Karlsruher Institut für Technologie

Big Data Architectures and Technologies

Marcel Kunze

Research Group Cloud Computing - Steinbuch Centre for Computing

Karlsruhe Institute of Technology (KIT)

Largest European scientific Institution

Main Topics: Energy, Nanotechnology, Astrophysics, Engineering

Mission:

- Research

- Education

- Innovation

www.kit.edu

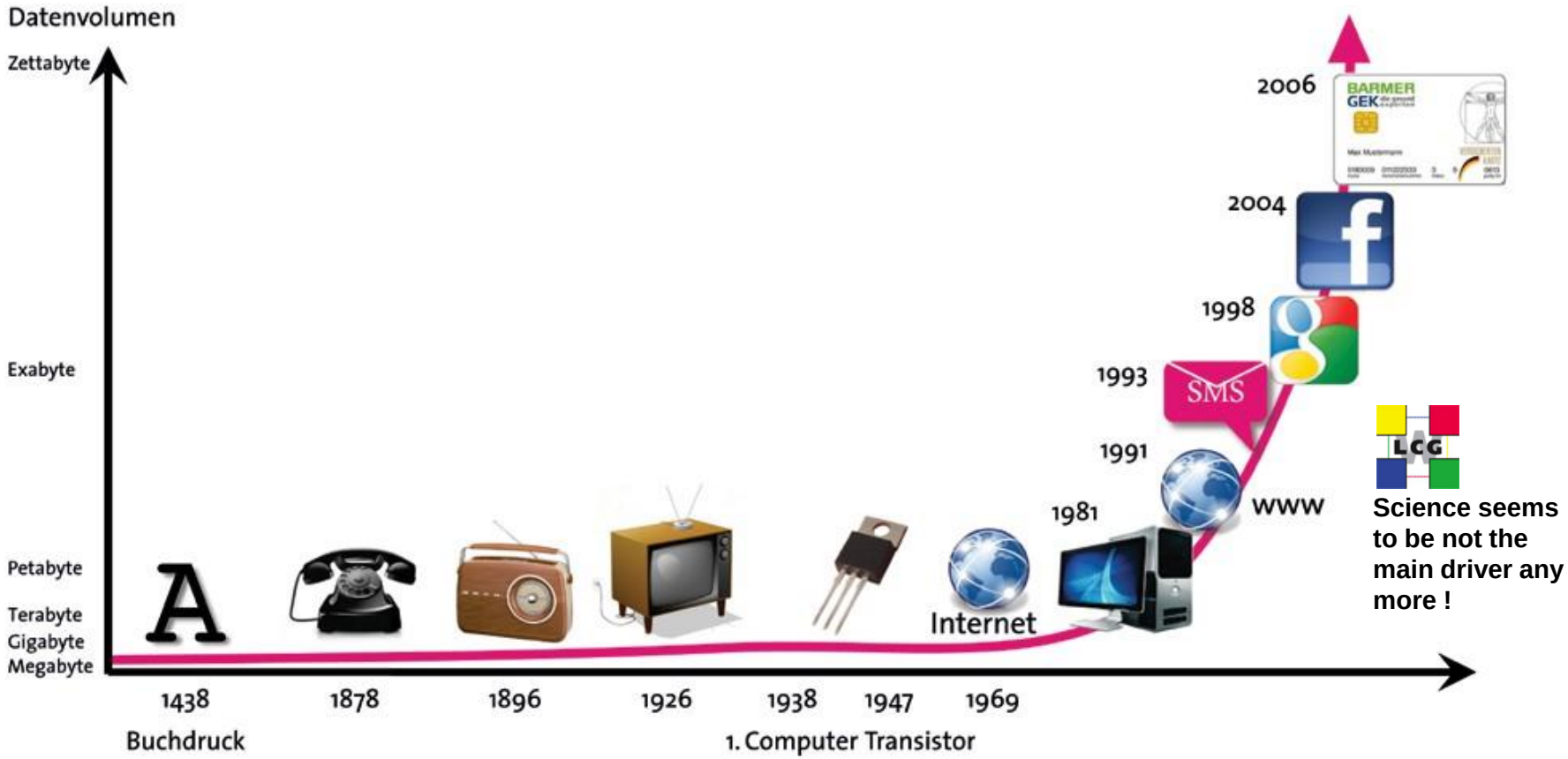
Agenda

- What is “Big Data” ?
- Big Data Management
- Big Data Toolbox
- R&D Projects

What is “Big Data” ?

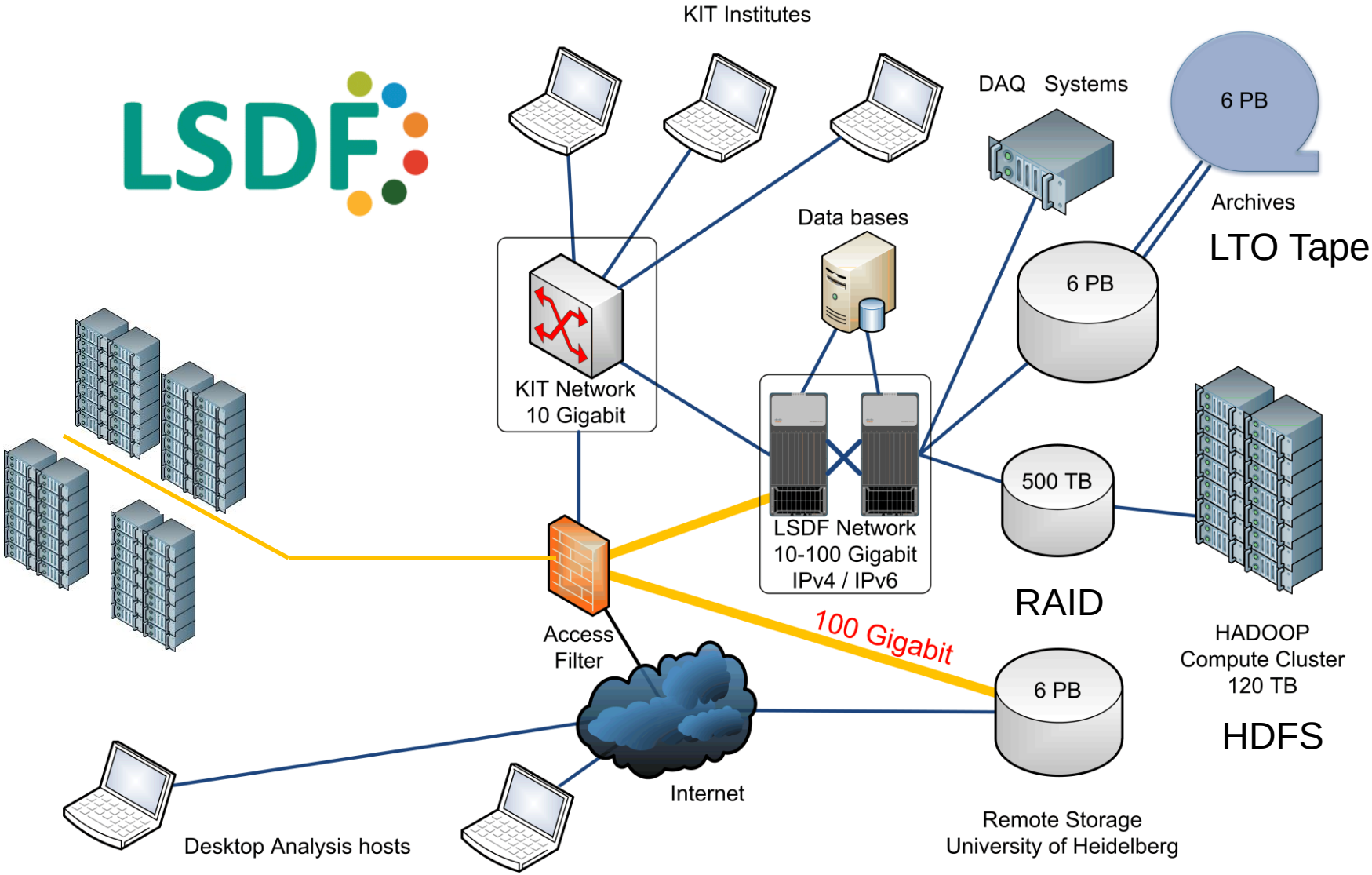
Big Data Volume

Million Peta Bytes = Zetta Bytes

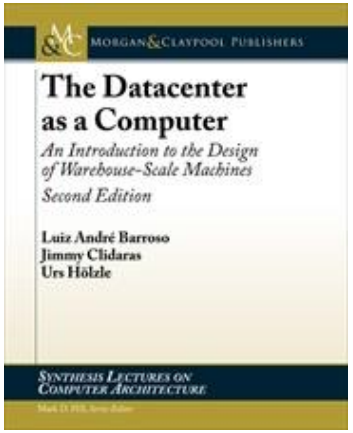


Source: „Big-Data im Praxiseinsatz, Leitfaden", BITKOM 2012

Large Scale Data Facility (LSDF)



