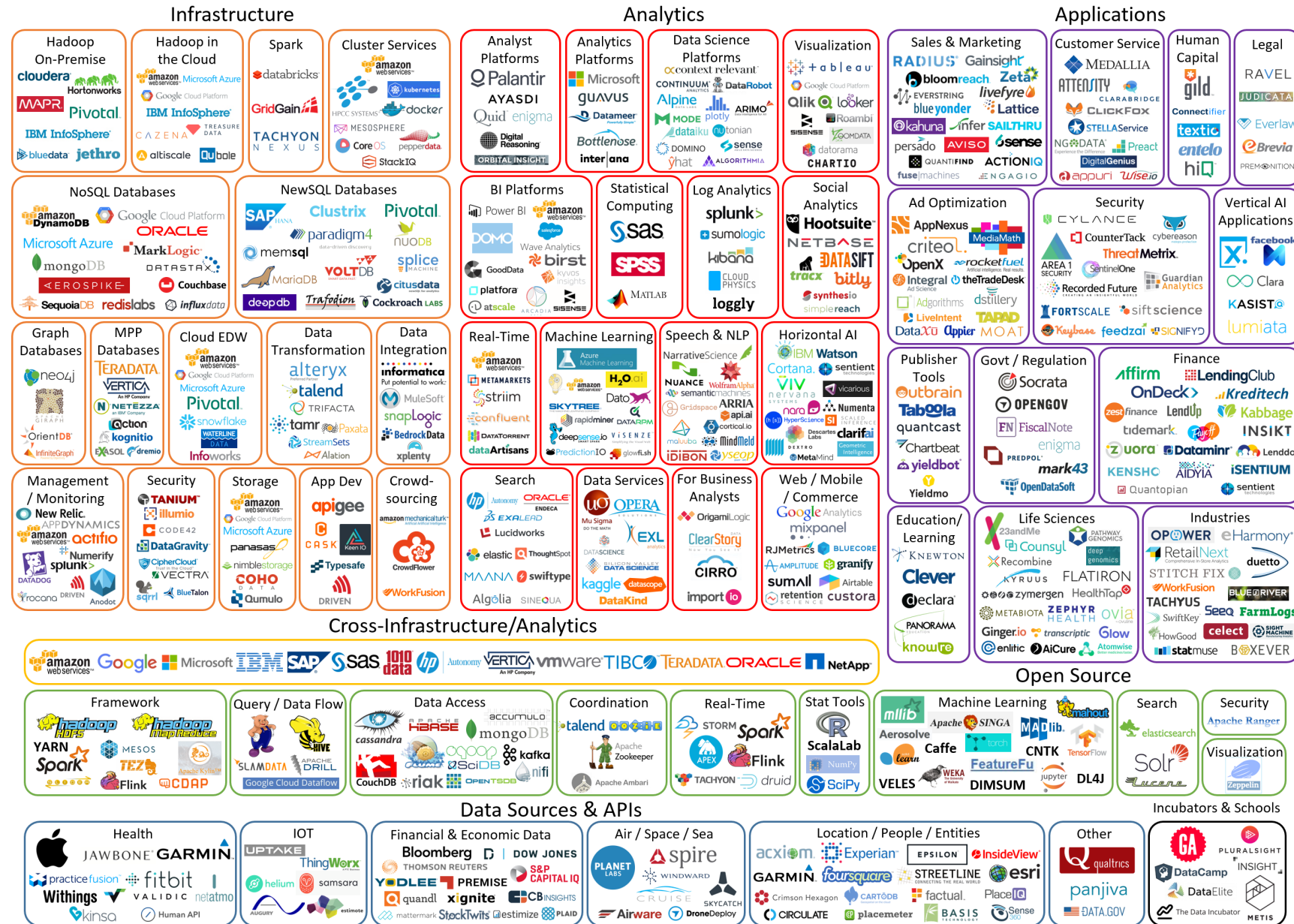- A lot of ad-hoc computation instead of global, aging information

Outcome

- Better results by ad-hoc computation 👍
- Incremental maintenance of aging linguistic information 👍
- Massive reduction of storage requirements (est. up to 70-80%) 👍
- True horizontal scalability by service 👍

- Complete removal of Hadoop batches 👍

Conclusions

- Hadoop is seen as the role model of Big Data (our observation)

- Investment in Hadoop one of the reasons to stick with Hadoop
- We observe that problems are crafted to be solvable by Hadoop

- Expectations for Big Data evolved beyond batch processing

- Allow rethinking of the business goals and solution requirements without technology in mind

Big Data Landscape 2016 (Version 3.0)

THANK YOU FOR YOUR ATTENTION!

Franz-Schubert Straße 75
70195 Stuttgart
+49 711 699 475 60
info@dibuco.de
www.dibuco.de

dibuco
solutions for the digital age

Sources

- Cheese theme: "Who Moved My Cheese?: An Amazing Way to Deal with Change in Your Work and in Your Life" (Spencer Johnson) G. P. Putnam's Sons; 1 edition (September 8, 1998)

- Cheese picture, Wikimedia Commons (Christian Bauer)

- Big Data Landscape 2016 (C) Matt Turk, Jim Hao, FirstMark Capital

- Hammer Malene Thyssen, http://commons.wikimedia.org/wiki/User:Malene

- Pocketwatch, Wikimedia Commons (No user listed)