Big Mining

Die Suche nach der Nadel im Datenhaufen

Data Mining Vorlesung HS Offenburg WS 2018/19

Dr.-Ing. Alexander Schätzle

13.12.2018

Data Center & Managed IT-Services Provider

Who we are !?

Shareholder
badenova AG & Co. KG (100%)

Employees
104 contact persons

Founded
1997

Managed Services

Big Data Services

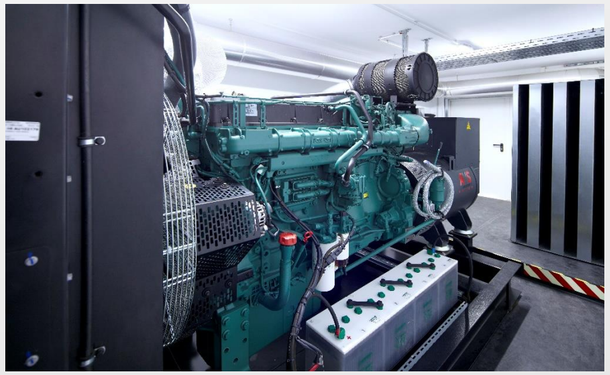
Internet / MetroNet

Communication Services

Data Center

Your data. Our business.

- ISO / EMAS certified
- Secured and supervised company premises with 24/7 on-site staff
- Logged access control
- Water signaling devices
- Fire alarm system with early fire detection
- Fire extinguishing system (F90 Room-in-Room solution)
- Redundant power supply
- Redundant UPS, own redundant transformer station
- Diesel generator
- Multistage redundant firewall systems
- Tier3+ standard



D Z G M Deutsche Zertifiziergesellschaft Mittelstand GmbH zertifiziert nach DIN EN ISO 9001:2008

EMAS GEPRÜFTES UMWELTMANAGEMENT DE-126-00119

TÜV SÜD ISO 20000

TÜV SÜD ISO 27001

Fachfirma Brandmeldeanlagen TÜV SÜD DIN 14675

Our Customers

Who I am !?

Dr.-Ing. Alexander Schätzle

Big Data Architect

badenIT GmbH
Tullastr. 61
79108 Freiburg

+49 (0)761 5035-4838
+49 (0)151 17652542

alexander.schaetzle@badenIT.de
www.badenIT.de

- **Informatik-Studium (M.Sc.)**
  Albert-Ludiwgs-Universität Freiburg
- **Promotion (Dr.-Ing.)**
  Thema: Distributed RDF Querying on Hadoop
  Albert-Ludwigs-Universität Freiburg
- **Seit 2017 Big Data Architect**
  badenIT GmbH

What comes into your mind when you think about "Data Lake"?

pollev.com/alexandersch099

What comes into your mind when you think about "Data Lake"?

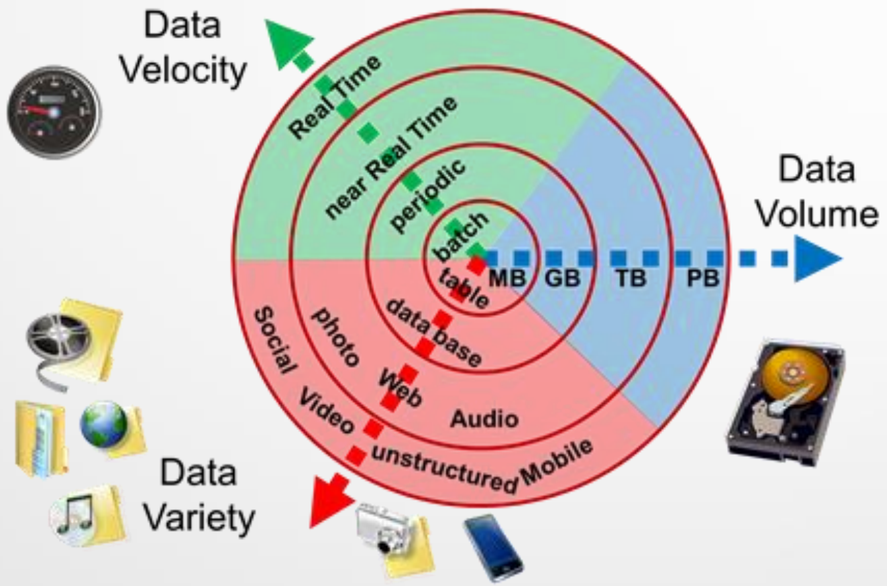
Big Data

Size does matter. But not only.

What is Big Data?

There is no simple definition of Big Data

- Often "defined" using the **3V-model** of Gartner



[Source: https://gi.de/informatiklexikon/big-data/]

Volume

Data size exceeds the capabilities of traditional data processing systems

Variety

Data exists in many different shapes (unstructured | semi-structured | structured)
There is no single valid data format

Velocity

Data is produced very short intervals and might also be outdated in short time

Requirements for Big Data

So what are the most important design properties of a Big Data platform?

