RAVE
RESEARCH AT ALPHA VENTUS
Eine Forschungsinitiative des Bundesumweltministeriums

Data Warehousing for Distributed R&D Overview and Insights gained

Arno Claassen
Stefan Gudenkauf
Dr. Ulrike Steffens
OFFIS – Institute for Information Technology

International Conference
RAVE 2012
May 8-10, 2012
Bremerhaven, Germany

Gefördert auf Grund eines Beschlusses
des Deutschen Bundestages

Projektträger

Koordination

Bundesministerium
für Umwelt, Naturschutz
und Reaktorsicherheit

Projektträger Jülich
Forschungszentrum Jülich

Fraunhofer
IWES

OFFIS
INSTITUT FÜR INFORMATIK

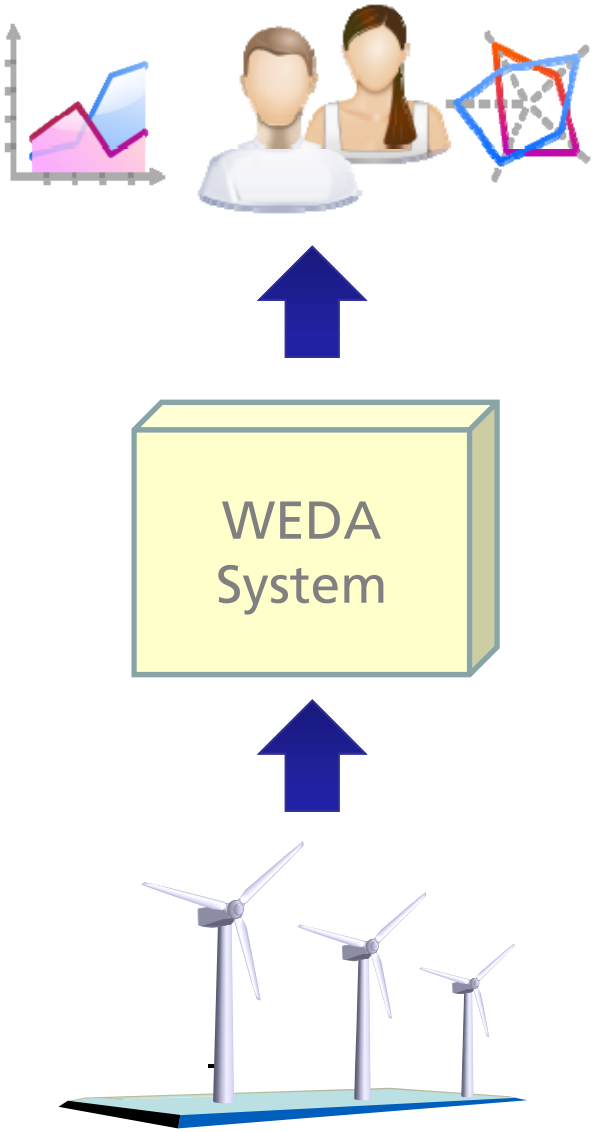
R&D Institute for ICT

- Associated institute of the Carl von Ossietzky University in Oldenburg
- More than 290 employees (~150 research assistants)
- Established in 1991

R&D Divisions

Project Overview	Data Warehouse	Use in Transition	Current Measures	Insights Gained
0	1	2	3	4

Project Overview



Icons © Oxygen Team via iconarchive.com

Data Warehouse (DWH)

- Collection and harmonization of engergy data
- Data provisioning for research accompanying the alpha ventus offshore wind park
  - Wind turbine (WEA) optimization
  - Environmental impact analysis
- Data access policy enforcement

