Grid Technology for Supporting Health Education and Measuring the Health Outcome

Nitin Sukhija
Slippery Rock University of Pennsylvania
Slippery Rock, PA
nitin.sukhija@sru.edu

Arun Datta
National University
La Jolla, CA
Arun.Datta@natuniv.edu

Sonny Sevin
Slippery Rock University of Pennsylvania
Slippery Rock, PA
srs1030@sru.edu

John E. Coulter
Indiana University
Bloomington, IN
jecoulte@iu.edu

ABSTRACT

In this paper, we present our developed health-IT solution that address these challenges: developing the strategies not only to store such a vast amount of data but also making those available to the researchers for further analysis that can measure the outcome on the participant's health. The developed community health grid (C-Grid) solution to store, manage and share large amounts of these instruction materials and participant's health related data, where the remote management and analysis of this data grid is performed using iRODS, the Integrated Rule-Oriented Data System is discussed and presented in this paper.

CCS CONCEPTS

• **Information systems** → **Database Management Systems**; • **Applied computing** → *Life and medical sciences*;

KEYWORDS

iRODS, cyberinfrastructure, data grid, XSEDE, community health, grid computing, data integration, portal, virtual, health education, mobile technology

ACM Reference Format:
Nitin Sukhija, Arun Datta, Sonny Sevin, and John E. Coulter. 2018. Grid Technology for Supporting Health Education and Measuring the Health Outcome. In *PEARC '18: Practice and Experience in Advanced Research Computing, July 22–26, 2018, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3219104.3229247

1 INTRODUCTION

Earlier studies support the notion that health education is a critical component of preventive medicine. Several school health providers routinely offer health education to school age children on a variety of topics ranging from obesity intervention to prevention of sexually transmitted diseases. Similarly, community health organizations operating in various counties provide health education to both adults and school age children. Training is also provided to the social- and healthcare providers. These educational and training materials require relatively large digital storage space. Moreover, to measure the effect of such health education on the outcome on the participant's health, data from demographic, socioeconomic, personal health records, genomics, etc. are also needed to be stored that require very large digital storage space for further analysis. In addition, these health education and training materials are needed to be available 24/7 through the web and be accessible using any devices including Smartphones and tablets whenever participant's desire to access. Our group is engaged in developing health-IT solution to these challenges: developing the strategies not only to store such a vast amount of data but also making those available to the researchers for further analysis that can measure the outcome on the participant's health.

Data grid utilizes grid technologies and provides huge archiving places to store and manage large amount of user collaborative data objects. Data or files are stored into a huge distributed structure that is fabricated using many geographically dispersed heterogeneous storage instruments. A number of largescale data grid projects have been developed that uses data grid to discover, transfer, store and manage distributed heterogeneous data. Some of them are Southern California Earth-quake Center (SCEC), the Biomedical Informatics Research Network (BIRN), SIO Explorer, GEON, and Real-time Observation, Application, and Data management Network (ROADNet) [9]. The Southern California Earthquake Center (SCEC) aims to gather terabytes of earthquake related data and to capture the knowledge about Geophysics models. The Biomedical Informatics Research Network (BIRN) enables storage, retrieval, analysis and documentation of biomedical data using data grid technologies. The SIO Explorer project is an online oceanography digital library for inquiring driven learning and provides resources for science, technology, engineering and mathematics to its users. The GEON is a collaborating project for developing cyberinfrastructure for supporting sharing and integration of 3D and 4D data among the earth science community. The Real Real-time Observation, Application, and Data management Network (ROADNet) facilitates use of geophysical, oceanographic, hydrological, ecological, and physical data to advance understanding and management of coastal, ocean, riparian and terrestrial earth systems. These projects utilize and cooperate with iRODS to transmit and manage the data, such as

PEARC '18, July 22–26, 2018, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-6446-1/18/07.
https://doi.org/10.1145/3219104.3229247

images, videos, and other data generated in the heterogeneous environments [5]. The iRODS enables collaborative data sharing and maintenance of distributed, aggregated, and integrated data storage collections.

