Data in Motion: a new paradigm in Research Data Lifecycle Management

Princeton Workshop

July 18 – 20, 2011

Nicholas F. Tsinoremas, Joel Zysman, Christopher Mader and Jay Blaire

Center for Computational Science, Clinical Research Building, 1120 NW 14th Street, Miami, FL 33136

Computational Science is widely recognized as a critical component in solving many of today's most challenging scientific problems. The tremendous explosion of data being produced by modern experiments, observations and computer generated models presents a number of new challenges for research informatics organizations. Data-Intensive discovery (the fourth paradigm of scientific research), and Multi Scale Interdisciplinary science approaches are becoming more prevalent in the way that Science and Engineering is generating knowledge. The speed at which scientific disciplines advance depends in large part on how effectively researchers can communicate and collaborate with one another. In many cases the limiting factor to these advances is not the generation of data, but rather its effective management and analysis. This type of data management is not only highly dependent on robust and flexible data and network architectures; but also on the policies and procedures used to foster distribution of the data in a secure yet usable fashion. In this paper we will be describing some of the challenges that we have identified along with some thoughts on how to address these challenges.

1. Technology Challenges to meet explosive data growth

    The immediate challenges faced by any research informatics organization are those associated with simply storing and moving the ever increasing volumes of data produced by the modern scientific discovery process. We have progressed beyond the point where upgrading individual parts of the data management ecosystem is enough. Now hardware architects need to evaluate their entire operations and look holistically at the data and user needs. Until data finally reaches the archiving step of the data life cycle it must be viewed as being in motion. During this movement, the data must still be secured both physically and logically.

    In this paper we will discuss in more detail an example in the field of Genomics to demonstrate the challenges of the size and nature of modern data driven science.

In the 1990's dataset sizes in bioinformatics/genomics ranged from several to tens of gigabytes. However, in the last five years and with the introduction of new chemistry and instrument technologies a typical Next Generation Sequencing (NGS) instrument generates 4+ terabytes of data in a matter of a few days. Full production levels can reach in excess of 150 TB/year per instrument (40 week X 4 TB). Furthermore, processing this data can produce interim sizes of ten times that amount. At the University of Miami, we currently operate eight such instruments with plans to double the number in the next few years. These instruments are located in specialized wet lab facilities, far removed (relatively speaking) from the University Data Center. Like most places we use Ethernet networks for data movement. Given the location of the instruments and the size of data needing to be transferred we were forced to reengineer not only our campus network, but our inter-campus backbone as well in order to accommodate data transfer and the network service interruption during the daily moves of this massive data. Utilizing a DWDM ring between campuses, we are now able to assign 10Gb wavelengths for different needs. We have isolated the Lab and HPC/Research networks from the rest of the University's traffic to minimize the impact in all directions. But campus networks are only a part of the equation. Serious thought needs to be given to the architecture of the Local Area Network as well. Many functional data operations share common traits. Identifying the needs of the consumers of these data and identifying the common features can help optimize a network design for data movement and availability. The traditional hub and spoke design of most datacenters may need to shift to other paradigms such as top of rack switches and multiple network access points for different servers. Server connections themselves are another area that need to be analyzed. While Gigabit Ethernet is now ubiquitous in research facilities, the adoption of 10GbE and higher needs to be analyzed not only at the edge, but within LAN's as well.

Other technical challenges are faced when storing and processing this data. Data analysis and capacity planning now needs to be centered on the consumers of the data, the motion of data, and the utility derived from the data; rather than just the size alone. That being said, the size of data is still formidable. Instead of looking at file-systems ranging in the terabyte range, we must now look at data stores occupying petabytes of space, with hundreds if not thousands of clients needing access at any point in time.

One way we have addressed this issue is to look at how data is used within different analysis/modeling and production pipelines. After observing that a core data set from observation or experimentation may undergo many transformations during its "lifecycle" we determined that taking a multi-tier

approach to storage would deliver the best price/performance compromise. We have implemented and recommend a tiered structure as follows:

Tier 1: High-Speed storage designed for pure processing and highly parallel data manipulation (Data in motion)

Tier 2: Mid-Range storage designed for data presentation and mid-range parallel data manipulation (Data in motion)

Tier 3: Deep storage designed for long term storage, presentation of data, and single thread data manipulation (Data still in motion)

Archive: Near-line or off-line storage of past data (DATA AT REST, not really at rest but rather not accessed as frequently)

As can be seen from the above tiers, most of the data until archived is considered in motion and can move easily between all three layers. Data within the archive component can still be moved to the faster tiers, but with higher latency than motion between the other tiers. The size of each tier needs to be customized at each facility, but we have found that the ratio between tiers occurs roughly at an order of magnitude between them. Having data pools of this size requires new patterns of design in file-systems and archive solutions. We have utilized different forms of parallel file-systems for Tiers 1 – 3 (depending on data patterns and access requirements) and traditional archive methods for archive storage. But with increasing archive sizes, new approaches will be needed as that size approaches multi-petabyte ranges and even higher in the future.

