e-Infrastructure Requirements for Big Data Science

Yuri Demchenko, Paola Grosso, Cees de Laat

University of Amsterdam, Institute for Informatics, Science Park 904, 1098 XH Amsterdam, NL

Keywords: Big Data Science, Scientific Data e-Infrastructure, Scientific Data Lifecycle

1. Introduction

Modern e-Science infrastructure allows targeting new large scale problems which solution was not possible before, e.g. genome, climate, global warming, etc. e-Science typically produces a huge amount of data that need to be supported by a new type of e-Infrastructure capable to store, distribute, process, preserve, and curate these data: we refer to this new infrastructures as Scientific Data e-Infrastructure (e-SDI) [1].

This paper analyses the new challenges imposed to modern e-Science infrastructures by the emerging big data technologies and proposes general approach and architecture solutions that introduce the new Scientific Data Lifecycle (SDLC) management model and summurises requirements to SDI architecture model that provides a basis for heterogeneous e-SDI components interoperability and integration, in particular based on cloud infrastructure technologies.

2. Data Management in Big Data Science

Emergence of computer aided research methods is transforming the way how research are done and scientific data are processed/used. The following types of scientific data are defined [1, 2]: (1) raw data collected from experiment; (2) structured data and datasets after initial processing, classification and filtering; (3) published data that supports one or another scientific hypothesis, research result or statement; (4) data linked to publications to support the wide research consolidation, integration, and openness.

New approach to data management and handling in e-Science is reflected in the Scientific Data Lifecycle Management (SDLM) model (see Fig. 1) proposed by the authors as a result of analysis of the existing practices in different scientific communities. The proposed model is compliant with the data lifecycle study results presented in [3].

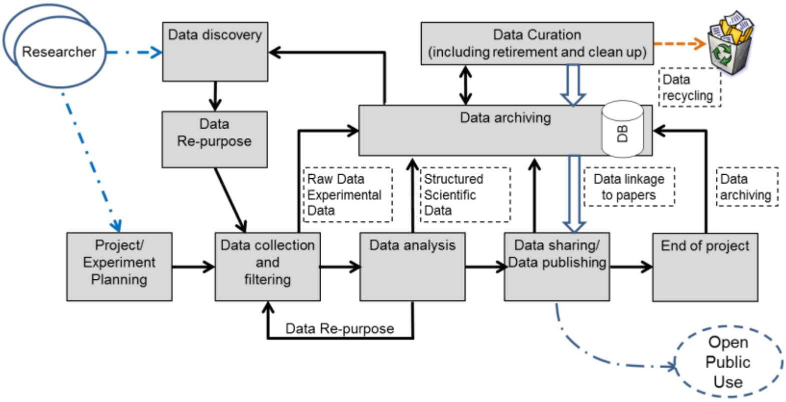

Fig. 1. Scientific Data Lifecycle Management in e-Science

The generic scientific data lifecycle includes a number of consequent stages that include: research project or experiment planning; data collection; data integration and processing; publishing research results; discussion, feedback; archiving (or discarding). New SDLM requires data storage and preservation at all stages what should allow data re-use/re-purposing and secondary research on the processed data and published results. However, this is possible only if the full data identification, cross-reference and linkage are implemented in SDI. Data integrity, access control and accountability must be supported during the whole data during lifecycle. Data curation is an important component of the discussed SDLC and must also be done in a secure and trustworthy way.

3. General SDI requirements

The emerging e-SDI should allow different groups of researchers to work on the same data sets, build own (virtual) research and collaborative environment, safely store intermediate results, and later share the discovered results. New data provenance, security and access control mechanisms and tools should allow researchers to link their scientific results with the initial data (sets) and intermediate data to allow future re-use/re-purpose of data e.g. with the improved research technique and tools.

The future e-SDI should support the whole data lifecycle and explore the benefit of the data storage/preservation, aggregation and provenance in a large scale and during long/unlimited period of time. Important is that this infrastructure must ensure data security (integrity, confidentiality, availability, and accountability), and data ownership protection. With current needs to process big data that require powerful computation, there should be a possibility to enforce data/dataset policy that they can be processed on trusted systems and/or complying other requirements. Researchers must trust the e-SDI to process their data on e-SDI facilities and be ensured that their stored research data are protected from non-authorised access. Privacy issues must be also addressed at infrastructure level.

4. Future research and development

The future research and development will include further SDLM definition, e-SDI components definition, including access control and accounting infrastructure that should allow trustworthy use of the shared infrastructure by researchers. Special attention will be given to defining the whole cycle of the provisioning e-SDI services on-demand specifically tailored to support instant scientific workflows using cloud IaaS and PaaS platforms. This research will re-use the results from other infrastructure oriented projects: GN3 Composable Services developing GEMBus (GEANT Multidomain Bus) cloud PaaS platform, and GEYSERS Intercloud architecture for on-demand infrastructure services provisioning.

Acknowledgments. The research leading to these results was partly funded by the EC FP7 project GEANT under Grant Agreement No. 238875.

References

1. Global Research Data Infrastructures: Towards a 10-year vision for global research data infrastructures. Final Roadmap, March 2012. [online] http://www.grdi2020.eu/Repository/FileScaricati/6bdc07fb-b21d-4b90-81d4-d909fdb96b87.pdf
2. Advancing technologies and Federating communities: Study on Authentication, Authorization and Accounting (AAA) Platforms For Scientific data/information Resources in Europe. [online] https://confluence.terena.org/download/attachments/30474266/AAA-Study-Report-0907.pdf
3. Data Lifecycle Models and Concepts. [online] http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v8.docx