Tracking large-scale simulations through unified metadata handling

Jose Villamar[1,2], Matthias Kelbling[3], Dennis Terhorst[1], Heather More[1,4], Tom Tetzlaff[1], Johanna Senk[1], Stephan Thober[3]

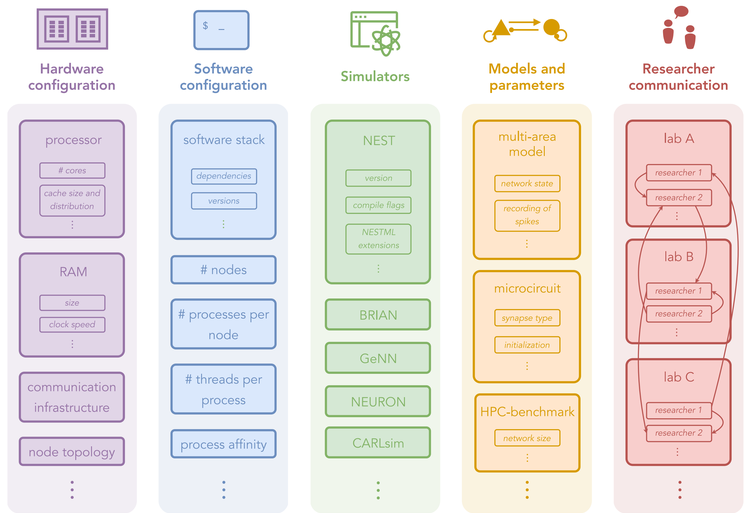
1. Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany
2. RWTH Aachen University, Aachen, Germany
3. Department of Computational Hydrosystems, Helmholtz-Centre for Environmental Research, Leipzig, Germany
4. Institute for Advanced Simulation (IAS-9), Jülich Research Centre, Jülich, Germany

Contact: j.villamar@fz-juelich.de Webpages: https://www.fz-juelich.de/de/inm/inm-6 and https://www.ufz.de/index.php?en=34211

Summary

- Metadata management framework for HPC simulation workflows to assist with:
  - Reproducibility of simulation experiments
  - Efficient organization, exploration and visualization of simulation data
- Address all components of simulation research and corresponding metadata types
- Cope with modularity and flexibility demands of rapidly progressing science
- Applicable to diverse simulation based research fields, example use cases from:
  - Computational Neuroscience
  - Earth and Environmental Science

Complexity of HPC simulations with examples from Computational Neuroscience [1]



User stories

Story 1 (Model reproducibility):
Scientist X cannot reproduce simulation results of scientist Y due to lack of information on software dependencies and inconsistencies between the article and the code published by Y. Even personal communication with Y does not resolve these inconsistencies. [2]

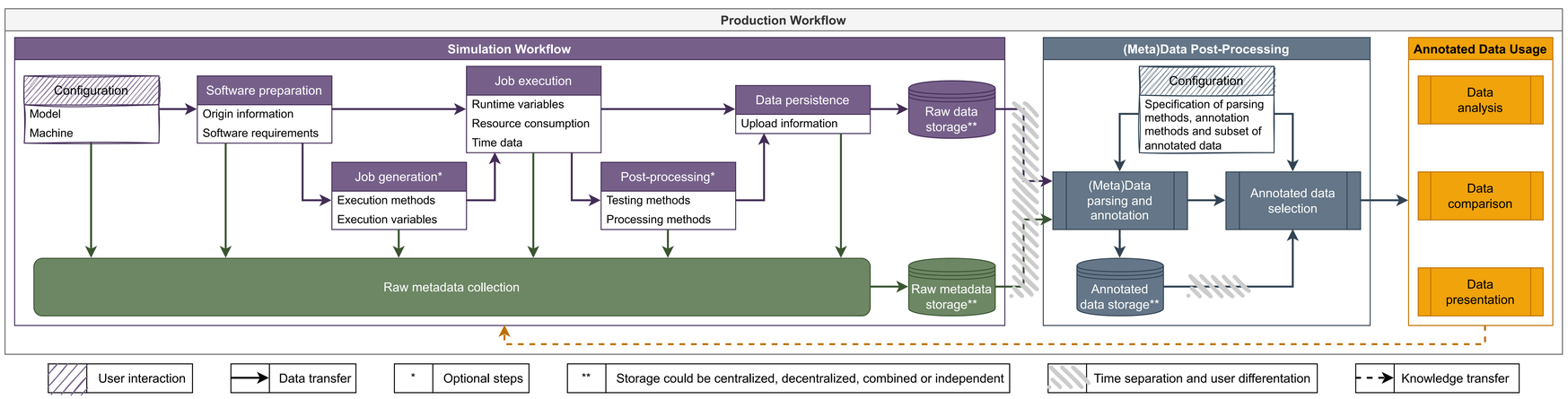
Story 2 (Hardware reproducibility):
Scientist X cannot reproduce their previous simulation performance results even though they are using the same model implementation, software stack, and hardware. Only after personal communication with the IT department, X finds out that the system was actually running at higher clock speed.

Story 3 (Data exploration):
A team of developers is regularly running validation experiments with different configurations and models to continuously monitor software performance. After years of development the group has accumulated large amounts of validation data for each software version with no means of efficient exploration.

Story 4 (Data re-usability):
A group of scientists regularly runs simulations of a particular mathematical model. The data generated during the simulation is often identical and could potentially be used by several scientists. However, each of them is interested in a very different aspect of the simulation outcome and runs a different type of analysis. Although sharing the simulation data would be trivial, the scientists have no efficient way of sharing the underlying information necessary to understand the meaning and structure of the data.

...

Concept of metadata management framework



References

- Albers et al. (2022) A Modular Workflow for Performance Benchmarking of Neuronal Network Simulations, Front. Neuroinform. 16:837549
- Pauli et al. (2018) Reproducing Polychronization: A Guide to Maximizing the Reproducibility of Spiking Network Models. Front. Neuroinform. 12:46

Acknowledgments: The authors would like to thank Jan Bumberger, Helen Kollai, Michael Denker, Rainer Stotzka, Guido Trensch, and Stefan Sandfeld for ongoing fruitful discussion. This project was funded by Helmholtz Metadata Collaboration (HMC) ZT-I-PF-3-026, EU Grant 945539 (HBP), Helmholtz IVF Grant SO-092 (ACA), and Joint lab SMHB; compute time was granted by VSR computation grant JINB33, Jülich. The work was carried out in part within the HMC Hub Information at the Forschungszentrum Jülich.