TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

Andreas Knüpfer
Center for Information Services and High Performance Computing (ZIH)

Data Volume Considerations for NHR and NFDI

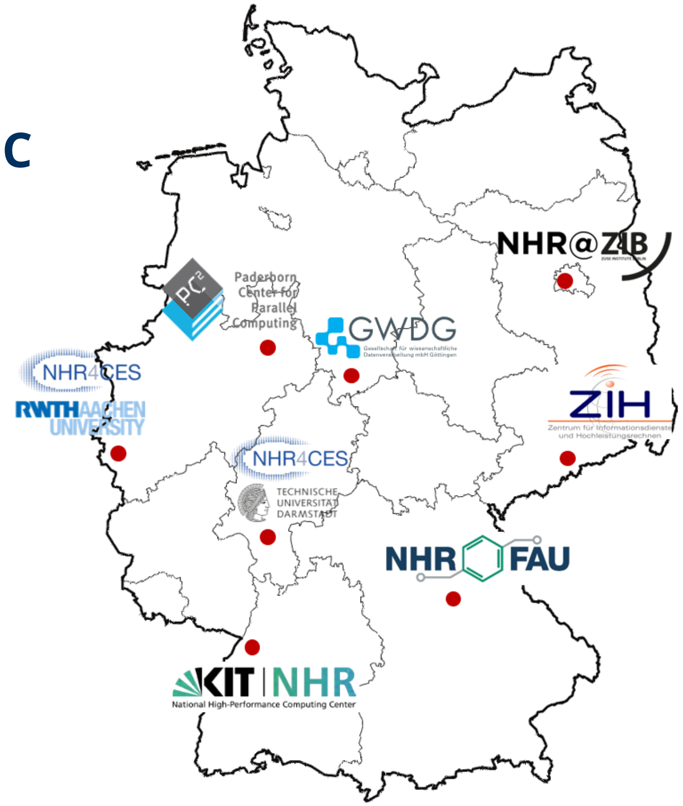
GWDG Data Lakes Workshop
2021-12-17

ZIH
Center for Information Services &
High Performance Computing

Vantage point: NHR and NFDI

TU Dresden host NHR center with focus on data intensive HPC

- — Will host HPC projects with their data
- — Invests large budget shares into storage



NHR@ZIB
Paderborn Center for Parallel Computing
PC²
GWDG
NHR4CES
RWTH AACHEN UNIVERSITY
ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen
NHR4CES
TECHNISCHE UNIVERSITÄT DARMSTADT
NHR FAU
KIT NHR
National High-Performance Computing Center

TU Dresden is NFDI partner

- — No Hardware funding for NFDI
- — Instead explicitly referred to NHR centers (among others)
- — ... synergy is good but doesn‘t buy you extra hard drives

Is the Data Lake the solution to all our data problems?

- Storage volumes?
- Data transfers between storage tiers? Entirely transparent?
- Data management?

https://pxhere.com/en/photo/1358249

HPC Data Management … old style

Research group perspective:

- Have their data or know how to get from public/community/whatever sources
- Use some group internal data management conventions (if they are advanced)
- Need to store their data after the project

HPC center perspective:

- Compute time allotment + storage quota
- Feel free to copy your data here
- Please take your data with you after the compute time project ended

https://pxhere.com/en/photo/912

HPC Data Management … old style

Research group perspective:

- Have their data or know how to get from public/community/whatever sources
- Use some group internal data management conventions (if they are advanced)
- Need to store their data after the project

HPC center perspective:

- Compute time allotment + storage quota
- Feel free to copy your data here
- Please take your data with you after the compute time project ended

Consequences

- **Challenge!** Data Lake may help though

(We don’t even ask for meta data standards)

- **Challenge!** Can they store it reliably?

- **Challenge** partly solved by Data Lake
- **Impossible** situation for very large data sets*

HPC Data Management … new style

Research group perspective:

- Local Data Lake at HPC center or national NFDI services offer community data sets
- There are basic data management and meta data standards per science community

HPC center perspective:

- Compute time allotment for limited periods
- Allow to keep large data sets at HPC center
- Long term data archiving/sharing/publication

https://pxhere.com/en/photo/948354

HPC Data Management … new style

Research group perspective:

- Local Data Lake at HPC center or national NFDI services offer community data sets
- There are basic data management and meta data standards per science community

HPC center perspective:

- Compute time allotment for limited periods
- Allow to keep large data sets at HPC center
- Long term data archiving/sharing/publication

Consequences

- NFDI to the rescue
- See below why this becomes important
- Why not let it flow back to the Data Lake?
  No, we couldn’t buy enough disks/tapes if handled casually

Redundancy

How much redundant data is there?

- Some*
- So we could reduce redundancy within research groups. Can we go even farther?

Joint data management conventions

- Separate data sets
  - Raw input data sets (large, more or less immutable)
  - Result data sets (often much smaller but that is not the point here)
- Allow to reuse / share / publish them independently. Incentivize reuse / sharing!
- Keep only one copy per HPC or data center. Or even all across Germany?

Summary

Don‘t make your Data Lake entirely transparent after all.

Require some swimming aids for everyone – a.k.a. joint data management.

(This metaphor is actually misleading.)

https://pxhere.com/en/photo/1358249