Optimization of data life cycles

2014 J. Phys.: Conf. Ser. 513 032047

(http://iopscience.iop.org/1742-6596/513/3/032047)

Optimization of data life cycles

C Jung[1], M Gasthuber[2], A Giesler[3], M Hardt[1], J Meyer[1], F Rigoll[1], K Schwarz[4], R Stotzka[1] and A Streit[1]

[1] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[2] Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany
[3] Forschungszentrum Jülich, Jülich, Germany
[4] GSI Helmholtz Centre for Heavy Ion Research, Darmstadt, Germany

E-mail: christopher.jung@kit.edu, martin.gasthuber@desy.de, a.giesler@fz-juelich.de, marcus.hardt@kit.edu, joerg.meyer2@kit.edu, fabian.rigoll@kit.edu, k.schwarz@gsi.de, rainer.stotzka@kit.edu, achim.streit@kit.edu

Abstract. Data play a central role in most fields of science. In recent years, the amount of data from experiment, observation, and simulation has increased rapidly and data complexity has grown. Also, communities and shared storage have become geographically more distributed. Therefore, methods and techniques applied to scientific data need to be revised and partially be replaced, while keeping the community-specific needs in focus.

The German Helmholtz Association project ”Large Scale Data Management and Analysis” (LSDMA) aims to maximize the efficiency of data life cycles in different research areas, ranging from high energy physics to systems biology. In its five Data Life Cycle Labs (DLCLs), data experts closely collaborate with the communities in joint research and development to optimize the respective data life cycle. In addition, the Data Services Integration Team (DSIT) provides data analysis tools and services which are common to several DLCLs. This paper describes the various activities within LSDMA and focuses on the work performed in the DLCLs.

1. Introduction

1.1. Motivation

Data play a central role in the scientific life cycle (see figure 1). In fact, they have their own life cycle, ranging from data acquisition via analysis to archival. For researchers, the main goal is data analysis.

As scientific data can usually be reproduced not at all (e.g. observational data) or under (high) costs, their value is obvious. With data deluge hitting modern science, their exploration is recognized as the fourth pillar in modern science [1]. The challenges of data exploration can be described by the three Vs of Big Data: volume, velocity and variety [2].

1.2. Idea and concept of LSDMA

The Helmholtz Association portfolio extension ‘Large Scale Data Management and Analysis’ (LSDMA, [3]) is a project focusing on research and development for the whole scientific data life cycle. Its dual approach covers (see figure 2):

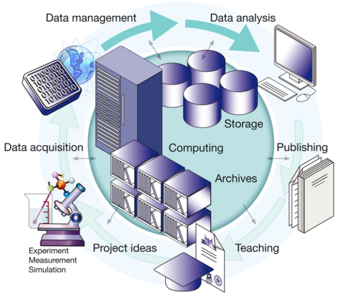

Figure 1: The scientific life cycle (outer circle) includes the data life cycle (ranging from data acquisition via analysis to archival).

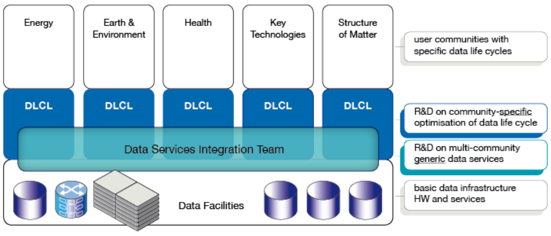
Energy
Earth & Environment
Health
Key Technologies
Structure of Matter
user communities with specific data life cycles
DLCL
DLCL
DLCL
DLCL
DLCL
R&D on community-specific optimisation of data life cycle
Data Services Integration Team
R&D on multi-community generic data services
Data Facilities
basic data infrastructure HW and services

Figure 2: The LSDMA project structure.

- Five Data Life Cycle Labs (DLCLs) engage in joint R&D with their respective communities to optimize their data life cycles and develop community-specific data analysis tools and services.
- The R&D of the cross-sectional Data Services Integration Team (DSIT) focuses on data analysis tools and services common to several DLCLs. Also, it serves as an interface between federated data infrastructures and the DLCLs/communities.

