Big Data Challenges for e-Science Infrastructure

Yuri Demchenko[1], Zhiming Zhao[1], Paola Grosso[1],
Adianto Wibisono[1], Cees de Laat[1]

[1] System and Network Engineering Group, University of Amsterdam
Science Park 904, 1098XH Amsterdam, The Netherlands
{y.demchenko, z.zhao, p.grosso, a.wibisono, C.T.A.M.deLaat}@uva.nl

Abstract. This paper discusses the challenges that are imposed by the Big Data Science on the modern and future Scientific Data Infrastructure (SDI). The paper refers to different scientific communities to define requirements on data management, access control and security. The paper introduces the Scientific Data Lifecycle Management (SDLM) model that includes all the major stages and reflects specifics in data management in modern e-Science. The paper proposes the SDI generic architecture model that provides a basis for building interoperable data or project centric SDI using modern technologies and best practices. The paper explains how the proposed models SDLM and SDI can be naturally implemented using modern cloud based infrastructure services provisioning model. The paper also addresses issues with the federated access control to the SDI resources that provides a flexible access control and identity management model for scientific and research communities.

Keywords: Big Data Science, Scientific Data Infrastructure (SDI), Scientific Data Lifecycle Management (SDLM), Cloud Infrastructure Services.

1 Introduction

Modern e-Science infrastructure allows targeting new large scale problems that were not possible before like genome, climate, global warming, etc. e-Science typically produces a huge amount of data that need to be supported by a new type of Scientific Data e-Infrastructure (SDI) to store, distribute, process, preserve, and curate these data [1, 2].

In e-Science, the scientific data are complex multifaceted objects with the complex internal relations, they are becoming an infrastructure of their own and need to be supported by corresponding physical or logical infrastructures to store, access and manage these data.

The emerging SDI should allow different groups of researchers to work on the same data sets, build their 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 will 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 presented paper analyses new challenges imposed to modern e-Science infrastructures by emerging big data technologies and proposes general approach and

architecture solutions that introduce the new Scientific Data Lifecycle Management (SDLM) model and the generic SDI architecture model that provides a basis for heterogeneous SDI components interoperability and integration, in particular based on cloud infrastructure technologies.

The paper is organised as follows. Section 2 provides of the main research communities and summarises requirement to future SDI. Section discusses challenges to data management in Big Data Science, including SDLM discussion. Section 4 introduces the proposed e-SDI architecture model that is intended to answer the future big data challenges and requirements. Section 5 discusses SDI implementation using cloud technologies. Section 6 discusses specific requirements and provides suggestions about building federated access control infrastructure for modern and future SDI.

2 General Requirements to Big Data e-Science Infrastructure

2.1 Paradigm change in Big Data Science

Big Data Science is becoming a new technology driver and requires re-thinking a number of infrastructure components, solutions and processes to address the following general challenges [2, 3]:

- Exponential growth of data volume produced by different research instruments and/or collected from sensors
- Need to consolidate e-Infrastructure as persistent research platform to ensure research continuity and cross-disciplinary collaboration, deliver/offer persistent services, with adequate governance model.

The recent advancements in the general ICT and big data technologies facilitate the paradigm change in modern e-Science that is characterized by the following features:

- Automation of all e-Science processes including data collection, storing, classification, indexing and other components of the general data curation and provenance
- Transformation all processes, events and products into digital form by means of multi-dimensional multi-faceted measurements, monitoring and control; digitising existing artifacts and other content.
- Possibility to re-use the initial and published research data with possible data re-purposing for secondary research
- Global data availability and access over network for cooperative group of researchers, including wide public access to scientific data.
- Existence of necessary infrastructure components and management tools that allows fast infrastructures and services composition, adaptation and provisioning on demand for specific research projects and tasks.
- Advanced security and access control technologies that ensure secure operation of the complex research infrastructures and scientific instruments and allow creating trusted secure environment for cooperating groups and individual researchers

The future 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 SDI to process their data on SDI facilities and be ensured that their stored research data are protected from non-authorised access. Privacy issues are also arising from distributed remote character of SDI that can span multiple countries with different local policies. This should be provided by the Access Control and Accounting Infrastructure (ACAI) which is an important component of SDI [4, 5].

2.2 Research communities and SDI requirements

In this section we provide short overview of the research infrastructures and communities, in particular defined for the Europe Research Area (ERA) [3] and analyse their specific requirement to future SDI to address Big Data challenges.

