Development of a Cloud-based Data Management Infrastructure

Marius Dieckmann, Alexander Goesmann
Bioinformatics & Systems Biology, Justus Liebig University Giessen, Germany

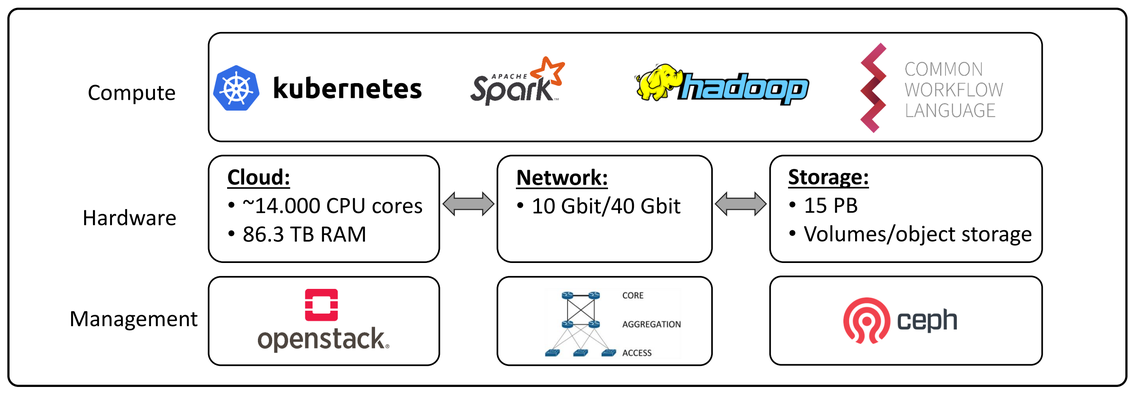
The research group of A. Goesmann at JLU Giessen.

The de.NBI BIGI Competence Center Giessen

The German Network for Bioinformatics Infrastructure (de.NBI) consists of eight service centers with 40 bioinformatics groups from all over Germany joining forces to provide bioinformatics services and training for researchers in the life sciences. As part of the Bielefeld/Giessen (BiGi) service center, the Goesmann Lab at Justus Liebig University Giessen (JLU) has a strong focus on high-performance computing services including an OpenStack infrastructure for cloud computing and a repository of reusable workflows suitable for high-throughput sequence data analysis. Together with our project partners from different scientific disciplines from all over the world, we analyze tremendous amounts of biological sequence data. For this task, we provide bioinformatics software solutions for automated genome assembly (ASA³P) [1], high-throughput automatic genome annotation (GenDB) [2], large-scale comparative genomics (EDGAR) [3] and metagenomics (GMX) [4] as well as RNA-seq data analysis and visualization (ReadXplorer) [5]. Furthermore, we provide support and user training in the field of microbial bioinformatics to the life science community.

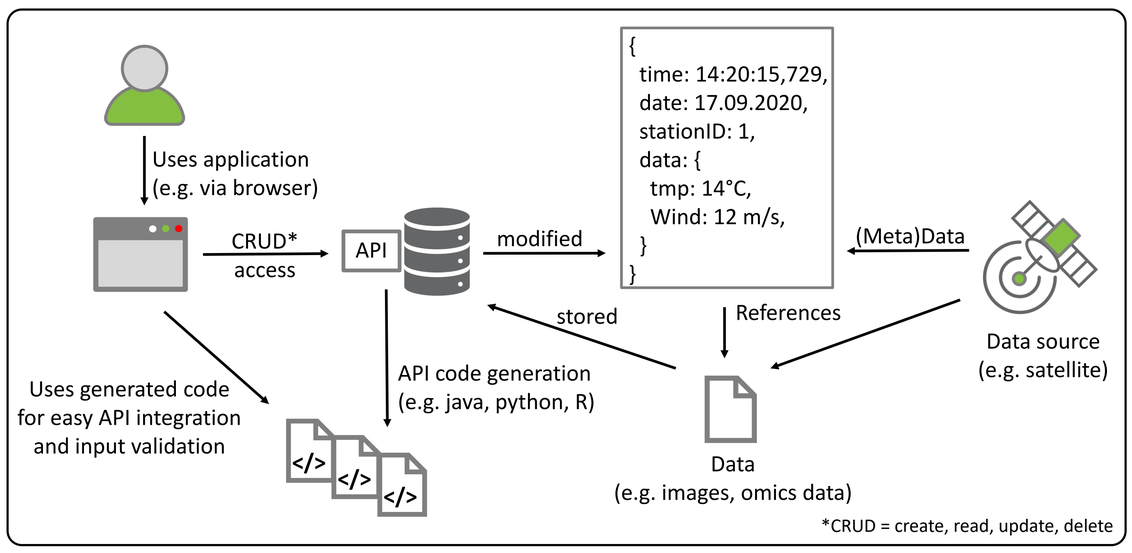
de.NBI Cloud Activities at JLU Giessen

Besides the ready-to-use software applications, we have direct access to a large-scale cloud computing infrastructure that is operated by the Bioinformatics Core Facility (BCF) at JLU. With more than 130 external users and round about 100 projects, this system was used with a total of almost 40 million vCPU hours over the last 12 months.