The project started in 2012 and is initially funded until the end of 2016. Its partners comprise four Helmholtz Association research centers (DESY, FZ Jülich, GSI and KIT) and seven external partners (HTW Berlin, TU Dresden, University of Frankfurt, DKRZ Hamburg, University of Heidelberg, University of Hamburg and University of Ulm).

The activities of LSDMA will be included in the sustainable program-oriented funding of the Helmholtz Association in 2015 as a ‘cross-program initiative’.

2. DSIT

The activities of DSIT are subdivided into six work packages, which are described in the following subsections.

2.1. Federated identity management

This work package focuses on providing a secure, simple to use, and at the same time flexible and open authentication and authorization infrastructure (AAI). In an initial workshop, relevant solutions were presented. For a good security standard and flexibility, the decision was taken to expose only Security Assertion Markup Language (SAML [4]) based interfaces to the user. For integration of non-SAML based services, such as services based on X.509 [5], OpenID [6] or OAuth2 [7], activities focus on credential translation. This credential translation service, which is currently being implemented, will allow authentication via SAML and the optional creation of authentication tokens for users. Especially for X.509 special care is taken to ensure the existing high trust level.

2.2. Federated data access

The work package focuses on the challenges arising from data-intensive experiments, which have specific performance requirements. This work involves performance tuning of services, e.g. dCache [8] and other scalable object stores. The supported protocol for object stores will be the open Storage Networking Industry Association [9] standard Cloud Data Management Interface (CDMI) [10].

Another task is the simplification of data access. This is done on two levels. On the lower level, federated file system approaches like FedFS with http/WebDAV or NFS back ends are considered, while on the higher level graphical and web based user interfaces like Globus Online [11] and the KIT Data Manager [12] are taken into consideration.

Moreover, security aspects are being investigated. A Linux kernel module for Lustre has been developed which allows access control based on network protocols as well as on user identifier (UID) and group identifier information (GID) of the accessing user, providing control mechanisms for federated data access and metropolitan area networks which so far were only applicable in LAN environments.

2.3. Metadata catalogs and repositories

Since metadata requirements are very community specific the main goal is to develop and offer common services. The focus is on federated search capabilities on two levels. This is integrated in the UNICORE middleware [13] to allow for search over an arbitrary number of UNICORE metadata instances. Besides the UNICORE integration, The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, [14]) infrastructures are being designed and implemented. OAI-PMH allows for reading of metadata without knowledge about the actual structure or content; it is used for so-called harvesting possibly across a variety of communities. The collected data are stored and can be further accessed by the Apache Solr project[15] for searching in the previously harvested data.

2.4. Archive service

The focus is on two topics. One is the creation of policies and service levels which are needed to decide where, whether and how long datasets are to be archived. This is important, since archiving costs are a significant factor for given data volumes. Also, different data require different qualities of service, since simulation data can be reproduced, while observed data cannot.

The other topic is moving towards building services required to provide an archive service. These are bitstream preservation, content preservation and data curation. All three represent different technologies and responsibilities. While bitstream preservation is typically provided by IT centers, content preservation and curation fall into the responsibility of the user community. The bitstream preservation is addressed within DSIT, the remaining two are envisaged to be pursued in cooperation with the communities. In order to provide bitstream preservation, different archival middlewares are evaluated. There is close contact with the library community, which has established sophisticated workflows.

2.5. Monitoring, modeling, optimization

Profiling tools are developed and analyze of the performance of specific software components relevant to LSDMA are performed. These include methods for monitoring the I/O performance for analysis of performance loss between subsystems as well as monitoring in a heterogeneous and distributed environment. In order to implement this, the SIOX [16] framework is used as a basis. On top, Java components that can create tracing information are developed. Such trace files can be visualized using the VAMPIR [17] framework, known from Message Passing Interface (MPI) trace visualization. Another activity is the use of monitoring information within LUSTRE for the automatic detection of performance degradation in order to find, analyze and understand bottlenecks better.

2.6. Data-intensive computing

User-friendly tools for the analysis of measured data are developed. The focus is on automatic and policy based triggering of analysis workflows. One example is the automatic metadata extraction, after successful data ingest. This is supported by the new data oriented capabilities of the UNICORE middleware. The LAMBDA [12] tool, an execution framework for large scale applications based on metadata control and data processing events, such as data uploaded and time expired, is developed.