What's used to analyze the data is as important as how it's stored. Commodity clusters are now used at most sites. These clusters have become easy to install and operate, but design is still often neglected. The traditional design of commodity clusters (and in fact Supercomputers) has focused on numbers of cores and Flops (Floating Operations per Second). This design trend is slowly starting to move to a more balanced approach with I/O being recognized as a high priority along with processing. However this trend needs to be adopted more widely in order for data processing to keep up with the data deluge. While most vendors are happy to provide statistics about processing, very few have gone as far with I/O operations. As customers of these vendors we need to be cognizant that Clusters and Supercomputers can no longer be viewed in isolation, but as a normal part of the data lifecycle.

The final piece to the data intense computational ecosystem is the consideration of the datacenter for the equipment. Datacenters need to be

designed and priced to house the advanced technologies required for data processing and presentation. Some aspects to be considered are:

1) Colocation Facilities: Unless datacenter design and operations are a core competency at an institution, co-location should be evaluated
2) Power requirements: All this equipment is very power hungry. Modern facilities should be equipped to handle much higher densities of equipment. Power ranges up to 20kw per rack are common for solutions
3) Cooling: Hand in hand with power requirements are cooling. Air flow studies in datacenters are critical to maintaining equipment
4) Connectivity: Connectivity is critical to any data intensive operation. Data must be presented to many clients (both computer and human) which necessitates high speed redundant access to systems

2. "The State of Data"

Data in motion means that data exists in a number of states from its generation to its eventual archiving. Normally, the scientific data life cycle can be described in the following process:

- Data acquisition
- Management and storage
- Processing/modeling
- Post-processing analysis and data mining
- Integration
- Decision Support and Knowledge generation and preservation
- Archiving

How this data is treated during these various stages is often different. Information and knowledge of how the data is understood and used is critically important. There is often a large difference in the nature of this data during these stages of development. Following our example, in NGS genomics, the 4-5 terabytes of raw data produced by a single experiment needs to go though an intense pipeline of several dozen steps including multiple QA/QC processes, assembly of all data, mapping them on to a given genome, to be able to call and understand the differences at the nucleotide level (polymorphism) and to conduct studies at the population level.

How long to maintain the data is also an important issue from validity, usefulness, and economic points of view. Also, who is to decide this is an important question that will need to be reviewed periodically. Continuing our NGS example from above, we work closely with research personnel to determine what data are primary for purposes of retention and further analysis. This type of definition is absolutely critical from both a technological and financial perspective. From a technological perspective, the correct storage architecture must be used to accommodate the processing needs of this data while in motion. From a financial perspective, defining a minimum acceptable performance profile is critical to make sure that the right cost storage is used in the right case. In our case using the Tier definitions in section 1 have enabled us to stage the data to the appropriate levels. In our case we support roughly 100TB Tier-1 storage, which costs roughly $2,000/TB, so it is important to use this space wisely. By staging data to Tier-2 storage which costs roughly $600 - $700/TB, we are able to keep far more data online for researchers at a much more reasonable cost. By extending this to Tier-3 ($300/TB) we can keep data sets online much longer than we could by using only one tier. Given the latest efforts by NIH and NSF to implement consistent data management plans and programs across all grants, keeping data online and usable for as long as possible will be very important. It is critical for systems architects to work not only with research personnel but also with IT administration to look for long-term solutions within existing IT strategies. How technologies and their costs evolve will impact the data preservation strategy. For example at some point generation costs vs. storage costs may shift for even very large data sets and produce changes in preservation strategies.

At the University of Miami for example, we utilize high performance NAS servers (using NFS) to present our data to our different computational clusters. High-performance storage is defined in this case as being able to provide in excess of 100,000 I/OPs, and 40Gb/sec of bandwidth. This performance is required in order for data files to be read and written to by over 1,000 simultaneous cluster nodes at any given time. While this storage is very high performing, access to it is typically restricted to NFS mounting. Adding additional protocols and access controls slows down the storage and can extend runtimes to unacceptable levels.

Once the primary analysis of data has taken place, the datasets are then moved from the high-performance file-systems to lower performing ones. We do this for technical reason as well as financial. Our Tier2 infrastructure is designed specifically for secondary analysis and ease of user access. This access is provided by several mechanisms. We use CIFS for remote access to the data from clients of all types (Linux, Windows, MacOS X). Researchers are able to perform different forms of secondary analysis this way. The data can also be presented to websites and informatics applications using protocols like WebDAV, SFTP as well as CIFS.

Once a project or dataset is no longer being actively developed, but still needs to be viewed, we move it to our Tier 3 storage, which has a lower performance profile than our Tier 2. While Tier 3 does not have the performance profiles of the faster gear, it is remarkably dense. This allows us to store, search and even visualize large data sets more economically than that of the other tiers. Access controls at this level are extremely fine grained. At this tier we utilize parallel file-systems across commodity hardware. While parallel file-systems are typically used for distributed access to data, we use them for redundancy and scalability; as well as for fine-grained access control.

At the end of the day, the value of the data is clearly in its productive use. Access to the data requires careful consideration at the earliest stages of designing the data management architecture, and planning for flexibility to accommodate a broad base of uses is key.