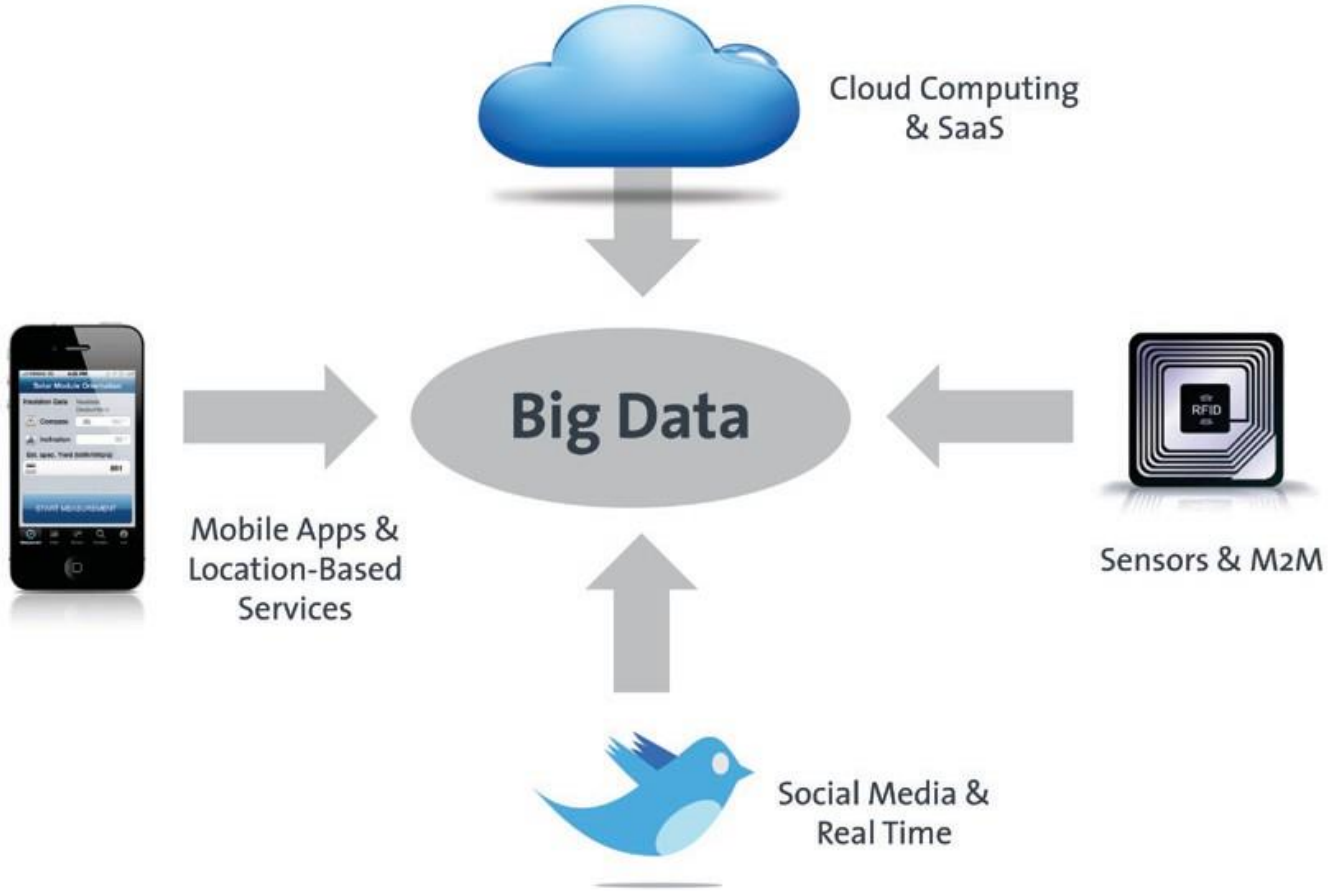
The Data Center as a Computer



- Apple data center in Maiden, NW Carolina: Exa-Scale data center (iCloud)

Comparison:
SCC@KIT,
LSDF, LHC Tier-1

Big Data Drivers



Source: „Big-Data im Praxiseinsatz, Leitfaden", BITKOM 2012

Big Data Dimensions (4V)

Karlsruher Institut für Technologie

Volume

- Number of Data Sets
- Number of Files
- Number of Objects
- Tera, Peta, Exa, Zetta, Yotta Bytes (10^{24} Bytes)
- Brontobyte (10^{27} Bytes)
- Geopbyte (10^{30} Bytes)

Variety

- Data Sources (Biz, Web)
- Structured / unstructured
- Poly-structured
- Text, Video, Pictures, Tweets, Blogs, Meters
- Machine Communication

Veracity

- Digital certificates
- Data quality
- Uncertainty quantification
- Reference data

Velocity

- Internet of Things
- Data Generation in (near) real-time
- Milliseconds, Seconds, Hours

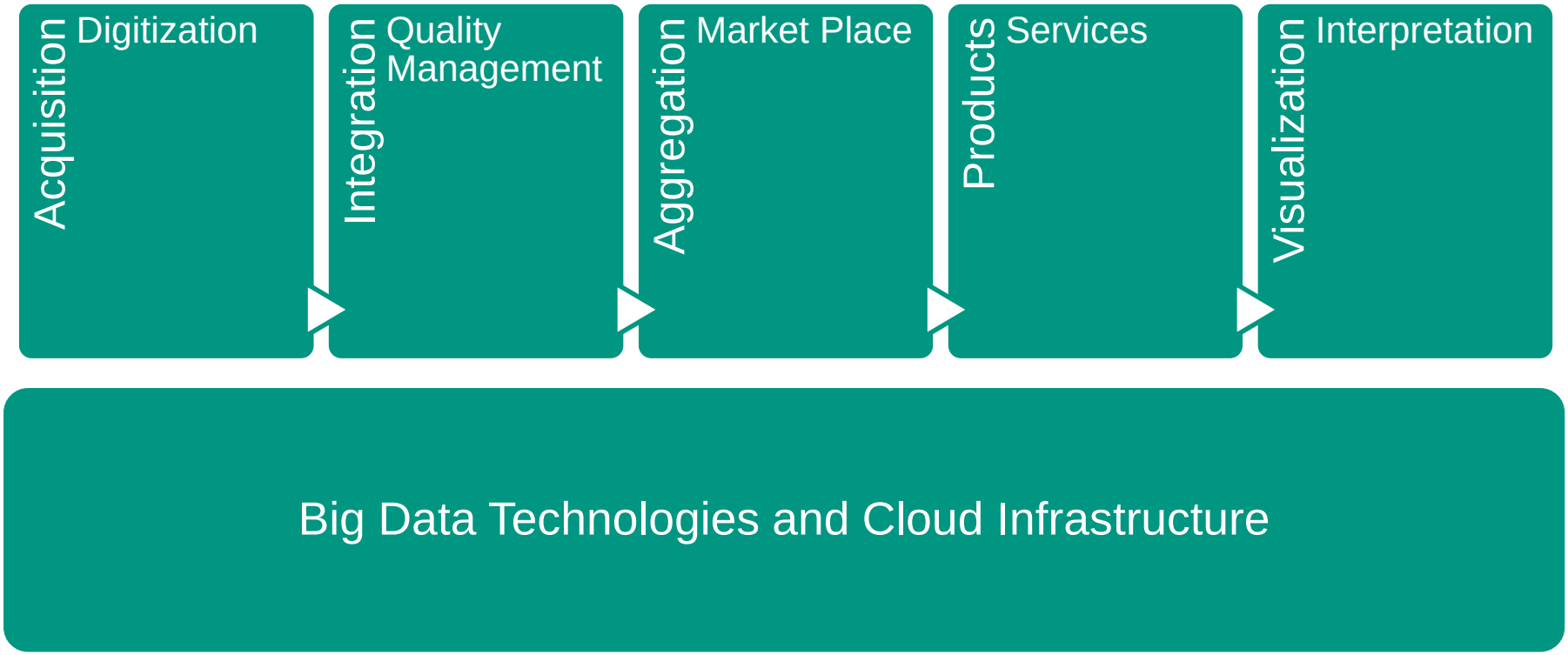
The 5th Dimension: Value

- A major new trend in information processing will be the trading of original and enriched data, effectively creating an information economy
  - Data mining
  - Descriptive analytics (Past)
  - Predictive analytics (Future)
  - Prescriptive analytics (Actionable insight)
    - Correlation of data
    - Intelligence of patterns, relations, etc.
    - …

„When hardware became commoditized, software was valuable. Now that software is being commoditized, data is valuable.“ (TIM O‘REILLY)

„The important question isn’t who owns the data. Ultimately, we all do. A better question is, who owns the means of analysis?“ (A. CROLL, MASHABLE, 2011)

Added Value in the Data Economy



- Example: Weather data - heterogeneous data/simulation sources – forecast/probability – map/app

Big Data Management

Data Management

- How can we manage billions of files/objects?
- How can we manage exabytes of data?
  - Data security
  - Data usage control
  - Data movement
  - Data availability
  - Data preservation
  - Data publication

Data Security

- There is no absolute security, neither in the cloud nor on your own premises. Spies are everywhere…

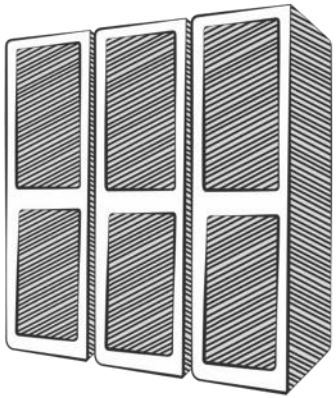
- Risk management:
  - Risk = probability of disaster * cost of disaster
  - Given a lower cost, a higher probability of data loss could be tolerable
  - A web server and clickstream analyses might be safely migrated into a public cloud

- Strong encryption helps to treat compliance and legal issues. Control of:
  - Keys
  - Algorithms
  - Data

“Encryption helps…”