3. DLCLs

In this section, selected highlight activities of the DLCLs are presented.

3.1. DLCL Earth and Environment

The work group for satellite-borne remote sensing of trace gases at KIT stores satellite data from various research satellites on KIT's Large Scale Data Facility. There are data from about 40 satellites including data from the Michelson Interferometer for Passive Atmospheric Sounding (MIPAS, [18]), one of the instruments of the Envisat satellite for the detection of limb emission spectra in the middle and upper atmosphere. The goal of the DLCL is to design and setup a database for metadata of various measurements. This database should simplify analysis workflows by providing a common database instead of data in various database and/or file formats. As there are millions of geolocations of satellite measurements the database needs to be able to scale out, i.e. it should be possible to increase the number of parallel read queries that can be handled in a reasonable amount of time by adding more server instances. Performance measurements showed that this requirement is fulfilled by the NoSQL database MongoDB [19].

The analysis of climate data often requires the transfer of data from one site to another. For example, researchers at KIT who want to use computing resources at DKRZ need to transfer their input data to DKRZ and then need to transfer their results back to KIT. Therefore, DKRZ and KIT set up a storage federation that allows a policy-based replication of data between sites. The idea is that users specify via an easy-to-use web interface where they want their data to be located. The actual transfers are executed by iRODS servers [20] in the background.

3.2. DLCL Energy

In the project Solar Irradiation System Karlsruhe (SISKA, [21]), two-dimensional aerial images are the basis for calculations of three-dimensional point clouds. From these results efficiency estimations of photovoltaic plants on roof tops are calculated. The workflow comprises a chain of tools written in different programming and scripting languages. During the data life cycle analysis and optimization, parts of the calculations were ported to the Large Scale Data Facility (LSDF, [22]) cluster to speed up the calculations. Besides the increase in CPU power, the increase in available space enables calculations of larger areas.

The results from data life cycle analysis in various projects show that privacy concerns are a very common issue. Energy data such as measurements of energy consumption in a smart building, or tracking data from an electric vehicle almost always yield information that can be linked to a person. Standard anonymization or pseudonymization techniques often cannot be applied and there is always a trade-off between usability of data and anonymization. Therefore, data privacy has become the new focus of this DLCL.

3.3. DLCL Health

Special emphasis is placed on the collaboration with the Institute of Neuroscience and Medicine (INM-1) of the Forschungszentrum Jülich GmbH. This community tries to understand the anatomical connections of cortical areas and subcortical nuclei which are interconnected via short- and long distance fiber tracts. The three-dimensional polarized light imaging (PLI) technique [23] represents a novel neuroimaging way to map nerve fibers and their pathways in human postmortem brains with a resolution at the sub-millimeter scale. This modern imaging method produces a huge amount of raw data and post-processed image data demanding a wide ranging data management in the areas of data acquisition, data transfers, storage management, and also metadata. The cooperative work with the community starts for the DLCL by establishing a secure and performant transport mechanism for data, generated in microscopic images. Tools and services like the UFTP [24] daemon, developed at the Jülich Supercomputing Centre (JSC), are deployed to enable a high throughput image transfer on wide area network connections. Compared to standard tools like scp, up to four times higher transmission rates are obtained. The produced data are stored on a fast General Parallel File System (GPFS, [25]) cluster at JSC designed with consideration for a long-term storage strategy. For efficient I/O operations the data will be stored in an HDF5 data format [26] which maps the deep structure of the image data and their dependencies amongst themselves and allows queries based on metadata attributes.

3.4. DLCL Key Techologies

The synchrotron light source ANKA [27] requires many resources in the fields of data-intensive computing and large-scale data management. The tomography beam line is used for testing of new materials and investigation of tiny biological structures. Novel extensions allow ultra-fast imaging with spatio-temporal observations even in living species. Topo-Tomo is the most data-intensive ANKA beam line capable to produce 30-100 TB in a week, where the size of the raw data of a single experiment may exceed 500 GB. Most data storage and access operations are directly steered by the user, resulting in a data flow that is hardly transparent and reproducible. A new concept to manage the data life cycle including transparent user workspaces was required and is being implemented [28]. It allows to steer the data flow and to ingest the data and its descriptive metadata into the LSDF. Ultra-fast and time resolved imaging requires vast computing resources for image reconstruction. Firstly, GPGPU systems with specially adapted back-projection algorithms allow a quick preview of the recorded volumes in near real-time. Secondly, in ultra-fast tomography only a few hundred projections are recorded requiring new reconstruction methods. At KIT a new algorithm has been developed producing high quality images using only 150 projections. Since the algorithms are very demanding in computation and data throughput, the LSDF HADOOP computing environment and tools of the KIT Data Manager [12] are used to reconstruct a volume in approximately two minutes [29].