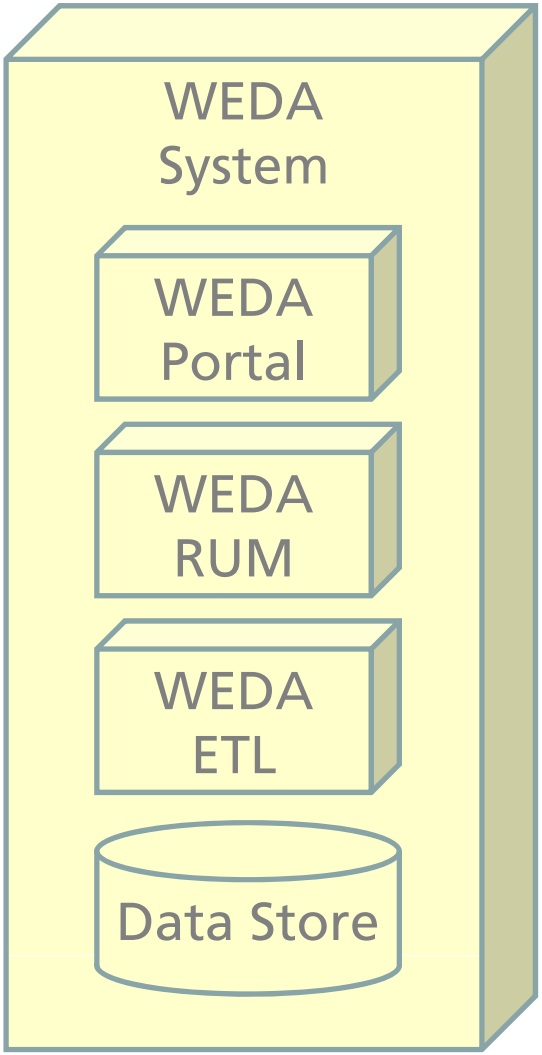
Storage-Relevant Energy Data

- Secondary data (calibrated sensor data)
- Tertiary data (statistically aggregated data)
- Metadata (data on sensors and tertiary data)

RAVE
RESEARCH AT ALPHA VENTUS
Eine Forschungsinitiative des Bundesumweltministeriums


Project Overview
Data Warehouse
Use in Transition
Current Measures
Insights Gained
0