Do not enforce a specific schema or format on time of data ingestion

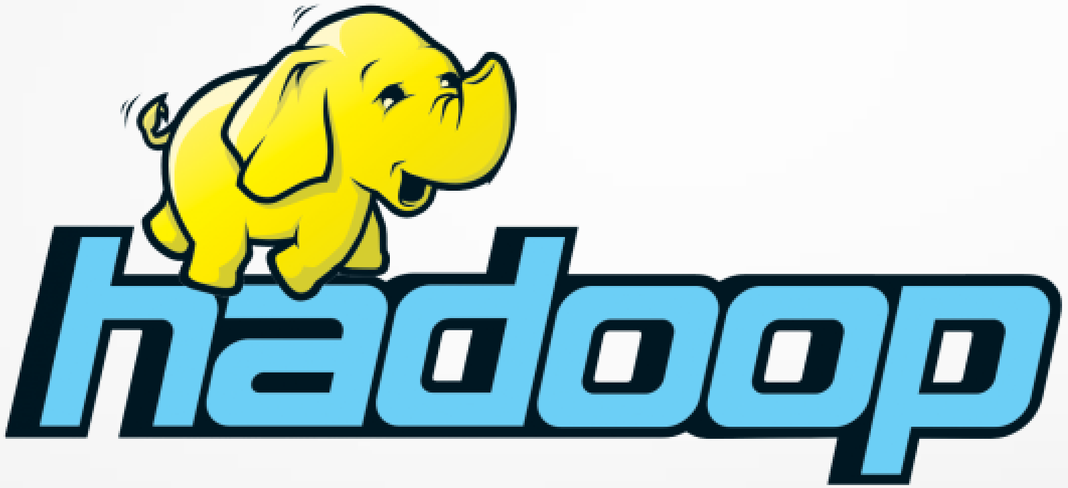
for cost and time efficiency

Assume hardware failures to be normal

3 Schema on read, not on write

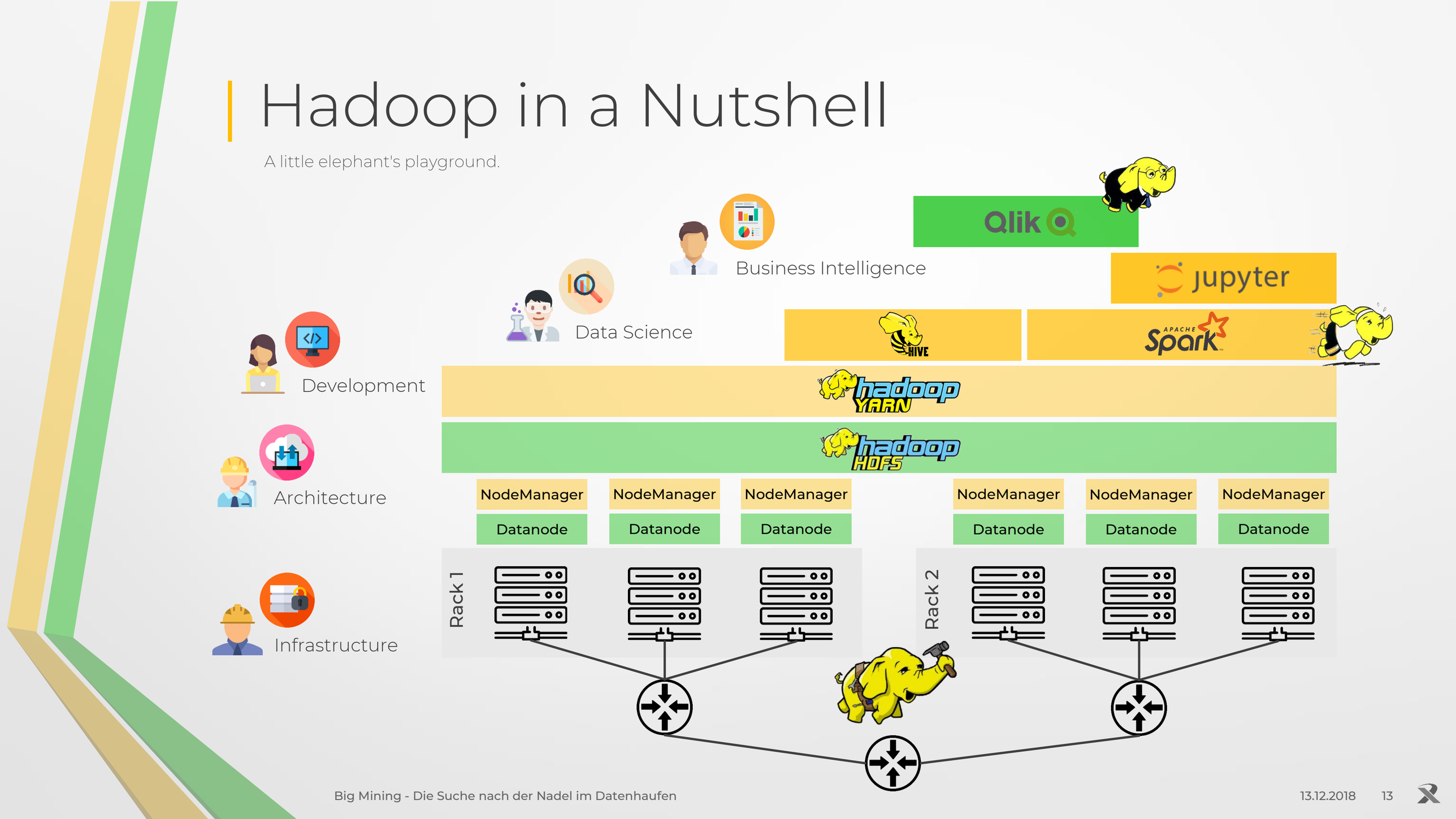
2 Rely on commodity hardware

1 Scale out, not up!


hadoop

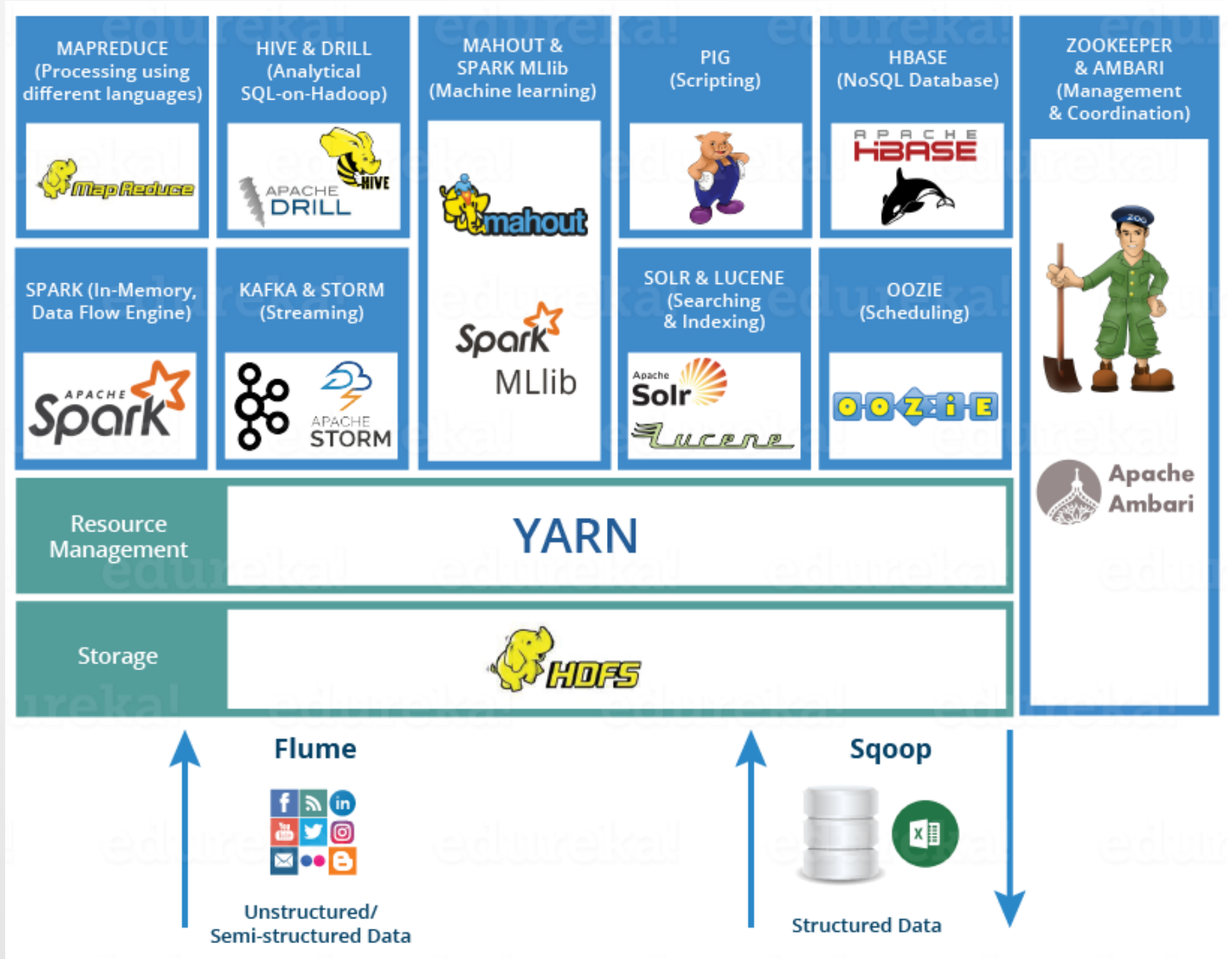
Hadoop in a Nutshell

A little elephant's playground.



Hadoop Ecosystem

The Madness beyond the Gate.



[Source: https://www.edureka.co/blog/hadoop-ecosystem]

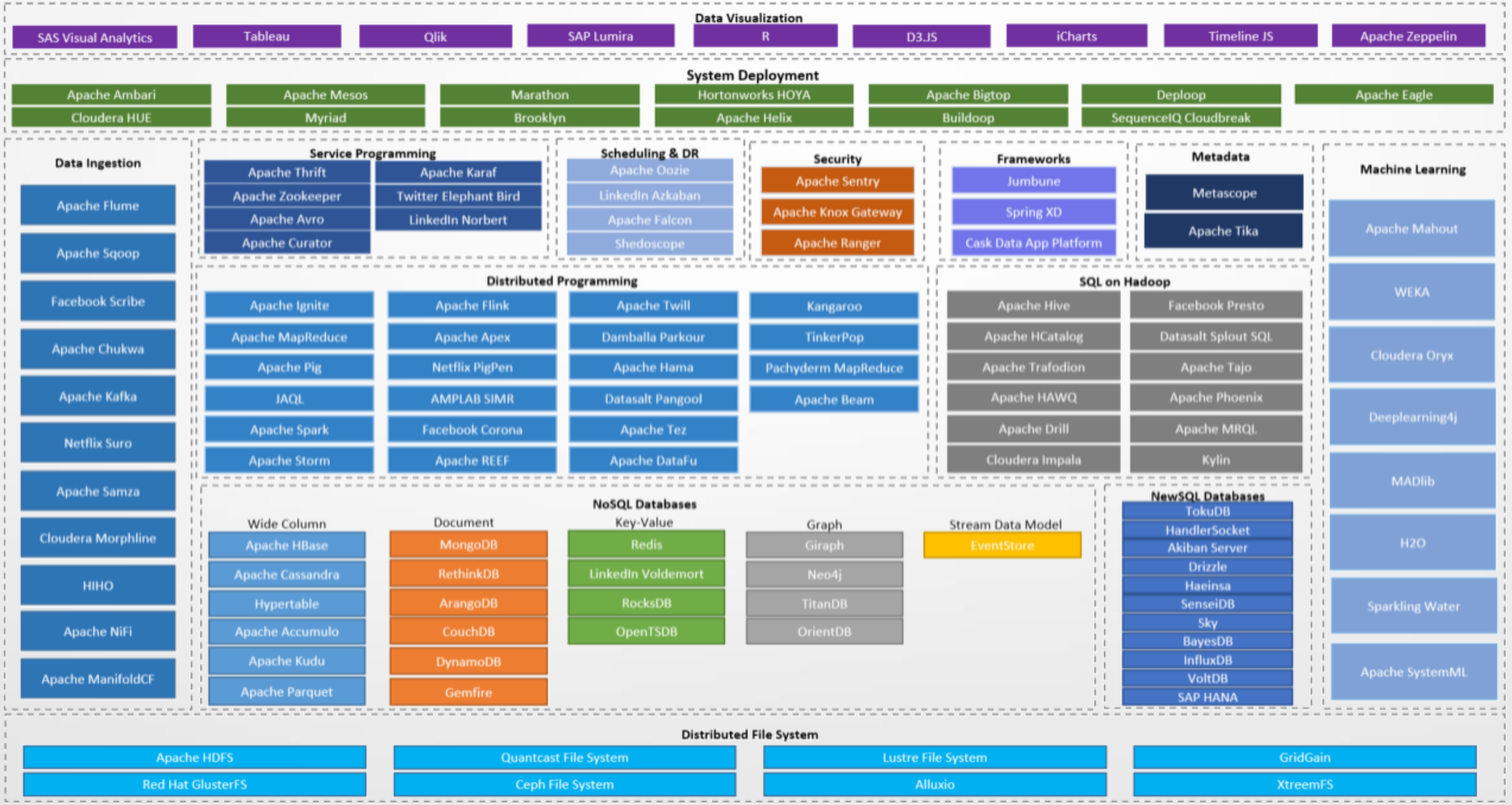
But actually that's just the ...

tip of the eisberg

Hadoop Ecosystem

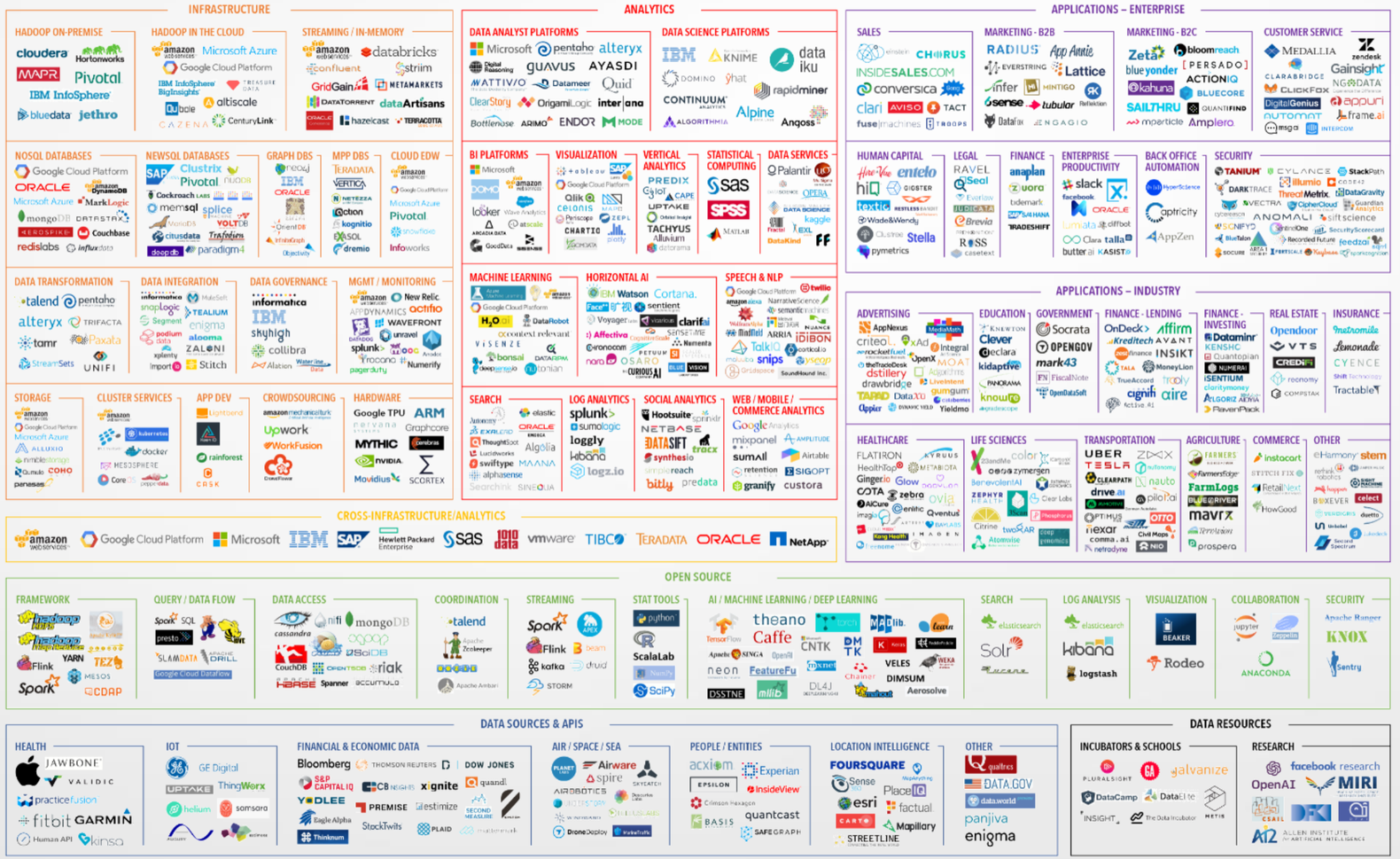
The BIG Madness beyond the Gate.