Data Usage Control

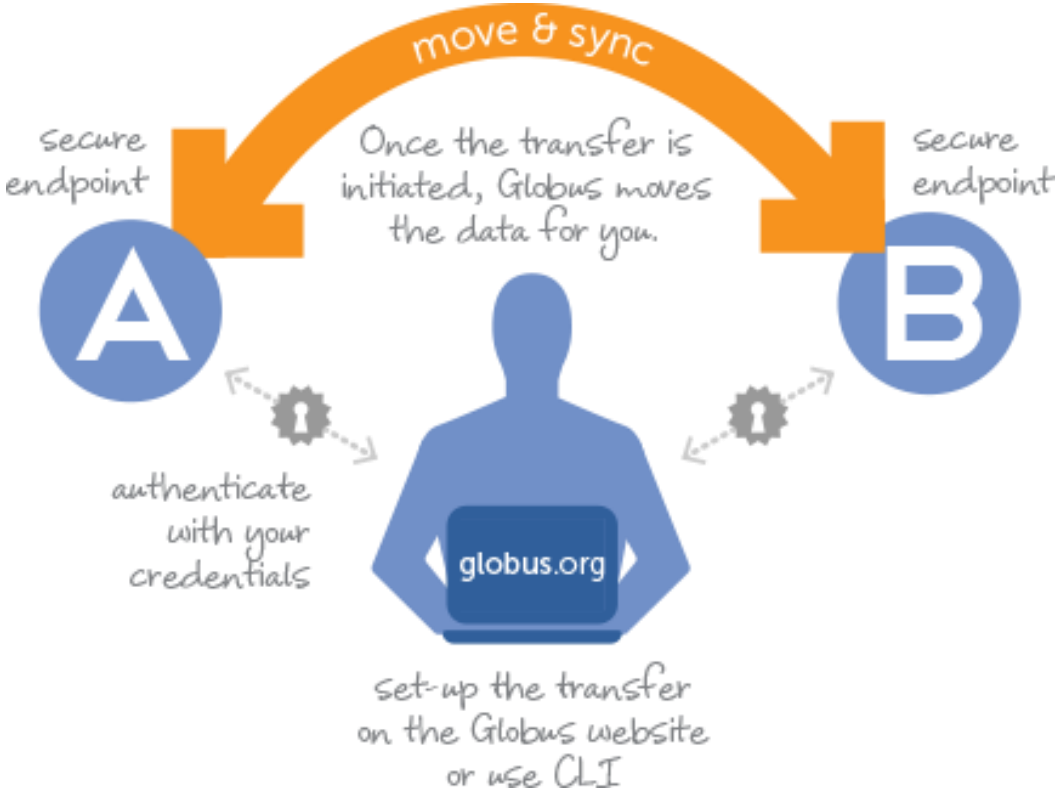
- Serious cloud providers are offering corresponding services that even comply with the highest legal standards
  - Regional concepts to define the data location
  - Trusted data stores with encryption
  - Identity and access management
  - Data usage control via transactional trails logging



Who had What access to Which resources When

Source: Amazon CloudTrail

Data Movement



Source: https://www.globus.org/

- Globus Online: Fully managed and automated Big Data file transfer
- SaaS: Separation of control plane and data plane (Third party transfer)

Data Publication

- What does it mean to publish?
- Data is:
  - Identified
  - Described
  - Curated
  - Verifiable
  - Accessible
  - Preserved

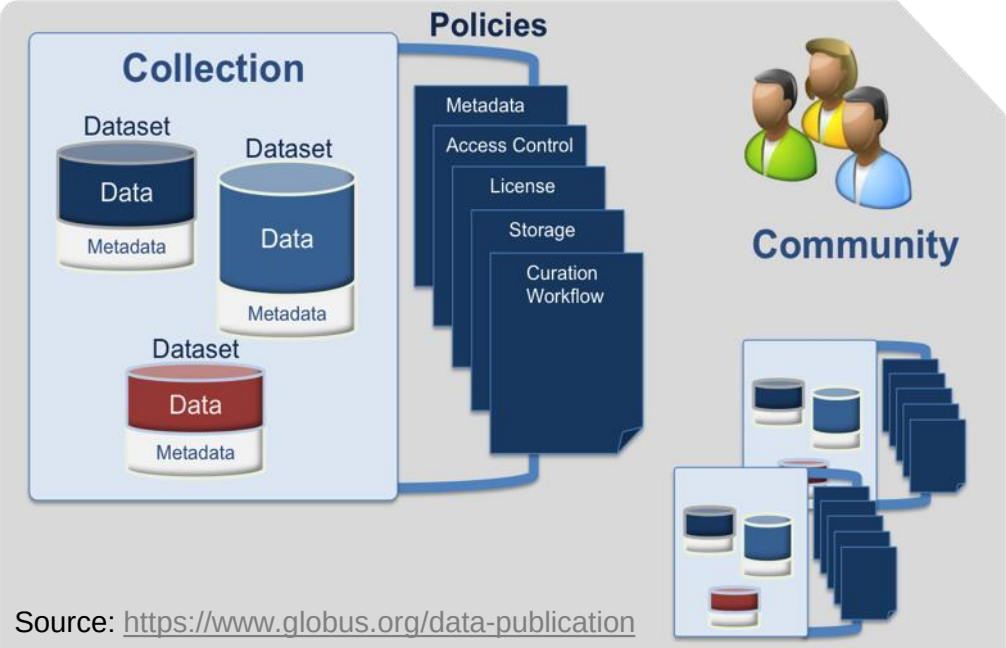
Data Discovery

- What does it mean to discover?
- Data can be:
  - Searched
  - Browsed
  - Accessed

Globus Data Publication Services

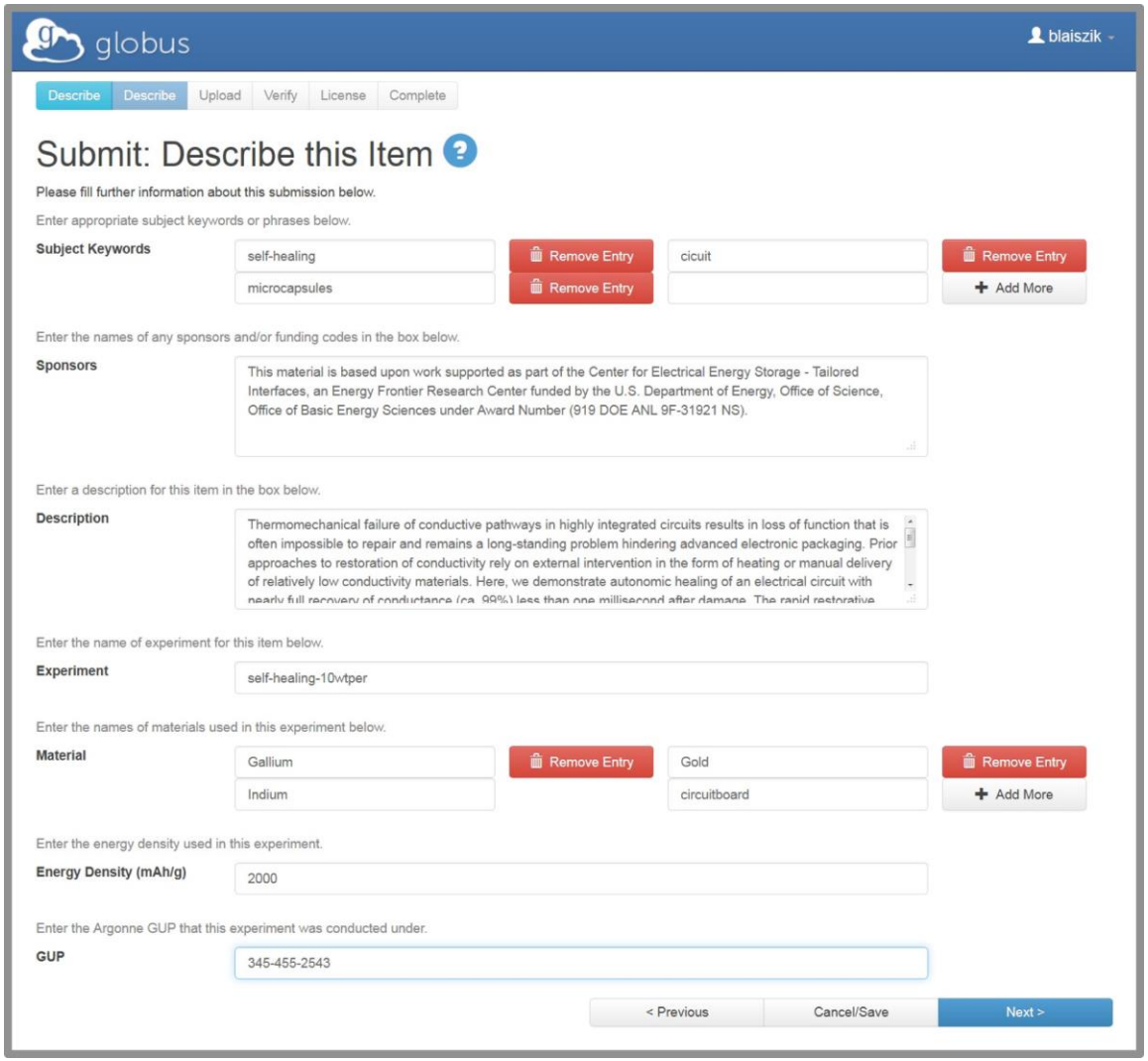
- SaaS for publishing Big Data
- Bring your own storage
- Extensible metadata
- Publication and curation workflows
- Public and restricted collections
- Rich discovery model

- Architecture:
  - Communities create collections of datasets
  - Describe the metadata
  - Define policies for data access
  - Define processes for publication



