DECEMBER 29, 2017

NYSPI STORAGE SERVICES OVERVIEW

A REVIEW OF DATA STORAGE OPTIONS

NEW YORK STATE OF OPPORTUNITY.

PsyIT
New York State
Psychiatric Institute

CONTENTS

Abstract

Historic proliferation of data storage solutions is inefficient given modern and emerging data needs in the Psychiatric Institute. A more strategic approach must be employed to reduce costs associated with administering, maintaining and operating storage solutions; to limit risks of both inappropriate data exposure and data loss; and to enable advanced and rapid access to data for collaboration and research indicatives, inside and outside the walls of the Institute.

Storage solutions must be more than just hardware storing our most vital asset – our data. Storage solutions must support data of various access methods, classifications, performance, retention requirements and user bases.

This paper attempts to describe the available and emerging storage solutions at NYSPI and how these solutions address these crucial factors.

Administrators and investors must use the most appropriate storage solution not only for their own direct needs, but also with the overall Institute mission in focus, for NYSPI to advance in its core mission.

The Need for a Data Storage Evolution

The New York State Psychiatric Institute continually endeavors to be THE world-wide leader in psychiatric research. As such, not only must our faculty and research staff continue innovation in research methods and approaches, but as an Institute we must also continually assess the tools used in these efforts. While we often focus on the latest "cool" consumer technologies, our foundation is storage. Data is the raw material from which the Institute's outstanding faculty can press us forward. Handling and processing that material must be increasingly efficient, not only because of ongoing fiscal overhead constraints but also because of the opportunities that exist as we manage more and larger data sets.

It's not sufficient to just expand current approaches. Often, how data is handled today introduces real risk to our work, or even to the Institute as a whole. While generally NYSPI has been spared, we all have heard stories from colleges impacted by localized disasters. Computer failures further threaten data loss or, minimally, the availability of our life-blood – our data. Existing and traditional robust backup or disaster recovery solutions cut fiscally, through direct costs in staff who performing these coarse data management services or indirectly through expensive redundant servers and storage platforms.

Traditional NYSPI Approaches to Storage

The most traditional approach to storage is ad hoc, "local" storage; datasets and documents are stored on PI/RA workstations, copied to others as needed. While this provides very easy access, it introduces numerous liabilities. First, without administratively expensive efforts, this provides no routine and reliable backup of the data, resulting in risk for data loss. Second, it is not conducive to sharing of data directly, relying instead on copying files for other's contributions. This, of course, leads to issues merging changes in data, overhead maintaining current copies of data for all who require access and makes it nearly impossible to remove access from those who may no longer require the data. This later issue is particularly problematic since the Institute supports use of properly-configured personal devices.

A slight extension of this local model, ad hoc local sharing attempts to address some of the issues with copying data, particularly larger files such as imaging data. In this scenario, a workstation is configured to act like a file server so that others may access the data directly from that workstation. This suffers from other drawbacks of the aforementioned model – risk of failure and data loss due to device failures and typically a lack of efficient backup or good access control capabilities but leans towards a much more accepted approach – network file services.

In the traditional NYSPI network file services model, dedicated storage servers are used to keep the data secure. These servers are typically configured to provide protection against hard drive failures and to ease backup operations. When configured with the NYSPI directory, access control is easy to maintain either through requests to psyIT or through direct administration of

access groups. This eliminates duplicate data issues – costs associated with storing the same data repeatedly, overhead keeping copies of data in sync and handling the issues of recalling data from those individuals or devices which should not have the data in the future. Data is also much easier to physically secure, providing both protection from theft as well as environmental issues, such as power and cooling problems. On the flip side, this approach traditionally limits access to NYSPI networks and does require dedicated hardware which can be costly and time consuming to procure and configure. Use of shared storage services, rather than storage installed exclusively for a program or Division, can limit the later issues; use of hybrid storage options emerging in NYSPI provides some options for the former.

Modern and Emerging NYSPI Data Storage Models

Revising our approach to data storage can minimally provide efficiencies, resulting in cost controls and savings, and potentially enable research advances through newer analytic approaches.

At the basest level, NYSPI has for many years leveraged consumer-level cloud sharing services, such as consumer-grade Dropbox or Google Drive, to help facilitate collaboration, particularly outside the Institute. While this does little to eliminate duplicate storage impact, it does provide some level of redundancy and backup. Diligently managed, this can also help limit who has access to data. On the flip side, consumer-grade services often cannot provide protection for non-public information; PHI, including de-identified PHI, cannot be stored in these services. Microsoft OneDrive, introduced to NYSPI with Office 365 in June 2017, supports stronger data confidentiality controls and does allow PHI. Sharing typically remains ad hoc and is explicitly limited to those within the NYS Office365 environment.