[Source: https://mydataexperiments.com/2017/04/11/hadoop-ecosystem-a-quick-glance/]



Big Data Landscape

The GIANT Madness beyond the Gate.



Data Lake

Diving into your hidden data treasure.

"If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples."

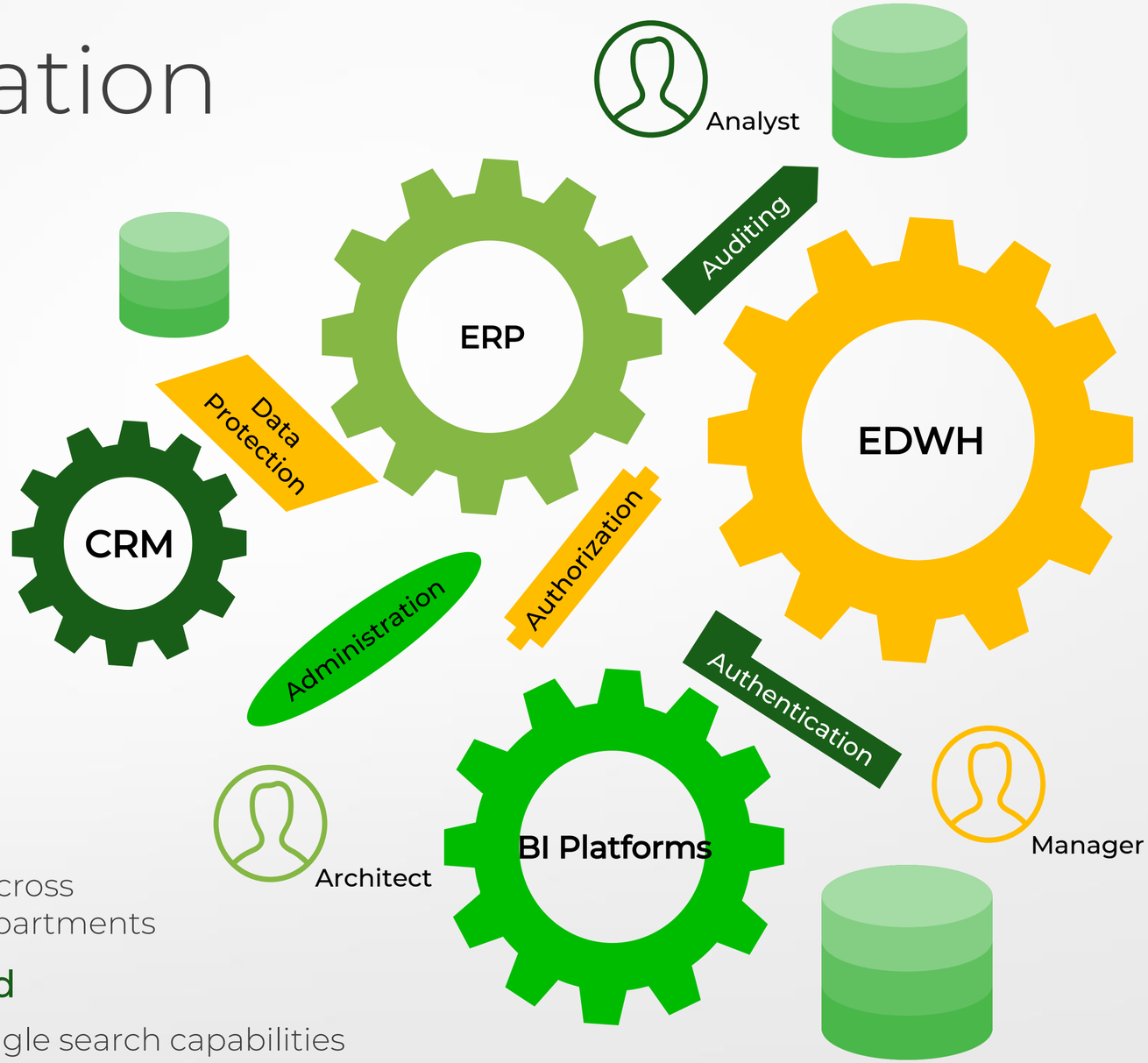
(James Dixon, CTO of Pentaho, 2010)

Data Orchestration

Or why scattering is the best data protection.

Big companies have big / diverse IT landscapes

- Data orchestration becomes more and more ineffective and inefficient



Diverse

- no single version of truth
- many meanings in many contexts

Scattered

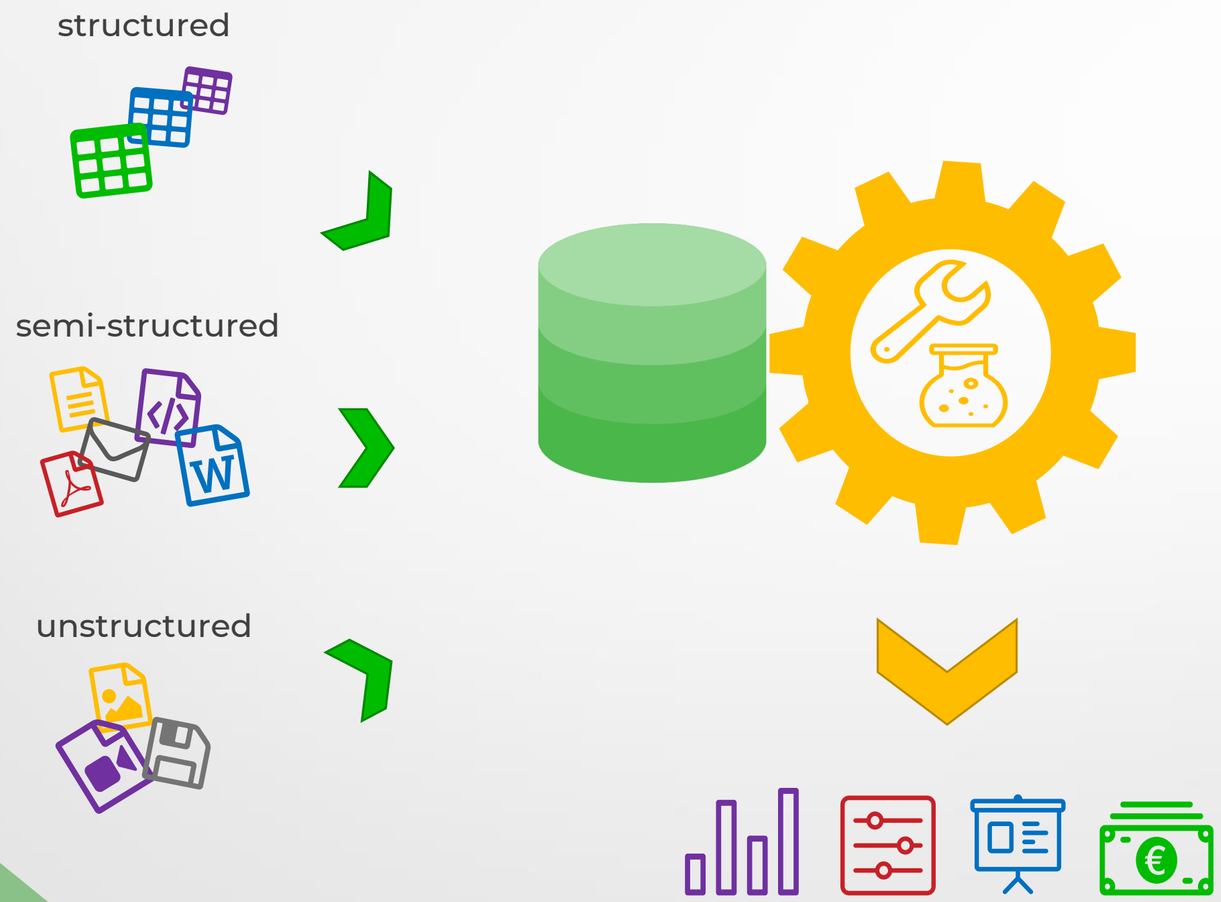
- no single data source
- scattered and fragmented across different systems, apps & departments

Ungoverned

- no single search capabilities
- no clear ownership

Data Lake - Concept

The BIG picture.



Key characteristics

- Flat architecture, centralized repository
- any kind of data in its native format
- massive scale-out storage
- scale-out processing power / ability
- Multiple applications
- In-place analytics

Data Lake = Data Warehouse 2.0 ?

Old wine in new bottles. New wine in old bottles.