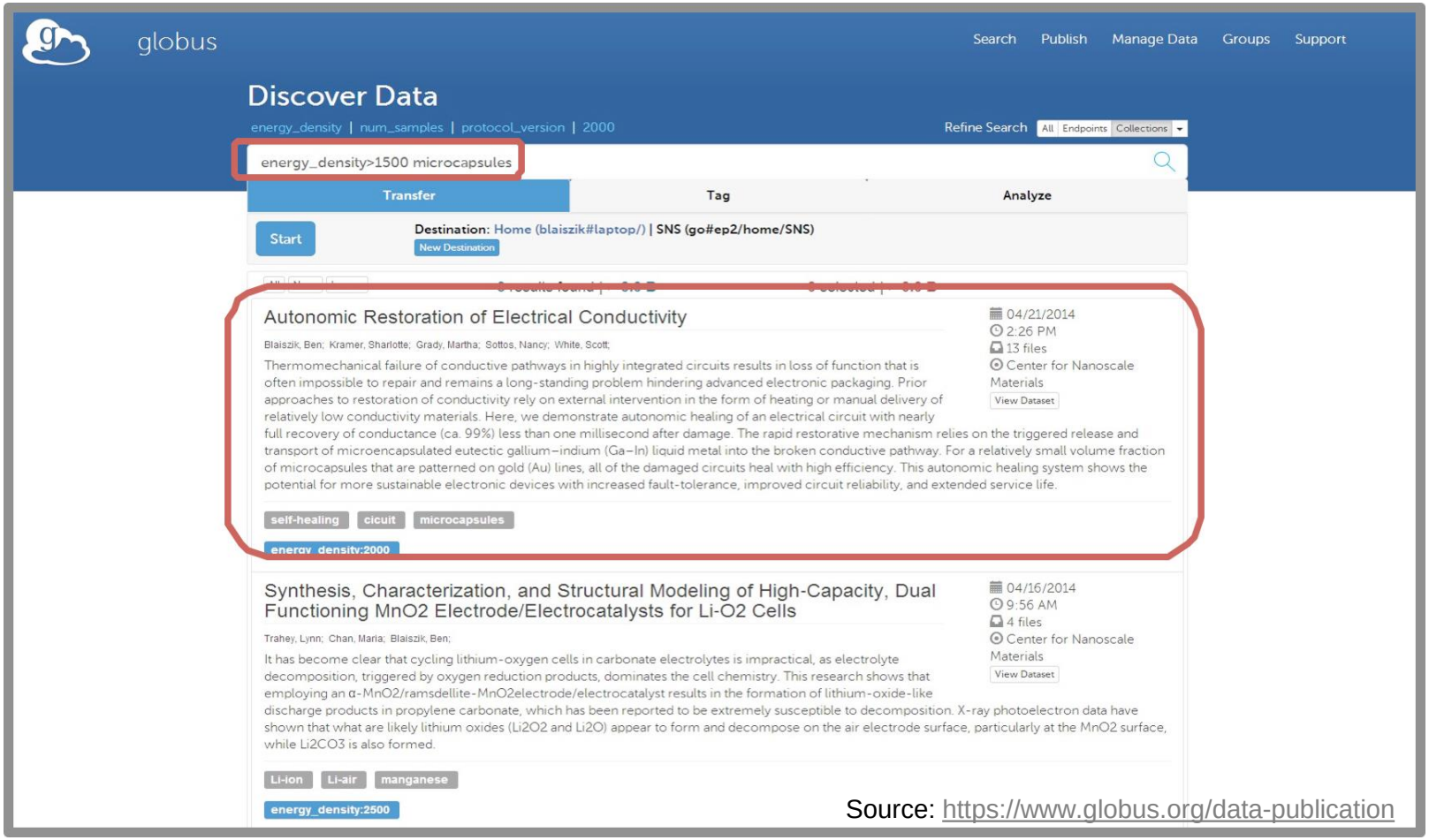
Source: https://www.globus.org/data-publication

Supply Domain specific Meta Data



Source: https://www.globus.org/data-publication

Search within and across Collections



Source: https://www.globus.org/data-publication

- Globus may be a pathfinder project to create open data markets

globus genomics

About Us Publications Technologies Sign Up

End-to-end sequencing analysis.
Flexible, scalable, simplified.

sequencing centers

elastic compute resources

powered by amazon web services

scalable storage repositories

Galaxy workflows

A solution as cutting-edge as your research.

Globus Genomics combines state-of-the-art algorithms, data management tools, a graphical workflow environment, and an elastic computing infrastructure. We take care of IT complexity so you can focus on your research.

Researchers

Easily leverage advanced tools to scale your work.

learn more

Core Labs

Provide customers with cost-effective services.

learn more

Scalability and flexibility for big genomics data.

Genome sequencing is notoriously data-intensive. We make it easy to manipulate, store, and share your data, no matter how big it gets.

Your workflows, simplified.

We help you easily connect and configure all the tools you need to automate your pipelines.

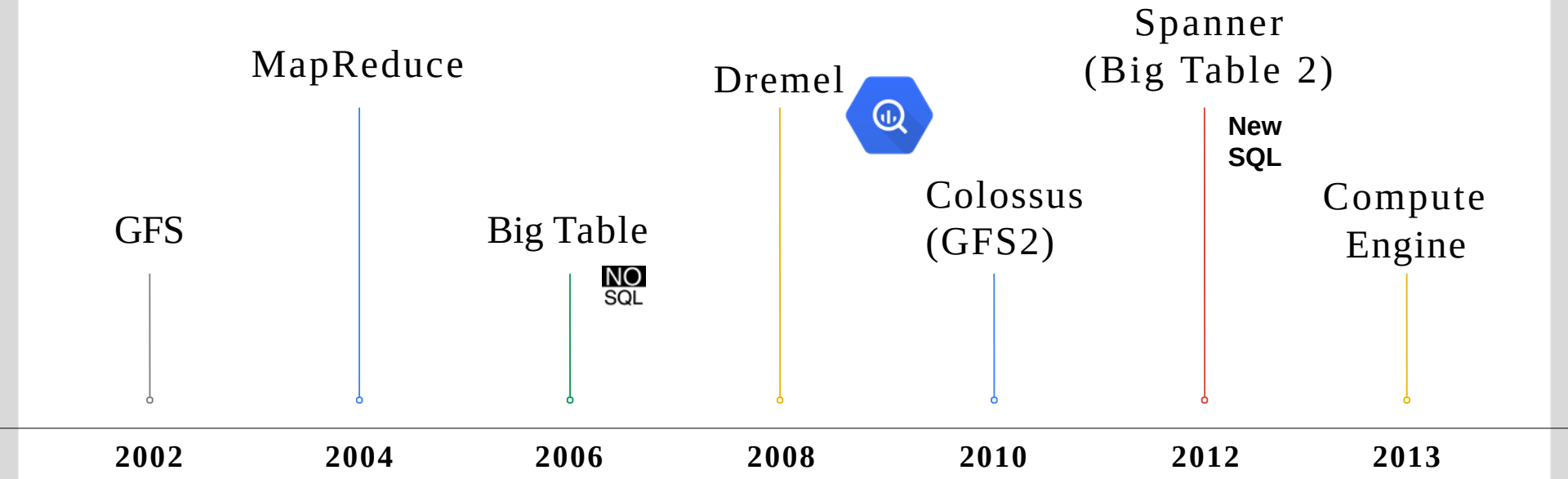
Let the cloud power your analysis.

Globus provides the data management solution using Amazon's compute cloud. The efficiencies are all yours.

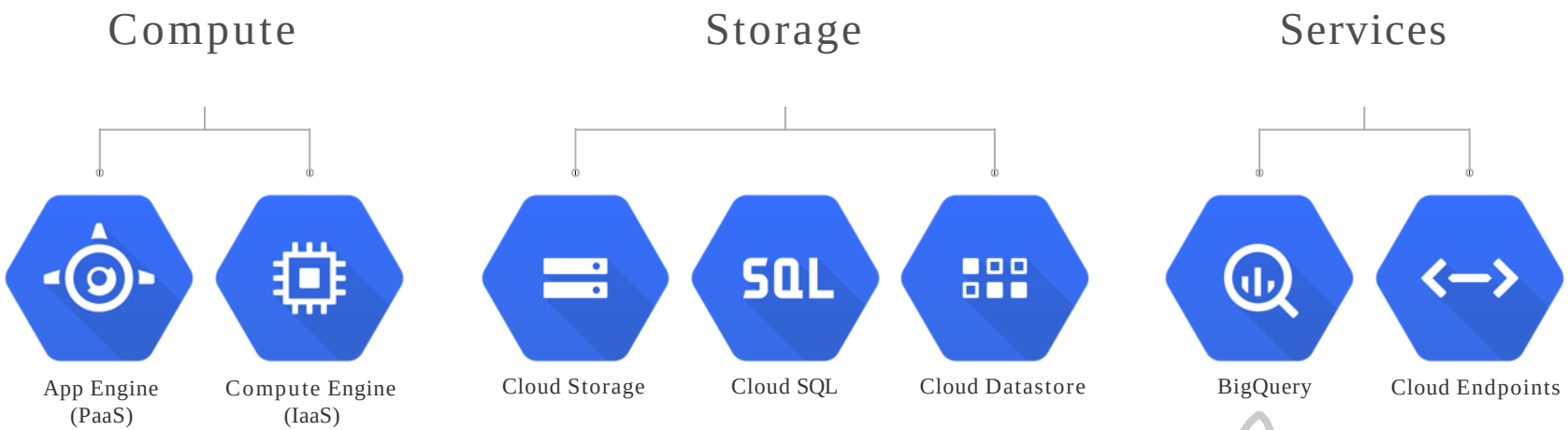
Source: https://www.globus.org/genomics/