1
2
3
4
© jeff_w_brooktree via flickr

Architecture Overview



WEDA ETL

- Central component for sensor data storage and retrieval
- Monitors data input delivery and quality
- Guarantees technical data integrity

WEDA RUM

- Role-based user access control

WEDA Portal

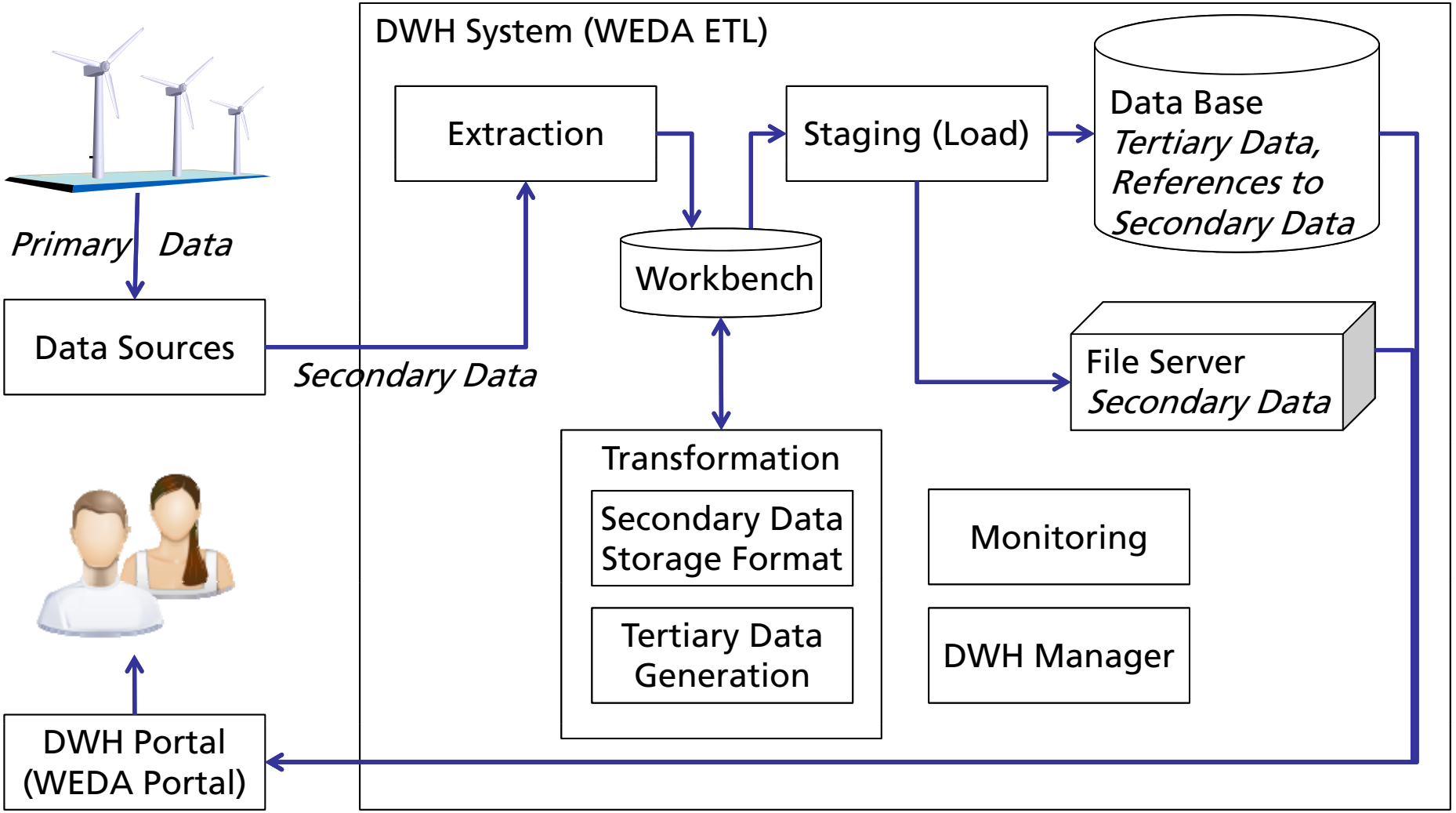
- Access portal for research partners
- Data exploration and download