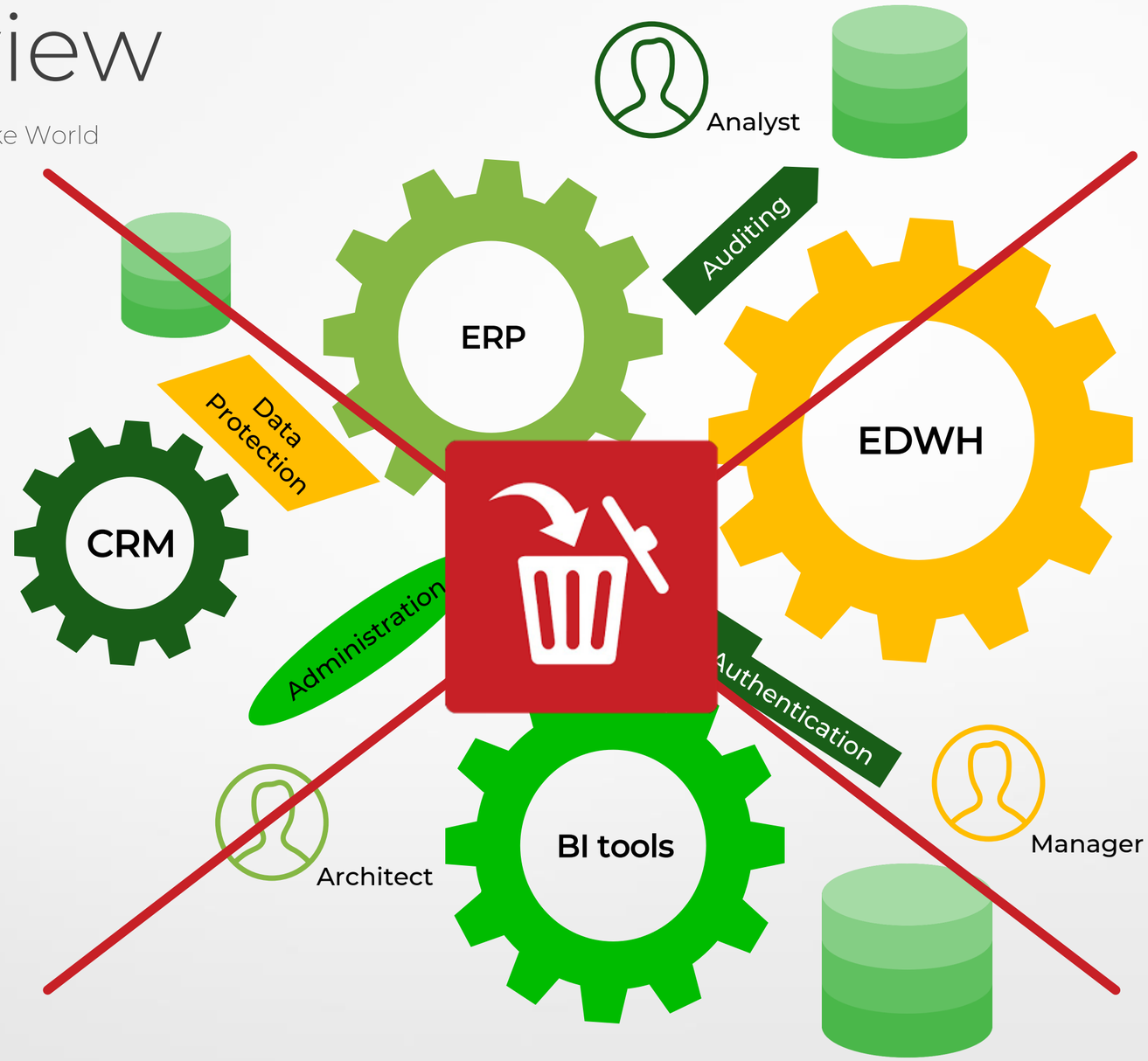
Some people say a data lake is just a data warehouse revised

Data Warehouse	vs.	Data Lake
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et al.

[Source: https://blogs.sas.com/content/customeranalytics/tag/big-data-cheat-sheet-on-hadoop/]

Review

Pre-Data-Lake World



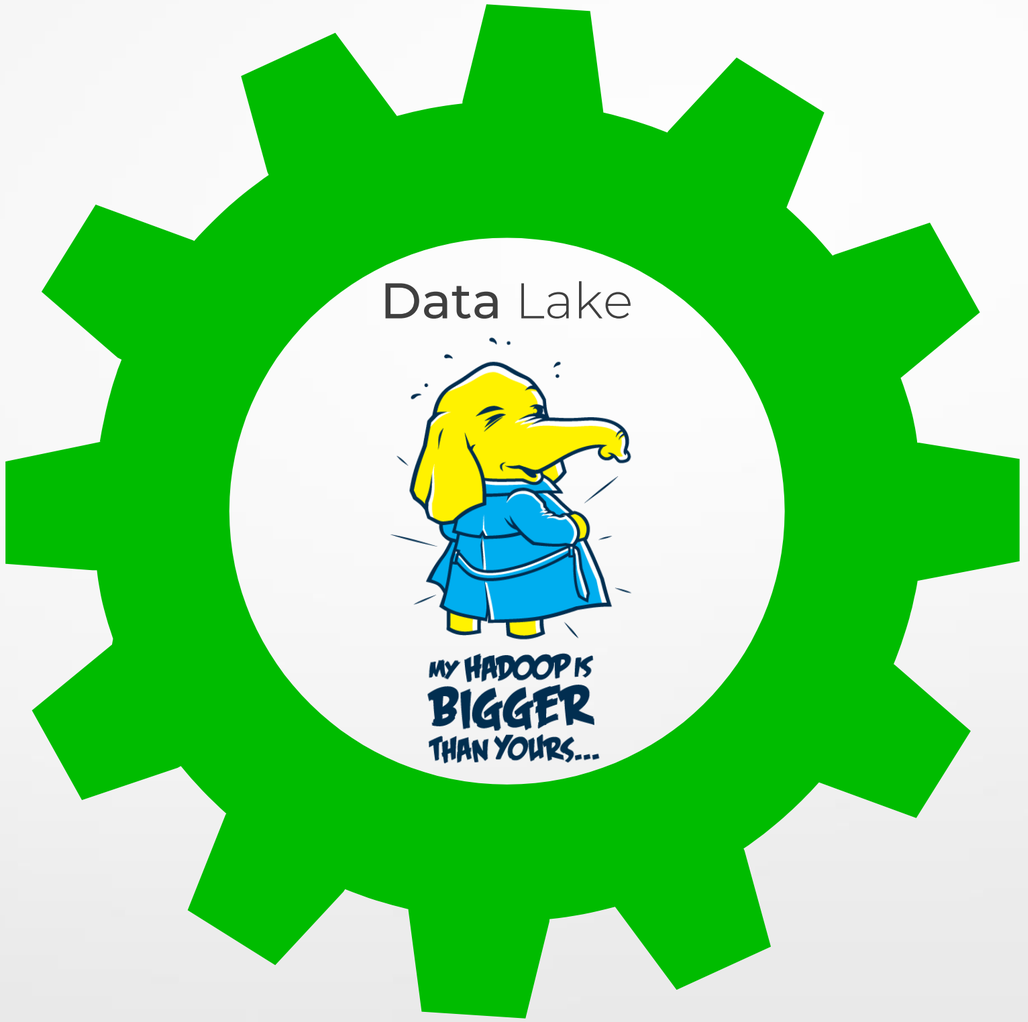
Brave new ~~World~~ Wonderland

Data Lake. ONE size fits ALL.

Can you imagine

any project in an

IT landscape …



that works

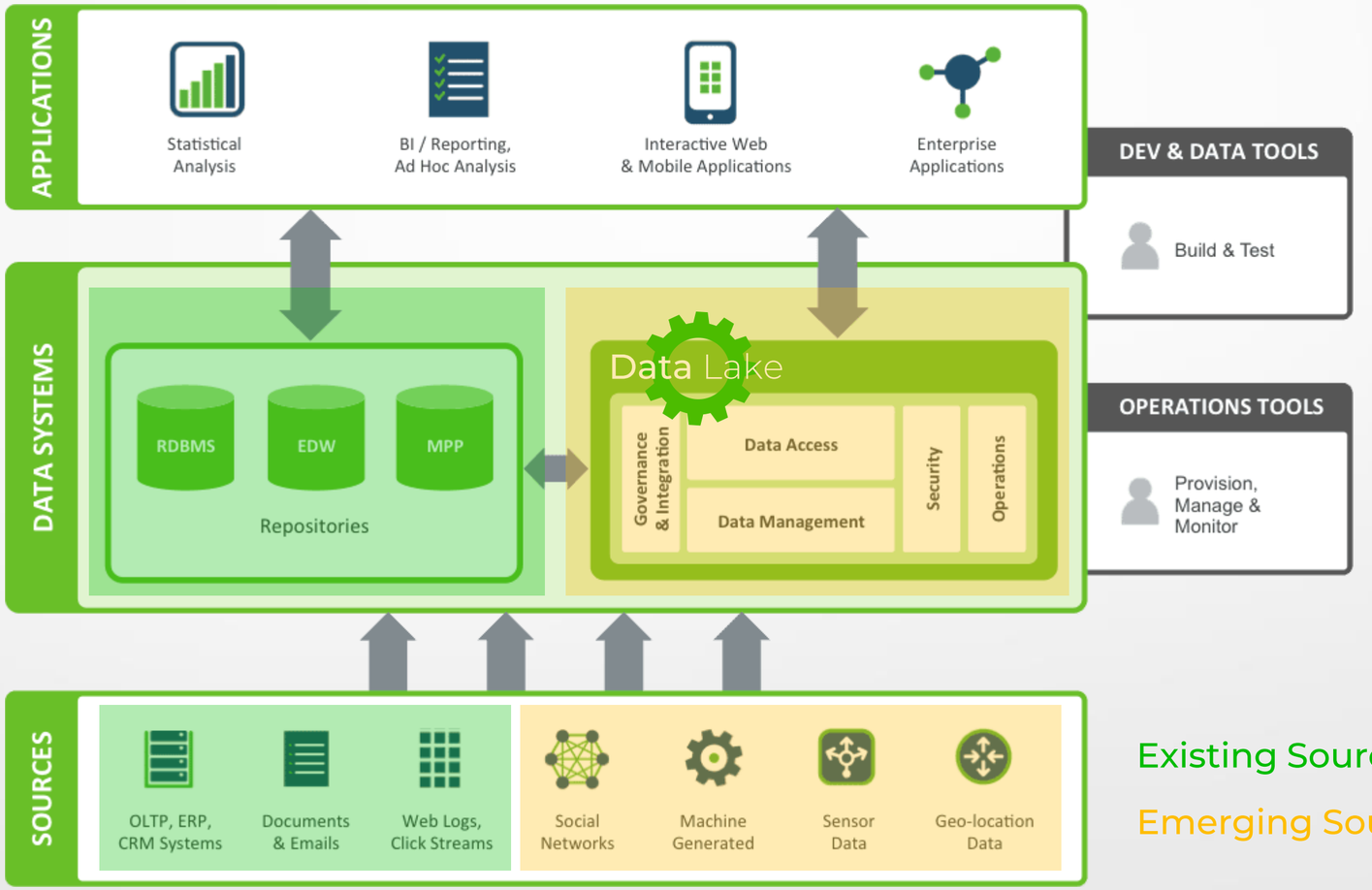
like that !?

Enterprise Data Lake

Not a REVOLUTION but an EVOLUTION.