Big Data Toolbox

Google Innovations over the last twelve Years



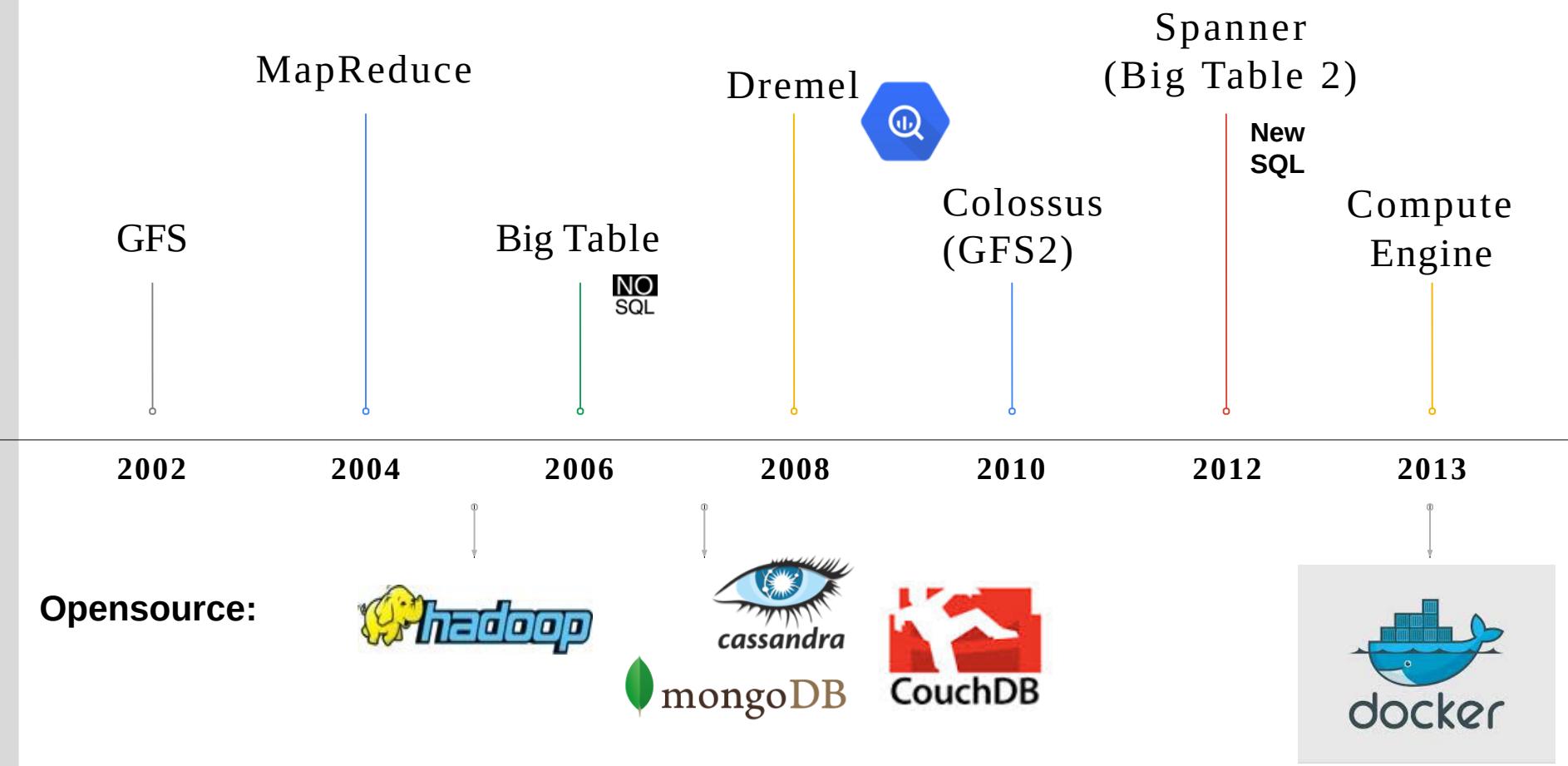
Google Cloud Platform



Big Data Analysis Tool

- Analyze terabytes of data in seconds
- Data imported in bulk or using streaming
- Supports CSV and JSON
- Browser tools, command line tool, or REST-API

Google Innovations over the last twelve Years

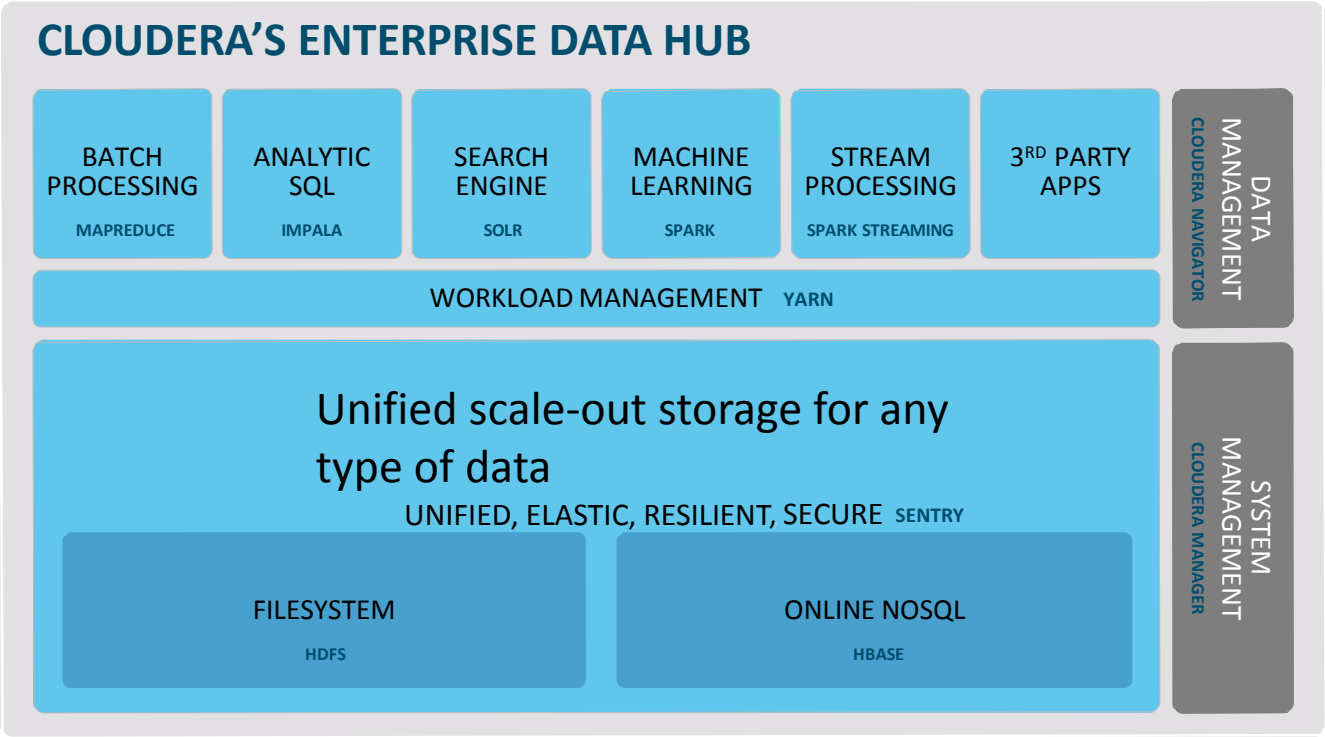


Hadoop

- Hadoop is a Big Data ecosystem that implements
  - Hadoop core utilities
  - HBase: A scalable, distributed database for large tables.
  - HDFS: A distributed file system.
  - Hive: A data warehouse, data summarization and ad hoc querying.
  - MapReduce: distributed processing on compute clusters.
  - Oozie: Workflow management.
  - Pig: A high-level data-flow language for parallel computation.
  - Spark: Ultra-fast in-memory computing.
  - ZooKeeper: coordination service for distributed applications.
  - And much more …

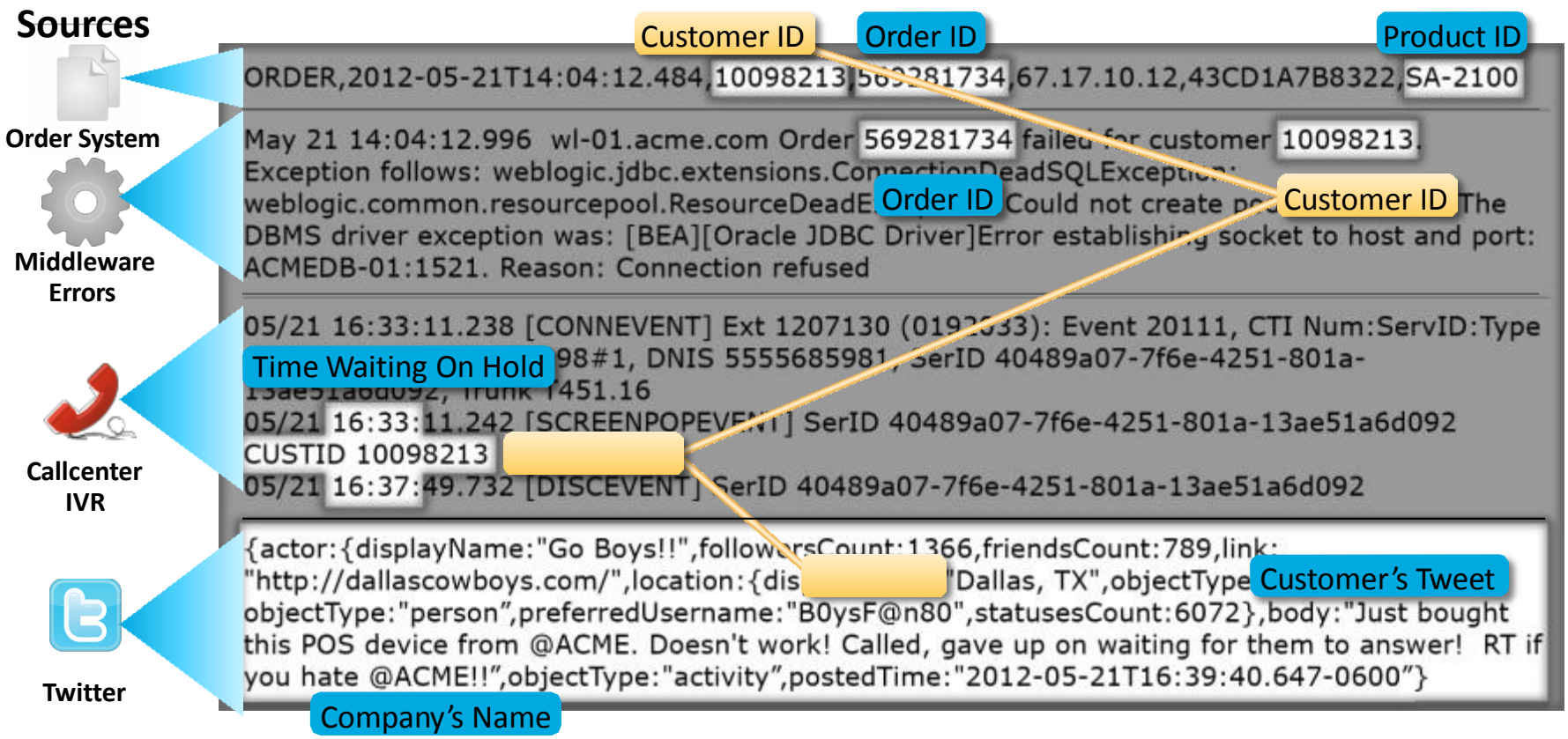
From Hadoop to the Enterprise Data Hub

Opensource Scalable Flexible Cost effective	✔
Managed	✔
Open Architecture	✔
Secure	✔



http://www.cloudera.com/content/cloudera/en/products-and-services/cloudera-enterprise.html

How do we conveniently treat Machine Data?