Import Data Volumes

Survey „Storage Strategies for Sensor Data in Wind Energy“	alpha ventus DWH	Other (non-wind energy) Projects
20 to 250 sensors	1,500 sensors	*Data sizes comparable or even greater than those in alpha ventus*
5 min to 20 Hz	approx. 75% of the sensors operating at 50 Hz	
45,000 to 100,000 metered values per day	4,860,000,000 metered values per day	
	~18 GByte per day (data value size approx. 4 Byte) and 6.6 TByte per year	
Up to 45,000 values file system-based storage or data base; over 45,000 only file system-based storage	Hybrid (see next slide)	KIWI-concepts (*Kill it with Iron*) e.g., [Yuen et al., 2007; Ghemawat et al., 2003]
Volume reduction by compression	Volume reduction by compression	

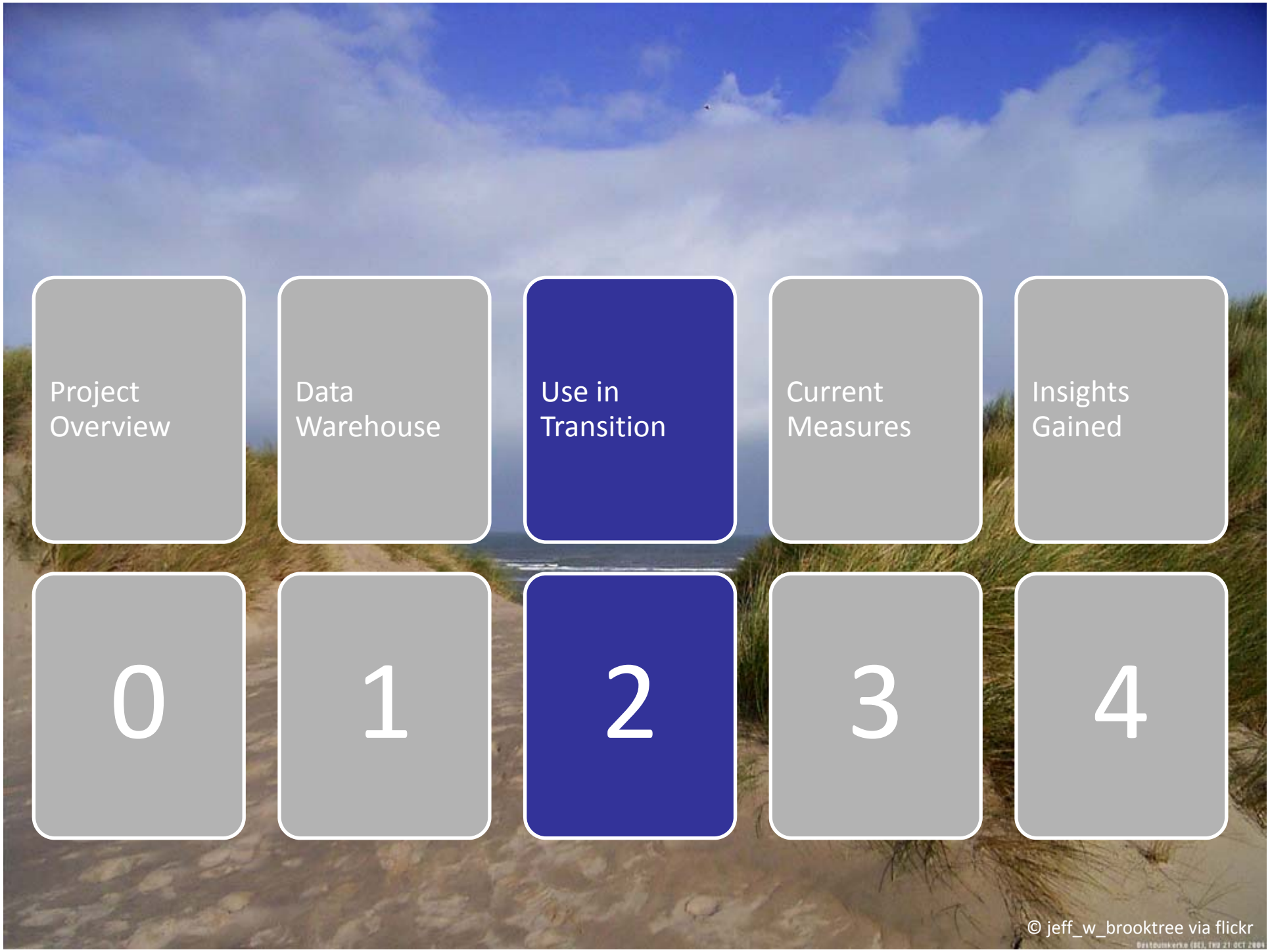
Evaluation of Storage Strategies and Design choices for alpha ventus documented in [Beenken et al., 2009]

