EUDA T
Towards a Pan-European Collaborative Data Infrastructure

Dr. - Ing. Morris Riedel
Head of Research Group ‘High Productivity Data Processing’, Juelich Supercomputing Centre, Germany
Adjunct Associated Professor, University of Iceland, Iceland

VSC User Day, Brussels
16th January 2014
Mitglied der Helmholtz-Gemeinschaft

Increasing complexity and variety

Gigabytes
Terabytes
Petabytes
Exabytes
Zettabytes

Exponential growth

Data Analysis & Analytics face ‘Big Data Waves’

- Where to store it?
- How to find it?
- How to make the most of it?

Zettabytes
Exabytes
Petabytes
Terabytes
Gigabytes

Increasing complexity and variety
HPC Simulation Pre-/Post-processing

Data results need to be analyzed and understood
Computed data must be stored and re-located
Subsets of data might be referenced in publications
Sampling vs. whole ‘big data’ sets (serial/parallel)
Pre-/Post-Processing & visualizations as new data

Data Analysis & Analytics in HPC
facing limits & challenges

HPC Simulation & Computational Science

Increasing complexity and granularity: data $\rightarrow \infty$
How data is organized has impact on performance
Multi-physics simulations & multi-model ensemble
E.g. physical processes in climate science (land, atmosphere, ocean, sea ice) & observation validation

[7] DOE ASCAC report
Mitglied der Helmholtz-Gemeinschaft

Selected Juelich Supercomputing Centre (JSC) Research Examples

Alpha Magnetic Spectrometer (AMS) @ ISS
What are building blocks of the Universe?

Search for Cosmic Antimatter
JUROPA++
1-2 P flop/s + Booster
~400 TB
~20 TB
JSC is main facility for AMS computing in Germany

Selected Juelich Supercomputing Centre (JSC) Research Examples

Simulation Labs (e.g. Climate SimLab)...
...one example out of many JSC SimLabs

~10 years
500 HDF Files
~50 TB

~20 TB

Applications with combined characteristics of simulations and analytics...

Towards Exascale
Better Prediction Accuracy...  
... means ‘Bigger Data’

We are unable to store the output data of all computational simulations/users'

[4] F. Berman
Summarizing Big Data Waves & Surfboards
How to engage in the rising tide of scientific data?

Unsolved Questions:
Scale
Heterogeneity
Stewardship
Curation
Long-Term Access and Storage

Research Challenges:
Collection, Trust, Usability
Interoperability, Diversity
Security, Smart Analytics,
Education and training
Data publication and access
Commercial exploitation
New social paradigms
Preservation and sustainability
A framework for the future?

Collaborative Data Infrastructure

- **User functionalities, data capture & transfer, virtual research environments**
- **Data discovery & navigation, workflow generation, annotation, interpretability**
- **Persistent storage, identification, authenticity, workflow execution, mining**
Breakwaters – Offer Concrete Solutions for Researchers

Is there a common set of services often needed by scientists?

Identified Common Data Services

Persistent Identifiers for Research Data
Safe Replication of Scientific Data
Transfer of Data to/from Computing
Simple Sharing of Research Data
Metadata Catalogue

‘Concrete’

Next Steps →
Track the Origins and Characteristics of Information

Domain of Registered Research Data

Persistent Identification of Scientific Datasets

- Provides PID for each data/digital object
- Based on the ‘Handle System’

modified from [5] Handle System
Providing a robust, safe, and highly available...

Data Replication Service

...to guard against data loss in long-term archiving & preservation

- Realized in registered data domain
- Enables reliable data curation
- Optimize data access for users
- Provides adaptable policy mechanisms

EUDAT CDI Domain of registered data
Bringing research data closer to powerful computers with a...

Data Staging Service

...for compute-intensive scientific data analysis

- Realized in registered data domain
- Enables easy access to execution services
- Offers CPU-intensive data transformations
Offering an easy data deposit and upload via the Simple Store Service...

...to share data & collections with other researchers

- Think as ‘YouTube for Scientific Data’
- Realized in registered data domain
- Upload ‘long tail’ research data
Find and access research data collections via the...