In 2017, NYSPI introduced not only OneDrive, but also enterprise-level Dropbox and Google services, the later added in late December with rollout expected in early 2018. These provide some advantage to their consumer-grade cousins in terms of data rights protections – limiting how long someone may have access to a shared document or folder, for example, and, in the case of Google, offering increase storage capacity and a broader set of services than mere storage. There is inherent inefficiency in managing multiple similar solutions, so this will be evaluated in early 2018 and feedback on these competing solutions is welcome.

At a larger scale, NYSPI's central IT organization – psyIT – expanded ITS shared storage service. Supporting a clustered and tiered-storage model, file sharing within the Institute can now occur with little or no direct cost to programs. Significant storage needs, such as those associated with new grants, can be planned and storage procured, but work can begin immediately, without need to wait for procurements and configuration of new equipment. Physical and logical controls support efficient operations and sharing of data. Backups are centrally managed, limiting program-specific administrative overhead. Delegated administration of data access can be provided where needed, to ensure local control of data access, without the delay of request processes.

For small team collaboration, including with a limited set of users outside the Institute, the Office 365 SharePoint platform provides not only secure file storage, but richer collaboration approaches for team-based efforts. Calendaring, task lists, and information sharing are all native strengths of SharePoint. Real-time document collaboration is also possible. SharePoint can work similar to document synchronization services (e.g. OneDrive) or can be used for secure file and information exchange.

In 2018, NYSPI expects to greatly expand the use of cloud hybrid storage solutions, particularly with the introduction of Amazon Web Service storage solutions. This will further increase agility and efficiency, as storage is dynamically expandable and data backup / archive costs less than a $0.01 per gigabyte. Perhaps even more exciting, use of these hybrid storage models will allow the Institute to achieve hybrid data processing – where data is collected at high speed on site but synchronized to cloud storage, where cost efficient processing services can analyze the data in the cloud – eliminating network latency and providing for the dynamic processing power not affordable in on premise models. This further has the ability to eliminate duplicate data, as data can be exposed to multiple processing sources transparently.

Storage Selection

To efficiently select and implement data storage, information owners, such as a principle investigators, clinical and operational department heads, and administrators, must evaluate their needs – both immediate, day-to-day needs as well as compliance requirements, emergency access and longer-term audit and archive demands.

Details vary greatly by use case, but below are some questions to help frame storage requirements.

Data Confidentiality	• What is the CUMC and/or NYS CONFIDENTIALITY rating • Is the data covered by confidentiality regulations (e.g. HIPAA, 42 CFR Part 2) • What is the impact if there is unauthorized access to the data
Data Integrity	• What is the CUMC and/or NYS INTEGRITY rating • How bad would it be if my information were unexpectedly or inappropriately altered
Data Availability	• What is the CUMC and/or NYS AVAILABILITY rating • How long can I operate without access to the information (recovery time objective – RTO) • How much data can I lose if my system fails (recovery point object – RPO)

File Access	• Will information need to be updated outside the Institute (e.g. from the Internet) • Will data processing occur externally (e.g. cloud-based analytics)
User community	• What is the user community who needs to be able to see or modify the information • What is the federal/NYS Identity Assurance level • How sophisticated is the user base • How do the users access other stored information • Who will provide level 1 support for the users
Performance	• What is the impact of access latency • How frequently and rapidly is the data updated
Size	• How much data do I have • How is this likely to change over time (scalability)
Data format	• Am I managing individual files or large relational data • Does my system require raw (block) storage
Costs	• What is the purchase cost • What are the recurring costs (maintenance) • What is the overhead for managing the solution and access to the data • Are there ways to share the costs or leverage available resources • Who is responsible for CUMC or other oversight fees • What are the overhead costs (e.g. cooling, network infrastructure, physical security, power, space)

Many storage solutions can support various degrees of service. As such, tools and training which can support quick review of needs and guide storage decision making will empower wise data storage solutions and MUST be developed.

For example, a system may support high availability and rapid disaster recovery but only if configured to do so. Additional components which are required to meet full availability needs introduce additional costs. These configurations also may not be available by default or, available at all in certain shared solutions. Conversely, solutions which over-deliver may introduce unnecessary fiscally burden in many use cases.

A collection of storage solutions may be appropriate in many cases. For example, data collection by instrumentation or through online applications may require high-performance/low latency transaction processing but RTO/RPO may support much lower cost backup and archiving solutions.

While one size will not fit all, we must make the evaluation, selection and implementation of the right storage solution simple.

