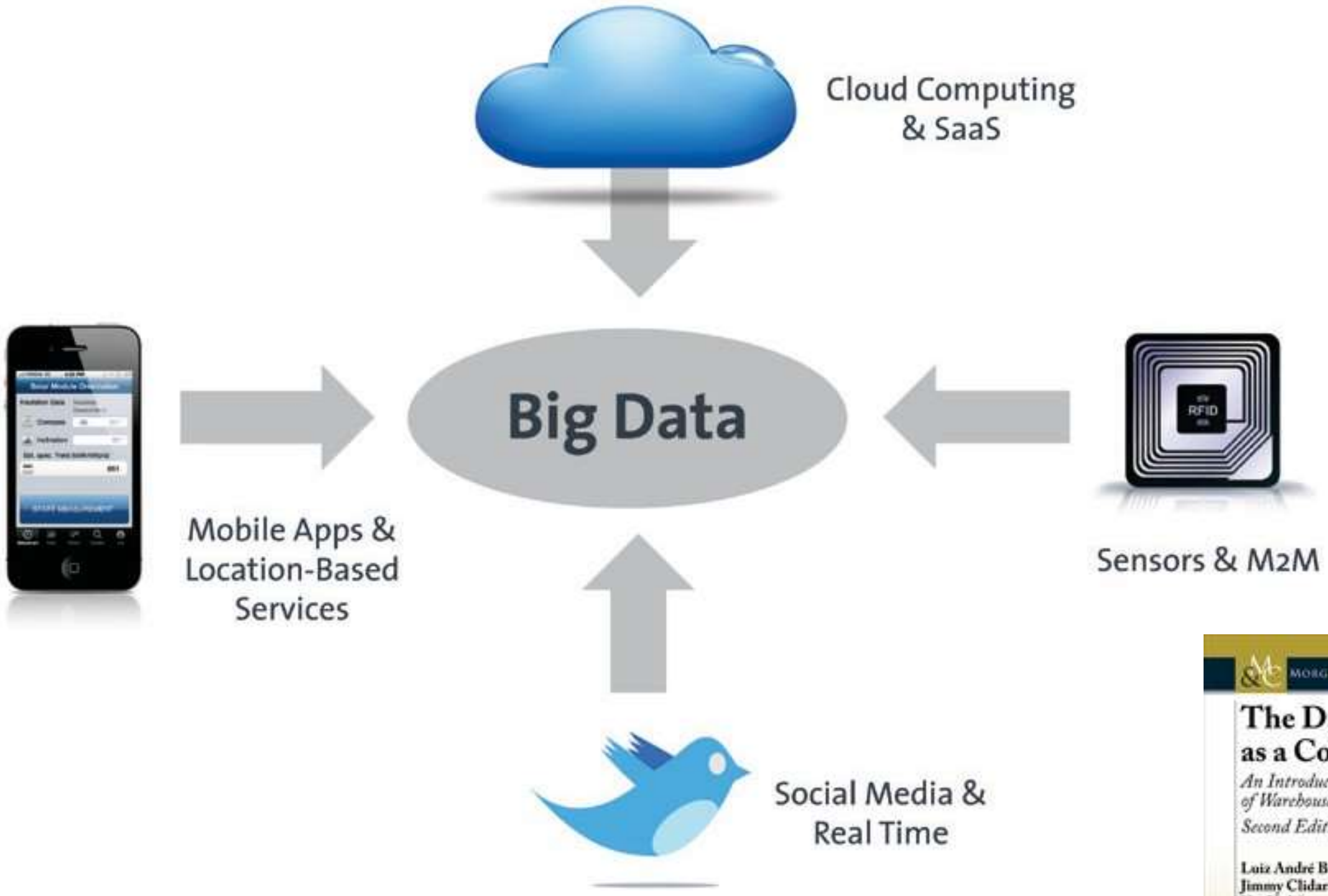
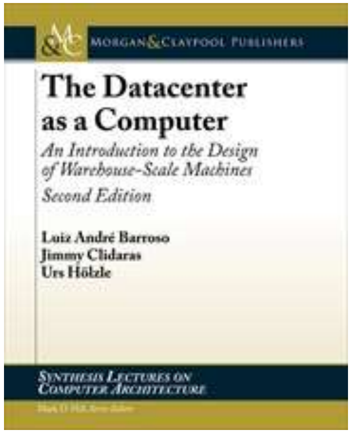
Big DataTechnologies