Production

Research, Development & Adanced Analytics



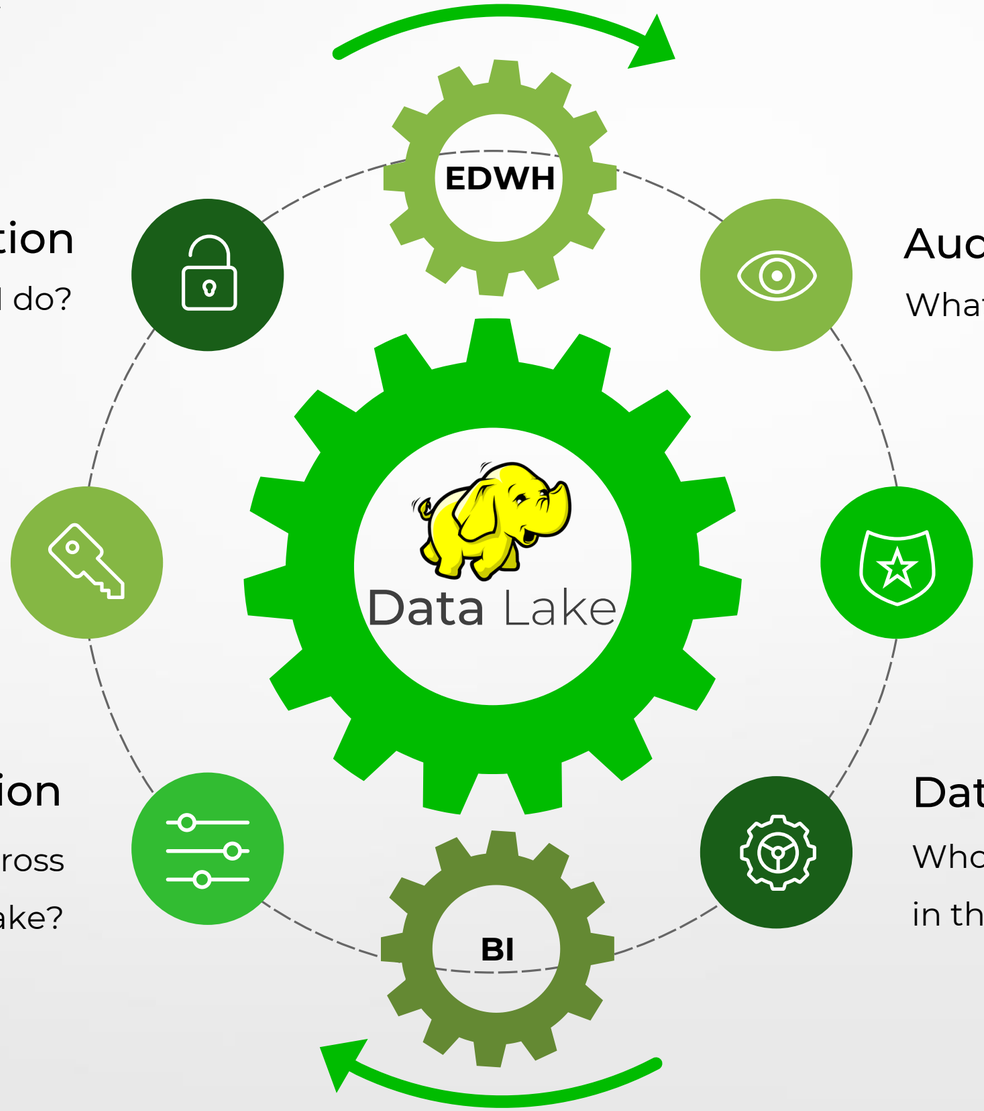
[Source: Hortonworks]

Existing Sources

Emerging Sources

Data Lake – 6 Pillars

Enterprise-ready architecture.



Authorization
What can I do?

Audit
What did I do?

Authentication
Who am I / prove it?

Data Protection
How can I protect my data from unwanted access?

Administration
How do I set policies across the entire data lake?

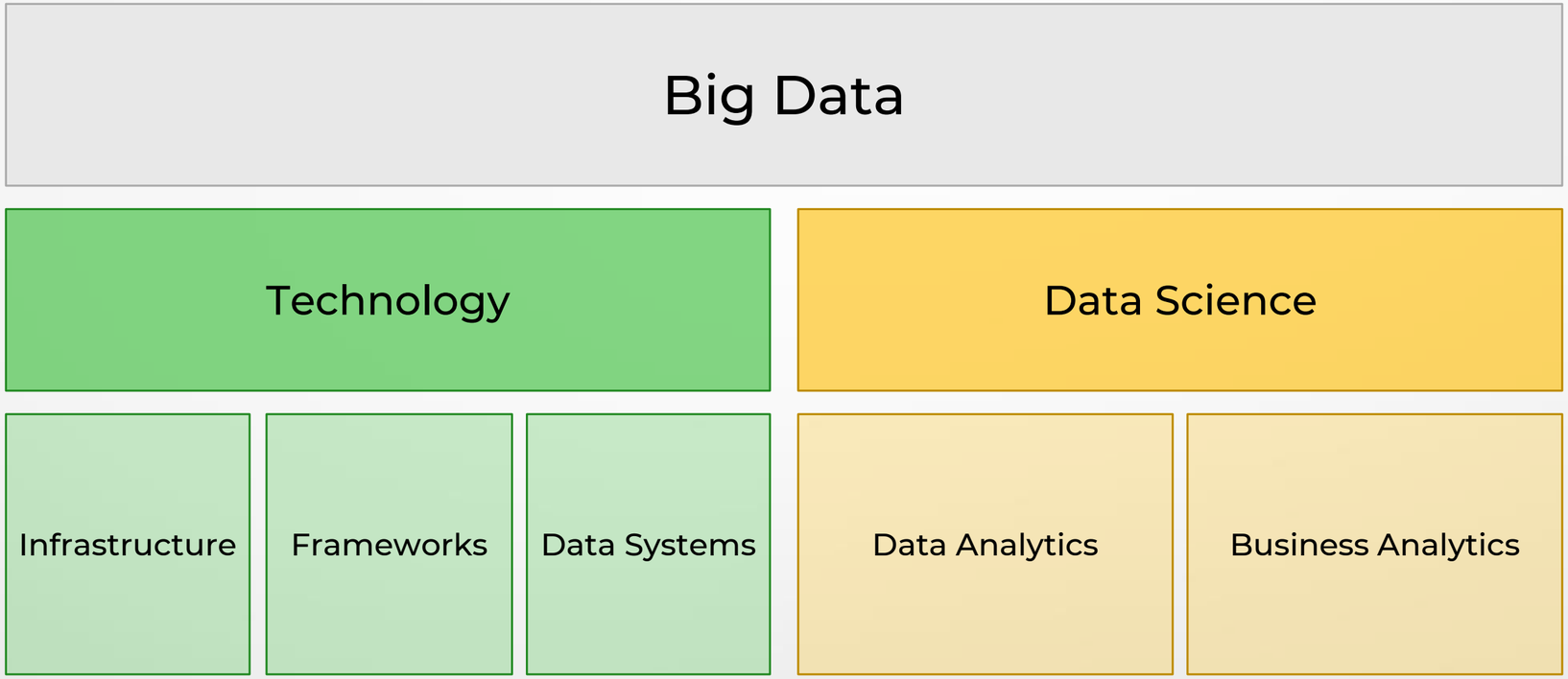
Data Governance
Who is responsible for the data in the lake?

Application Fields

Data is like oil. It's may more useful if processed.

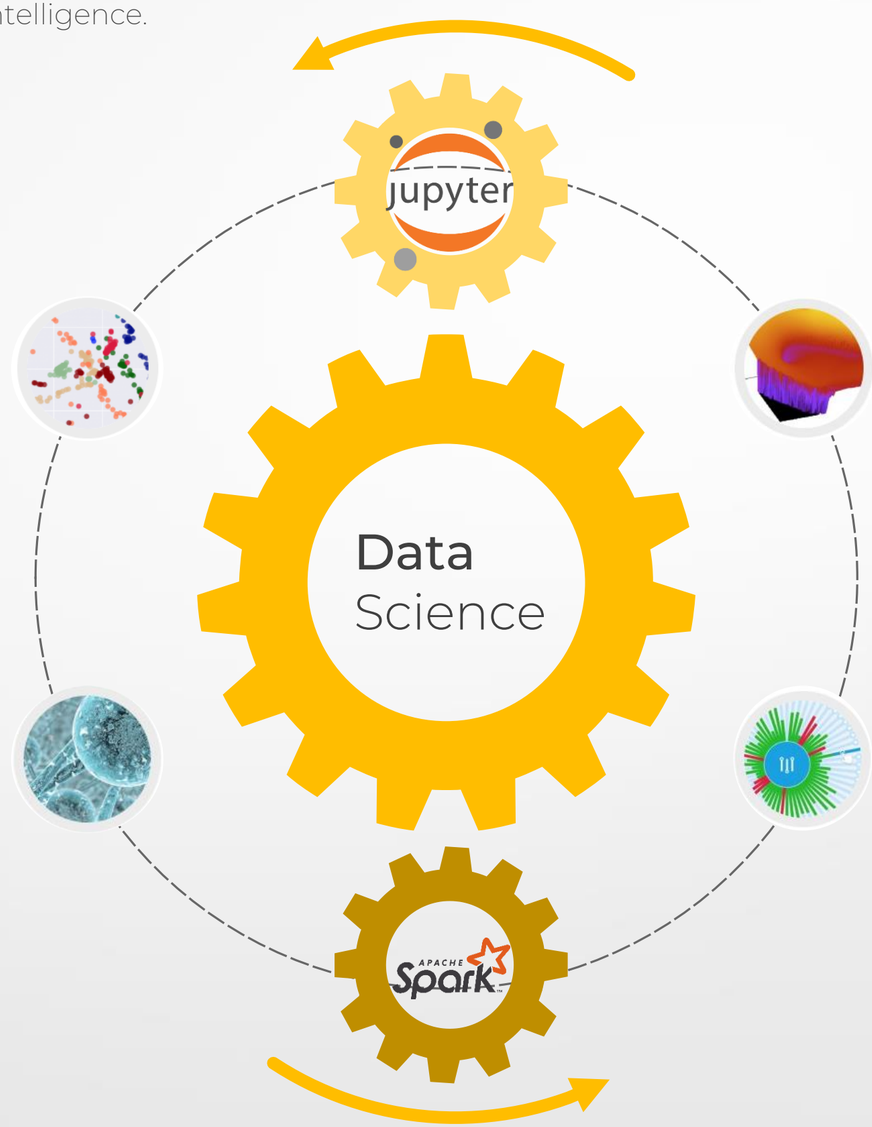
Big Data Categorization

Big Data and its descendants.