3. "Process as Data"

An increasing amount of data managed as part of the modern scientific discovery process is derived data generated by computational and analytical methods. The tools used to generate this derived data, as well as the versions and configurations of these tools at the time this data was produced, must be preserved along with the datasets. Absent this requirement, the data and any errors which might have been introduced cannot be fully understood and/or replicated.

This requirement should also extend to observational and experimental data which is collected and stored by the use of computational tools for the same reasons.

These requirements raise additional issues, especially in cases where proprietary tools and/or commercial software have been used in the generation or storage of the data, and need to be addressed.

In our NGS example the intense pipelines mentioned above include a number of instrument specific software and third party commercial software in combination with open source community based software. Using such a hybrid is really the only way that one can unlock the value of the data. Instrument companies develop their own software for primary and secondary analysis (closer to the technology of the instrument). However, tertiary, additional analysis and data mining is often done with a combination of open source, commercial and in house developed tools.

4. "Layers of Responsibility"

Data in motion creates layers of responsibility for data security and ownership. Who is responsible for what? Who has rights to access and use the data produced in the different parts of the discovery process? Data collections often have multiple

individuals or groups involved in the acquisition, generation, organization, curation, interpretation and use of these collections. Assembly of these collections may take place over time spans that can range from days to years to generations and possibly longer. How to "manage" this data with respect to these diverse interests and over these timespans presents complex challenges including technical, interpretive, legal, ethical, and economic to name a few.

The most formidable challenges here initially are likely to be organizational. What are the appropriate roles of central organizations including individual and teams of scientists, disciplinary societies, universities, libraries, government and private laboratories, for managing this data? Where should the lines of responsibility be drawn? Any comprehensive organization-wide approach will almost certainly require a substantial commitment of institutional funds to establish both the required governance and technical infrastructure to address these management issues in a thorough fashion.

Returning again to the NGS genomics example, it's very likely that any NGS groups would be operated as a core facility, and would centrally process samples from different departments within the same organization as well as, possibly, samples from external collaborating organizations. In the case of clinical samples (e.g., tissue samples from patients) being used for research purposes, there will be a whole set of regulatory (HIPAA and others) requirements that must be adhered to when handling this data. Patients must consent to the use of their samples for research purposes. These patients must also have the ability to rescind their consent at anytime, requiring the removal of their data from any future or ongoing studies. Any study involving patients will require a study protocol to be approved by an Institutional Review Board (IRB). This protocol must list who is allowed to see and use the data, as well as for what purpose. If the protocol is amended, investigators and key personnel may be added or removed. All of this information needs to be effectively communicated between the governing organizations, the investigators and the research informatics groups. Derivative datasets must be tracked and secured, and appropriate access privileges updated and maintained for the life of these datasets. From a purely technical perspective it is certainly possible to secure the data. However, in order to truly meet the combined regulatory and research requirements it is essential that the organization conducting the research has the requisite governance structure and policies to meet these requirements effectively. To do this, the design of the management structure and organizational polices should be done in close collaboration with the research informatics organizations who will be required to help enforce these policies.

It is very clear that these issues are extremely complex and challenging for any organization. It requires development of policies and a common understanding to best accommodate such critical needs. It also requires dialog with multiple Institutional

organizations across groups such as academic departments, administrative organizations, compliance organizations, IT and of course investigators.

5. “The Ubiquitous Use of Data” (from DATA to KNOWLEDGE)

Finally, the goal must be to enable the Ubiquitous Use of Data from any part of the collection, gathered or generated at any point in time, to be made available to researchers for the discovery of new knowledge, while still observing all regulatory and security requirements. The generation and organization of datasets by researchers working in discipline specific, or even interdisciplinary studies often cannot imagine how this data might be used in the future, especially by researchers in remote fields of study. Researchers in these remote fields may not be aware of the existence of data well known in the “generating” field, and may be completely unaware of the applicability of this data to the study at hand.

Making the entire collection visible in a meaningful way, while still respecting all of the relevant security requirements, will require the development of new software tools. These tools will very likely incorporate semantic, text and data mining, and computational linguistics (e.g., NLP) technologies to build easily searchable and accessible catalogs and indices of these collections. For example, a “Smart Collaborator Tool” might be developed that could “understand” the study at hand and discover data previously generated by research studies in other fields that should be considered as informative by the current study. There are indeed currently funded projects that are developing the underlying technologies to develop these tools, but fortunately for those of us interested in this subject, there remains much interesting work to be done to apply these technologies and create novel systems to help optimize the use of these collections as resources for the ongoing discovery of knowledge.

CONCLUSION

We put forward the notion that in addition to the unprecedented increase of the data volume we experience in science and engineering we also need to consider the constant data movement that is required in order to extract information and ultimately knowledge. In summary, this paper presents a number of issues for consideration, but those are just a subset of the complex set of interrelated challenges of an integrated data management system. There will naturally be multiple approaches to meeting these challenges, yet it is more than worthwhile to produce examples of best practices as they emerge.