Marcel Kunze

Research Group Cloud Computing - Steinbuch Centre for Computing

Big Data Drivers and the Industrialization of IT




Morgan & Claypool Publishers

The Datacenter as a Computer

An Introduction to the Design of Warehouse-Scale Machines

Second Edition

Luiz André Barroso
Jimmy Clidaras
Urs Hölzle

Synthesis Lectures on Computer Architecture

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

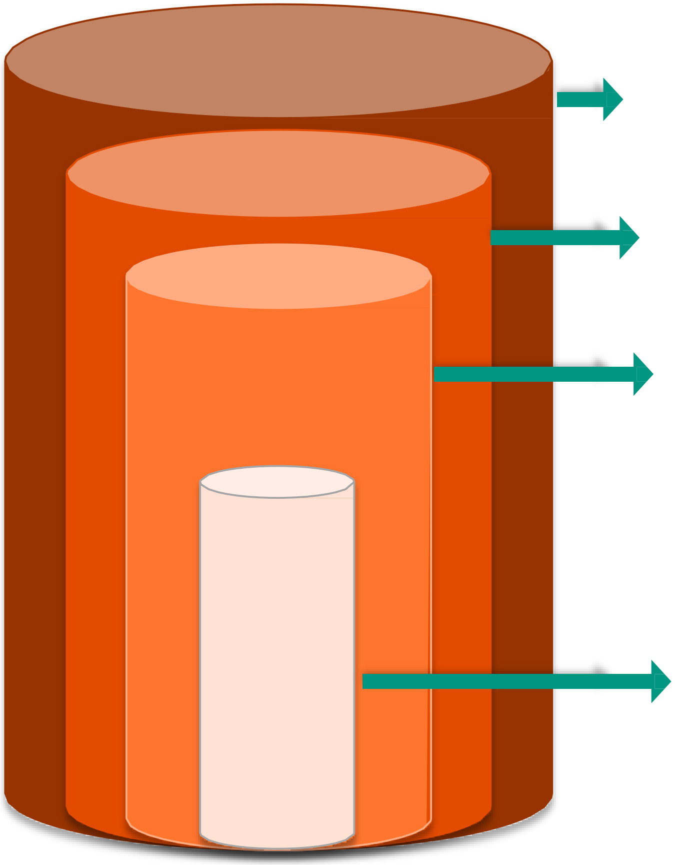
The Data Center as a Computer

PROJECT DOLPHIN
SITE PLAN · PHASE 2

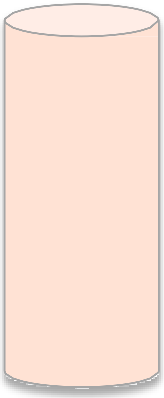
For Comparison:
SCC@KIT,
LSDF, LHC Tier-1

- Apple data center in Maiden, NW Carolina: Hosting ExaBytes for 450+ million iCloud users

