ISSN No. 0976-5697

Volume 3, No. 4, July- August 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Towards an Approach of Agricultural Virtual Organizations (AgriVOs) for Sharing and Management of Scientific Research Data

Mukhtiar Memon*, Akhtar A. Jalbani, Mansoor H. Depar, Gordhan D. Menghwar
Information Technology Center,
Sindh Agriculture University, Tandojam, Pakistan
mukhtiar.memon@sau.edu.pk, akjalbani@sau.edu.pk, mansoor.hyder@sau.edu.pk, gdas@sau.edu.pk

Abstract: Since long-time, scientists are using different modes to create, store and present the data in the form of structured tables, images, documents and animations. However, the scientists still rely on the conventional software tools and applications, which are slow and unreliable at the same time. Moreover, the data centers are confined within the organizational boundaries due to policy restrictions and unavailability of proper environment, where data can be shared in a secure and reliable collaborative way. On the other hand the presenttechnological innovations in ICT have brought a dramatic change in sharing, management and presentation of data. These technologies have more power for sharing, analysis and presentation of scientific research data. This paper presentsa solution of using Grid Computing (GC) Technologies for sharing and management of scientific data. GC-enabled networked application environment provides the facility to the scientists not only instantly share the scientific data in research collaborations, but also share the technical infrastructure such as high performance computing facility, huge storage and software application licenses. More importantly, the communication in such an application is secure and reliable to respect the security and privacy of sensitive research data.

Keywords: Research Data Sharing, Grid Computing, Virtual Organization, Resource Sharing, Data Security.

I. INTRODUCTION

The scientific contributions and results obtained by the research organizations play a vital role in the industrial and economic development of the country. Most of the research activities are knowledge-oriented, where knowledge is encapsulated in the form of data (Feerria, 2005). The data is the hub of research, as most of the research activities rover around data, which is created, processed, shared and stored during and after the field and laboratory experiments. In other words, data management and sharing plays a key role in expansion of research irrespective of the type, size and method of research.

Agriculture is the economical backbone of developing countries like Pakistan. In particular to agricultural research, in the present scenario, scientists are working in research and development collaborations, which span among multiple organizations. They perform variety of field and laboratoryexperiments and want to share the empirical data to each other during and after the experiments to take further necessary measures based on the experimental results. The scientists use a variety of software tools for data storage, analysis and presentation. The problem is that in current setup the scientists are not able to share the research data instantly to convey timely research results to other research fellows. In addition, they don't have network-based systems in their organizations for real-time sharing of data, software tools, computing power and memory. They rely on conventional approaches which are slow, unreliable and insecure at the same time [2]. Moreover, these approaches confine the scientists within the organizational perimeters and do not provide them the rights to administer the access rights in the collaborative agricultural research.

It is also pertinent to mention that research data is security and privacy-sensitive. For example, the medical data of patients is very sensitive and the patients would never like to expose their medical history with public.

Similarly, the data related to scientific organizations such as nuclear research, defense and strategic research data is very sensitive and requires appropriate security and reliability both for storage and in communication. This necessitates that only authorized users with specified permissions should be allowed to access the data based on the roles assigned to them in the organizations. The permissions are assigned to them based on the ownership or access rights [3]. In this connection, it is essential to provide such network-based communication application environment, where scientists are not only able to manage and share the research data, but also work without fear of unauthorized disclosure of the sensitive research data.

A. *Problem Identification:*

The physical and administrative boundaries in the organizations have restricted the access to the scientific information. Presently, scientists working in different organizations rely on automated or semi-automated information systems, emails or storage media. These conventional methods provide sharing of information in traditional request-response paradigm. However, this does not ensure instantsharing of research data to the authorized users and also the scope of sharing of scientific resources is very limited. The reason being, that the current approaches use static technologies for sharingscientific data in research collaborations. Moreover current approaches are limited only for sharing the core research data. However, the scientific research requires sharing a variety of additional resources including computing power, memory storage, software licenses, business policies and software tools. This necessitates a high-level collaborative network environment, which enables the scientists within research organizations to create their own projects and instantlyshare scientific data, relevant software tools, computing power and memory to provide facilities for research collaboration.

