iRODS and the RENCI Data Working Group (RDWG): Selected Case Studies

Howard M. Lander and Michael Shoffner, et al
RENCI Data Working Group
Renaissance Computing Institute, The University of North Carolina
Chapel Hill, North Carolina
{howard, shoffner}@renci.org

Abstract

The Renaissance Computing Institute (RENCI) is a multi-institute collaborative research center established in 2004. RENCI develops and deploys advanced cyber technologies with the goal of addressing critical issues such as the results of climate change and the genetic basis of alcoholism. RENCI maintains a Data Working Group (RDWG), which is responsible for providing leadership and strategic guidance for RENCI in the data technology area. In this paper we present the current status of several projects at RENCI for which RDWG and the project implementers believe iRODS will be a significant portion of the data architecture.

1. Introduction

The Renaissance Computing Institute focuses on the development and deployment of advanced cyberinfrastructure technologies and techniques as well as the creation and encouragement of collaborative projects and techniques. Our resources encompass a diverse group of people including domain scientists in oceanography, meteorology, chemistry, informatics and computer science. We also have a diverse set of projects spanning these domains and more, several computing clusters with an aggregate peak computing power in excess of 30 Teraflops and more than one Petabyte of spinning disk.

The RENCI Data Working Group (RDWG) was chartered in May 2010 as an outgrowth of discussions that began in late 2009. The creation of a persistent working group was motivated by the realization that RENCI had a number of ongoing projects with significant data challenges. Existing projects, knowledge and techniques were being confined to project specific stovepipes.

RDWG is responsible for providing leadership and strategic guidance for RENCI in the data technology area. Its responsibilities include data architecture, technology research, development and dissemination as well as education. RDWG focus areas include research-based data challenges such as very large-scale data sets, distributed data sets, multi-institutional data collections and novel analysis and visualization approaches.

As a part of its duties, RDWG often provides an informal forum for the discussion of data architecture and technology usage for projects in various stages of the planning, proposal or implementation processes. Because of its flexibility, extensibility and wide acceptance, as well as the close relationship between RENCI and the DICE Center, it is not unusual for iRODS to be investigated and included as part of a project plan. The following paragraphs present two examples of the usage of iRODS in RENCI projects in different phases of the project life cycle.

2. The National Climatic Data Center

The National Climatic Data Center, located in Asheville, North Carolina is one of the world's major archives of historical weather data. As such, it maintains collections of data going back over 150 years and includes data collected by Benjamin Franklin and Thomas Jefferson.

Included in these collections is a nationwide archive of radar precipitation estimates. RENCI and NCDC are currently collaborating on a data and compute intensive pilot project to construct a repeatable, scalable workflow utilizing this data set. The data processing pipeline involves several computational steps and several data management steps.

The computational steps are executed on RENCI's Blue Ridge, an MPI enabled cluster consisting of 2.8Ghz quad core Nehalem chips. The computation begins by combining 9 overlapping radar precipitation estimates to produce a single mosaic of precipitation estimates over a 10-year period. This data is then adjusted using external "truth on the ground" resources such as rain gauges. The end product of the computational portion of the workflow is a historical record of precipitation estimates gridded at a high resolution. The portion of the result set in which we are most interested is called Q2. The computational portion of the workflow is currently managed by hand with the aid of several perl and sh scripts by our collaborators at NCDC.

In addition to the computational portion of the workflow there is a significant data management portion of the workflow. This is the area of the workflow in

which we have begun to use iRODS and where we feel iRODS can make a critical contribution. As noted above the computational portion of the workflow is executed at RENCI in Chapel Hill, NC. The input data, however, is resident in Asheville at NCDC. The input data for the pilot project is in the low 10s of Terabytes. The output data is considerably smaller, but must still be managed. While it is true that 20 to 30 Terabytes of data is not an enormous load for our storage infrastructure, we cannot expect to move all of the data at once over the network between NCDC and RENCI.

There are also several other issues we would like to address. The number of distinct jobs that must be started and monitored in the computational portion of the workflow is in the hundreds if not thousands. Blue Ridge is a resource shared by many users and attempting to queue all of these jobs at once would overwhelm our computational resources and inconvenience many collaborators on other RENCI projects. In addition to returning the Q2 data set to NCDC, we would like to retain a copy here both for our own uses and to share with other collaborators of ours such as the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI). In the next paragraphs, we discuss how we expect to use iRODS to overcome the challenges in the data management portion of our workflow.

The first usage so far of iRODS has been to transport the data between the NCDC and RENCI sites. In the current configuration, NCDC has an iRODS client installed on a machine named icicle inside of their firewall; this client is attached to an iRODS server at RENCI. The files are ~~inserting~~ moved to RENCI using iput commands ~~–~~and appear in the iRODS server at RENCI in the normal way. The interesting thing about the setup is the file transfer speed. Using a standard scp transfer~~install~~ from icicle to RENCI yields a transfer speed of about 2.8 MB/s. Switching to the iRODS based transfer mechanism boosted the transfer speed to 32.8 MB/s, a speedup of almost 12 times.