Following a careful consideration of various robust features offered by iRODS, we have designed and implemented a community health grid, C-Grid as a solution to store, manage and share large amounts of these instruction materials and participant's health related data. This system will now be deployed for our partner organizations to participate in this study.

The rest of the paper is organized as follows. A review of the data grids, together with a description of iRODS and related work are presented in Section 2. The design and organization of the community grid is described in Section 3. The data analysis using the developed grid is discussed in Section 4. The conclusions and possible future directions are summarized in Section 5.

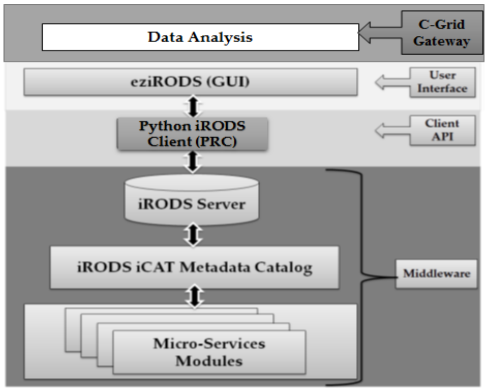


Figure 1: The architecture of C-Grid.

2 BACKGROUND AND RELATED WORK

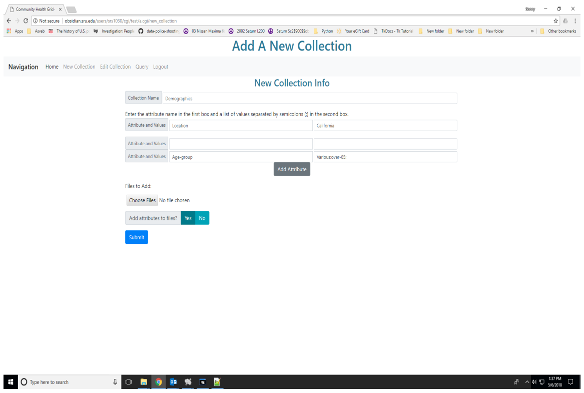
2.1 Overview of Data Grids

Data analytics is a rapidly growing field that has many new problems that need to be addressed as the field matures. Being able to efficiently store, retrieve, and analyze substantial amounts of data, on the order of petabytes, is a huge obstacle. Data grids are a popular solution; they allow data to be accessed and analyzed across a distributed system. Integrated Rule-Oriented Data System (iRODS) is a particularly powerful data grid system; it allows for metadata to be applied to all parts of its system, collections, users, files and entire zones. This metadata can be used for data discovery and workflow automation within the system. iRODS is equipped with a rule engine system that when certain conditions are met, background processes are automatically triggered [9]. This allows for automatic data discovery as users input data into the grid as well as data discovery by more manual means. iRODS provides methods to search through the metadata stored on the system.

There exist many systems that use iRODS to analyze and store data. One example of its use is in the DataNet Federation Consortium (DFC). The DFC grid is a Nation Science Foundation (NSF) funded project; the goal is to maintain long term access to data as well as provide computing services to the scientific community. Using the rule system in iRODS they were able to both collect relevant data and do prepressing on the data in order to feed the data into more specific data analyzing software. The rule system used the metadata attached to the collections and files within the system to collect the necessary data to be processed in to a form usable by the modeling systems. Specifically various rules were used to convert temperature, precipitation, wind speed, etc. into gridded datasets that were then fed into the models used to analyze the data [2]. Modern-Era Retrospective Analysis for Research and Applications Analytic Services (MERRA/AS) use of iRODS in their Virtual Climate Data Server (vCDS) stack. Here the microservice and rules sit between the application layer and the iRODS installation; this allows for a greater flexibility in tuning these rules to application specific workflows. For instance, this layer could be used to transfer existing metadata from another system to iRODS [9],[2]; this was used to transfer metadata for NetCDF objects into the example vCDS [2]. The initial analysis of the data in vCDS is also performed in this microservice layer as well; a wide scaled ordering and structuring of the data is performed [4]. While iRODS has been shown to be tremendously useful already, there is more to iRODS then just being a data grid. The usability of iRODS can be increased to include more data analytics. Since all of an iRODS installation's metadata is stored in a SQL database (either MySQL, PostreSQL, or Oracle)[9], there is opportunity of analyzing that data independently of iRODS. There is also opportunity to take more advantage of the rule system that exists in iRODS; the pluggable rule engine allows for rule sets to be written in other languages, like python and C++ for example[9]. Using this feature set we can start to more fully analyze the data, rather than just structuring and organizing the data to be fed into another system. iRODS has shown to be a very useful tool in data analytics but it has potential beyond being a data management system.