We refer to existing studies of the European e-Infrastructures that analyse the scientific communities practices and requirements such as those undertaken by the SIENA Project [6], EIROforum Federated Identity Management Workshop [5], European Grid Infrastructure (EGI) Strategy Report [7], UK Future Internet Strategy Group Report [8].

The **High Energy Physics** community represents a large number or researchers, unique expensive instruments, huge amount of data that are generated and need to be processed continuously. This community has already the operational Worldwide Large Hadron Collider Grid (WLCG) [9] infrastructure to manage and access data, protect their integrity and support the whole scientific data lifecycle. WLCG development was an important step in the evolution of European e-Infrastructure that currently serves multiple scientific communities in Europe and internationally. The EGI cooperation [7] manages European and worldwide infrastructure for HEP and other communities.

Material science, analytical and low energy physics (proton, neutron, laser facilities) is characterised by short projects, experiments and consequently highly dynamic user community. It requires highly dynamic supporting infrastructure and advanced data management infrastructure to allow wide data access and distributed processing.

Environmental and Earth science community and projects target regional/national and global problems. They collect huge amount of data from land, sea, air and space and require ever increasing amount of storage and computing power. This SDI requires reliable fine-grained access control to huge data sets, enforcement of regional issues, policy based data filtering (data may contain national security related information), while tracking data use and keeping data integrity.

Biological and Medical Science (also defined as **Life science**) have a general focus on health, drug development, new species identification, new instruments development, it generates massive amount of data and new demand for computing

power, storage capacity, and network performance for distributed processes, data sharing and collaboration. Biomedical data (healthcare, clinical case data) are privacy sensitive data and must be handled according to the European policy on Personal Data processing [10].

Social Science and Humanities community and projects are characterised by multi-lateral and often global collaboration between researcher from all over the world which need to be engaged into collaborative groups/communities and supported by collaborative infrastructure to share data, discovery/research results and cooperatively evaluate results. Current trend to digitise all currently collected physical artifacts will create in the near future a huge amount of data that must be widely and openly accessible.

The following are general infrastructure requirements to SDI for emerging Big Data Science:

- Support long running experiments and large data volumes generated at high speed
- Multi-tier data distribution and replication
- Support of virtual scientists communities
- Trusted environment for data storage and processing
- Data integrity, confidentiality, accountability
- Policy binding to data to protect privacy

3 Data Management in Big Data Science

Emergence of computer aided research methods is transforming the way how research are done and scientific data are used. The following types of scientific data are defined [4]:

- Raw data collected from observation and from experiment (according to an initial research model)
- Structured data and datasets that went through data filtering and processing (supporting some particular formal model)
- Published data that supports one or another scientific hypothesis, research result or statement
- Data linked to publications to support the wide research consolidation, integration, and openness.

Once the data is published, it is essential to allow the other scientists to be able to validate and reproduce the data that they are interested in, and possibly contribute with new results. Capturing information about the processes involved in transformation from raw data up until the generation of published data, becomes an important aspect of scientific data management. Scientific data provenance becomes an issue that also needs to be taken into consideration by SDI providers [11].

Another aspect to take into consideration is to guarantee reusability of published data within the scientific community. Understanding semantic of the published data becomes an important issue to allow for reusability, and this had been traditionally been done manually. However, as we anticipate unprecedented scale of published data that will be generated in Big Data Science, attaching clear data semantic becomes a necessary condition for efficient reuse of published data. Learning from best practices

in semantic web community on how to provide a reusable published data, will be one of consideration that will be addressed by SDI.

Big data are typically distributed both on the collection side and on the processing/access side: data need to be collected (sometimes in a time sensitive way or with other environmental attributes), distributed and/or replicated. Linking distributed data is one of the problems to be addressed by SDI.

The European Commission's initiative to support Open Access to scientific data from publicly funded projects suggests introduction of the following mechanisms to allow linking publications and data [12, 13]:

- PID - persistent data ID
- ORCID – Open Researcher and Contributor Identifier [14].

New approach to data management and handling in e-Science is reflected in the Scientific Data Lifecycle Management (SDLM) model (see Figure 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 [15].

The generic scientific data lifecycle includes a number of consequent stages: research project or experiment planning; data collection; data processing; publishing research results; discussion, feedback; archiving (or discarding).

