Deployment of an HPC-Accelerated Research Data Management System: Exemplary Workflow in HeartAndBrain Study

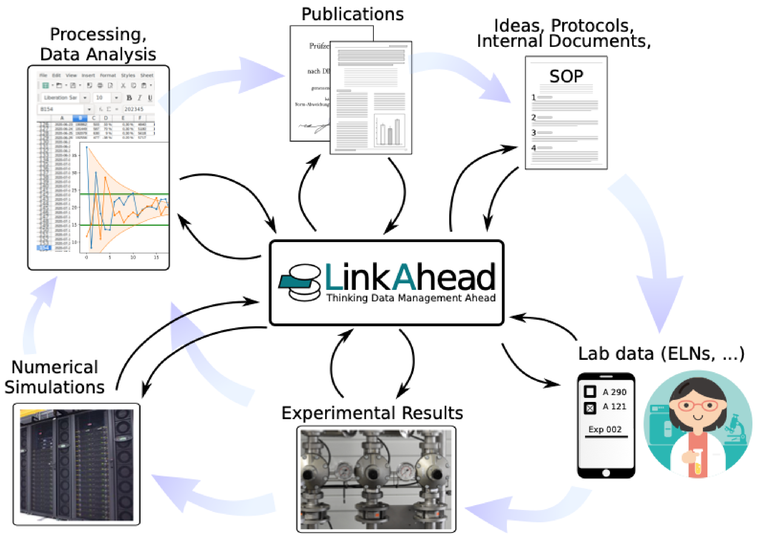
V. Telezki, H. tom Wörden, F. Spreckelsen, H. Nolte, J. Kunkel, U. Parlitz, S. Luther, M. Uecker, M. Bähr

Abstract

We present our research data management sytem (RDMS) used to facilitate reserach of the brain's waste clearance mechanisms. In this research project,, we collect (longitudinal) data from multiple sources, in particular from MRI, ECG, SpO2, breathing belt, laboratory analysis of blood and urine. Our RDMS allows us to integrate these inhomogeneous data sources in one data base where it is accessible via structured queries either via API or GUI. Furthermore, we developed (semi-) automatic post-processing pipelines that take care of routinely used post-processing steps.

Computationally demanding tasks were set up to utilize high-performance computing (HPC) infrastructure, with automatic job submission and re-integration into the data base.

Core Tool: LinkAhead

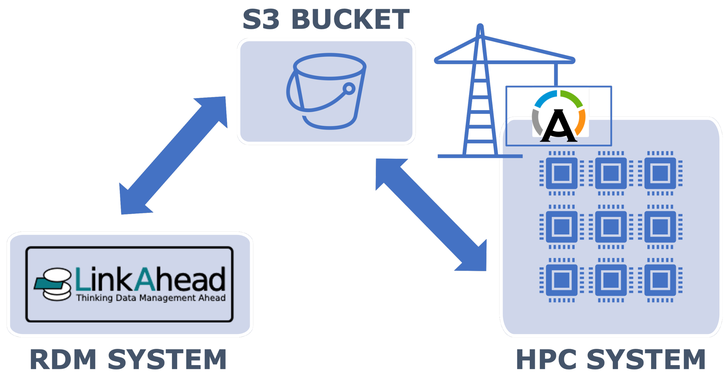


LinkAhead is an open source tool-kit, which is highly adaptable, allows integration of automated analysis pipelines, enables management of (meta) data relevant to this project with ist flexible data model and its integrated semantic search lets us access data conveniently and precisely.

Workflow

- All acquired data is anonymized and integrated automatically into LinkAhead data base, based on custom data model. Key aspects of data model are adopted from BIDS. Important meta-information from MR images are automatically extracted.
- Integrated data can be accessed and retrieved via structured queries (e.g. "FIND T1WeightedImage") within the LinkAhead GUI) or its API either individually or in custom defined groups for batch processing.
- Integrated data typically needs to be post-processed. Some analyses (e.g. volumetric segmentation, image correction) require computer-intensive algorithms and tools.
- we utilize resources of high-performance computing (HPC) infrastructure to use these tools efficiently

Connection to HPC system



Communication layer between RDMS and HPC system is realized via REST interface implemented with Simple Storage Service (S3), where only pre-configured jobs can be executed. Communication layer is therefore secured via HTTPS protocol.

Details of Communication Layer

- communication between RDMS and HPC system within user space of individual who triggered computation by assigning prefix
- job specification file in JSON format includes all variables relevant for execution like path to data, output path, environment variables, etc.
- status file tracks progress, monitor script reports job status back to RDMS so that finished jobs can be fetched and integrated within RDMS
- **requierments**: user has access to S3 bucket and all functions that shall be executed are already available in users *HOME* directory as containerized images.

Method	Security	User Space	Transaction Safe	Scalability	Portability
SSH Keys	-	+	+	+	+
SSH Force Commands	+	-	+	+	-
our workflow	+	+	0	-	+
HPC API	+	+	+	+	0

Comparison of different communication layers qualitatively indicating advantages and disadvantages of different methods.

Outlook

In the future, we plan to use a dedicated service which provides a REST interface for HPC system. It can be used as a drop-in replacement for steering the control flow, i.e., submitting commands from the RDMS to the HPC system, because it also uses similar JSON files to communicate job specifications, while being transaction-safe and scaling almost without any overhead.

Else Kröner Fresenius Stiftung
UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG
HLRN NORDDEUTSCHER VERBUND FÜR HOCH- UND HÖCHSTLEISTUNGSRECHNEN
DFG
MBExC Multiscale Bioimaging Cluster of Excellence