3 IRODS BASED COMMUNITY HEALTH GRID

In this research, we have implemented and deployed a web-based C-Grid to collect and archive high volume of geographically dispersed data including large video files and other data objects related to healthcare and educational training that can be useful for treating patients at an individualized level. These video files are generated from multiple educational and training events. Similar to ezSRB that was developed earlier [3], we have developed PYTHON-based ez-iRODS to access and interact with iRODS server of C-Grid with 10 TB size and RAID 5 configuration. The ez-iRODS is developed as a direct wrapper of the iRODS Client API, Python-iRODS Client (PRC)[9]. The C-Grid web portal interface provides input forms for users to interact with iRODS-server through ez-iRODS. The server for C-Grid web portal and the iRODS-server, is located at the Slippery Rock University, SRU's Obsidian server, and runs with Linux operating system (Red Hat Enterprise). C-Grid server is set up with PYTHON and PostGres database applications. This server provides functions for storing user accounts, verifying user accounts, communicating with the iRODS-server, and transmitting data objects. The architecture of C-Grid a web-based data grid system is shown in Figure 1, which is developed as a portlet of myCHOIS [3] [8] utilizing grid technology to manage and store health related data. Among the various data management systems, we will be using the iRODS because of multiple advantages. It has been implemented to serve as a distributed computing environment and data management system for sharing resources, data and computing power with the collaborators. It provides the collaborative data sharing and maintenance of distributed storage resource collections. The eziRODS (Graphical User interface) of C-Grid conveniently control and interact with the iRODS (Middleware) of the data grid located at the Obsidian server of SRU's computer science department. This system utilizing the Data Grid Technologies provides a long-term data preservation and allows user community to access valuable data objects conveniently through the user-friendly intuitive user interface from anywhere. Furthermore, we will utilize the newly acquired LAVA supercomputing system [1] to perform the data analysis of the health data stored and managed by the iRODS server hosted on the Obsidian server at Slippery Rock University.



Add A New Collection

Navigation Home New Collection Edit Collection Query Logout

New Collection Info

Collection Name: Demographics

Enter the attribute name in the first box and a list of values separated by semicolons (;) in the second box.

Attribute and Values: Location | California

Attribute and Values:

Attribute and Values: Age-group | Various;over-65;

Add Attribute

Files to Add:

Choose Files No file chosen

Add attributes to files? Yes No

Submit

Figure 2: Portal operating menu interface for viewing all collection names in C-Grid. When a user clicks a collection name in the list, the user will be relocated to the subcollection. Then all data objects of the subcollection are displayed on the main menu page for applications. The user also can go back to the current collection or the home collection, which is the starting collection when the user logins to the C-Grid web portal.

4 DATA ANALYISIS USING COMMUNITY GRID

The C-Grid web portal allows a user to define metadata for any file or directory stored in the iRODS-server or elsewhere in the network. In addition to the 'upload' file functionality, the GUI (shown in Figure 2) also provides users with the fields to input userdefined metadata that includes information about users, groups, collections, and locations of the data objects. A query in the interface assists users to search a file or a collection based on the user-defined metadata. Metadata search increases the possibility of locating the desired data with the finest recall and accuracy. This search functionality allows users to lookup files using terms or keywords recorded in metadata for the files. Hence, this interface acts as a metadata search engine that receives a metadata query from a user and sends that query to the iCAT server (a PostgreSQL server) for searching a matching file or collection and then returns the file that matches the queried metadata parameters.