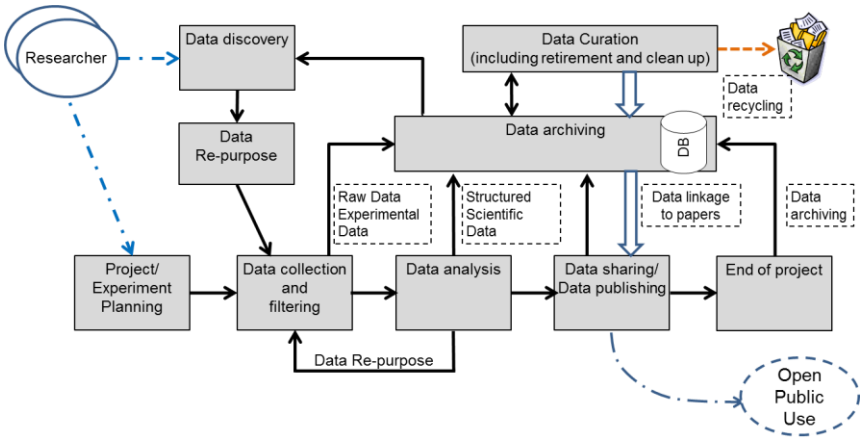

Figure 1. Scientific Data Lifecycle Management in e-Science

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 SDLM and must also be done in a secure and trustworthy way.

Support data security and access control to scientific data during their lifecycle: data acquisition (experimental data), initial data filtering, specialist processing; research data storage and secondary data mining, data and research information archiving.

4 Proposed SDI Architecture Model

The proposed SDI Architecture for e-Science (e-SDI) is illustrated in Figure 2. It contains the following layers:
Layer D1: Network infrastructure layer represented by the general purpose Internet infrastructure and dedicated network infrastructure
Layer D2: Datacenters and computing resources/facilities
Layer D3: Infrastructure virtualisation layer that is represented by the Cloud/Grid infrastructure services and middleware supporting specialised scientific platforms deployment and operation
Layer D4: (Shared) Scientific platforms and instruments specific for different research areas
Layer D5: Federation and Policy layer that includes federation infrastructure components, including policy and collaborative user groups support functionality.
Layer D6: Scientific applications and user portals/clients

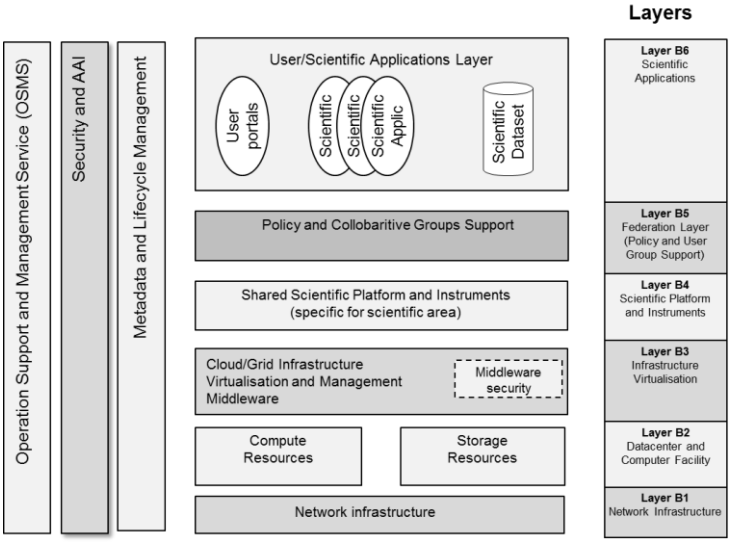

Figure 2. The proposed SDI architecture model

The three cross-layer planes are also defined: Operational Support and Management System; Security plane; and Metadata and Lifecycle Management. •

The dynamic character of SDI and its support of distributed multi-faceted communities are supported by the dedicated layers: D3 – Infrastructure Virtualisation layer that typically uses modern cloud technologies; and D5 – Federation and policy layer that incorporates related federated infrastructure management and access technologies [4, 16, 17]. Introducing the Federation and Policy layer reflects current practice in building and managing complex SDI (and also enterprise infrastructures) and allows independently managed infrastructures to share resources and support the inter-organisational cooperation.

5 Cloud Based Infrastructure Services for SDI

Figure 3 illustrates the typical e-Science or enterprise collaborative infrastructure that is created on demand and includes enterprise proprietary and cloud based computing and storage resources, instruments, control and monitoring system, visualization system, and users represented by user clients and typically residing in real or virtual campuses.

