LASDA: an archiving system for managing and sharing large scientific data

JEONGHOON LEE
Korea Institute of Science and Technology Information
Scientific Data Strategy Lab.
245 Daehak-ro, Yuseong-gu, Daejeon
Republic of Korea
jhoon@kisti.re.kr

Abstract: ICT Technology has cultivated data-intensive and interdisciplinary research paradigm. In various scientific domain, data became an essential mean to advance research and, furthermore, it is considered as great asset to have latent value. Archiving helps researchers or institutes managing their information assets. For managing and sharing data effectively, it is required to provide functionality considering the characteristics of scientific data. We define core functions of an archiving system by analysis of the characteristics of scientific data and compare representative data systems in the view of the functions and performance evaluation. As a result, we present an archiving system design named LASDA for large scientific data using a base repository system that is chosen by functional analysis. Our system is designed to provide enduring facilities for efficient and reliable management of large scientific data. It is planned to implement for Korea national research data service.

Key–Words: Archiving system, Scientific data, Big data, Data management, Data sharing

1 Introduction

As decision making in business depends on data rather than intuition and theory, data-intensive research paradigm is pervasive across research domains. In recent years, the development of information and communications technologies (ICT) has made it possible to generate various types of data and materialize large volume of it. This has reshaped the scientific enterprise and introduced new look seeking to promote long-term research as well as innovation [8]. For this new paradigm of research, it is required to manage data effectively and to share it for further collaboration.

An archiving system is as specialized repository used to preserve, protect, control, maintain authenticity and integrity, accommodate physical and logical migration, and guarantee access to information and data objects [4]. However, the existing data processing systems have limitations to handle volume and variety of scientific data generated from various science disciplines such as astronomy, cosmology, biology, humanities, etc. A data processing platform to support research scientists should consider distributed storage, load balancing and data analysis capabilities in order to handle volume and variety of data. Furthermore, it should have archiving functions to manage and share the data.

In this paper, we present the design of an archiving system design for large scientific data, which is named LASDA. First, we define core functions of an archiving system by characteristics analysis of scientific data and functional analysis of representative data systems. Second, we compare functionality and performance of the systems and choose a desirable base one. Finally, we design an archiving system in order to provide enduring facilities for effective and reliable management of scientific data.

2 Types of Scientific Data and Core Functions for an Archiving System

In this paper, we define scientific data as quantitative information or qualitative contents collected and generated by experiments, observations, measurements, and investigations that are activities of scientific research. Generally, scientific data is categorized into two types, experiment data and associated data [11]. Experiment data is from experiments as measurements of some physical phenomena and simulations as complex computations. There are found various features in experiment data according to regularity, density, and time variation. Associated data is to support the experiment or to be generated from the experiment data such as configuration data, instrumentation data, analyzed data, summary data, and property data.

Table 1: The list of scientific data analyzed in DCP

Discipline	Subject
Life Science	Biochemistry Movement of Protein Human Genomics Plant Genomics
Cosmology	Astrophysics Astronomy
Geology	Soil Ecology Carbonate Sedimentology
Chemistry	Chemical Kinetics (Aerospace Engineering)
Arts & Humanities	Architectural History / Epography Linguistics Sociology / Demographics History, Sustainable Development

Based on these type categories, we analyzed characteristics of 13 scientific data listed in Data Curation Profiles[1]. Table 1 shows the list of scientific data analyzed in DCP. These data demonstrates the subjects of Korea national data service in the future.

From analysis of selected data, the nine core functions of a desirable archiving system were defined. Then we surveyed the functionality of six representative repository systems: Fedora commons, iRods, DataOne, Dataverse, Figshare, Open Science Data Cloud (OSDC) [2][1][7][6] as shown in Table 2. We focus on the storage management and data collection and transmission in order to manage and share large scientific data. The core functions for storage management are data partitioning and categorizing for distributed storage, work-load and data-load balancing, versioning, scale-out and extensiblility. Data import, export and transmission are additional functions for data collection and transmission [9].

3 Design of an Archiving System for Large Scientific Data

For in-depth analysis of performance, three systems are chosen: Fedora commons, Dataverse, and iRods that have some suitable storage management policies and good extensibility, which are necessary to manage large data. The additional performance evaluation for these three systems was carried out. For a DNA sequence data set of 1.08TB, we compared query response time with respect to different number of files and system nodes. The result of performance for the number of files is shown in Figure 1. Fedora commons shows better response time than other systems but the difference could be negligible because it is too small (less than 2s/TB). The performance result for the number of system nodes is shown in Figure 2. This evaluation is to show the scale-out and Dataverse was excluded since it does not provide the distributed storage. As a result, Fedora commons shows the better performance but the difference is also too small (less than 0.1/TB).