Industrial Cloud Storage Prices vs. Inhouse



Cloud Storage:
Only actual
usage is charged

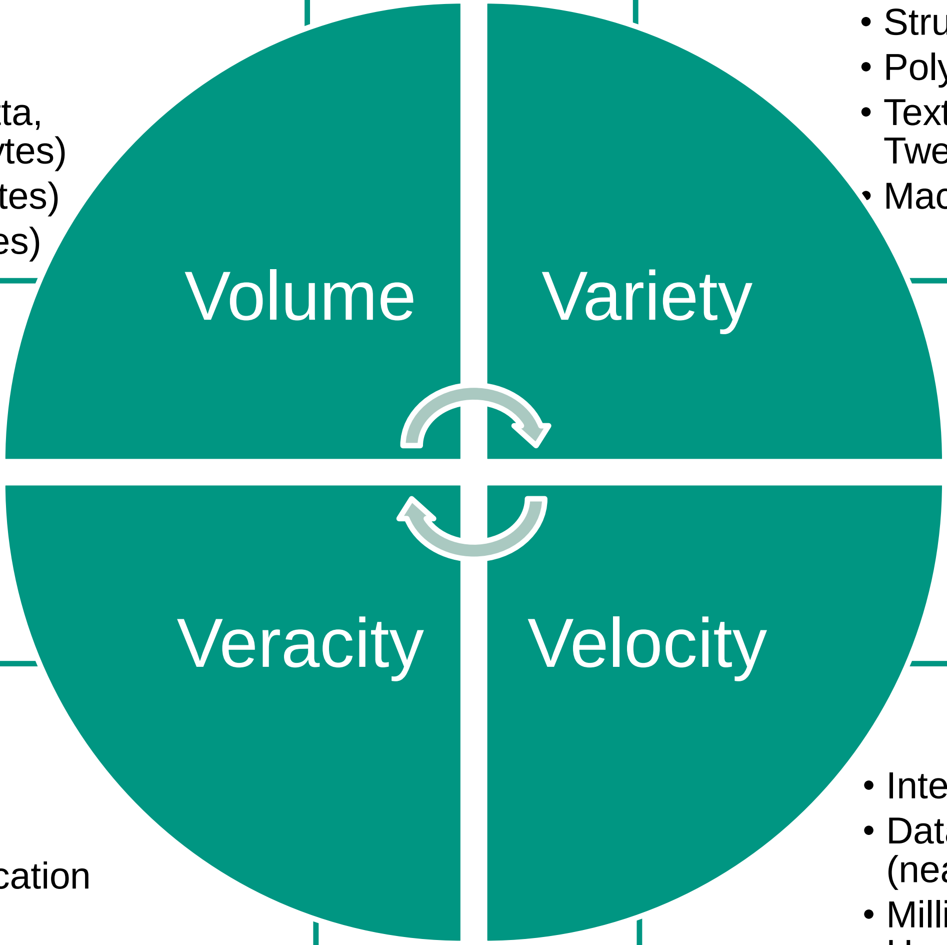


- TCO comparison has to take into account the cost of on-premise RAW capacity
- Enterprise data plans:
  - Microsoft Onedrive: **$2.99** per month and user for **1 TB**
  - Google Drive: **$10** per month and user **unlimited storage**

Big Data Dimensions (4V)

- 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)

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



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

- 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)*

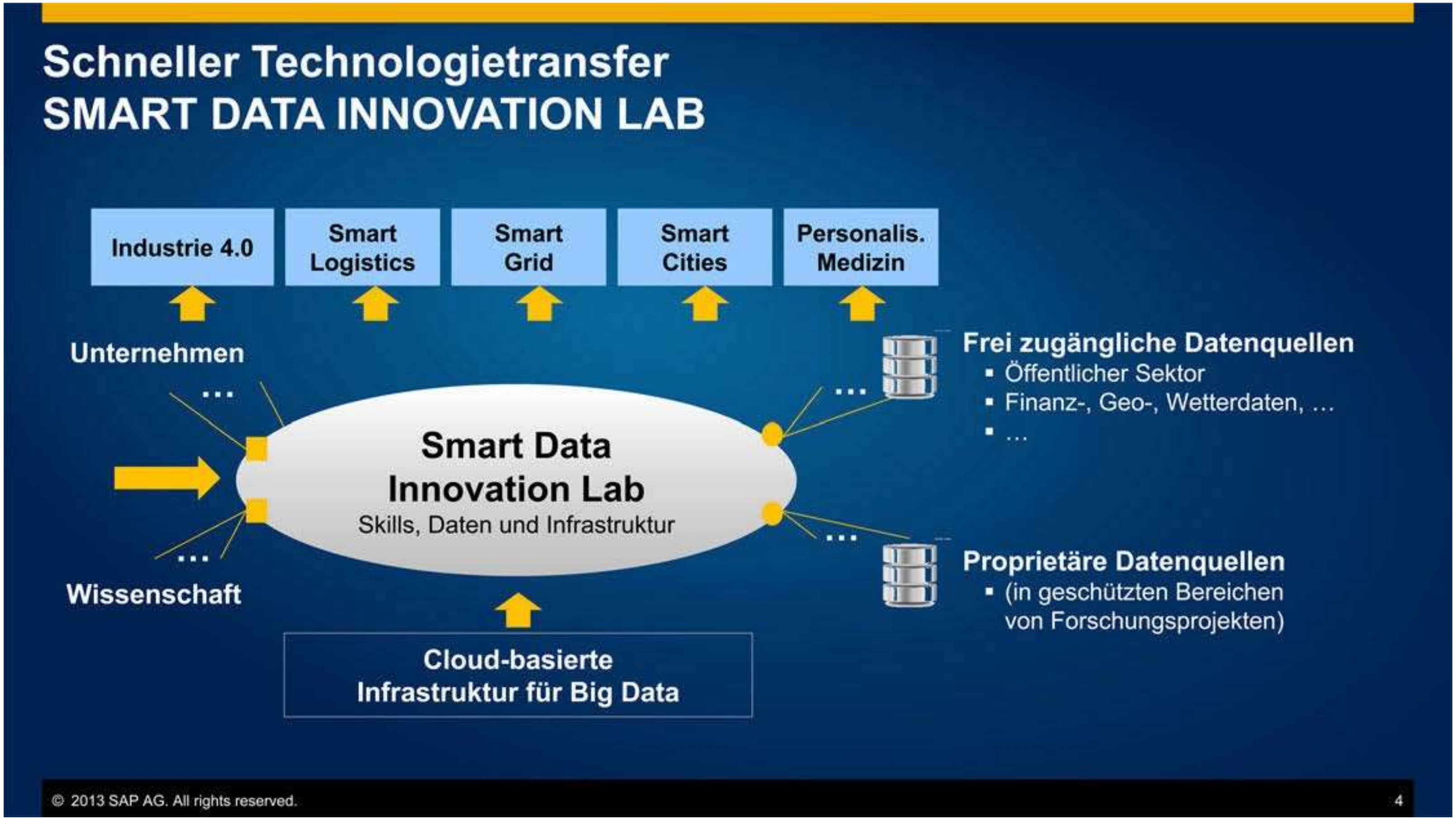
Ingredients of a successful Big Data Project