- Variety: Data are "poly-structured" and live in lots of formats

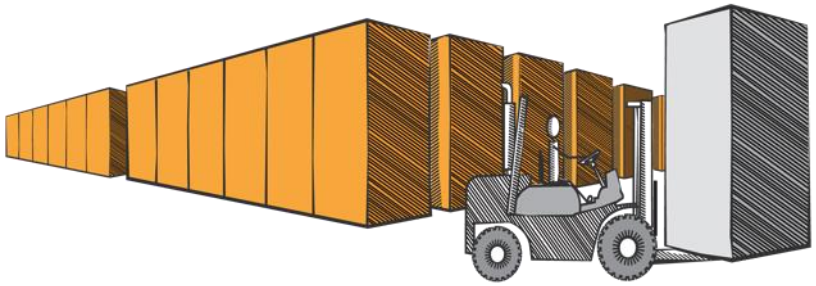
The Data Warehouse

Machine Data

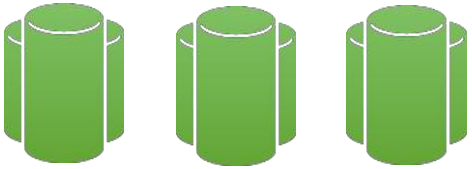
Online Services
Web Services
Servers
Security
GPS Location
Networks
Packaged Applications
Desktops
Storage
Custom Applications
Messaging
RFID
Telecoms
Online Shopping Cart
Energy Meters
Databases
Web Clickstreams
Call Detail Records
Smartphones and Devices

Operational Intelligence

Search and Analytics | Proactive Monitoring | Operational Visibility | Real Time Business

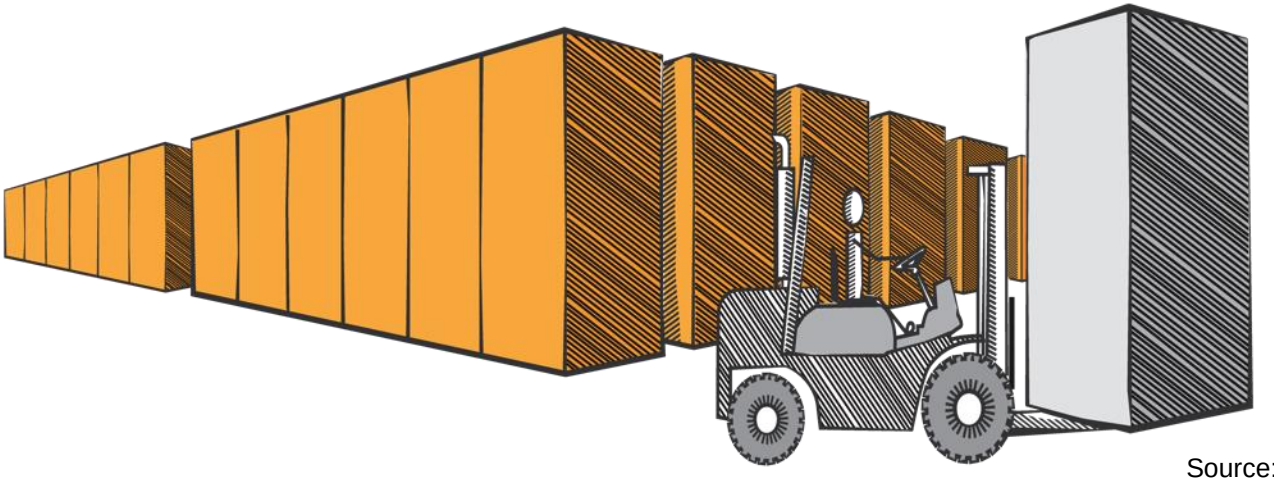


HA Indexes and Storage



Commodity Servers

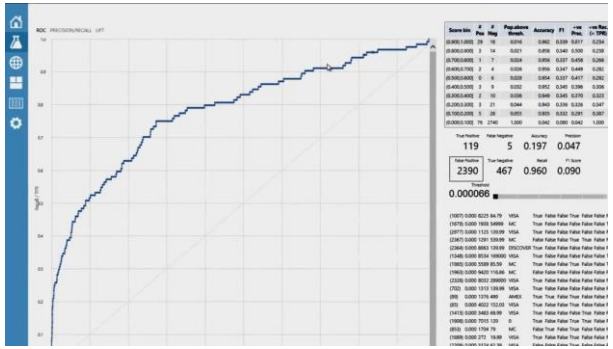
Big Data Analytics in the Cloud: Amazon Redshift

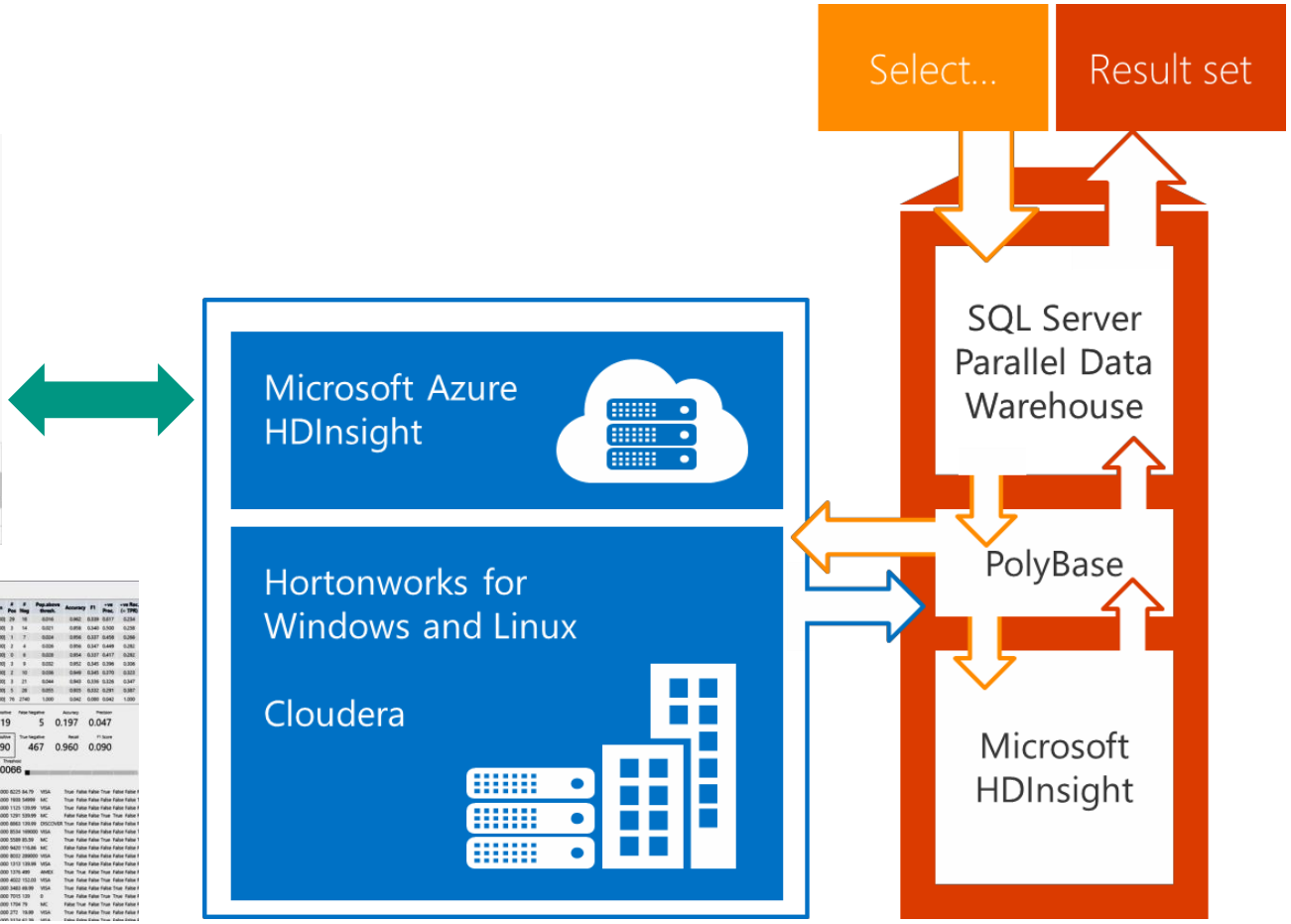


Source: Amazon Redshift and DynamoDB

- Fully managed, massively parallel relational data warehouse
- Takes care of cluster management and distribution of data
- Optimized for complex queries across many large tables
- Use standard SQL & standard BI tools
- Can be combined with Hadoop on-demand (Elastic MapReduce)

Microsoft Analytics Platform System





- Bringing together Hadoop with the data warehouse (Windows Azure)
- New: Microsoft machine learning service (Azure ML)

