REPORT

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Storage requirements for large engineering data sets

Author(s):	Thomas Parsons & Gary Smith
Audience:	Project stakeholders & JISC MRD
Published:	14/06/2013

Contents

1. Introduction

The ADMIRe project started in October 2011 and aimed to complete research data management pilots in all five Faculties at The University of Nottingham (Nottingham).

The Faculty of Engineering were approached and a number of potential pilots identified. Of these, the Applied Optics group which is part of Electrical and Electronics Engineering[1] were considering the processes of migrating data from legacy equipment to secure infrastructure provided by IT Services.

ADMIRe approached IT Services support staff and sought to analyse and document the needs of the researchers. It was hoped that this pilot would be a good use case of the current practice of researchers and identify the challenges faced when considering IT Services infrastructure. The following report outlines key comments and findings that should be taken into account before the migration of their data from local to networked storage.

2. The use of research data

The views of 12 researchers from the group were sought and categorised accordingly. As this pilot took place after the main ADMIRe requirements gathering phases, this in-depth pilot into working practices was deemed a useful exercise with which to validate the overall ADMIRe RDM requirements and system models.

Research data principally takes the following forms:

1. Raw data from experiments (microscopy data)
2. Images
3. Documents

Size of data varies significantly and there were range of estimates based upon the type and complexity of the experiment:

- *"5~10 GB of data a day"*
- *"500Mb a month"*
- *"50GB of data per experiment day"*
- *"My experiments generate ~4GB of data per week in total, spread across several machines, stored on the local hard drive of each."*
- *"The TIRM microscope used to eat 3Tb drives for breakfast. Using about 20G per hour when it was imaging - we've since removed the high res cameras from the setup and now it's using about 5 gig per hour."*

These comments demonstrate that certain experiments can generate significant amounts of data, but that data collection does not happen all the time and is limited. However, other members of the group are collecting data continuously:

[1] http://www.nottingham.ac.uk/engineering/research/electricalsystemsandoptics/index.aspx

"I also analyse images from motorway CCTV cameras - every day I collect about 200M of images. I have about 75G of these so far. Each image is tiny - maybe 15K tops"

Demands such as these mean that the majority of the group are using portable storage devices to store their data:

- *"The data is stored locally on external USB drives at the acqusition PC"*
- *"For my PhD I have approx. 100GB raw data files stored on my own HDD and backed up to one other computer in the office."*
- *"I consolidate the aggregate of each experiment to a USB disk every day, and transfer it to my laptop where it is stored straight into Dropbox"*
- *"stored on the local experiment machine and backed up to an external USB drive after each experiment"*
- *"Stored on the local hard drive and transfer it to my office PC by a USB hard disk."*

These comments are not atypical and indicate a strong reliance upon portable storage; particularly if as in this case, the machines used to run the experiments are legacy hardware or run Linux. What is of interest is that researchers are aware of the need to back up their data, with the majority specifying that data should be moved or at least backed-up from the experimental machine after each experiment. This indicates a level of good practice or an awareness of the risk associated with the machines used to generate the data.

Others utilise a shared departmental Linux server if their research data volumes are relatively small:

"I save the data in /home/scan/ where anyone on the system can access it…I generate about 500M a month."

The department Linux server provides a shared data service to members of the group through a SAMBA networked drive. Others choose a number of methods:

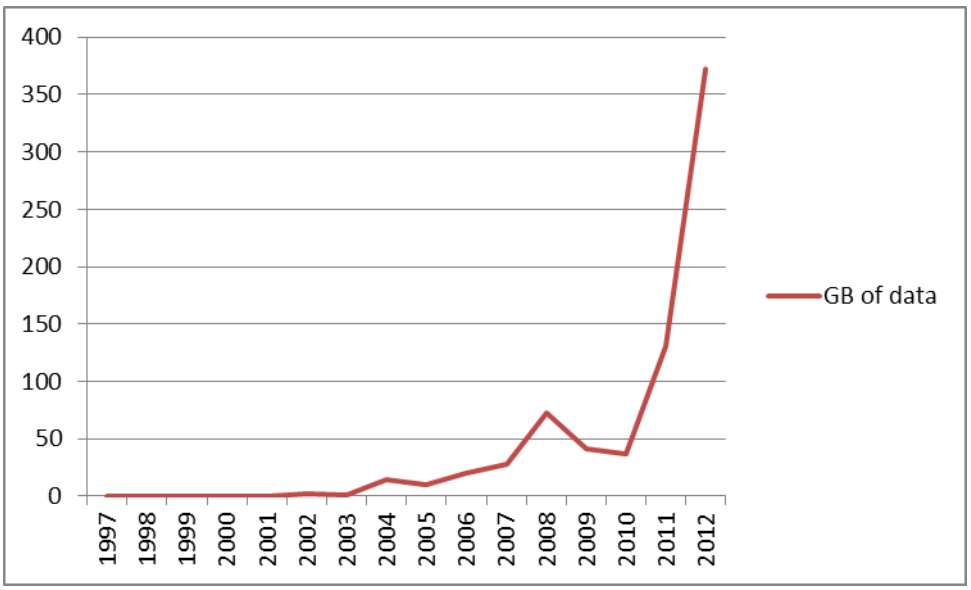
- *"For the project I'm working on now there's approximately 20GB data with the central repository on SharePoint"*
- *"We have used SCP/SFTP to share data with people off campus before"*
- *"I pay for my own Dropbox facility and would love to have a similar facility (a local directory, stored on each machine I use, which is mirrored automatically to a remote server which is also internet-accessible with settable permissions and backup which I don't have to care about until something goes wrong."*

These more technical methods are supplemented by simply sharing hard drives for larger data sets, as one group member commented:

"Large hard drives or pen drives are used depending on the size of the data."

The total size of the data generated by one research group within the wider group was estimated at 20TB each year, with a request for 30TB of networked space as the group projects and data output grows. This figure does not include legacy data or backups that a stored elsewhere, but rather represents working data requirements rather than archive requirements.

Therefore the amount of data held overall by the Applied Optics group is very significant and >30TB. Compiling a graph of the data held by the Linux server alone, reveals the extent of how data is growing within the group:



It is also worth noting that although the volume of data held on the Linux server is approximately 350GB, the backup server to this holds nearly 5TB of daily backups and recovery data, should this server ever fail.

The working practices of the group do take into account good RDM practice, yet the size of the data held and the potential for new data generation suggests that consideration should be given as to what can be provided. Researchers acknowledged that *"some projects individually account for one fifth of the space utilisation."* One example highlighted that a change in instrumentation to a less precise set-up did not harm the research, but significantly reduced the data collected. Other commented that they never delete data:

"Quite a lot of the data in a stack isn't usable (out of focus for example) but 3Tb drives are cheaper than our time in working out which images to keep - and also saves the worry of erroneously deleting useful images, that we thought were useless, by mistake."

This raises a serious cost element to providing storage that is both scalable and cost-efficient long-term[2], particularly if the data is unlikely to be used again. Thus suggesting the options of tiered storage would be welcome:

"I'd like to add that I don't need quick access to all of my data – some of it could be stored offline somewhere - so long as there's an easy way (and not too slow - say within a day) of bringing it back online so I can access it again!"

Clearly experimental data which is being analysed should be as close to the analysis machines as possible, and the issue of network lag was touted as a negative against bot having local storage available. Yet, once analysis has finished then the majority of data could be archived, providing an index function was provided:

"Being able to access an index of what I've got stored offline would be dead handy though - imaging being able to go through a directory structure to see what the files are and where they are, but having to wait a while if you want to read a file."

This level of indexing does suggest the creation of a manual data catalogue or indexing via a search engine, but ties in well with the main findings of the proposed ADMIRe RDM system.

3. Further work

The results of this study by IT Services will form part of an offering to the group to replace their legacy equipment. This pilot will be looked on with interest by both IT Services and other research departments across Nottingham – many of whom will require the management of similar amounts of data.

What is clear from this pilot is that each department when engaged has the potential to unleash TBs of data on a centralised storage system, as one researcher remarked:

"[I] Can't really tell how much data I generate, have to try and figure out some average. Sometimes a lot, sometimes little."

So scoping and deciding who should be able to access and utilise expensive storage will become a growing problem in itself.

[2] http://datapool.soton.ac.uk/2013/03/21/cost-benefit-analysis-experience-of-southampton-research-data-producers/