IT infrastructure for data management, bioinformatics and collaboration hosted by the BCF at JLU Giessen.

Project Data Management System



Generic project data management system at JLU Giessen.

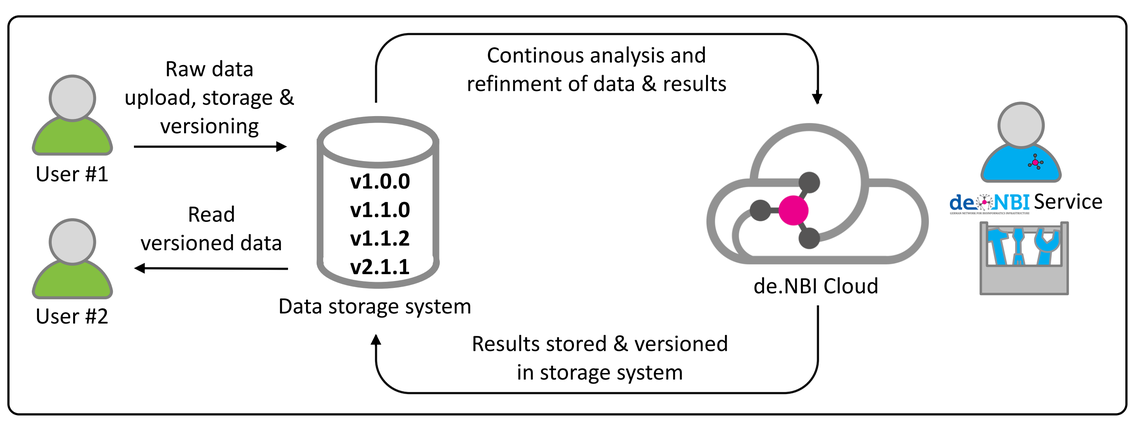
The development of a generic data management system is part of the recently funded NFDI4BioDiversity project to handle data generated within this initiative. It is designed to support a state-of-the-art data lifecycle management that can deal with petabytes of data and millions of data objects. The data stored can be subdivided into two types, (i) the data itself which is treated as binary objects and stored in our object storage and (ii) its corresponding metadata that will be stored in JSON format flat files.

The metadata files will be stored in a document database and indexed to allow efficient querying at any time. The individual data objects are versioned using an adapted variant of semantic versioning. Once data is placed in the system, it can be stored as immutable records. The system is designed to accommodate data from a wide variety of different backgrounds.

We will also build an additional component that connects the individual layers of this system to make data searchable across multiple instances and to allow site-local caching to avoid data transfer overhead. Login is provided by implementing OAuth2 for various SSO providers.

Data life Cycle Management

Modern data management requires a defined data lifecycle management that will be supported implicitly by our system. Within this lifecycle, data is not only generated, analyzed and stored once, but can be part of a continuous refinement process. Within this process the data is analyzed multiple times and the results are typically constantly improved. Depending on the requirements, all or subsets of the analysis output can be stored along with well-defined version numbers to support traceability of results. Changes at different positions in the version number will be used to indicate the impact of each update, ranging from (i) small bugfixes and (ii) major changes in the dataset due to the incorporation of additional data, (iii) to breaking changes in the data format. Within the de.NBI network, major steps of such analyses can be performed by using storage and compute resources of the de.NBI cloud, and by applying tools that are built to support and utilize cloud computing resources.



Overview of the data life cycle management.

Contributions to NFDI Consortia

The Goesmann group is a major partner in two NFDI initiatives: NFDI4Biodiversity and NFDI4Microbiota. For these research consortia, we (i) provide cloud resources via the de.NBI cloud, (ii) operate a Kubernetes cluster on these resources and we (iii) provide and support a variety of use cases. We are also responsible for the software architecture and the development of the core storage systems as described above. Besides that, we also support other partners to use the provided infrastructure, both the core storage layer and the cloud infrastructure.

Publications

[1] Schwengers *et al.* (2020) ASA³P: An automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. Microbial genomics, 6(10, DOI:10.1099/mgen.0.000398
[2] Meyer F *et al*. (2003) GenDB – An open-source genome annotation system for prokaryote genomes. Nucleic Acids Research 31(8): 2187-2195, DOI:10.1093/nar/gkg312
[3] Blom et al. (2019) EDGAR: a versatile tool for phylogenomics. Bergey's Manual of Systematics of Archaea and Bacteria.
[4] Jaenicke *et al.* (2018) Flexible metagenome analysis using the MGX framework. *Microbiome* 6(1): 76
[5] Hilker R *et al*. (2016) ReadXplorer 2 – detailed read mapping analysis and visualization from one single source. Bioinformatics, 32(24): 3702-3708, DOI:10.1093/bioinformatics/btw541

More than 150 (co)-authored peer reviewed publications and more than 50 acknowledgements have been published involving the BIGI de.NBI Giessen scientists.

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
Projektträger Jülich Forschungszentrum Jülich
DFG
www.denbi.de