Hybrid Storage Concept

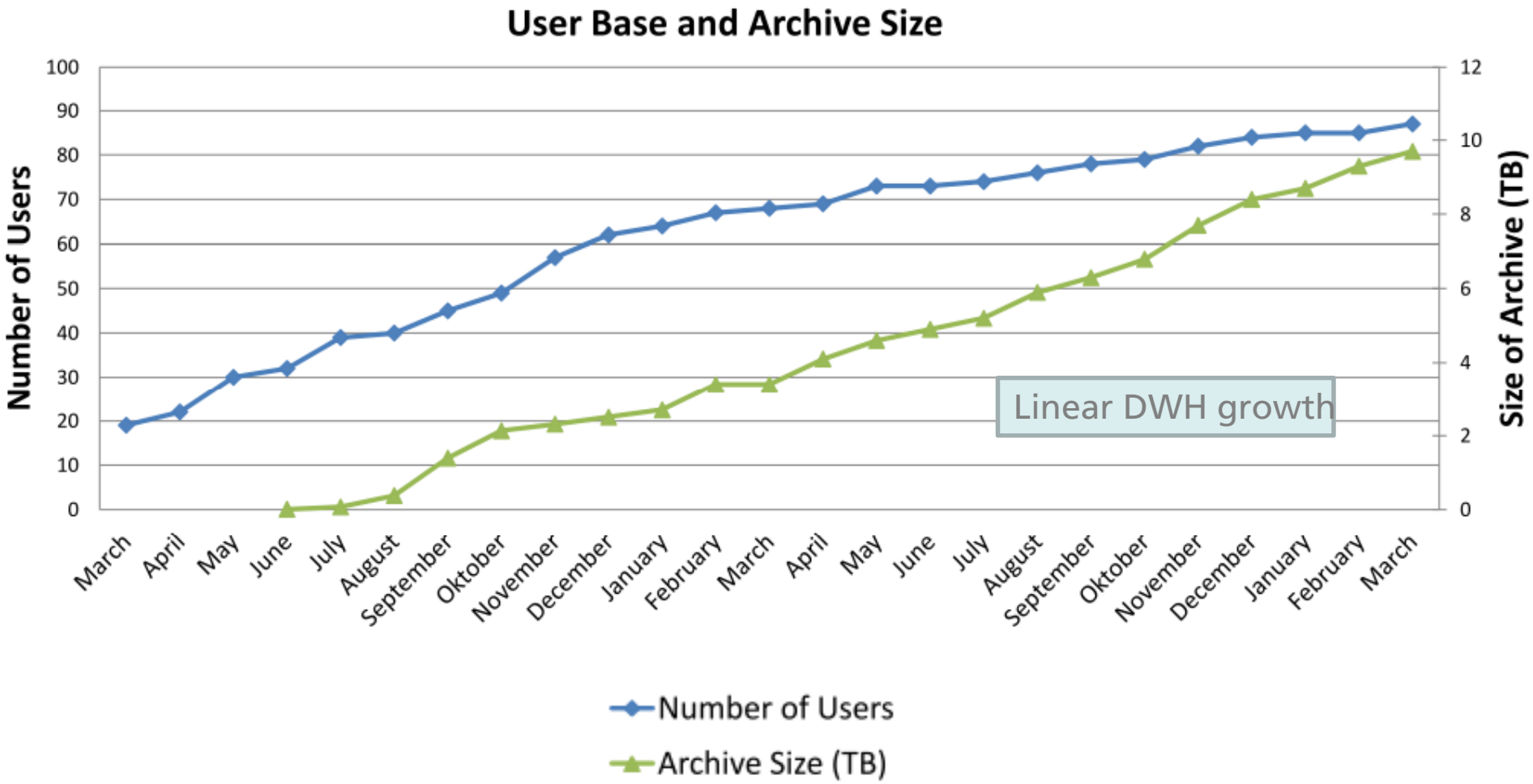


Extended DWH Reference Model in accordance with [Bauer and Günzel, 2004]

RAVE
RESEARCH AT ALPHA VENTUS
Eine Forschungsinitiative des Bundesumweltministeriums