- **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)



Schneller Technologietransfer
SMART DATA INNOVATION LAB

Industrie 4.0

Smart Logistics

Smart Grid

Smart Cities

Personalis. Medizin

Unternehmen

…

Wissenschaft

Smart Data Innovation Lab
Skills, Daten und Infrastruktur

Cloud-basierte Infrastruktur für Big Data

Frei zugängliche Datenquellen
- Öffentlicher Sektor
- Finanz-, Geo-, Wetterdaten, …
- …

Proprietäre Datenquellen
- (in geschützten Bereichen von Forschungsprojekten)

4

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

Source: SDIL

Research and Development Areas

Applications

- Industry 4.0
- Logistics
- Smart Grids
- Smart City
- Personalized Medicine

Methods

- Data mining
- Machine Learning
- Statistics Analysis
- Predictive Analytics
- Tools

Storage

- Data Warehouses
- NoSQL Databases
- Column Stores
- In memory DBs

Processing

- Hadoop Engines
- Real time Analytics
- Software Defined Data Center

Representation

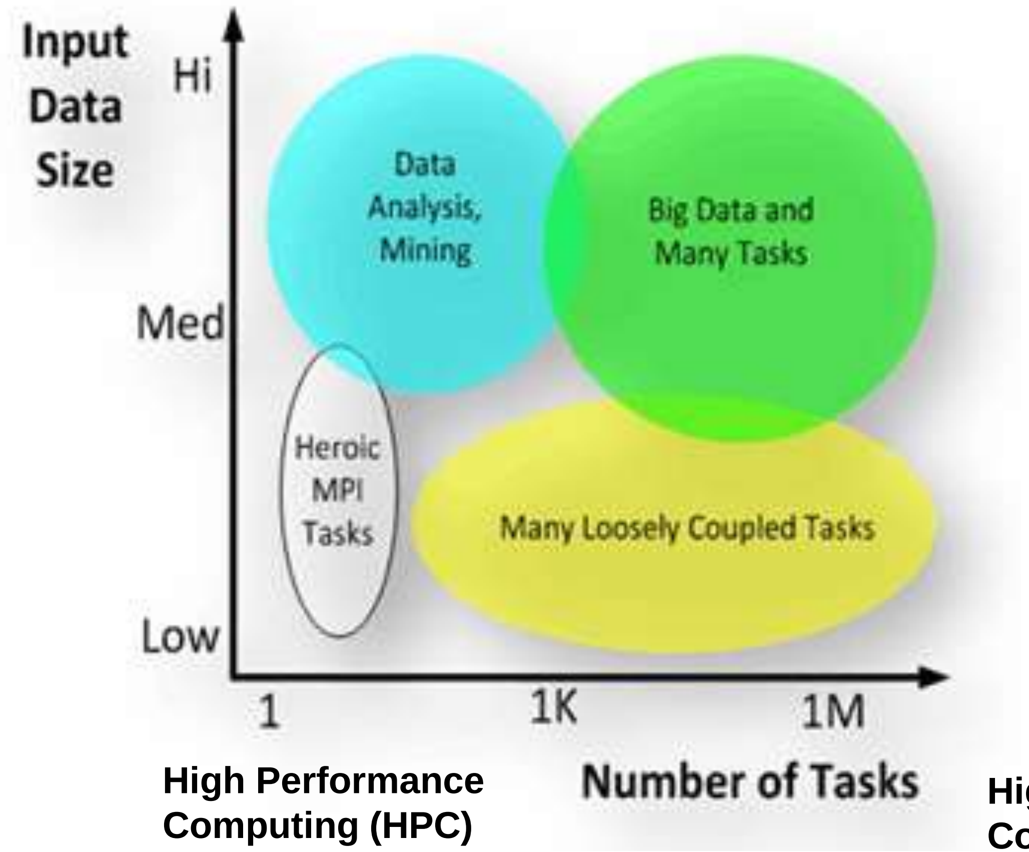
- Dashboards
- Visualization
- Rich Clients
- Collaboration Platforms