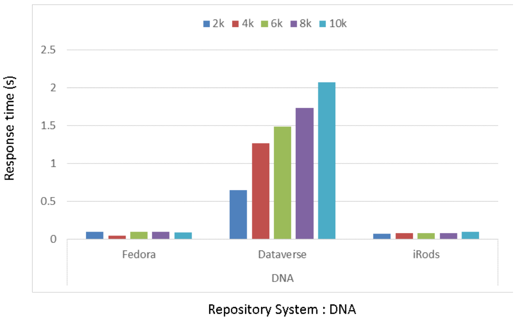

Figure 1: Response time with respect to the number of files

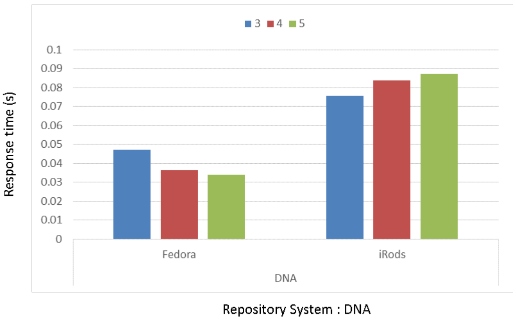

Figure 2: Response time with respect to the number of system nodes

Table 3 shows the result of functional analysis of the three systems. iRods system supports customizable cluster structure, functional extensibility with programmable rules, SQL-like query, and linkage with various file formats and data systems [3][5][9][10]. According to the result functional analysis and performance evaluation, we design our archiving system for large scientific data (LASDA) based on iRods. LASDA supports the 9 core functions. We classify the core functions into the function modules provide by iRods and the extended rule modules to implement separately. LASDAs functional diagram is shown in Figure 3.

[1] Digital Curation Profiles (DCP) was developed through a research project conducted by the Purdue University and the University of Illinois Urbana-Champaign, US, http://datacurationprofiles.org

Table 2: Functional comparison of data systems

	Function	**Fedora**	**iRods**	**DataOne**	**Dataverse**	**Figshare**	**OSDC**
Storage Management	Data Partitioning	x	x	x	x	x	x
	Data Categorizing	x	x	x	x	x	x
	Workload Balancing	o	o	x	x	x	x
	Data load Balancing	o	o	x	x	o	o
	Versioning	x	x	o	o	o	x
	Scale-out	o	o	o	x	o	x
	Extensibility	o	o	x	o	o	x
Collection & Transmission	Data Import / Export	o	o	o	o	o	o
	Data Transmission	o	o	o	o	o	o

The functional modules are of distributed storage management, load balancing, data integrity, versioning, data collection and transmission. Distributed storage management is to store and manage data in distributed storage according to the characteristics of data and usage patterns, to determine policies for collating of related data, partitioning of large data, and duplication and migration of data, and to perform scheduled archiving. Load balancing is to manage distributed storage dynamically considering usage and access patterns of data and nodes' status. The activities of load balancing management include dynamic data migration, data duplication, and data partition. Data integrity is to preserve integrity of data loaded in the archive system and to provide security methods such as access and authority control and user authentication. Versioning is to manage the version history of data efficiently so that reliability of data is enhanced and historical data gains accessibility. Data collection and transmission is to support various formats of scientific data and to provide interface for efficient transmission of large data including replication and validation.

Storage management and data collection and transmission are designed being based on iRods basic functions. Load balancing and data transmission consists of rule extensions added on iRods functions. Versioning, distributed storage management, and data staging are not supported by iRods so they are to implemented individually and connected to the system through rules.

4 Concluding Remarks

An archiving system to manage and share data should have functionalities to support the characteristics of the data. We draw the core functions of an archiving system by the characteristics of scientific data and present the plan of LASDA to support them. From functional analysis and performance evaluation of some representative repository systems, we could conclude proper functionalities based on a proper system. Our evaluation is for the data subject we are planning to manage and it is not to provide a benchmark of the systems in general view. Each system has a *sui generis* architecture and strength for its own purpose. LASDA is designed to provide enduring facilities for efficient and reliable management of large scientific data. It is planned to implement for Korea national research data service in the future.