The main goal of the enterprise or scientific infrastructure is to support the enterprise or scientific workflow and operational procedures related to processes monitoring and data processing. Cloud technologies allow to simplify building such infrastructure and provision it on-demand. Figure 3 illustrates how an example enterprise or scientific workflow can be mapped to cloud based services and next deployed and operated as an instant inter-cloud infrastructure. It contains cloud infrastructure segments IaaS (VR3-VR5) and PaaS (VR6, VR7), separate virtualised resources or services (VR1, VR2), two interacting campuses A and B, and interconnecting them network infrastructure that in many cases may need to use dedicated network links for guaranteed performance.

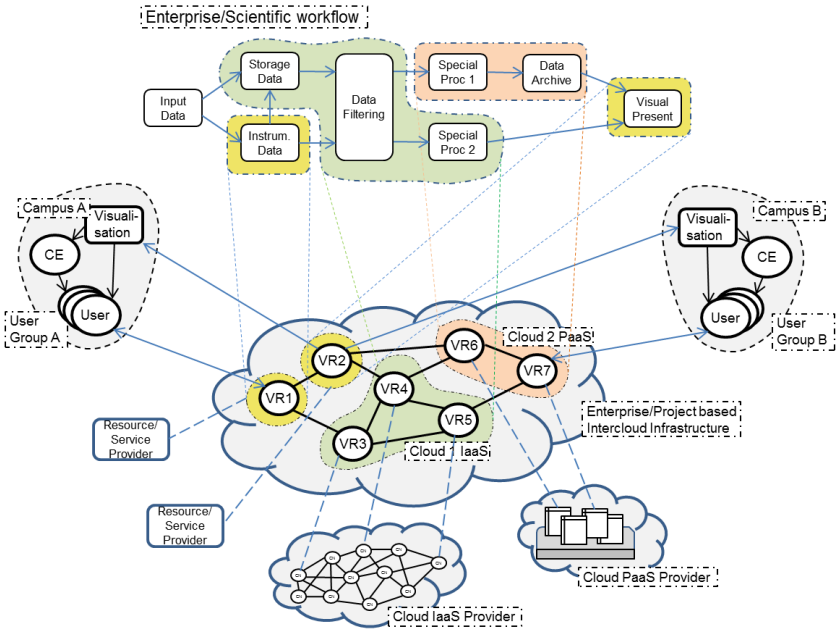

Figure 3. From scientific workflow to cloud based infrastructure.

Efficient operation of such infrastructure will require both overall infrastructure management and individual services and infrastructure segments to interact between themselves. This task is typically out of scope of the existing cloud service provider models but will be required to support perceived benefits of the future e-SDI. These

topics are a subject of another research by authors on the InterCloud Architecture Framework [18, 19].

6 Access Control and Accounting Infrastructure for SDI

6.1 General Requirements to SDI and Access Control Infrastructure

To achieve the suggested functionalities the future Scientific Data e-Infrastructure (SDI) should be supported by a corresponding Access Control and Accounting Infrastructure (ACAI) that would ensure normal infrastructure operation, assets and information protection, and allow user identification/authentication and policy enforcement in distributed multi-organisations environment.

Moving to Open Access [12] may require partial change of business practices of currently existing scientific information repositories and libraries, and consequently the future ACAI should allow such transition and fine grained access control and flexible policy definition and control.

Taking into account that future 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, the future ACAI should also support all stages of the data lifecycle, including policy attachment to data to ensure persistency of the data policy enforcement during continuous online and offline processes.

The required ACAI should support the following features of the future SDI:

- Empower researchers (and make them trust) to do their data processing on shared facilities of large datacentres with guaranteed data and information security
- Motivate/ensure researchers to share/open their research environment to other researchers by providing tools for instantiation of specialised/customised pre-configured infrastructures to allow other researchers to work with existing or own data sets.
- Protect data policy, ownership, linkage (with other data sets and newly produced scientific/research data), when providing (long term) data archiving. (Data preservation technologies should themselves ensure data readability and accessibility with the changing technologies)

6.2 Federated Access Control and Identity Management

Big Data Science communities should explore federation of existing Authentication and Authorisation Infrastructures (AAI) (organisational, community, and national) that allows sharing responsibility and load of managing such federated infrastructure between member communities and research organisations, on one hand, and common federation infrastructure services and policy, including attribute and trust management authorities.

Federated access to e-infrastructures is the recent and most attractive concept – in particular for the users. This approach will eliminate the barriers of e-infrastructure access: the users can use their institutional account to login into a SDI applications

(typically accessed via portal) and access related resources and processes. The portal takes care of identifying the users with their institutional account, mapping this account to a local identity/credentials which are recognised by the cooperating/federated SDI sites and using this credentials to access other sites or resources inside federation on behalf of the user.