1. Hadoop

- Hadoop is a Big Data ecosystem that implements
  - Hadoop core utilities
  - Avro: A data serialization system with scripting languages.
  - Chukwa: Managing large distributed systems.
  - HBase: A scalable, distributed database for large tables.
  - HDFS: A distributed file system.
  - Hive: Data summarization and ad hoc querying.
  - Mahout: Machine learning
  - MapReduce: Distributed processing on compute clusters.
  - Pig: A high-level data-flow language for parallel computation.
  - ZooKeeper: Coordination service for distributed applications
  - And much more …

Classification of Computing Applications

Many Task
Computing (MTC)



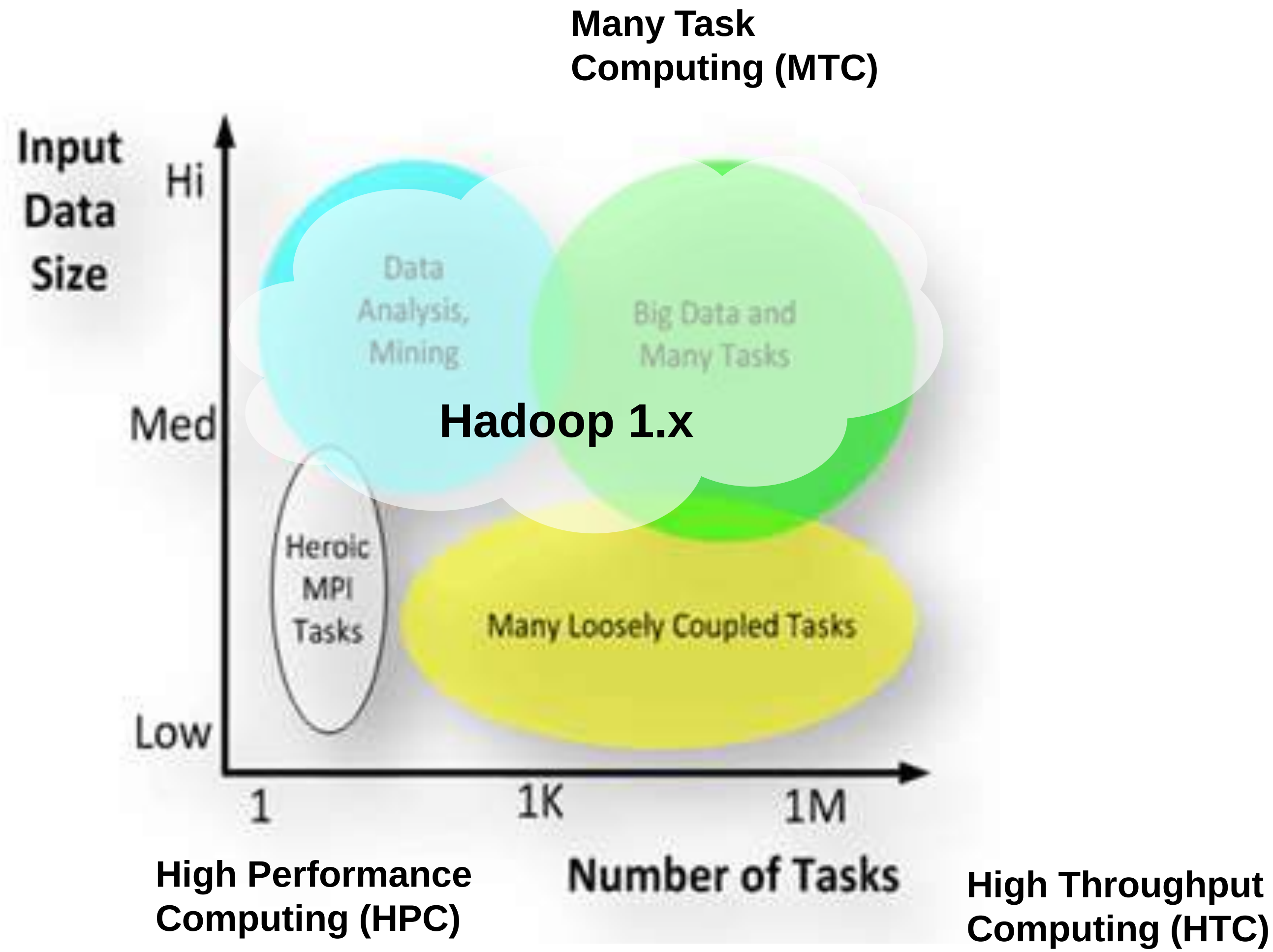
High Performance
Computing (HPC)