References:

[1] S. Allard, DataONE: Facilitating eScience through collaboration, *Journal of eScience Librarianship*, 1.1:3, 2012.

[2] M. Bertazzo and D. Angela, Preserving and delivering audiovisual content integrating Fedora Commons and MediaMosa, *Journal of Digital Information*, 13.1, 2012.

[3] GT. Chiang, et al., Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute, *BMC bioinformatics 12.1*, 2011, pp. 361.

[4] Contoural (Inc.), Seven Essential Strategies for Effective Archiving, *EMC, Analyst Reports,* 2012.

[5] D. Hnich and R. Muller-Pfefferkorn, Managing large datasets with iRODS-A performance analysis, *Computer Science and Information Technology (IMCSIT)*, Proceedings of the 2010 International Multiconference on. IEEE, 2010.

[6] R.L. Grossman, et al., An overview of the open science data cloud. Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, 2010, pp. 377-384.

[7] G. King, An introduction to the Dataverse Network as an infrastructure for data sharing, 2007, pp. 173-199.

[8] OECD, Making Open Science a Reality, *OECD Science, Technology and Industry Policy Papers,* No. 25, OECD Publishing, 2015.

[9] A. Rajasekar, R. Moore, F. Vernon, iRODS: A Distributed Data Management Cyberinfrastructure for Observatories. *InAGU Fall Meeting Abstracts*, Vol. 1, 2007, pp. 1214.

[10] A. Rajasekar, et al., iRODS Primer: integrated rule-oriented data system, *Synthesis Lectures on Information Concepts, Retrieval, and Services 2.1*, 2010, pp. 1-143.

[11] A. Shoshani, F. Olken and HK. Wong, Characteristics of Scientific Databases, *No. LBL-17582-REV.*, Com. Sci. Research Department, University of California, Lawrence Berkeley Lab., 1984.

Table 3: Functional analysis of Fedora, Dataverse and iRods

Function	**Data System**		
	Fedora	**iRods**	**Dataverse**
Performance	Excellent	Excellent	Good
Extensibility	By modification of source code	By utilization of rules	By additional java code
Scale-out	No limitation of the number of nodes	100 of nodes per iCat server per iCat server	Not provided
Search	Not provided (needs Solr)	SQL-like query support	File / Metadata search
Cluster Construction	Inflexible	Flexible	Not provided
Linkage with other data systems	Limited possible	Possible (includes Fedora, Dataverse)	Almost impossible
Linkage with File & Database system	Redis, Cassandra, S3, Hbase MongoDB, BerkeleyDB, etc.	PostgreSQL, MySQL, Neo4j (Hbase, MongoDB prearranged)	PostgreSQL, MySQL

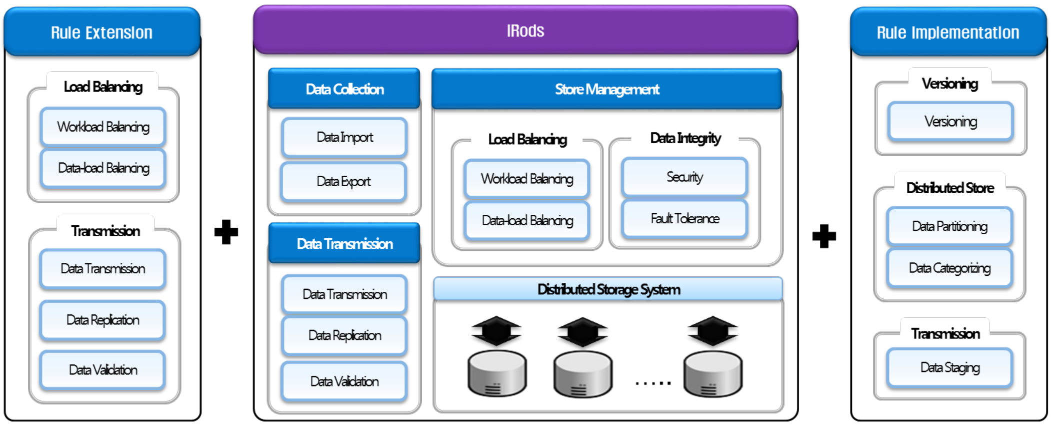

Figure 3: The functional diagram of LASDA