Data Science - Toolbox

Machine Learning. Artificial Intelligence.



Data Mining
Analyze Data, Discover Potentials

Deep Neural Networks
Classification & Anomaly Detection

Reinforcement Learning
Self-optimizing Algorithms

Predictive & Prescriptive Analytics
Forecasts & Recommendations

Data Science - Processing Requirements

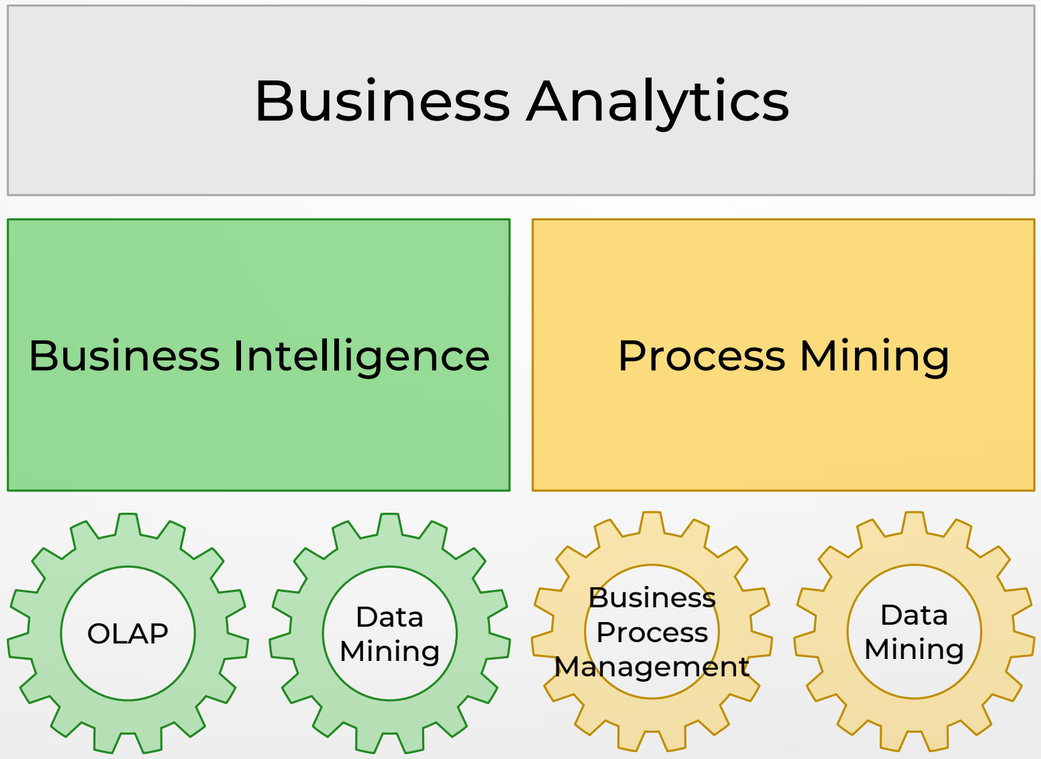
We need more POWER.

	Solution Development	Model Training	Deployment / Application
Purpose	**Select** model, **develop** and **validate** solution	**Train** Model on data	Model does **predictions**
Operation	Train model **several times** Data Scientist needs **Interactive Analytics**	Pass **all** data **several times** through model **and change parameters**	Pass **new** data **once** through the model
Complexity	100x 100x	100x	1x
Hardware	cluster	specialized hardware / cluster	standard server

Business Analytics

Data Science for Business.

Aufbereitung, Auswertung und Darstellung von Daten zur Ermittlung von KPIs (Key Performance Indicator)

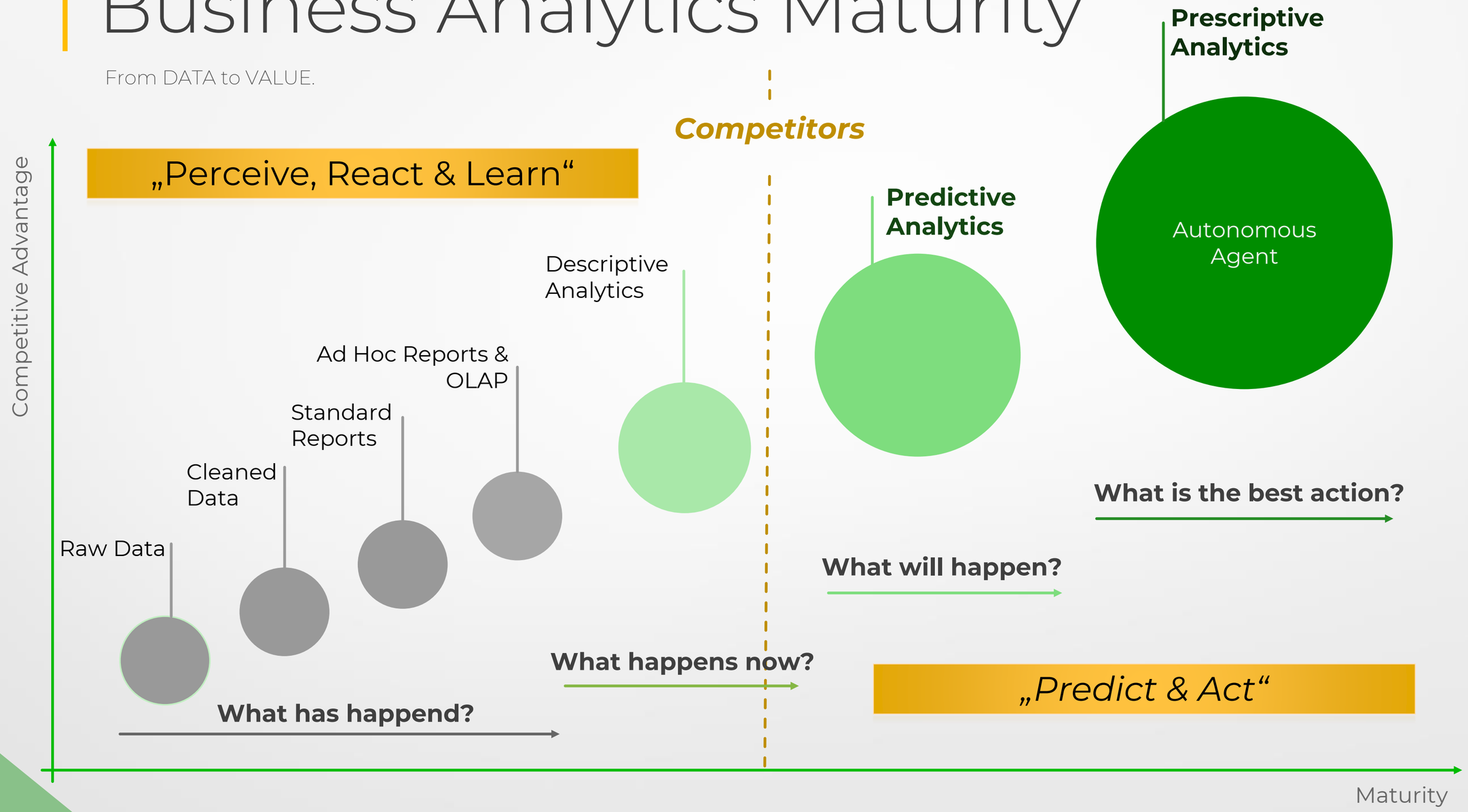


Umfassende Nutzung von Daten, statistischen und quantitativen Analysen sowie erklärenden und voraussagenden Modellen.

Prozessanalyse auf Basis digitaler Spuren in IT-Systemen. In den Daten enthaltenes, implizites und sonst verborgenes Prozesswissen.