High Throughput
Computing (HTC)

- HPC ≠ HTC ≠ MTC, each domain is different

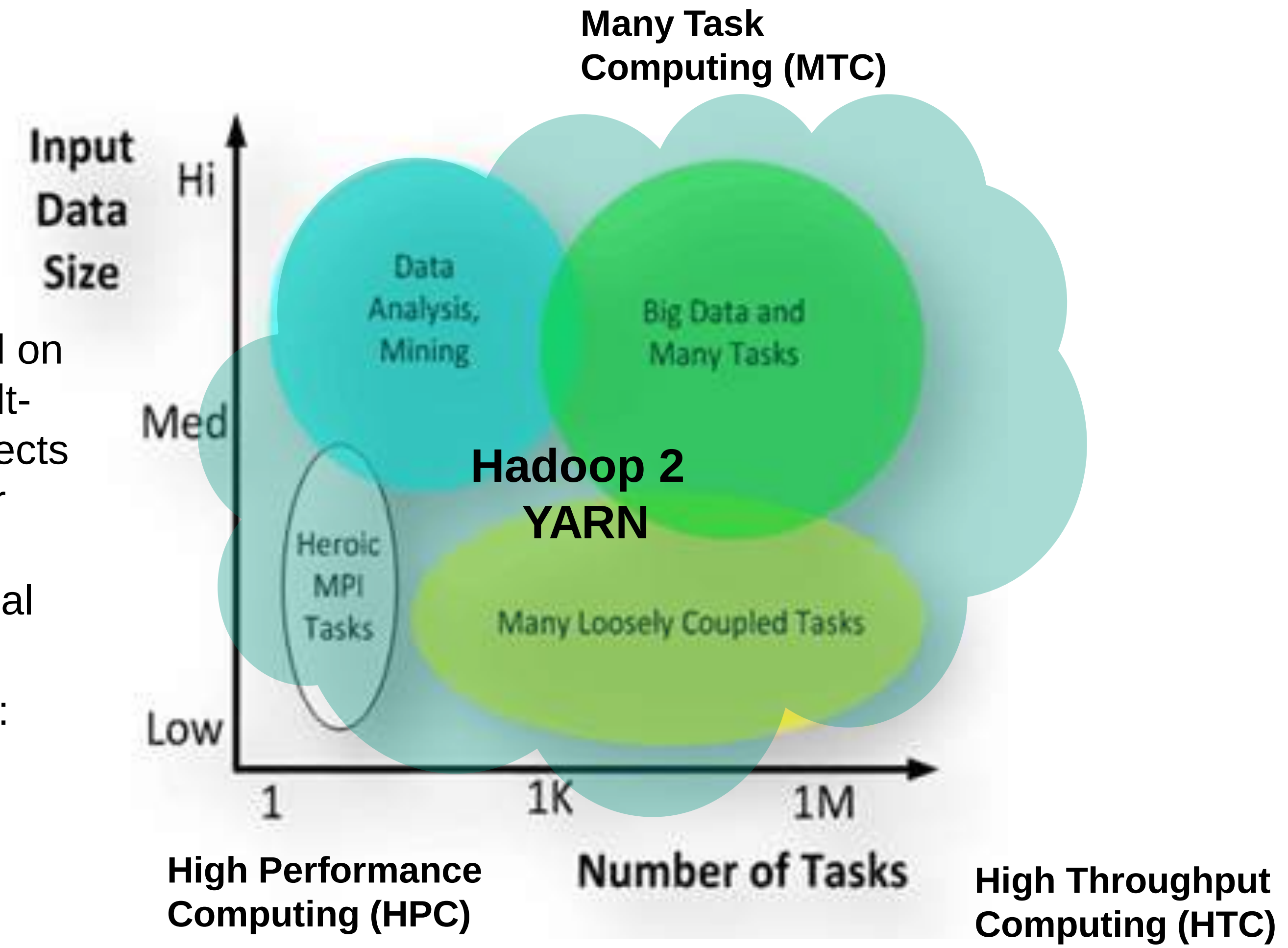
Source: I.Foster,I.Raicu 2008

Classification of Computing Applications


Many Task Computing (MTC)
Input Data Size
Hi
Med
Low
Data Analysis, Mining
Big Data and Many Tasks
Hadoop 1.x
Heroic MPI Tasks
Many Loosely Coupled Tasks
1
1K
1M
Number of Tasks
High Performance Computing (HPC)
High Throughput Computing (HTC)