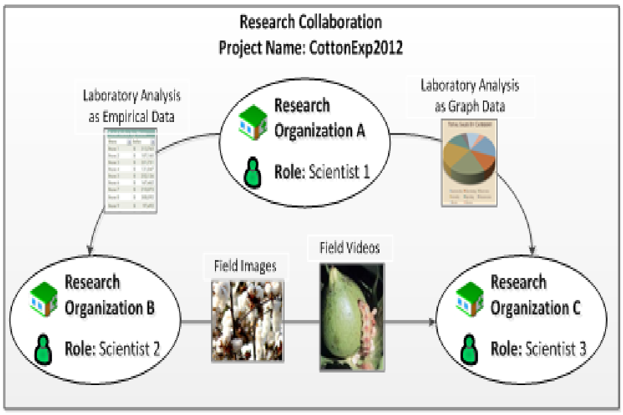

Figure 1: Scientific Data Sharing in Research Collaboration)

The scenario in Figure 1 illustrates the sharing of research data among the scientists working in different research organizations. Three scientists i.e., Scientist 1, 2 and 3 belong to three different organizations i.e., Organization A, B and C are working in a research collaboration related to cotton.In this scenario the scientists are sharing different types of data including Empirical Data, Graph Data, Images and Videos as shown in Figure 1. There are couple of very interesting issues concerned with sharing of this data. Firstly, the data elements are created by the scientist using different tools. For example, the data tables are graphs created using Microsoft Excel or SPSS. The graphics and videos are created as JPEG, AVI, WMA formats. This shows that there is a variety of tools required by every scientist in every organization to share and view the scientific data. So, every research organization has to purchase the software tools and technologies which support the different data representation types. Moreover, the scientist also needs knowledge to use those software tools. More importantly, the scientists have to rely on conventional data transfer methods such as emails or web-based systems to share the data with their fellows.

B. Challenges:

There are many challenges in the sharing of scientific data both at the design and implementation level. At the design level it is important to model the system in such a way that it should be possible to integrate the system with other systems without depending upon platform technologies. Such systems would enable easy enterprise integration of research organizations for sharing research data. Similarly, at the implementation level the systems should exploit the very powerful technologies based on latest tools which provide the dynamic way of data and resource sharing. Some of the main challenges in the area of scientific data and resource sharing are as under:

a. Currently scientist in most of research organizations share research data using heterogeneous file formats through emails, storage or plain information systems, which are insufficient for real-time data sharing during the field and laboratory experiments.
b. Security, Privacy and Reliability are the key requirements of research data sharing in the modern computer-based systems, which are open in the vulnerable Internet medium [4, 5]**.** The conventional data and resource sharing technologies do not provide a comprehensive model for authorized access to the shared and stored data, besides the confidentiality and integrityrisk of valuable research data.
c. The user-friendly environments targeting the domain of scientific research are not available for the scientists to work in collaborative projects.
d. The applications are standalone silo applications using variety of technological platforms; it is huge challenge to integrate such applications in the inter-organizational collaborations [6, 7].
e. Despite current technological developments, timely and long-term availability of research data is still an open issue, which needs considerable attention [8].
f. The stakeholders in scientific data sharing collaborations require sustainable data sharing and management systems, based on realistic goals and measures to align the research objectives with ICT. Presently there is a gap between research objectives and technology solutions used for data sharing.
g. Scalability and Extensibility of network-based data sharing and management solution is an open challenge in the ever-changing technology platforms.
h. The funding agencies for research have no control over the scientific data created during the experiments funded by them.
i. It is difficult to develop international collaborations due to unavailability of an environment for scientific data sharing and management.
j. The legal issues related to access, usage and dissemination of research data is a totally unexplored aspect in the developing countries like Pakistan.
k. Evaluation and Assessment of scientific projects is difficult for the funding agencies due to late and limited access to data.
l. Trainings to promote data sharing and management are not available for the researcher.