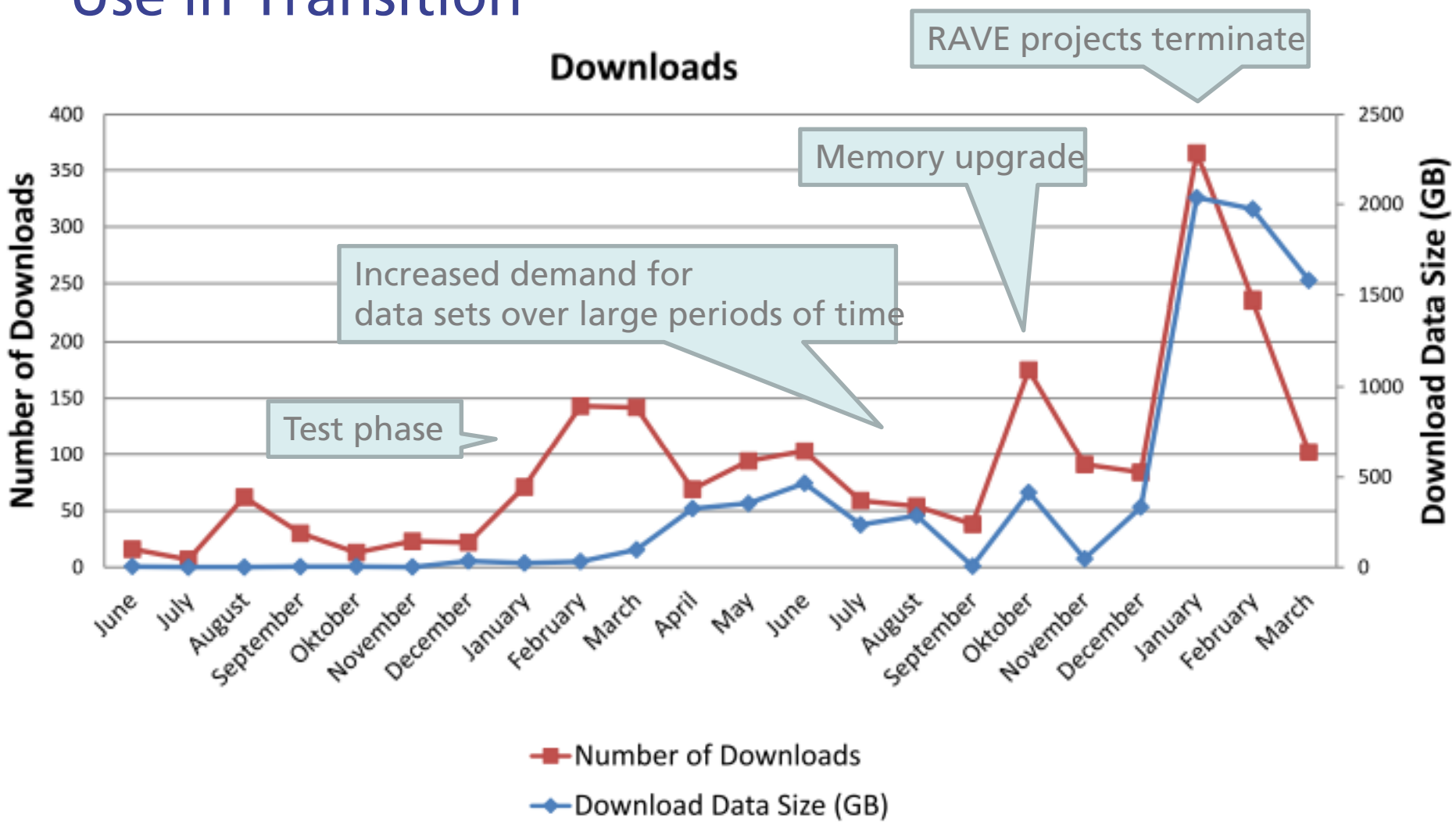

Project Overview
Data Warehouse
Use in Transition
Current Measures
Insights Gained
0
1
2
3
4
© jeff_w_brooktree via flickr