- Since 2008: Hadoop 1.x is a Swiss army knife for Big Data applications

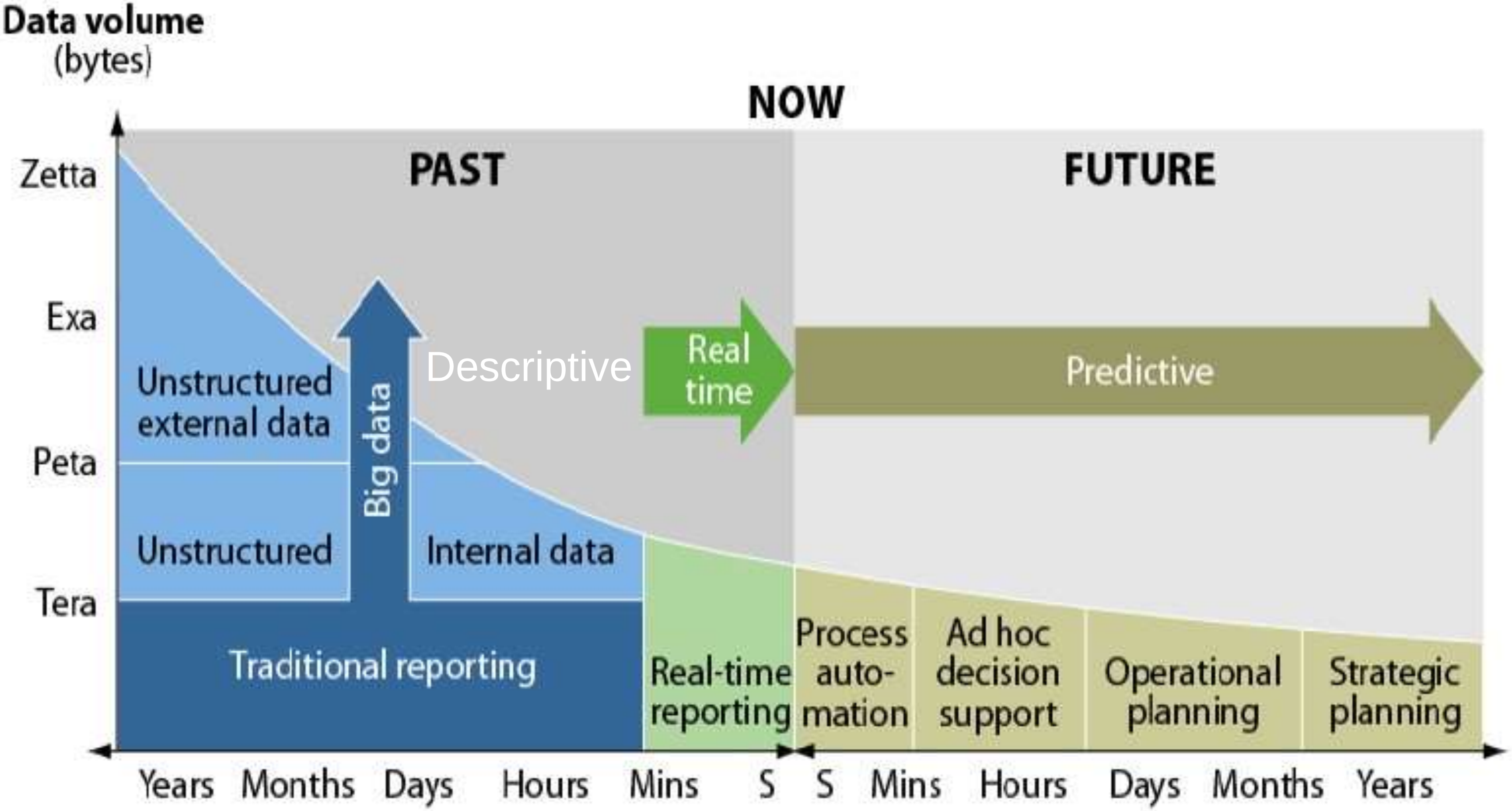
Classification of Computing Applications



- Hadoop is based on scalable and fault-tolerant data objects inside the cluster rather than traditional external file systems
- YARN scheduler: CPU and data
- Yahoo! runs a cluster of 32.000 nodes

