Fostering Data-Intensive Science in the German Helmholtz Association

Achim Streit
Karlsruhe Institute of Technology (KIT), Helmholtz Association, Germany
achim.streit@kit.edu

The challenge of Big Data in Science does not only comprise the volume of data, but also the velocity, variety, veracity and value of it. The exploration of data represents a revolution in how scientific discovery takes place and is generally considered as the 4th paradigm next to theory, experiment and simulation [Jim Grey, 2007]. The 4th of July 2012 gives an example how important the exploration of data is. On that day evidence of a Higgs Boson was publicly announced and the CERN Director General Rolf-Dieter Heuer pointed out in his conclusion that "the results today (were) only possible due to extraordinary performance of accelerators, experiments and Grid Computing", which is in the LHC case another word for large-scale, federated data management and analysis. One year later the physics Nobel Prize went to Francois Englert and Peter Higgs.

In the Helmholtz Association we address the challenge of Big Data with the following research questions on large-scale data management and analysis:

- How to address preservation and curation with tools and processes?
- How to advance visualisation algorithms for large-scale data?
- How to protect valuable data?
- How to easily share data in global collaborations?
- How to optimise meta-data handling?
- How to enhance generic techniques by orders of magnitude?
- How to archive and preserve data?
- How to learn from data with novel methods?
- How to shape curricula for future data scientists?

In data-intensive science collaboration is a key factor. On the national level we lead a multi-disciplinary initiative across the Helmholtz research fields to foster the exchange of knowledge, expertise and technologies. For technology transfer and connection with industry, KIT and SAP have launched – together with Bosch, EnBW, Siemens, Bayer, Microsoft, Software AG as well as FZJ, Fraunhofer and DFKI – the Smart Data Innovation Lab (SDIL), which is acting as a Data Life Cycle Hub for industry, and which is operated by KIT. It comprises a strong involvement of industry in four initial so called Data Innovation Communities on Industry 4.0, Energy, Smart Cities and Medicine.

On the international scale we are actively contributing to WLCG – the worldwide LHC Computing Grid – as well as to large EU initiatives like EUDAT and the FET Flagship Human Brain Project (HBP). We are actively contributing to the Research Data Alliance (RDA) in various working groups and interest groups.

On all levels we are actively and closely collaborating in joint R&D with the scientific communities ranging from elementary and astro particle physics, synchrotron research, systems biology, climatology, materials science, energy research to the humanities and social sciences.

To foster data-intensive science with innovative solutions, technologies, facilities and joint R&D several activities exist at KIT which are described in the following: we operate GridKa as a Tier-1 data and computing facility in the WLCG federation serving all four experiments of the LHC as well as Belle-II. It consists of > 10k cores, 12 PB disk and 17 PB tape capacity storing about 14% of the LHC data permanently.

Derived from more than 10 years of experience with GridKa, we have started the Large Scale Data Facility (LSDF) a few years ago, which provides storage and archival capacity (6 PB each) as well as analysis capability. General-purpose (e.g. iRODS, dCache, Hadoop) as well as specialized software solutions (e.g. DAWA) for data management and analysis are offered for a diverse set of scientific research disciplines. For universities in the state of Baden-Württemberg with close to half a million students and employees the LSDF offers drop-box like data storage solution based on a Shibboleth federation in the state.

The facility activities are enhanced with a profound R&D programme around our Large Scale Data Management and Analysis (LSDMA) initiative. It comprises Data Life Cylce Labs (DLCLs) on climatology, energy and key technologies, in which joint R&D with the respective communities is performed to optimize data life cycles in the communities working on community-specific tools and services. Generic methods and technologies applicable to more than one community are developed in the horizontal Data Services Integration Team (DSIT), which also acts as the interface between data facilities and infrastructures and the DLCLs resp. communities.

A first example scientific highlight is from structural biology: Light Sheet Microscopy is a novel method to image living zebrafish embryos in 3D – data rates of 16 TB per day are common depending on the number of samples. The data is generated in the camera and the data acquisition system and is ingested to a repository system managing data and metadata that is typically filled with 250 TB within 6 weeks. Finally, automatic processing of the images takes place for visual analysis. In joint R&D the data flow was defined, the repository system was set-up and all connections, interfaces and data formats were configured.

A second example scientific highlight is from climatology: spectrometer data from the GLORIA sensor mounted on a high-flying research aircraft is stored with the proper meta-data in a NoSQL database to allow fast search and processing of the raw data. Similarly, data life cycle management is needed for data from the MIPAS sensor that orbited on the Envisat satellite. Here the focus of joint R&D with the scientific community is on the development of a real-time response analysis framework. Furthermore in a joint project with the German Climate Computing Centre (DKRZ) a portal for data transport, quality control, persistent identifiers and publication of results and data is developed as well as an iRODS federation is established between DKRZ and KIT.