Use in Transition

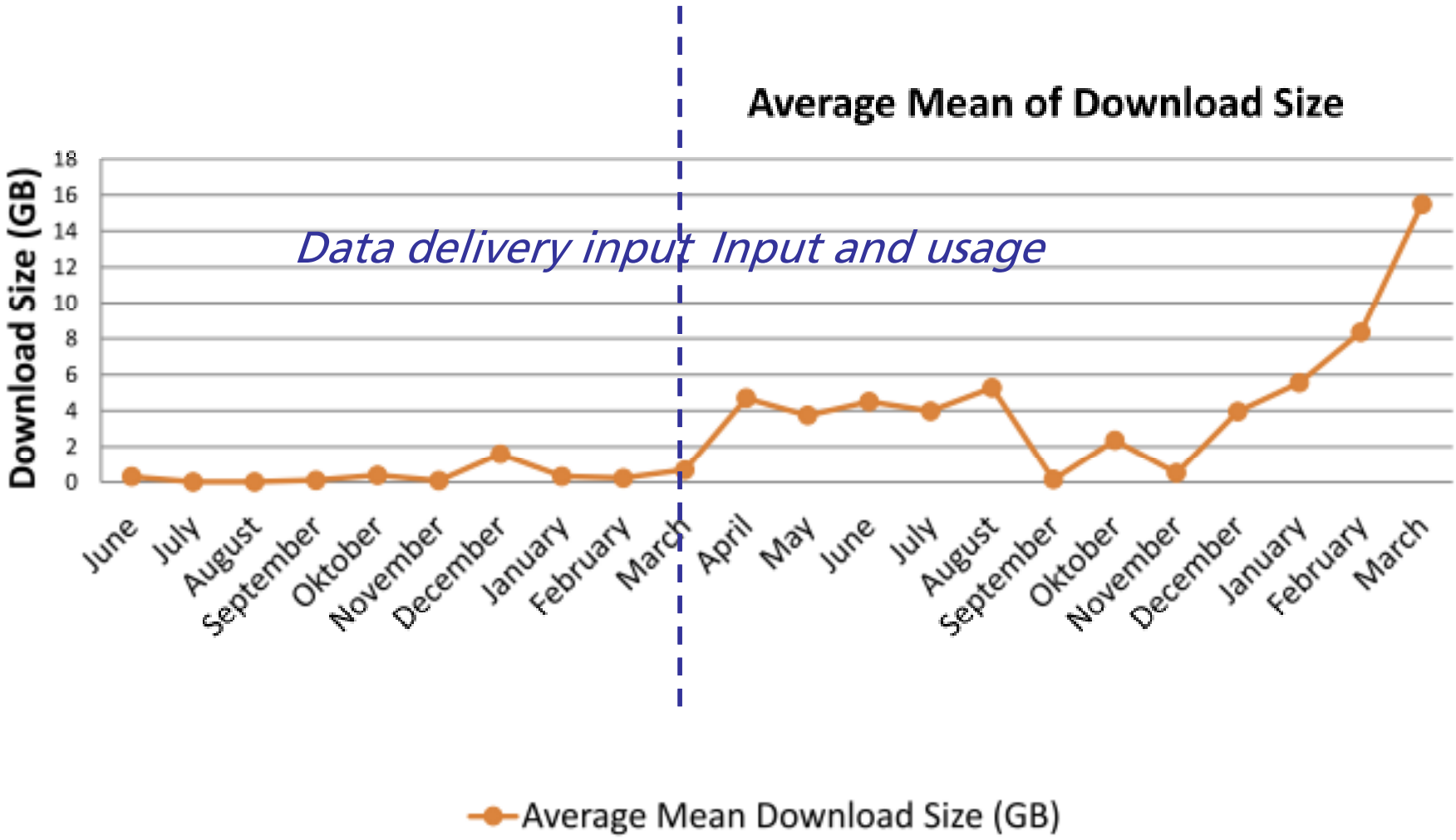


RAVE
RESEARCH AT ALPHA VENTUS
Eine Forschungsinitiative des Bundesumweltministeriums

Use in Transition



Use in Transition

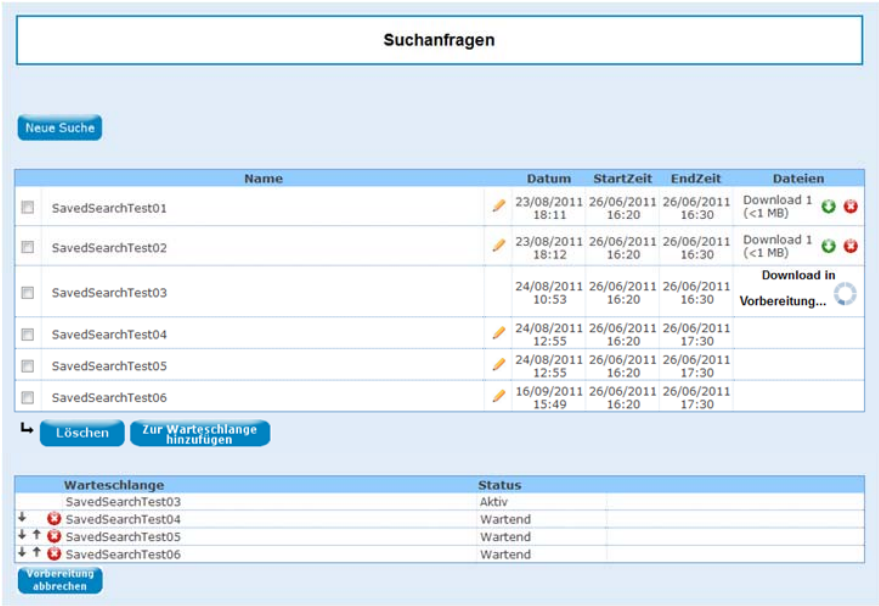


RAVE
RESEARCH AT ALPHA VENTUS
Eine Forschungsinitiative des Bundesumweltministeriums