II. RESEARCH METHODOLOGY

The main objective of this research is to exploit the use of modern state-of-the-art technologies for research collaboration. In this connection, we conducted some surveys, investigated the appropriate ICT technologies and developed a prototype. These main stages of our research methodology are defined below in more detail:

A. Survey Of Research Organizations:

We conducted some survey in few research organizations in Sindh to identify to identity the problems related to sharing of data.Besides, we investigated the ICT tools, which are presently used in the research organizations for sharing, management and analysis of data. In the surveys we found following ICT tools, which are used for sharing of data among the scientists.

a. Sending data through emails
b. Transfer of huge data through storage (USB, HD)
c. Data Analysis using SPSS
d. Sharing results through hard copies

As discussed earlier, that these methods of sharing scientific data are not enough and do not provide a secure and reliable way of data transfer. Moreover, these methods are slow and not sufficient for real-time data sharing during the research experiments.

B. Prototype Development:

In order to meet the challenges highlighted above we propose the use of latest ICT technologies for scientific data and resource sharing. We developed a network-based application environment using Grid Computing Technology. The prototype based on the grid technology enables the scientists in the research organizations to create Agricultural Virtual Organizations (AgriVOs) based on the research collaboration among multiple organizations. At the application level, the Content-based Management System (CMS) is used to integrate and configurethe AgriVO environment, which will provide a user-friendly interface to the scientists, who will create and share the research data. In addition, a prototype for an agricultural-information system was developed to check and validate the capabilities of grid-supported data sharing infrastructure.

III. AGRIVO GRID ENVIRONMENT

The proposed solution is based on Virtual organizations (VOs), which provide the facility to share the data in collaborative environment for business, governance, agriculture, education and development. The main idea of VOs is to create the organizations scattered among the existing (administrative) organizations [9, 10, 11]. The VOs are used to share a variety of resources such as data, computation capabilities, storage media and software applications. The resources in the real are permanently owned and administered by the individual administrative organizations. However, the resources are shared for the short period in the form of VOs for certain collaborative projects. The lifetime of the VOs solely depends upon the length of the project collaboration, where scientists can timely share the research data and other resources in a secure and reliable way [12]. The VOs can play a significant role in the agricultural research collaborations. As mentioned above, that the agricultural research relies on the multi-organizational research experiments and the scientific staff working in the projects needs timely access to the experiment results for better management of research activities [1].

An AgriVO will be based on a research project undertaken by an agricultural research organization such as Pakistan Agriculture Research Council (PARC), Central Cotton Research Institute or Wheat Research Institute. With the establishment of VOs the scientists in the collaborating organizations will be able to timely communicate the research findings and data in a reliable and secure way. This will enhance the capabilities of the research organizations, which will be beneficial for the target organizations and eventually for the people, who rely mainly on agriculture for their living and business. The subsequent part of this section presents the technical solution based on the grid computing technology. Grid Computing (GC) technology provides an infrastructure to create and share data and resources. GC is based on the idea not only to share data but also storage, computing power and software licenses through the applications, which offer services to the system users. To solve the problems for sharing the research data, we developed a grid environment. The deployment diagram in Figure 2 shows the architecture of the grid environment. The proposed grid environment for sharing resources among Agricultural Virtual Organizations (AgriVOs) consists of the following three layers:

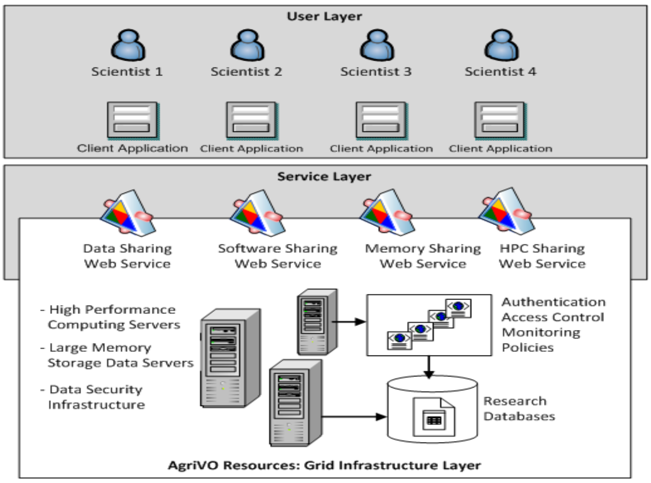
User Layer
Scientist 1
Scientist 2
Scientist 3
Scientist 4
Client Application
Client Application
Client Application
Client Application
Service Layer
Data Sharing Web Service
Software Sharing Web Service
Memory Sharing Web Service
HPC Sharing Web Service
- High Performance Computing Servers
- Large Memory Storage Data Servers
- Data Security Infrastructure
Authentication Access Control Monitoring Policies
Research Databases
AgriVO Resources: Grid Infrastructure Layer