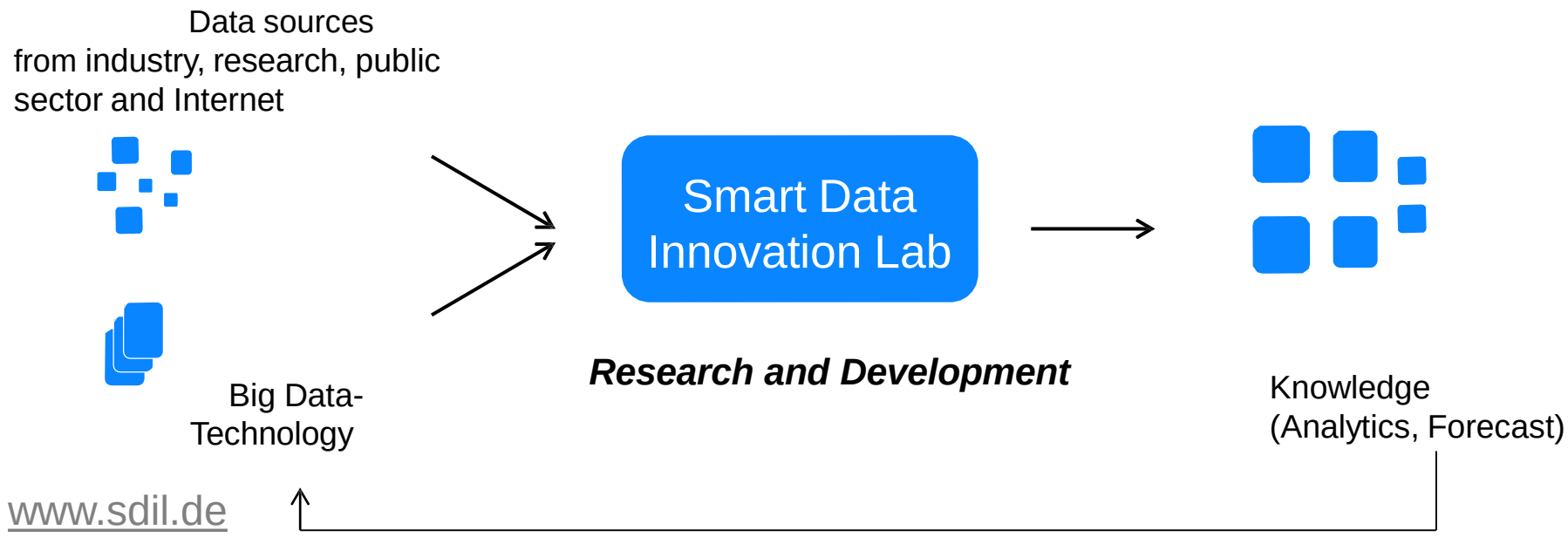
Research and Development

Ingredients of a Big Data Project

- Technology
  - Data preparation
  - Scalable processing
  - Scalable platform

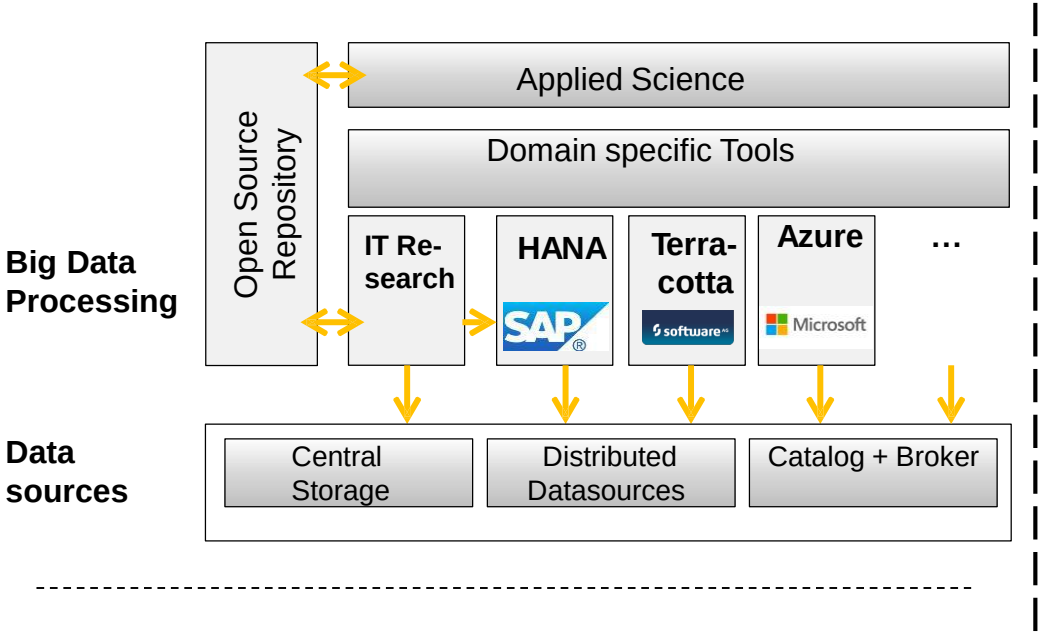
  } **Cloud Computing**
- Mathematical analysis methods
  - Machine learning
  - Statistics
  - Optimization
  - …
- Toolset
  - Natural Language Processing
  - Image processing
  - Visualization
- Application
  - Real-world analysis problem

Smart Data Innovation Lab (SDIL)



- Cooperation between industry and science to spur innovation
- Pilot R&D projects on dedicated Big Data infrastructure