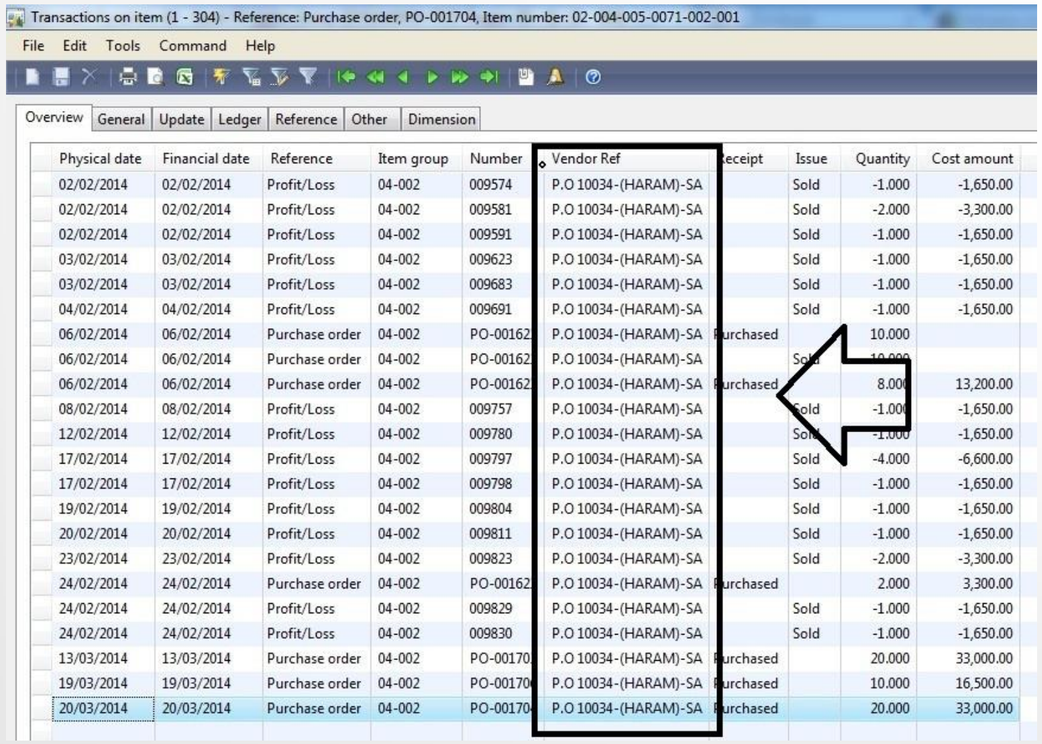
Business Analytics Maturity

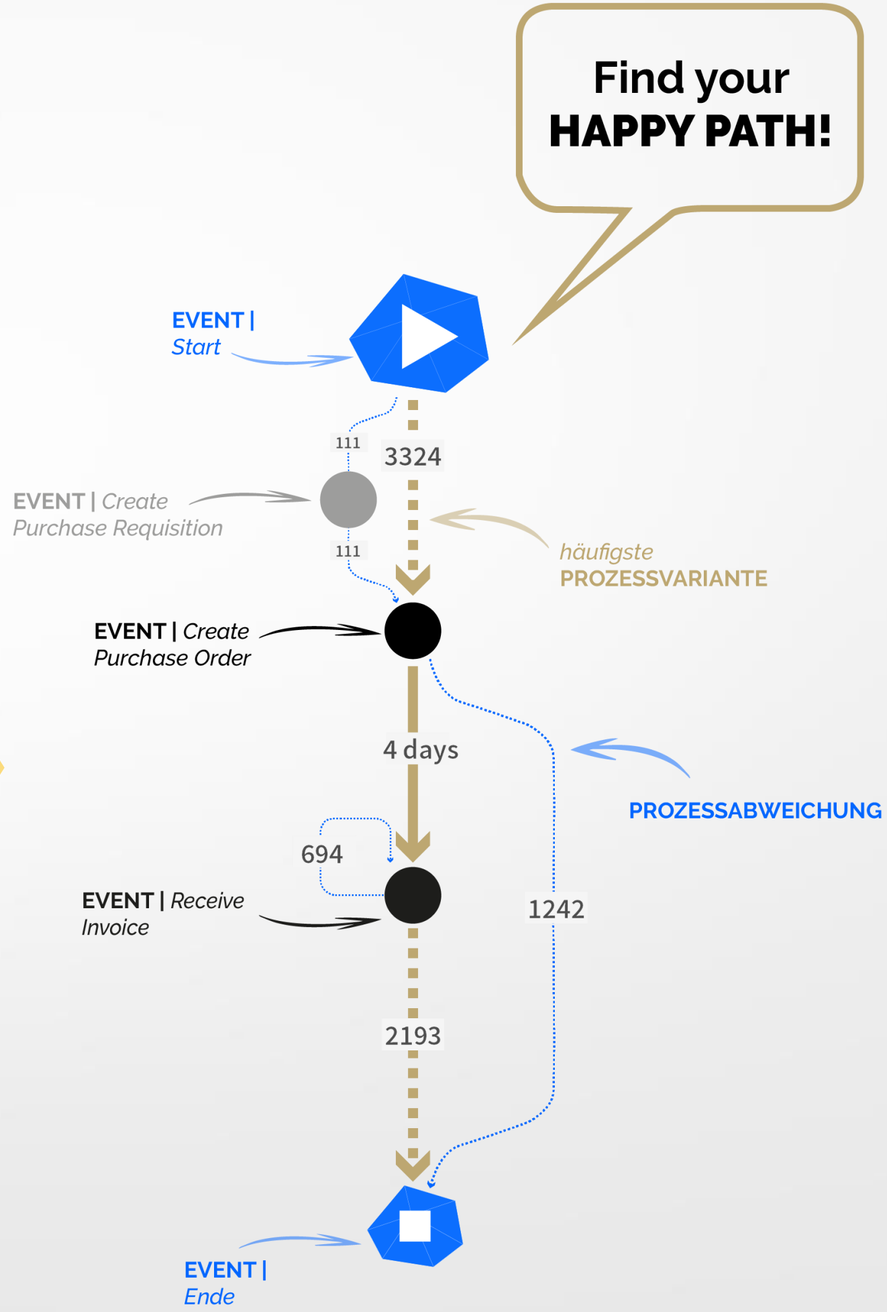
From DATA to VALUE.



Process Mining

Know your process. Know your business.



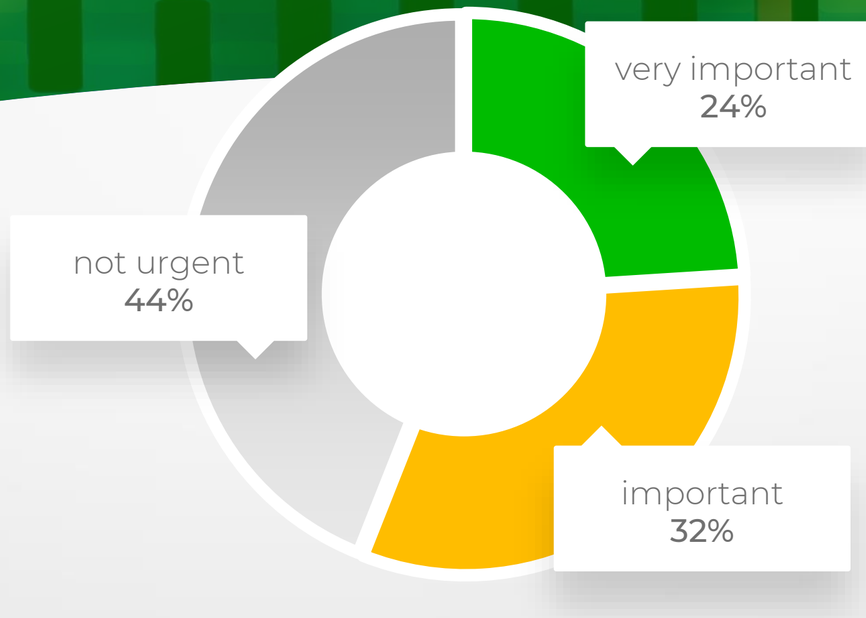


Demo

Get the hands dirty.

Some Numbers

A transition to the BI "language"

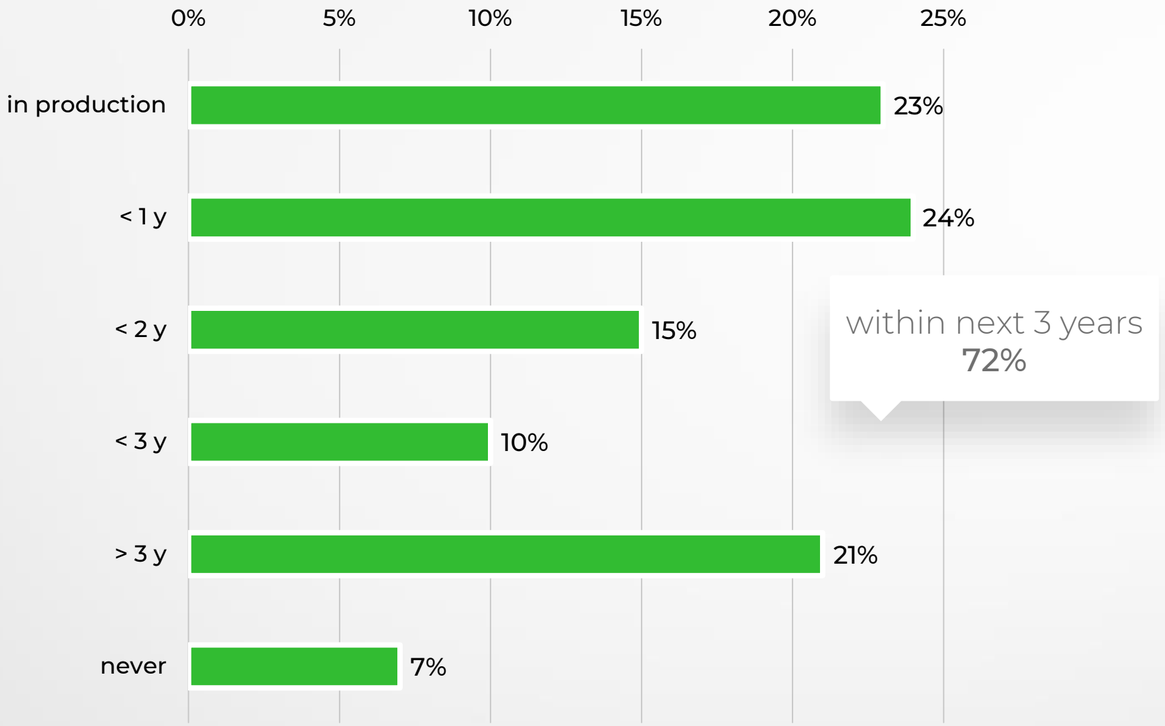


Importance of an (Hadoop-based) Data Lake for your own organization

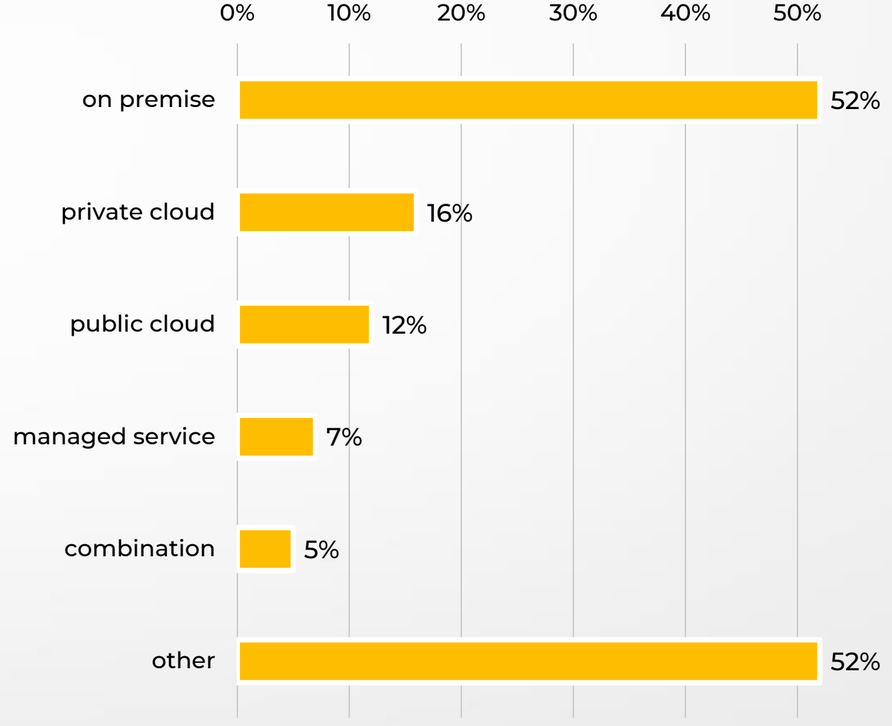
[Philip Russom: Data Lakes – Purposes, Practices, Patterns, and Platforms, TDWI Best Practice Report, Q1 2017]

Date Lake in Production

Start now. Move forward.



When do you expect to have a Data Lake in production?