Figure 2: Grid-enabled AgriVO Architecture)

A. User layer:

The user layer provides the applications with GUI-based user interfaces at the client side. In our AgriVO scenario the clients are the scientist who can use the web-based client applications to access a variety of web services offered by the agricultural grid.

B. Service layer:

The service layer defines the web services, which offer functionality to the systems users. Different types of services offered in the AgriVO grid environment are:

a. Data sharing web service:

The data sharing services are used to share agricultural research data. The scientist in a research organization can create a research project and upload the data into the AgriVO system. In addition, he can also assign permissions to other research fellows to access the data created by him. The data sharing and management is performed in a secure and reliable way. This means if the data is created by the Scientist 1, it cannot be accessed or modified unless the Scientist 1 assigns the permissions for that to the other research fellows. Similarly, the other scientists can also create and modify the data of their experiments and assign the access rights to each other.

b. Software sharing web service:

Sharing of software licenses is a strong feature of grid environment, which economizes the research activities by sharing the software licenses. In the normal practice every scientist has to buy the software license that he needs for his research. Most of the times the software costs are too high, which project funding cannot afford due to budget restrictions. As a result, the scientists are deprived of the very power software tools to conduct their research. Fortunately, grid computing provides the way to share software licenses. This means that every scientist does not have to buy the expensive software license, because he can share the software in the grid environment. However, the execution of the gird-enabled shared software is little bit different from the traditional software installation and usage. The software which isused in the grid environment is installed on the server machines and used at the client machines. The client will only send the data through the client applications as shown in Figure 2. The data is received by the grid facility and processed by the software requested by the user and the results are sent back to the user at the client application. This is also beneficial for the user in terms of performance, because he does not have very powerful machines at the client-side to execute the very sophisticated and computation-intensive software applications. Instead those are executed at the software sharing servers and results are retrieved by the user.

c. *Memory sharing web service:*

Besides software licenses, the memory storage for data is also one of the critical requirement in research experiments, which rely on analysis of huge amount of data. As already discussed the research data is not only plain text, but it comprises of the text, images and videos. These data formats consume lot of memory space and the scientists at their client machines don't have such high-storage facilities. Moreover, there is also security risk to data stored at the client side. For this the proposed AgriVO grid-enabled system provides huge storage facility up to Terabytes so that the scientists can store the large amount of data on the shared storage.

d. *Hpc sharing web service:*

Client computing machines are ordinary machines not capable of performing high computations required for data processing and analysis. If all the complex computations are executed at the client machines then the data processing and analysis becomes very slow and in long-term reduces the research productivity. To solve this problems the proposed AgriVO grid environment provides High Performance Computing (HPC) machines so that the client can only send the data, whereas the same is processed and analyzed by the powerful computer machines available in the grid. The HPC machine such as IBM's 256-core Power 795 (4.0 GHz) processes the data with parallel processors to deliver very fastresults to the users[13].

The services in the AgriVO grid infrastructure are offered asweb services, because, the scientists in research organizations use heterogeneous tools for creating data elements. It is not possible to integrate the system at the grid-level due to different data formats used by the scientists. The web services are created for the purpose to provide service interfaces in the web-based grid applications so that the functionality coded in different languages can be published as services irrespective of technology differences [14, 15]. Web services reduce coding by "write-once: use anywhere" paradigm. That is the reason we selected grid-based web services to share system functionality among the users.