Leveraging Efficient Storage Solutions

Historically researchers leverage or expand storage in the methods that have worked in the past. If there's a new project, staff get new PCs with bigger hard drives. Divisions may purchase a standalone NAS server for storing interviews or imaging data. These are easy answer; little thought is applied. NYSPI's growth depends on staff doing better.

Improvement begins with documentation of needs. The proper storage solution cannot be selected if information classification is not conducted, regulatory requirements considered, and true total cost of ownership evaluated. This does not need to be lengthy process – those who understand the project and data typically complete these evaluations in under 20 minutes.

From this evaluation, solution alternatives can be presented, and research and administrative staff can request the service that most effectively and efficiently addresses the needs. Partnering with the expanding Research Operations and Compliance initiatives regarding human subject research data governance will further open the options, as identifiers are decoupled from research data where appropriate and in a standard manner. These processes will help limit compliance demands, particularly ease external collaboration and cloud solutions.

Second, psyIT needs to become more agile, developing tools to request and provision storage and manage access to data with limited or no psyIT interaction, though delegated administration and join maintenance of day-to-day needs. Storage options and operations must be well described, role-based training provided as needed and services expanded appropriately. Funds historically supporting inefficient services must be shifted from aging, capital-centric models to operation-centric solutions.

Third, as the newer solutions are developed, both core and focused programs must be open to use of and support integration with these newer solutions. For example, imaging services, including NYSPI MRI, must explore synchronization through Amazon file gateway services. Epidemiological research data must explore retention requirements and pursue tiered storage services. Cloud-based backup must become the norm rather than outlier. As these newer storage solutions mature and demonstrate viability, less advanced and agile programs may need to plan for these transitions.

Potential Evolutions

Data availability and the small, workstation-based data need

In response to the World Trade Center tragedy, staff in the HIV Center developed procedures for backing up data on their workstations. This solution was intended to ensure, should there be an event which devastates a specific Institute locale, data would remain. Research staff kept their data in appropriate folders. Local IT staff would ensure backup servers and workstation

software was installed and functioning, handle the tapes used with the backup system and perform data restoration when needed.

Recently this backup solution failed. The server which formed the hub of the operation stopped working, as aging solutions will do. Staff reviewed the solution to restore using more robust infrastructure only to find that backups were not happening. The most recent data was over a year old.

Rather than simply reinstating that service, the HIV Center is now moving to leveraging Microsoft OneDrive for Business, a service included in the Office365 platform to which NYSPI migrated in mid-2017. This provides world-wide backup of data and versioning, with data synchronized automatically between the cloud and workstation. It even allows for limited, ad-hoc sharing of files between Institute team members and is fully HIPAA compliant.

Not only does it provide increased availability of the data, as noted, there is no additional cost for this solution, freeing staff who manage data for the Center to focus on more meaningful work.

Cloud-based backup and processing for imaging data

Some of the largest dataset processed in the Institute originates from the MRI suite. This reality, and it's associated challenges, will increase as the Institute upgrades the MRI platform to a high-resolution solution this summer.

Currently MRI staff encrypt and synchronize data backups to Amazon Web Services (AWS), using their "Glacier" backup solution. Archive costs are extremely low – less than a penny per gigabyte. On the data processing side, however, images are stored locally using an enterprise-grade storage platform from which researchers can then transfer their images for processing.

In the first half of 2018, we are hope to both expand the use of Glacier services and provide not only archival but real-time data processing and data synchronization to AWS. This will streamline the MRI backup processes slightly, but also provide this service to other imagine groups, such as those in Integrative Neuroscience who have been working to address animal imaging needs.

Further, it's hoped that this will allow staff to process data in the AWS cloud, not only improving performance through elastic processing capabilities which would be cost prohibitive locally, but also enabling cross-institute or even international collaboration, since the original-form data will be secured but no longer bound to the walls of NYSPI.

Conclusion

NYSPI's Strategic Plan for Research, presented in July 2017, establishes bold but realistic growth objectives. Among core missional goals, the report calls for investing in research infrastructure, specifically data storage, as a catalyst for faculty success. As psyIT provides this infrastructure,

faculty and staff must demonstrate boldness and innovation in not only the research we perform, but the methods by which we conduct that research. These venues will further support the envisioned collaborative partnerships and enable NYSPI to a become and remain a leader in strategic scientific priority fields.

Even where innovation is limited, these new models and services will help reduce costs and complexity while improving compliance and security. NYSPI, in particular psyIT as the provider of these storage services, must continue to support change to inefficient solutions and must partner with faculty, researchers and staff in the evolution of research and administrative services.