Project Overview
Data Warehouse
Use in Transition
Current Measures
Insights Gained
0
1
2
3
4
© jeff_w_brooktree via flickr

Current Measures

- Download cancellation and wait queue
- Complementary research documents
  - No restriction on file format
  - Upload is curated by OFFIS
- Time stamp calculation for download compilation
  - The frequency of the sensors could not always be met during data recording before delivery to DWH
- Data compression for volume reduction
  - Only data that is older than 1 year
  - On-the-fly decompression upon user queries
  - ~ 20% compression rate



RAVE
RESEARCH AT ALPHA VENTUS
Eine Forschungsinitiative des Bundesumweltministeriums


Project Overview
Data Warehouse
Use in Transition
Current Measures
Insights Gained
0
1
2
3
4
© jeff_w_brooktree via flickr

Insights Gained

Continuous Adaptations

- Query optimization and table index updates (cascading change effects)
- Changes to data format specification required changes to the portal and repeated downloads
- Increased demand for data sets over large periods of time required memory upgrade

Local DWH test instance

- Test with representative/real-life data after storage design phase

Tendency to local copies

- If data is available, researcher partners will also want them on their own systems

Insights Gained

Recurring technical data quality issues

- Delivered secondary input data to be stored in the DWH did not always conform to the agreed specification
- Communication overhead

Domain-specific data quality and additional information

- If delivered secondary input data is corrupted, it probably can be resolved automatically by batch processing
  - Batch job can be considered as a part of the sensor → must also be documented and archived
  - Unclear if processed data should replace the corrupted data or if it should be stored separately
  - In general, experimental documentation and data history is highly relevant for future research projects

Outlook

- Progress information for big download compilations
- Thorough use of multithreading
  - e.g., for uploads and download compilation
- Development of a performance laboratory based on the local DWH test instance
- Re-evaluation of DWH-internal storage structures
- Server-side data processing and analysis
  - e.g., standard interfaces for aggregated statistical data

Stefan.Gudenkauf@offis.de
Arno.Claassen@offis.de

References

[Bauer and Günzel, 2004] Bauer, A.; Günzel, H.: *Data Warehouse Systeme*, dpunkt-Verlag, 2004

[Beenken et al., 2009] Beenken, P.; Schwassmann, S.; Albrecht, M.; Appelrath, H.-J.; Heisecke, S.: *Speicherstrategien für die Sensordaten des Offshore-Windparks alpha ventus*. Datenbank Spektrum, Heft 28, February 9, 2009

[Yuen et al., 2007] Yuen, C.; Hopkins Dreyer, L.; Krneta, P.: *Performance Sizing Report – Petabyte Data Warehouse*. InfoSizing, Inc., August 20, 2007

[Ghemawat et al., 2003] Ghemawat, S.; Gobioff, H.; Leung, S.-T.: *The Google File System*. 19th ACM Symposium on Operating Systems Principles, 2003

Hybrid Storage Concept

Storage Efficiency

- Secondary data stored in file system
- Tertiary data calculated from secondary data and stored in data base
- Tertiary data references the corresponding secondary data in the file system

Data Handling

- Users can analyse tertiary data before secondary data must be considered and downloaded

Export Optimization

- Secondary data can be downloaded directly
- Users can select subsets of secondary data (queries) according to various parameters; server-side recompilation provides only the relevant data

Long-Term Technology-Independent Storage

- Guaranteed by file system-based secondary data storage in plain CSV format

Import Data Volumes

Some numbers

- 1,500 sensors, approx. 75 % operating in 50 Hz
- 4,320,000 data values per day and 50 Hz sensor
- 4,860,000,000 data values per day