The next significant challenge we hope to confront with iRODS is job scheduling. As mentioned above, the NCDC/iRODS computational workflow will require starting and monitoring a large number of distinct jobs. The naïve approach would be simply transfer all the input data down, and then queue up all the jobs. This would create several problems. On the data side, note from above that the size of the input data is 20 to 30 Terabytes. At the data rate of 32.8 Mb/s, this implies saturating the network continuously for 10 to 12 days. Even if this were technically feasible, it's not likely that the system administrators at either RENCI or NCDC would approve. This also implies a latency period of at least 10 days before the first job could run. On the execution side, this would involve saturating the Blue Ridge cluster with a large number of jobs. Since Blue Ridge is a shared resource, this is not a reasonable solution.

The iRODS based solution we are considering will tie the file transfer and job submission together. This will require the iRODS server to dispatch and monitor jobs. In the ideal scenario, the iRODS server would attempt to estimate the completion time of a running job: when the predicted ending time of a job approximated the amount of time required to download the input data of the next job, the iRODS server would begin transferring the data from NCDC to RENCI. Once the monitored job finished, the iRODS server would start the next job. In a simpler scenario, we could fetch the data for the next run from NCDC when result data from an existing run is inserted using iput into the local iRODS instance. The iRODS software would also maintain a queue for running jobs, so that all of the above discussion will apply to multiple concurrent jobs.

The other two data management issues that we will address with iRODS are well within existing common iRODS usage. Because we want to maintain a copy of the Q2 results here at RENCI, the existing computational workflow replicates the Q2 results as they are inserted. Sharing these results with collaborators such as CUAHSI will only require the use of the normal iRODS sharing mechanisms such as federation and user permissions.

3. The RENCI Sequencing Initiative

The RENCI Sequencing Initiative is made up of several projects involving genomic sequence data and analysis. Deep Sequencing Studies for Stimulant Dependence, with collaborator Kirk Wilhelmsen of the UNC School of Medicine Department of Genetics, is a primary project in the initiative. The Deep Sequencing project will identify genes of addiction using whole genome sequencing and genomic imputation using samples from multiple different populations and several complementary analytic approaches. The goal is to provide insights into the pathophysiology of drug dependence and inform prevention and treatment strategies. This project will require at least 2500 complete human genomes which, taken together, will constitute on the order of 100 TB of data assembled over five years.

Also part of the initiative is RENCI's work on the National Institutes of Health (NIH) Exome Project (http://www.nhlbi.nih.gov/resources/exome.htm) with Kari North from UNC Epidemiology and Ethan Lange from UNC Genetics. The overall goal of the Exome Project is to develop and validate a cost-effective, high-throughput sequencing application for sequencing all of the protein coding regions of the human genome.

The RENCI sequencing initiative involves data management and development of an analysis pipeline at

scale for these and other projects. One output of the initiative is a variants database generated against a reference genome. Annotations such as quality score, statements about the importance of the variant, and other metrics will be included in the database. Since this data set is relatively small, the current plan is to store this in a traditional relational database. For every single genome a consensus sequence, defined as the best estimate of the sequence, will be stored in a customized system based on the Hadoop[1] distributed file system. The size of this data is expected to be 3 billion bases per person with the base character and some metadata for each base.

The challenges RENCI Sequencing is facing are not unique to our initiative but are more appropriately understood as issues common to the discipline. Biological research is an increasingly data-driven field in which managing substantial raw data plus large numbers of analysis files is a task beyond the training and experience of most biologists. It is also beyond the scope of current data management technology. For instance, few biologists have either the background or the technology available to build or maintain a digital archive of their own data. Scientific journals increasingly require deposition of all data and analyses concomitant with publication of research articles, which places an additional burden on biologists who must now, without adequate technological support, track all relevant raw data sets and analyses and format them to meet a particular journal or repository's specifications.

RENCI is currently developing architectures for its sequencing systems and is planning to leverage iRODS where it can be of help. The first role for iRODS is to organize data flow using a grid for sharing among a number of collaborators. The architectures must support complex data flows; from a high sequencing throughput center to collaborators on campus to RENCI offices off campus to other collaborators at the Triangle Universities. iRODS can play a role guaranteeing that the data are~~is~~ where they~~it~~ need~~s~~ to be for computation and ensuring that correct policies for sharing are enforced.

The second role proposed for iRODS is as a data sharing/collaborating engine: leveraging the iRODS existing metadata collection abilities as files progress through the processing pipeline will allow users to locate files, understand the provenance of the data in these files and share both the files and their metadata with their collaborators. This functionality is in addition to what iRODS offers in conjunction with managing the lifecycle and policies on a large data set.

The third role under consideration is to utilize iRODS as a frontend to provide access to the consensus sequence for researchers. For this application the RENCI sequencing team would extend iRODS with a "Hadoop driver". Because we may store the consensus sequences in multiple independent Hadoop files systems, leveraging the ability of iRODS to federate distinct data resources could be crucial in providing seamless access to the entire data set. We may also eventually store useful metadata about each of the consensus sequences in iRODS. This would enable researchers to examine sets of consensus sequences that match user specified search criteria.

4. Conclusion

In this paper we have discussed several iRODS related projects at the Renaissance Computing Institute. As time goes on we expect iRODS to continue to be an important part of RENCI's strategy for addressing the data management and other data related requirements of projects in which we are participants. We also expect to continue to generate new iRODS requirements and to continue to develop innovative solutions using the iRODS system.

5. References

[1] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. "The Hadoop Distributed File System". In *Proceedings of the 26th IEEE Symposium on Massive Storage Systems and Technologies* (MSST '10), Incline Village, Nevada, May 2010.