Light Sheet Microscopy is a novel high throughput microscopy technology [30] resulting in huge image data sets when performing time-lapse studies to follow processes of zebrafish development. For example, two 3D volumes of a zebrafish embryo can be taken in less than one minute at highest resolution. Time-lapse imaging of one embryo takes ten hours and produces up to 16 TB of data. As measurements are performed on a regular basis, the data have to be processed within one to two days to avoid accumulation of open processing tasks. With tools developed within the KIT Data Manager, the data and descriptive metadata are ingested to the LSDF data center crossing KIT campus with an average transfer rate of more than 400 MB/s. In a single experiment 250 TB were produced within six weeks and successfully stored in the LSDF at the end of 2012.

3.5. DLCL Structure of Matter

This DLCL collaborates with researchers from high energy physics (incl. heavy ion physics) and photon science. These research fields have several similarities, such as triggering and filtering for data reduction, monitoring of data access and data popularity, need for scalable storage and fast access as well as multithreaded analysis code. A major difference is the time span during which the data are used and re-used for analysis (years vs. months) as well as the size of collaborations (hundreds to thousands vs. a few to tens).

In order to construct short-term solutions in photon science, i.e. data recording for high datarate beamlines, the development of a dedicated, yet unnamed simulation tool (for 'hammering' real storage and networks) helped to determine the critical aspects and further optimize the solution; this includes I/O tracing for detailed studies of storage systems used. A more mid- to long-term engagement is the upcoming baseline support for HDF5 (persistency model and data format selected by nearly all local photon science experiments) and parallel execution and programming.

At the Facility for Antiproton and Ion Research (FAIR, [31]), a novel triggerless detector read-out will be implemented. Data streams of up to 1 TB/s are expected. In order to enable the simulation and analysis framework [32] to perform high speed (online) data processing in real time, the messaging library ZeroMQ [33] has been integrated. Several close by sites will be connected with a high speed metropolitan area network via fibre link. By using an LNET router cluster with twelve nodes as interface between the Infiniband HPC system and the Ethernet WAN, a speed of 14 GB/s has been achieved.

For integrating the infrastructure into an international Grid/Cloud environment (e.g. PandaGrid [34]), storage interfaces have been created consisting of xrootd daemons running on top of the GSI Lustre cluster.

4. Experiences

The project structure of LSDMA was wisely chosen. The initial phase of the project emphasized the different characteristics of the communities. In particular, the communities differed in previous knowledge on data management, the level of specification of the data life cycle as well as tools, services and formats used. Their needs are not only driven by the three Vs of Big Data but also by the need of cooperations between different groups worldwide and by policies, such as Open Access/Open Data, long-term preservation of data used for publications (e.g. [35]) and data privacy issues.

As mentioned before, the communities' main focus is data analysis. In on-going experiments, researchers are reluctant to change their data life cycle, yet Big Data and data policies oblige them to. So optimizing the data life cycle usally is an evolutionary, not a revolutionary process.

Two of the policies represent major challenges. Data privacy involves several legal issues; as many collaborations are international, the situation becomes even more complicated. Though basically everyone agrees on the value of scientific data, the technical and organizational requirements for preservation are complemented by the open issue of funding.

For working in collaborations, AAI is an essential tool. There are different solutions being used by different communities, disciplines and projects. With collaborations growing and interdisciplinarity becoming increasingly important, it is crucial to make the different AAI solutions interoperable.

Acknowledgements

The authors wish to thank all people and institutions involved in LSDMA as well as the German Helmholtz Association for funding the project.