The eziRODS with iRODS act as a virtual data management system. It can successfully utilize the Data Grid Technologies to provide a long-term data preservation and allow user community to access valued data objects conveniently through the familiar and the intuitive user interface everywhere. The simple intuitive user interfaces offered by the eziRODS abridges the complicated operating steps and approaches of the iRODS services for the users. Thus, C-Grid acts as a passage for accruing and disseminating multifarious data sets related to health education. Also, C-Grid equips users with many services for managing and creating data collections, for creating, retrieving and viewing data files, and for handling user-defined metadata for datasets. Some of the operations which can be performed on each data set stored on data grid via C-grid portal are (shown in Figure 2, 3, 4) : (1) search for datasets; (2) view and/or update metadata record ; (3) view (visualize) dataset (plot, image, etc); (4) download or upload data sets; (5) transfer data to other locations or resources accessible to the authorized user. (6) The Community Grid web portal can be accessed at URL:

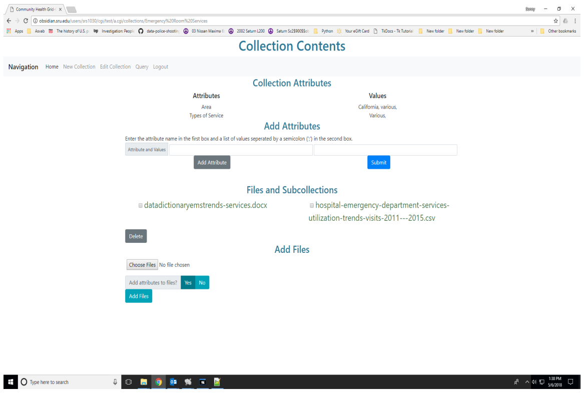
(http://obsidian.sru.edu/users/srs1030/cgi/test/a.cgi/).

The analysis of data stored on iRODS server is performed by triggering a job script, which ports the iRODS data onto the LAVA compute cluster to utilize Apache Spark MLlib library [6] to measure the effect of health education on the outcome on the participant's health, data from demographic, socioeconomic, personal health records, genomics, etc. The Lava cluster at Slippery Rock University (configured by XSEDE Cyberinfrastructure Resource Integration team) is theoretically capable of 9.8 TFLOPS, has approximately 2.25TB of RAM and 30TB of storage space.

We are currently utilizing the machine learning algorithms to find the correlation between the metadata stored on the iRODS server and the Health Outcomes. The prescriptive analytics performed using the iRODS data and the Apache Mllib can facilitate Doctors and administrators in using the critical data and information for supporting clinical, financial and operational decisions, which will aid them towards successful outcomes. As a result, prescriptive analytics can provide short-term and long-term answers to administrative and health concerns, such as efficient way of reducing health care costs as well as increasing the participant's overall health outcomes [7].

5 CONCLUSIONS AND FUTURE WORK

In summary, we have developed the user interface layer of C-Grid web portal, which act as multi-platform front-end GUI providing services (e.g., software tools at NCBI, CDC, WHO, etc.) to the users of the Community Health Organizations. Moreover, we have designed and implemented the C-Grid web portal that uses the PYTHON programming language and utilizes PRC, a PYTHON client API for iRODS which directly interacts with the iRODS-server. Furthermore, we have designed and developed three authentications to support system securities. The first authentication will be used for accessing the ez-iRODS (PYTHON wrapper), and the second authentication is for getting a permission to interact with the iRODS-server installed on obsidian database server and the third authentication will be used to access the LAVA computing cluster Finally, we have integrated the capability of iRODS as a middleware to facilitate flexibility and adaptivity in management of data workflows, and implemented user-defined data analysis functionality of the C-Grid



Collection Contents

Collection Attributes

Attributes	Values
Area	California, various,
Types of Service	Various,

Add Attributes

Enter the attribute name in the first box and a list of values seperated by a semicolon (';') in the second box.