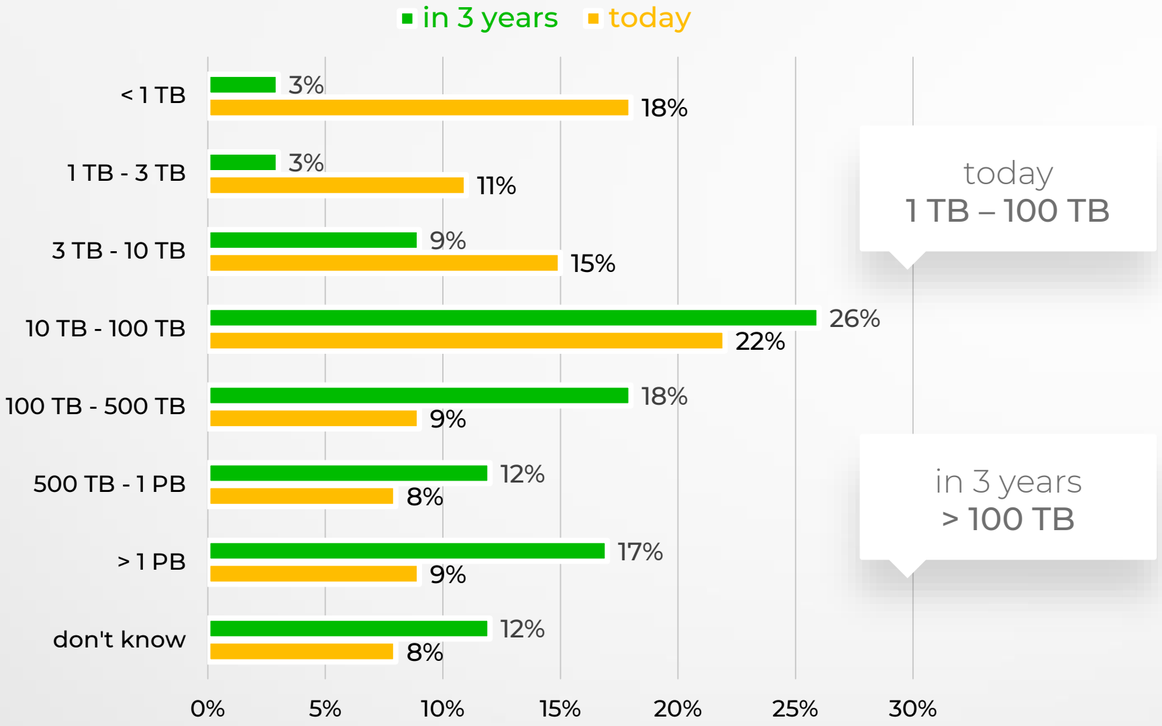


Where is your Data Lake deployed?

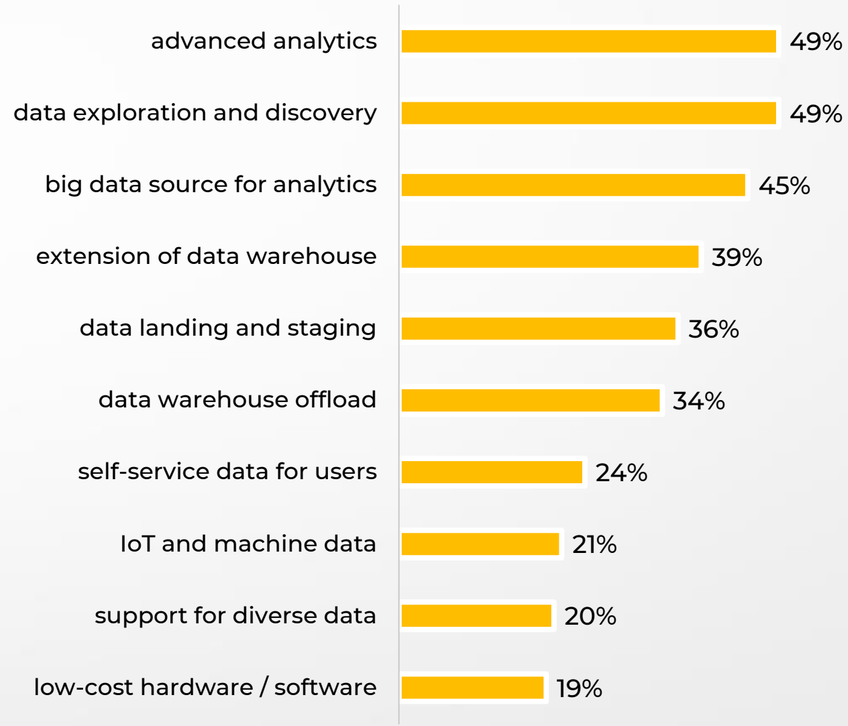
[Philip Russom: Data Lakes – Purposes, Practices, Patterns, and Platforms, TDWI Best Practice Report, Q1 2017]

Size and Use Cases

Big is beautiful. Analytics is key.



What is the approximate total volume of your Data Lake(s), today and in 3 years?

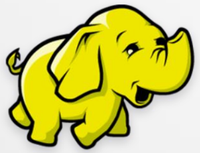


Top 10 use cases of an (Hadoop-based) Data Lake

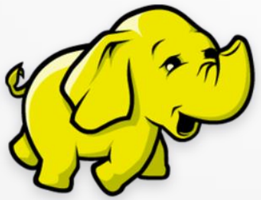
[Philip Russom: Data Lakes – Purposes, Practices, Patterns, and Platforms, TDWI Best Practice Report, Q1 2017]

What is the forecasted global market revenue of Hadoop in 2022?

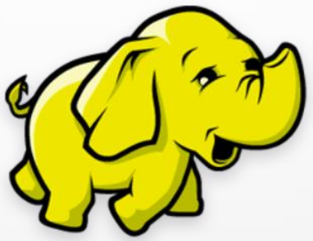
9 billion $



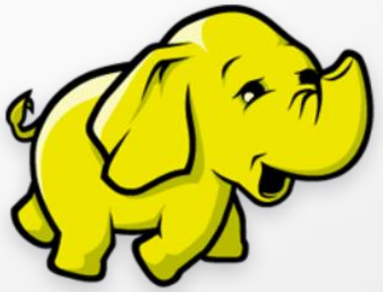
27 billion $



59 billion $

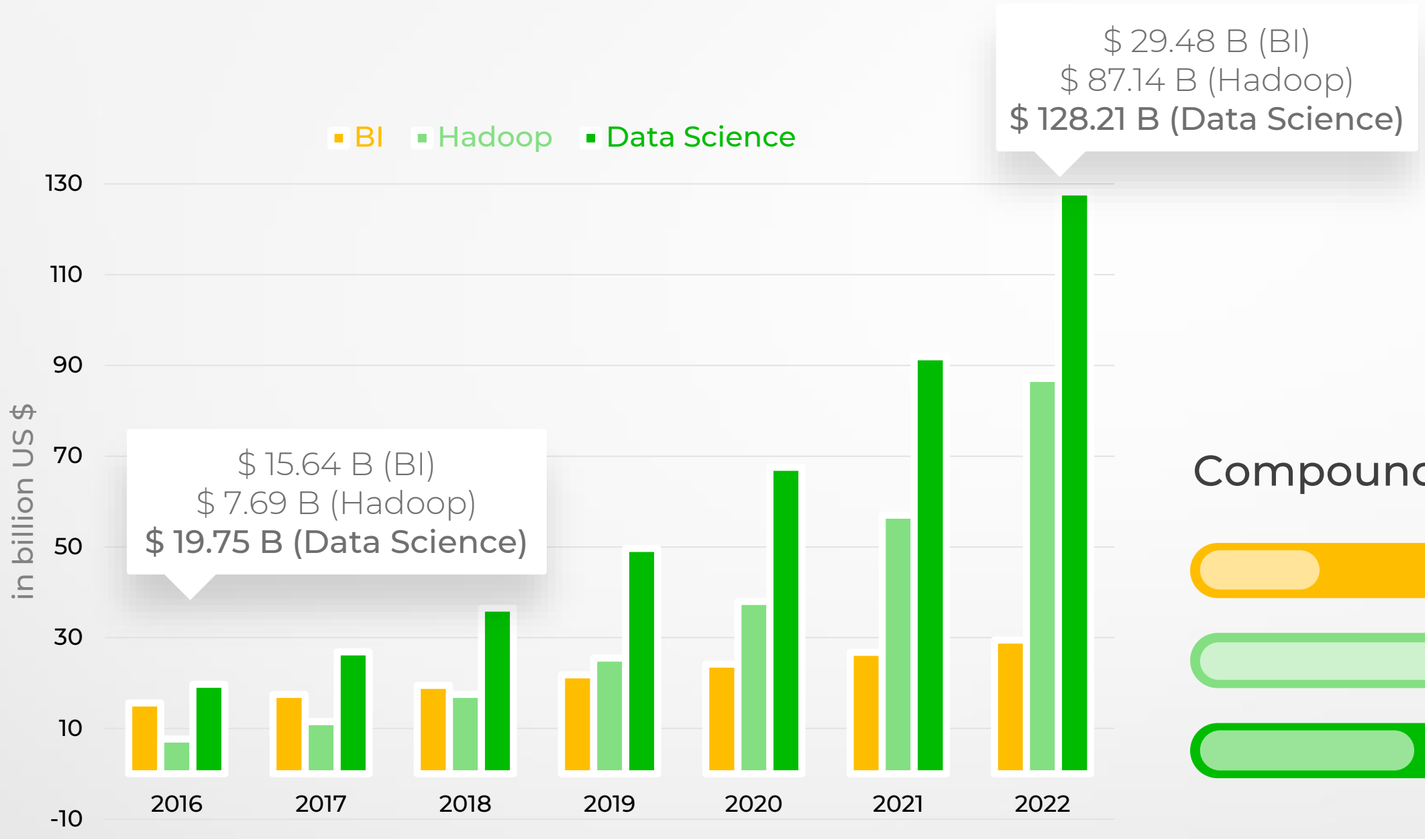


87 billion $



Global Market Revenues

It's all about money.



Compound Annual Growth Rate (CAGR)

[Business Intelligence (BI) – Global Market Outlook (2016-2022), Stratistics MRC, Q1 2017]
[Data Science Platform – Global Market Outlook (2016-2022), Stratistics MRC, Q1 2017]
[Hadoop Market by Type (Software, Hardware and Services), Trends and Forecast (2016 – 2022), Zion Market Research, Q1 2017]

You are Welcome!

Big Data Technology & Data Analytics

Dr. Alexander Schätzle

Big Data Architect

alexander.schaetzle@badenIT.de

Tel. +49 761 5035 4838

www.badenIT.de

www.exaprox.com

Business Analytics (BI & Process Mining)

Jan Birkholz

Sales Engineer

jan.birkholz@mehrwerk-ag.de

Tel. +49 174 164 01 68

www.mehrwerk-ag.de

www.mpm-processmining.com


badenIT