C. *Grid infrastructure layer:*

The Grid Infrastructure Layer is the core technology layer comprising the hardware, software, databasesstorage and network facilities. Together we call them as the *Grid Infrastructure*, which has following main components:

a. ***High Performance Computing Servers:*** HPC Servers have multiple processors to be used for computation-intensive tasks.
b. ***Large Memory Storage Data Servers:*** Storage Servers are used to provide large memory space to the user to store the scientific data in large size in secure and reliable way.
c. ***Data Security Infrastructure:*** Data Security Infrastructure is one of the main components for Quality of Service (QoS) in the grid. It is responsible to provide security services in the grid-based applications [16]. The grid security infrastructure is based on the Public Key Infrastructure (PKI), which is the de-facto standard for securing data in storage and communicationusing digital certificates, mutual authentication, authorizations policies and Single Sign-on (SSO).
d. ***Authentication, Authorization and Monitoring Policies:*** The grid security infrastructure relies on the security policies for authentication, authorization and monitoring. The policies are defined in the specified standards such as SAML, XACML and WS-Notification Standards.
e. ***Research Databases:*** The Research Databases store the data created by the scientists, which he wants to share with the research fellows based on the research collaboration projects. The data can be stored in different formats and permissions for access to the data can be granted easily by the project lead, so that the research team can instantly view the project results and analysis reports from the database.

Table 1

Parameters	**Conventional Tools**	**Grid Computing Technology Infrastructure**
Data Sharing	**Poor:** Data sharing is based on plain information systems, emails and through storage media, which are slow and do not provide instant sharing.	**Strong:** Provides the easy way to share research data of all formats and size instantly through easy-to-use web-based GUI interfaces.
Data Security	**Poor:** Data security is based on the web-based services of email servers, which are not in the control of research organizations.	**Strong:** Only authorized users can access data and resources depending upon the permissions assigned to them.
Data Monitoring	**Not Available:** Does not provide a way to monitor system activities.	**Strong:** Provides the way to monitor system and user activities through monitoring services.
Real-time Sharing	**Not Available:** Do not ensure real-time sharing.	**Strong:** Real-time research data sharing during and after empirical experiments.
Software Licenses Sharing	**Not Available:** Do not provide facility to share expensive software applications to reduce research expenditures.	**Strong:** Sharing of variety of software licenses used for data analysis, presentation and storage.
Storage Sharing	**Limited:** Very limited storage of maximum a few Gigabytes is possible.	**Huge Storage:** Storage of Terabytes of data is possible.
High Performance Computation Sharing	**Not Available:** Due to client-side execution of applications it is not possible to use HPC computation.	**HPC –enabled:** The shared HPC are available to execute the services on very fast computers.
Cost	**Moderate:** Not so expensive at deployment-time, but very expensive in long-run.	**Expensive:** The grid-enabled applications are expensive at the

		deployment-time but economical in long-run.

IV. RESULTS AND DISCUSSION

The main focus of this contribution is to enhance the capabilities of research organizations in terms of ICT usage. In the current scenario, most of the research organizations in Pakistan share research data using plain data processing systems. On the other hand the ICT technologies have developed a variety of tools and technologies and established methodologies and frameworks, which have brought revolution in information sharing, presentation and management. In this connection the main technologies are Grid Computing, Content Management Systems, Context-aware Networks and Knowledge-based Systems. These technologies can alter the research scenario dramatically with their power of data sharing and management. Moreover, this work mainly used Grid Computing technology and Content Management Systems (CMS) for sharing research data and resources among the scientists working in different agricultural organizations.

Grid Computing can provide network-aware application environment to the scientist for research, so that they can share the useful scientific results in collaborations during and after the experiments. In addition, they can also share the software licenses, hardware capabilities (Processing and Memory) and network. After implementing these technologies, we can compare them with the existing conventional technologies. The comparison for the evaluated parameters is shown below in the table.

V. REFERENCES

[1]. Ferreira L., F. Lucchese, T. Yasuda, C. Y. Lee, C. A. Queiroz, E. Minetto and A. Mungioli, 2005, Grid Computing in Research and Education, IBM Red Book: 37 - 46.

[2]. Crompton S., B. Aziz and M. Wilson, 2009, Sharing Scientific Data: Scenarios and Challenges, e-Science Centre, STFC Daresbury Laboratory, United Kingdom.

[3]. Bistline D., 2011, Protecting Sensitive Digital Research Data, University of Texas at Austin.