MetaData Service

...in a simple and user-friendly way

- Based on registered data domain
- Enables effective data understanding
- Provides metadata harvesting
- Offers comment functions on metadata
Mitglied der Helmholtz-Gemeinschaft

We need to ‘dive into data’

Long-term Data Preservation and Curation...

bears potentials to lower ‘Data Waves’
and supports data analytics & analysis

Addressed requirements of the High Level Expert Group on Scientific Data:

- **High reliability**, so data scientists can count on its availability
- **Open deposit**, allowing user-community centres to store data easily
- **Persistent identification**, allowing data centres to register a huge amount of markers to track the origins and characteristics of the information
- **Metadata support** to allow effective management, use and understanding
- **Avoids re-creation of datasets** through easy data lookups and re-use
- **Enables easier identification of duplicates** to remove them & save storage


[3] EUDAT Web Page

[1] HLEG Report
<table>
<thead>
<tr>
<th>#</th>
<th>Suggestions for Requirements of a Data Infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLR1</td>
<td>Open deposit, allowing user-community centres to store data easily</td>
</tr>
<tr>
<td>HLR2</td>
<td>Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years</td>
</tr>
<tr>
<td>HLR3</td>
<td>Format and content migration, executing CPU-intensive transformations on large data sets at the command of the communities</td>
</tr>
<tr>
<td>HLR4</td>
<td>Persistent identification, allowing data centres to register a huge amount of markers to track the origins and characteristics of the information</td>
</tr>
<tr>
<td>HLR5</td>
<td>Metadata support to allow effective management, use and understanding</td>
</tr>
<tr>
<td>HLR6</td>
<td>Maintaining proper access rights as the basis of all trust</td>
</tr>
<tr>
<td>HLR7</td>
<td>A variety of access and curation services that will vary between scientific disciplines and over time</td>
</tr>
<tr>
<td>HLR8</td>
<td>Execution services that allow a large group of researchers to operate on the stored data</td>
</tr>
<tr>
<td>HLR9</td>
<td>High reliability, so researchers can count on its availability</td>
</tr>
<tr>
<td>HLR10</td>
<td>Regular quality assessment to ensure adherence to all agreements</td>
</tr>
<tr>
<td>HLR11</td>
<td>Distributed and collaborative authentication, authorisation and accounting</td>
</tr>
<tr>
<td>HLR12</td>
<td>A high degree of interoperability at format and semantic level</td>
</tr>
</tbody>
</table>

Building a trusted collaborative data infrastructure with...

Strong and Sustainable Community & Generic Data Centers

...to enable federated data services together with users

[1] HLEG Report
Key Approaches:
Bridging National & EU Solutions
Not ‘one single data infrastructure’
Federated Network of Trusted Centers

Key Benefits for Scientific Users:
Trust, Sustainability, Interoperability,
Diversity, Extensibility (e.g. Belgium?),
New Social Paradigms & Sustainability
Preparing for new data challenges on the horizon …

The square kilometre array

... 1 PB in 20 seconds

New EUDAT Services in development with users:

- ‘EUDAT Box’ dropbox-like service
  - easy sharing
  - local synching

- ‘Semantic Anno’ checking & referencing

- ‘Dynamic Data’ immediate handling
Towards Exascale: Applications with combined characteristics of simulations & data analytics

‘In-Situ Analytics’

- In-situ correlations & data reduction
- Correlations & data reduction
- In-situ statistical data mining
- In-situ analytics
- In-memory
- Key-value pair DB
- Distributed archive
- Scalable I/O
- Interactive
- Computational simulation part
- Visualization part
- Analytics part

Exascale computer with access to exascale storage/archives

Inspired by a [7] DOE ASCAC report
Data Scientists with skills of various fields

Data Mining vs. Big Data

Training Data Scientists

Computational Scientist

Software Engineer

Data Scientist

Engineer

Parallelization!

Scientific Computing

Sampling

Applied Statistics

Data Mining

Machine Learning

Algorithms

new DBs

Statistical Data Mining Course

HPC – B(ig Data) Course

Reference Material for Data Scientists

EUDAT Services, Data Management Plans, Curation, ...

[8] RDA BDA Webpage

Data Scientists with skills of various fields
References

‘Thanks’

Talk available at: www.morrisriedel.de/talks

Contact: m.riedel@fz-juelich.de