Federated access control simplifies the virtual user groups management and should be supported by corresponding federation infrastructure as reflected in the proposed e-SDI architecture model. Federation and policy layer D5 should provide a number a number of functionalities, protocols and interfaces to support its operation:

- Service Registry and Discovery
- Trust and service brokers
- Identity provider (IdP)
- Trust manager/router
- Attribute/namespace resolver
- Intercloud gateway and/or attribute/namespace translator.

The Federated ACAI can leverage existing platforms for federated network access and federated identity management widely used for multi-domain and multi-provider infrastructure integration such as Eduroam [20], eduGAIN [17], Shibboleth [21], CILogon [22].

7 Future Research and Development

The future research and development will include further SDLM definition, e-SDI and ACAI components definition and development with focus on infrastructure components of e-SDI. Special attention will be given to defining the whole cycle of the provisioning SDI services on-demand specifically tailored to support instant scientific workflows using cloud IaaS and PaaS platforms. This research will be also supported by development of the corresponding Cloud and InterCloud architecture framework to support Big Data e-Science processes and infrastructure operation.

Acknowledgments

This work was motivated and partly supported by the special project “Study on Authentication, Authorization and Accounting (AAA) Platforms For Scientific data/information Resources in Europe” commissioned by the European Commission to the consortium of TERENA, LIBER, University of Amsterdam, University of Debrecen. The authors value wide discussions between the consortium members on the different aspects of the existing research infrastructures and AAA technologies which findings found further development in this paper.

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. Riding the wave: How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. October 2010. [online] Available at http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
3. ESFRI Roadmap Update 2010. [online] http://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2010/procedure_roadmap_update.pdf
4. Study on AAA Platforms For Scientific data/information Resources in Europe. Final report [online] https://confluence.terena.org/download/ attachments/30474266/AAA-Study-Report-0907.pdf
5. Federated Identity Management for Research Collaborations. Final version. Reference CERN-OPEN-2012-006. [online] https://cdsweb.cern.ch/record/1442597
6. SIENA European Roadmap on Grid and Cloud Standards for e-Science and Beyond. SIENA Project report. [online] http://www.sienainitiative.eu/Repository/Filescaricati/8ee3587a-f255-4e5c-aed4-9c2dc7b626f6.pdf
7. Seeking new horizons: EGI's role for 2020. [online] http://www.egi.eu/blog/2012/03/09/seeking_new_horizons_egis_role_for_2020.html
8. Future Internet Report. UK Future Internet Strategy Group. May 2011. [online] https://connect.innovateuk.org/c/document_library/get_file?folderId=861750&name=DLFE-33761.pdf
9. Worldwide Large Hadron Collider Grid (WLCG) [online] http://wlcg.web.cern.ch/
10. European Data Protection Directive. [online] http://ec.europa.eu/justice/data-protection/index_en.htm
11. Koopa, David, et al, A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers, International Conference on Computational Science, ICCS 2011. [online] http://vgc.poly.edu/~juliana/pub/vistrails-executable-paper.pdf
12. Open Access: Opportunities and Challenges. European Commission for UNESCO. [online] http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-handbook_en.pdf
13. OpenAIR – Open Access Infrastructure for Research in Europe. [online] http://www.openaire.eu/
14. Open Researcher and Contributor ID. [online] http://about.orcid.org/
15. Data Lifecycle Models and Concepts. [online] http://wgiss.ceos.org/dsig/whitepapers/Data%20Lifecycle%20Models%20and%20Concepts%20v8.docx
16. EGI federated cloud task force. [online] http://www.egi.eu/infrastructure/cloud/cloudtaskforce.html
17. eduGAIN - Federated access to network services and applications. [online] http://www.edugain.org
18. Demchenko, Y., C.Ngo, M.Makkes, R.Strijkers, C. de Laat, Defining Inter-Cloud Architecture for Interoperability and Integration. The 3rd Int'l Conf. on Cloud Computing, GRIDs, and Virtualization CLOUD COMPUTING 2012, July 22-27, 2012, Nice, France
19. Cloud Reference Framework. Internet-Draft, version 0.3, June 27, 2012. [online] http://www.ietf.org/id/draft-khasnabish-cloud-reference-framework-03.txt
20. eduroam. [online] http://www.eduroam.org
21. Shibboleth - Open Source Federated Identity Management System. [online] http://shibboleth.net/
22. CILogon Service [online] http://www.cilogon.org/