[4]. Kenneth P. S., L. Seligman, V. Swarup, 2008, Everybody Share: The Challenge of Data-Sharing Systems, IEEE Journal of Computer, Vol. 41, Issue 9: 54 – 61.

[5]. Larry D. Conrad and S. Waddell, 2011, Protecting the Security of Research Data, EDUCAUSE Center for Applied Research, http://www.educause.edu/ecar.

[6]. Hohpe G. and B. Woolf, 2004, Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions, Addison-Wesley.

[7]. Robert S. Chen, 2005, The GEO Data Sharing Challenge: Putting Principles into Practice, Center for International Earth Science Information Network, The Earth Institute, Columbia University, Palisades, New York, USA.

[8]. NSF Report, 2011, Digital Research Data Sharing and Management, National Science Foundation's National Science Board, USA.

[9]. Buyya R. and V. Srikumar, 2005, A Gentle Introduction to Grid Computing and Technologies, *CSI Communications*, Vol. 29, Nr. 1: 9 - 19.

[10]. Kristoffer Jacobsen, 2004, A Study of Virtual Organizations, Department of Computer and Information Science, Norwegian University of Science and Technology.

[11]. Les P., 2001, Understanding Virtual Organizations, Information Systems Control (online) Journal, Volume 6.

[12]. Cody E., S. Raj, R. Raghav and U. Shambhu, 2008, Security in grid computing: A review and synthesis, Journal of Decision. Support Systems, Vol. 44, Issue 44: 749 – 764.

[13]. Vecchiola C., S. Pandey and R. Buyya, 2009, High-Performance Cloud Computing: A View of Scientific Applications, Keynote Paper at 10th International Symposium on Pervasive Systems, Algorithms and Networks I-SPAN 2009, Kaohsiung, Taiwan.

[14]. Czajkowski K., D. F. Ferguson, I. Foster, J. Frey, S. Graham, I. Sedukhin, D. Snelling, S. Tuecke and W. Vambenepe, 2004, The WS-Resource Framework Version 1.0.

[15]. Srinivasan L. and J. Treadwell, 2005, An overview of service-oriented architecture, web services and grid computing, V02, 11/2005, Hewlett-Packard Software Global Business Unit.

[16]. Ian T. Foster, 2005, Globus Toolkit Version 4: Software for Service-Oriented Systems. In NPC, Vol. 3779:2-13.

Short Bio Data for the Authors

Dr. MukhtiarMemonstudied MSc. at the University Huddersfield, UK and PhD. at the University of Innsbruck, Austria. Dr. Memon is a researcher and a fulltime faculty member at the Information Technology Center, Sindh Agriculture University, Tando Jam, Pakistan.He has special interests in modeling security aspects of service-oriented systems and development of frameworks, methods and techniques based on the patterns and principles for model-driven development and service-oriented computing. On the technology side he excels in UML, Model Transformation, Service-oriented Architecture, Web Services, WS-Security, Enterprise Integration and Service-oriented Security Architectures.

Dr. Akhtar Ali Jalbaniis Assistant Professor, Information Technology Centre, Sindh Agriculture University Tando Jam Pakistan. He has obtained PhD (Computer Science) in 2011 from Institute of Computer science, University of Goettingen, Germany. His research interest includes Software Quality, UML, Web Services, Model Transformations and Data mining.

Dr. MansoorHyder has studied at University of Sindh at Jamshoro, Sindh, Pakistan and EberhardKarlsUniversitätTübingen, Germany for his Masters in Electronics and PhD in computer science. Dr. Hyder is currently working as a full time faculty member and researcher at Sindh Agriculture University Tandojam, Pakistan. His research interest areas are 3D audio, virtual reality, virtual acoustics, acoustic simulation, VoIP and distributed networks.

Dr. Gordhan Das Menghwar, Assistant Professor, Information Technologyis working as a Director/Assistant Professor at Information Technology Centre, Sindh Agriculture University, Tandojam, Sindh, Pakistan. He did his PhD from Vienna University of Technology, Austria. He was born in KotMirs, Sindh, Pakistan in 1978. His research interests include cooperative communications, space time codes, network coding and information theory.