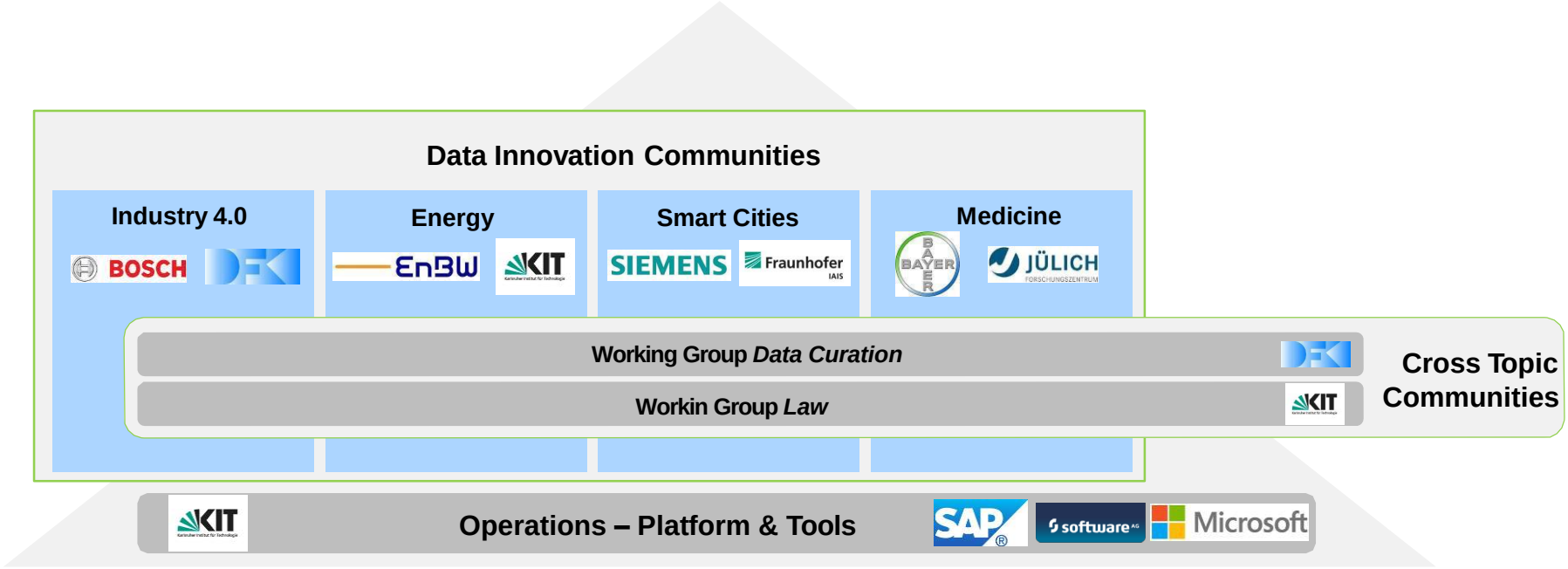
Architecture



IT and applied research on an open system

Data sources from industry, research, public sector and Internet

Application Area

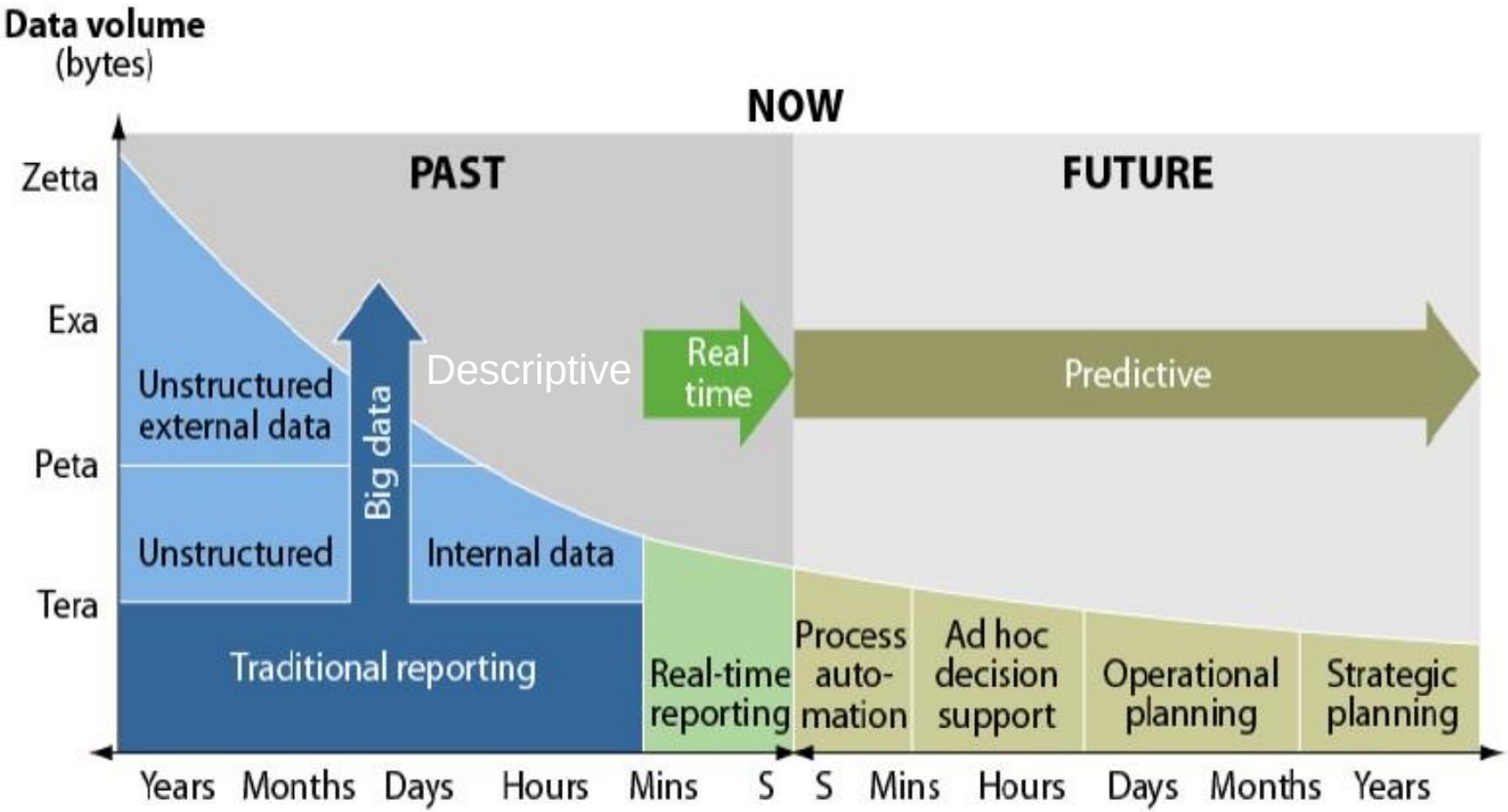


Partners

Research and Development Areas

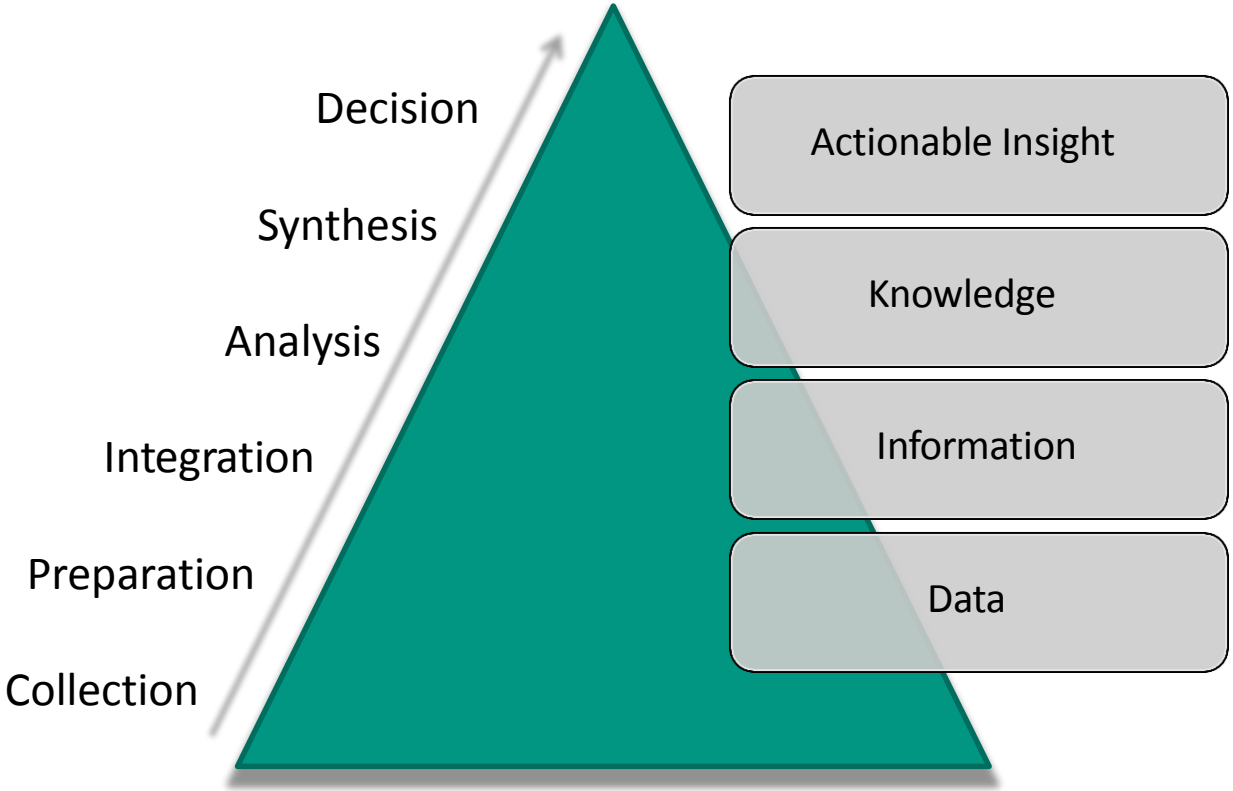
Applications	Methods	Storage	Processing	Representation
• Industry 4.0 • Logistics • Smart Grids • Smart City • Personalized Medicine	• Data mining • Machine Learning • Statistics Analysis • Predictive Analytics • Tools	• Data Warehouses • NoSQL Databases • Column Stores • In memory DBs	• Hadoop Engines • Real time Analytics • Software Defined Data Center	• Dashboards • Visualization • Rich Clients • Collaboration Platforms

Predictive Analytics



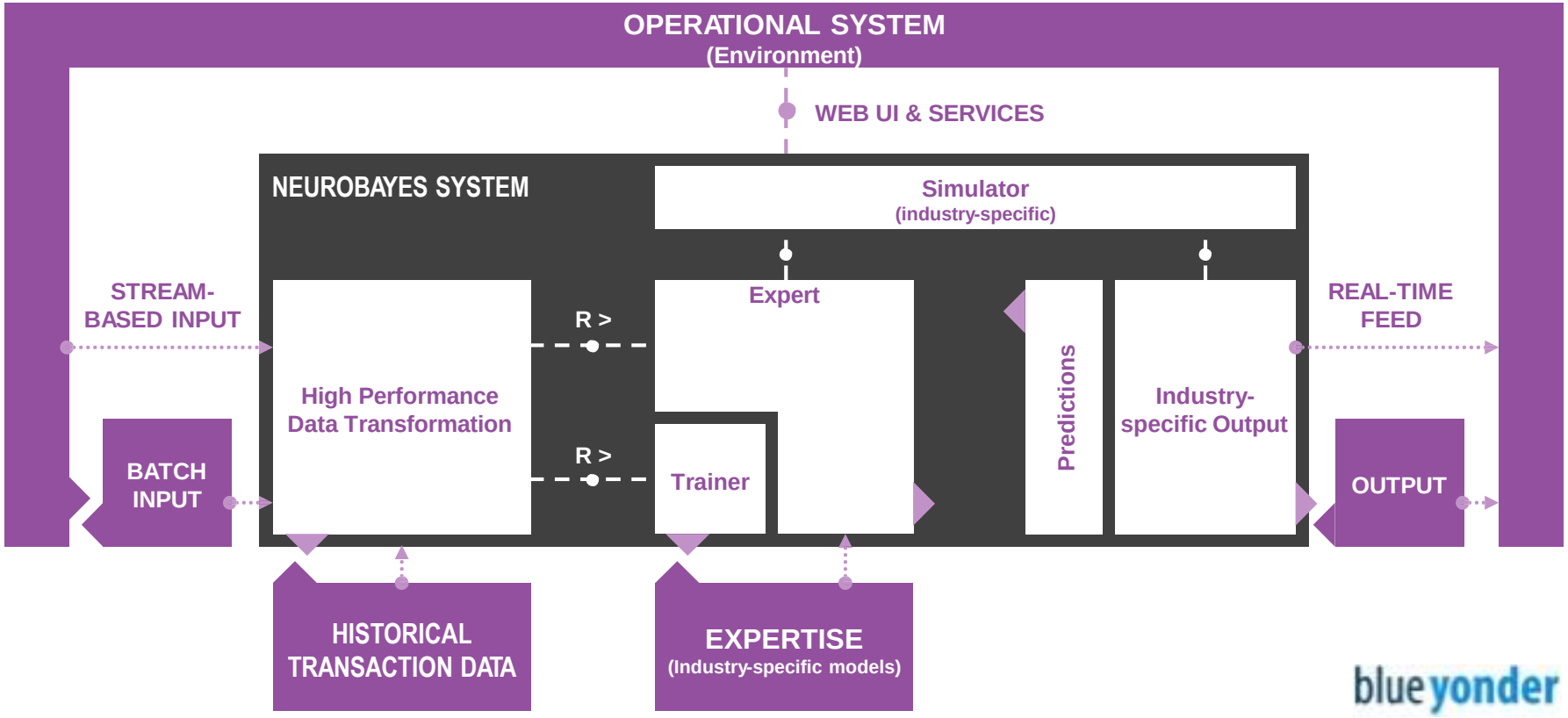
Source: blue yonder

Prescriptive Analytics



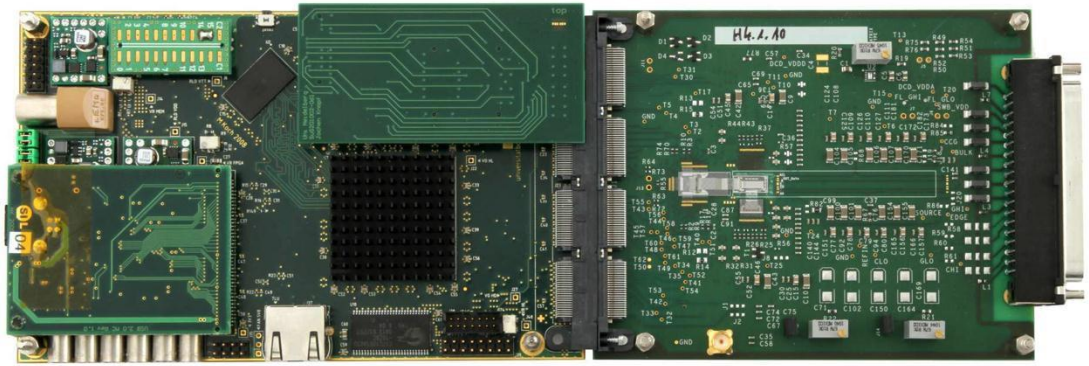
- From raw data to decision processes:
  - Data analysis takes time and often works with offline data only
  - Is it possible to improve the process and react in real time ?

Blue Yonder forward demand Architecture



- Machine learning utilizing modern in-memory database technology
- Direct integration into business processes (not just simple data-mining)

Future: Algorithm in Hardware



- NeuroBayes machine learning algorithm on FPGA
- Field Programmable Gate Array: (XILINX Virtex6 VLX75T)
- Clock frequency: 250 MHz
- Approx. 1 decision per clock cycle (fully pipelined architecture)
- 250 million decisions per second
- Throughput: 100 Gbit/s
- Interesting for real-time investigation of online streaming data

Summary

- Big Data depends on **scalable dynamic models** (Cloud is essential)
- Big Data is **interdisciplinary**: Computer Science, Mathematics, …
- **Hadoop** may assume the role of the **data hub**
- Big data is more about **value** than about volume

Karlsruher Institut für Technologie

Contact:

marcel.kunze@kit.edu

Dr. Marcel Kunze
Karlsruhe Institute of Technology (KIT)
Forschungsgruppe Cloud Computing
Steinbuch Centre for Computing
Hermann-von-Helmholtz-Platz 1
D-76344 Eggenstein-Leopoldshafen