Attribute and Values

Files and Subcollections

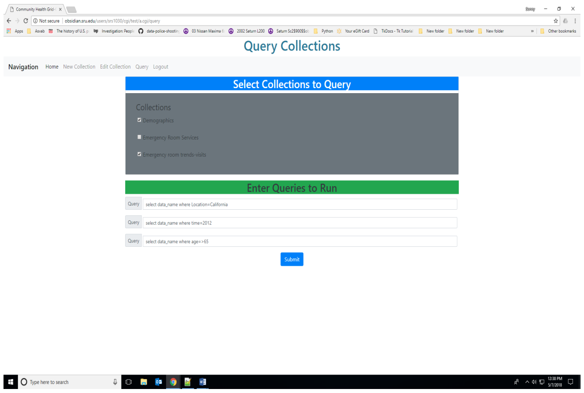
□ datadictionaryemstrends-services.docx
□ hospital-emergency-department-services-utilization-trends-visits-2011---2015.csv

Add Files

No file chosen

Add attributes to files?

Figure 3: The C-Grid web portal allows a user to define metadata for any file or directory stored in the iRODS-server or elsewhere in the network. In addition to the ’upload’ file functionality, the GUI also provides users with the fields to input user-defined metadata that includes information about users, groups, collections, and locations of the data objects.



Query Collections

Select Collections to Query

Collections

☑ Demographics
□ Emergency Room Services
☑ Emergency room trends-visits

Enter Queries to Run

Query: select data_name where Location=California
Query: select data_name where time=2012
Query: select data_name where age=>65

Figure 4: A query in the interface assists users to search a file or a collection based on the user-defined metadata. Metadata search increases the possibility of locating the desired data with the finest recall and accuracy. This search functionality allows users to lookup files using terms or keywords recorded in metadata for the files.

using the newly acquired LAVA supercomputing system to perform compute analysis of the health related data for prescribing outcomes.

REFERENCES

[1] [n. d.]. Slippery Rock University HPC with help from IU. https://itnews.iu.edu/articles/2017. Accessed: 2018-03-28.

[2] Mirza M Billah, Jonathan L Goodall, Ujjwal Narayan, Bakinam T Essawy, Venkat Lakshmi, Arcot Rajasekar, and Reagan W Moore. 2016. Using a data grid to automate data preparation pipelines required for regional-scale hydrologic modeling. *Environmental Modelling & Software* 78 (2016), 31–39.

[3] Arun K Datta, Victoria Jackson, Radha Nandkumar, and Weimo Zhu. 2010. Cyberinfrastructure for CHOIS-a Global Health initiative for obesity surveillance and control. *Proceedings in the PRAGMA* 18 (2010), 3–4.

[4] Daniel Duffy and John Schnase. 2014. Meeting the big data challenges of climate science through cloud-enabled climate analytics-as-a-service. In *Proceedings of the 30th International Conference on Massive Storage Systems and Technology. IEEE Computer Society.*

[5] Mark Hedges, Adil Hasan, and Tobias Blanke. 2007. Management and preservation of research data with iRODS. In *Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience.* ACM, 17–22.

[6] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. 2016. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research* 17, 1 (2016), 1235–1241.

[7] David B Nash. 2014. Harnessing the power of big data in healthcare. *American health & drug benefits* 7, 2 (2014), 69.

[8] Nitin Sukhija and Arun K Datta. 2013. C-grid: enabling iRODS-based grid technology for community health research. In *International Conference on Information Technology in Bio-and Medical Informatics.* Springer, 17–31.

[9] Srikumar Venugopal, Rajkumar Buyya, and Kotagiri Ramamohanarao. 2006. A taxonomy of data grids for distributed data sharing, management, and processing. *ACM Computing Surveys (CSUR)* 38, 1 (2006), 3.