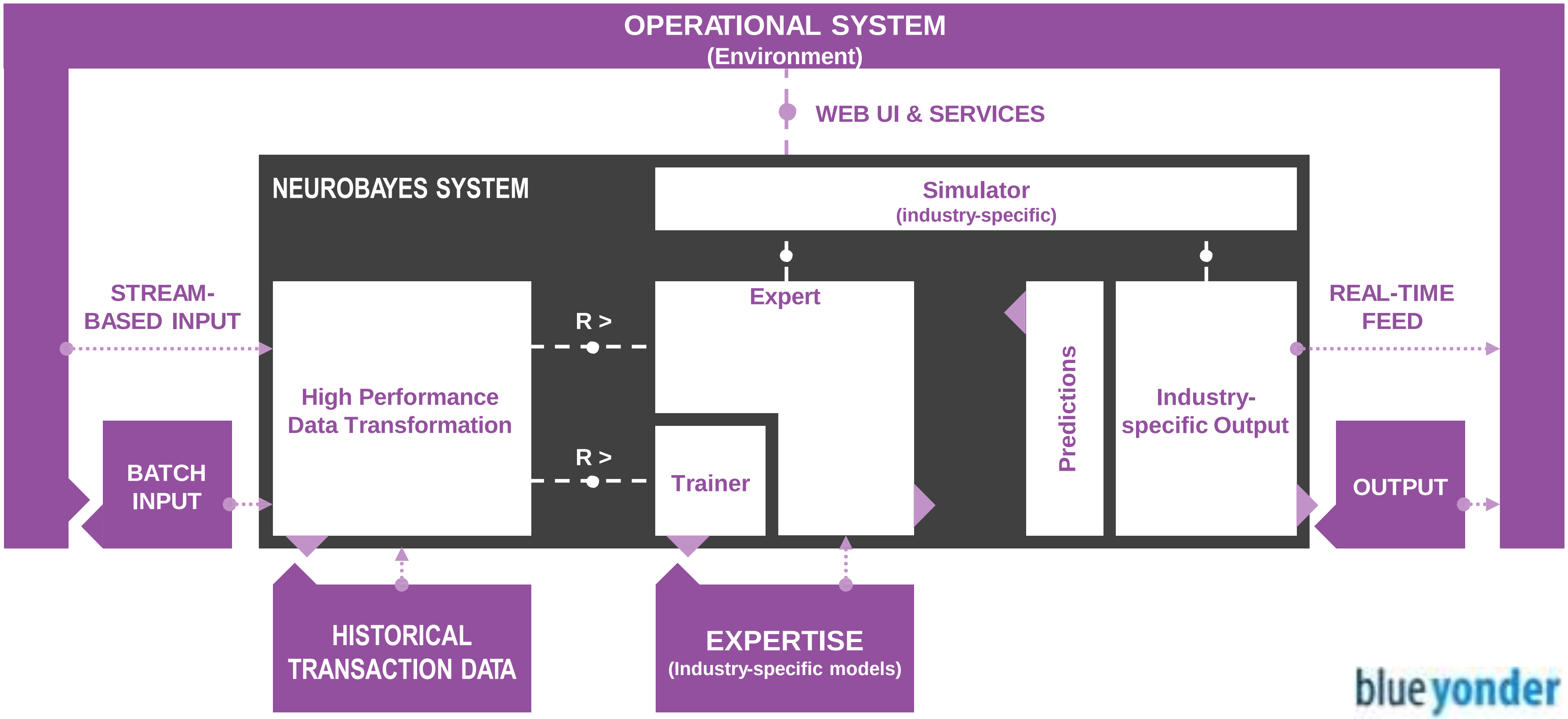
- 2014: Hadoop 2.x is a Swiss army knife for **Big Data and HPC** apps

2. Real Time Analytics



Source: blue yonder

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 streaming data

3. Software Defined Data Center

- **Trend: Software replaces (commodity) hardware functions**
  - Services vs. servers
  - Virtual machines vs. computers
  - Software Defined Networks (SDN) vs. switches and cables
  - **Object stores vs. traditional file systems**
  - …
- **Software Defined Data Center (SDDC)**
  - Data center as a software artefact
  - Configured out of resource pools
  - Checkpointed by version control (e.g. git)
  - Multi-tenant: A data scientist may have his own SDDC
  - Archival: SDDC may complement data publication services to preserve processing environment for reproduction of results

Summary

- Big Data depends on scalable models (Cloud is essential)
- Big Data is interdisciplinary: Computer Science, Mathematics, …
- Hadoop 2.0 offers interesting opportunities for combined BigData+HPC applications, especially by integrating storage and CPU

Contact:

marcel.kunze@kit.edu

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

Research Group Cloud Computing - Steinbuch Centre for Computing