References

[1] Hey T, Tansley S and Tolle K (eds) 2009 (Redmond, Washington: Microsoft Research) URL `http://research.microsoft.com/en-us/collaboration/fourthparadigm/`

[2] Laney D 2001 3D data management: Controlling data volume, velocity, and variety Tech. rep. META Group

[3] van Wezel J *et al.* 2012 Data Life Cycle Labs, A New Concept to Support Data-Intensive Science (*Preprint* `arXiv:1212.5596`)

[4] Cantor S *et al.* 2005 Assertions and protocols for the oasis security assertion markup language (saml) v2.0 URL `http://docs.oasis-open.org/security/saml/v2.0/`

[5] Internet X.509 Public Key Infrastructure: Certification Path Building (last visited on 2013-12-16) URL `http://tools.ietf.org/html/rfc4158`

[6] OpenID Foundation http://openid.net (last visited 2013-06-20)

[7] OAuth Protocol Homepage http://www.oauth.net (last visited 2013-06-20)

[8] Millar A P *et al.* 2012 *Journal of Physics: Conference Series* **396** 032077

[9] Storage Networking Industry Association Homepage http://www.snia.org/ (last visited 2013-12-16)

[10] SNIA webpage on CDMI (last visited on 2013-10-09) URL `http://www.snia.org/tech_activities/standards/curr_standards/cdmi`

[11] Foster I 2011 *IEEE Internet Computing* **15** 70–73 ISSN 1089-7801

[12] Jejkal T *et al.* 2012 *PDP* ed Stotzka R, Schiffers M and Cotronis Y (IEEE) pp 213–220 ISBN 978-1-4673-0226-5

[13] UNICORE Homepage http://www.unicore.eu/ (last visited 2013-12-16)

[14] Open Archives Initiative Protocol for Metadata Harvesting webpage (last visited on 2013-10-10) URL `http://www.openarchives.org/pmh/`

[15] Apache Solr Homepage http://lucene.apache.org/solr/ (last visited 2013-12-16)

[16] Wiedemann M C *et al.* 2013 *Computer Science - Research and Development* **28** 241–251 ISSN 1865-2034, 1865-2042

[17] Vampirtrace website http://www.tu-dresden.de/zih/vampirtrace, last visited on 2013-10-15

[18] Fischer H *et al.* 2008 *Atmospheric Chemistry and Physics* **8** 2151–2188

[19] MongoDB homepage (last visited on 2013-10-08) URL `http://www.mongodb.org/`

[20] iRODS homepage (last visited on 2013-10-09) URL `https://www.irods.org`

[21] SISKA webpage (last visited on 2013-10-09) URL `http://www.ipf.kit.edu/english/current_projects_SISKA.php`

[22] García A *et al.* 2011 *IPDPS Workshops* (IEEE) pp 1467–1474 ISBN 978-1-61284-425-1

[23] Axer M *et al.* 2011 *Frontiers in Neuroinformatics* **5** 1–13 ISSN 1662-5196

[24] UFTP User Manual (last visited on 2013-10-08) URL `http://www.unicore.eu/documentation/manuals/unicore6/files/uftp/uftp-manual.html`

[25] IBM General Parallel File System http://www-03.ibm.com/systems/software/gpfs (last visited 2013-12-16)

[26] HDF5 webpage (last visited on 2013-10-08) URL `http://www.hdfgroup.org/HDF5/`

[27] ANKA Homepage http://www.anka.kit.edu/28.php (last visited 2013-12-16)

[28] Yang X *et al.* 2013 *16th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP 2008)* **0** 86–93 ISSN 1066-6192

[29] Yang X *et al.* 2013 *16th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP 2008)* **0** 86–93 ISSN 1066-6192

[30] Mikut R *et al.* 2013 **10** 401–421 pMID: 23758125

[31] FAIR webpage (last visited on 2013-10-09) URL `http://www.fair-center.eu/`

[32] Al-Turany M *et al.* 2012 *Journal of Physics: Conference Series* **396** 022001

[33] ZeroMQ (last visited on 2013-10-09) URL `http://zeromq.org/`

[34] Protopopescu D and Schwarz K 2011 *Journal of Physics: Conference Series* **331** 072028

[35] Deutsche Forschungsgemeinschaft, Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme, Unterausschuss füur Informationsmanagement 2009 Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten