Long-Term Data Storage

Increasing data growth is outweighing the IT budgets of many biopharmaceutical and life sciences companies. In order to comply with legislation, the industry must address this issue and decide on a new data storage strategy

Jim Cook at Arkivum

The growth in the amount of data generated by biopharmaceutical and other life sciences organisations each year is far outstripping any increase – or more likely decrease – in the companies' IT budgets. Working within the highly regulated world of a GxP environment means that all such organisations must address the archiving of large portions of this data to comply with regulations. This article looks at the requirements for long-term data storage for life sciences organisations and some of the options for how to address them.

Data versus IT

In a 2011 report, McKinsey projected a 40 per cent growth in global data generated per year, while growth in global IT spending will manage only five per cent each year (1). Against this backdrop, McKinsey believes that the value of making use of big data to the US healthcare market could be $300 billion – more than double the total annual healthcare spending in Spain. It is also known that as little as 10 per cent of all data generated is accessed more than once.

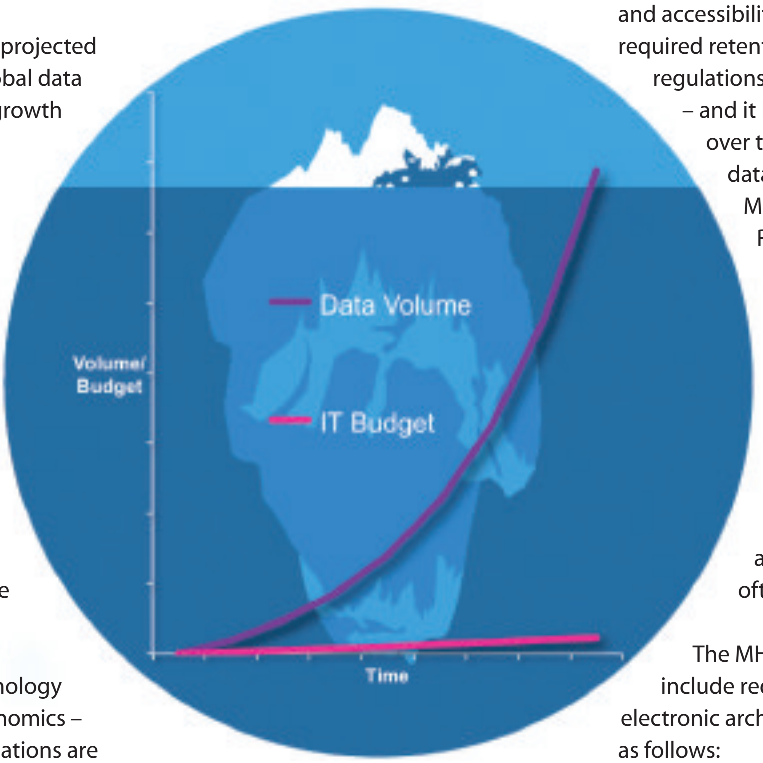
With the advances of technology – such as bioimaging or genomics – biopharmaceutical organisations are well aware of the data growth that they produce and many far outstrip the 40 per cent growth projected. Next generation sequencers alone can produce over one terabyte of data every two weeks.

GxP Compliance

Life sciences companies must comply with a large number of GxP regulations. There are many guidelines that stipulate extensive data retention and archiving requirements, including:

- (SI 2004/1031) the Medicines for Human Use Clinical Trials Act, UK
- (SI 2006 No. 1928), §31A Trial Master File (TMF) and archiving, UK
- European Union rules governing medicinal products, Volume 10, Chapter 5: Recommendation on the content of the TMF and archiving
- Regulations from the US Food and Drug Administration (FDA) and the International Conference on Harmonisation

Figure 1: Data volume versus IT budget



Source: Arkivum

All of these regulations take a similar viewpoint, in that archiving has to ensure the integrity, authenticity, confidentiality and accessibility of documents for the required retention period. For all of these regulations, the word 'access' is key – and it has had big implications over the timescales for which data has to be kept. The Medicines and Healthcare Regulatory Agency (MHRA) Good Clinical Practice (GCP) guidelines have retention times of at least five years, but often this can be 15 or 30 years. The FDA requires that sponsors and investigators retain records and reports for at least two years, but often longer.

The MHRA's GCP guidelines include requirements for how electronic archiving is addressed, as follows:

- More than one copy
- More than one location
- Different formats, media and manufacturers
- Access controlled
- Authenticity protected
- Validated and auditable migrations
- Periodic retrieval/restore to test access
- Demonstrate no loss or corruption

These requirements mirror best practice in data preservation, which focuses on holding multiple copies of the data, kept in different locations. This must be supplemented by using diverse technologies to reduce the risk of multiple failures. These copies must be actively managed by migrating to new storage or formats to address obsolescence, and by regularly checking and repairing any loss of data integrity. This ensures that if there is a problem with one of the copies, then it can be replaced by replicating one of the other working copies.

Total Cost of Ownership

Most organisations underestimate the long-term total cost of ownership of archiving, especially where the stringent data retention requirements of GxP compliance need to be met. Long-term archiving requires specialist expertise, active data management, the procurement and migration of systems to address obsolescence, and regular auditing to make sure retention metrics are being met.

With a steady advance towards a fully digital document lifecycle, many companies focus on the initial parts of the data lifecycle, rather than the entire lifecycle costs. For instance, when a business builds a business case for a 21 Code of Federal Regulations Part 11 compliant document management system, the focus is usually on putting proper digital signature processes in place to ensure authenticity. A deeper analysis, however, reveals that the main costs are typically in the long-term where integrity, authenticity, confidentiality and availability need to be delivered over multi-decade-level retention periods.

The IT system is also an important factor. It is not enough to write files onto a hard drive and put it on a shelf, as there is a strong likelihood that eventually data will be lost due to drive failures or become unreadable as technology becomes obsolete. Countering these problems requires an archive infrastructure that proactively manages data to ensure it remains intact and accessible. Many organisations are either unaware of these issues, or cannot afford to do this in-house, resulting in an increased risk of data loss.

It is clear that the complex and potentially costly requirements associated with compliance and long-term data retention can have a very large impact on IT budgets. However, as McKinsey and others have pointed out, we work in times when growth in data volumes is massively outstripping the growth in IT budgets.

Data Archiving Options

The most common approach to archiving data in corporate environments is to retain data on the same systems where it was first created – typically enterprise storage servers. The capacity of these servers then grows to match ever-increasing volumes of data and the need to keep ever more of it for compliance or reuse. This approach is expensive, difficult to manage, and can put data at risk – for example, through hardware failures, accidental data deletion or modification. Keeping infrequently accessed and static archive data on expensive enterprise storage servers is a luxury in today's challenging economic climate.

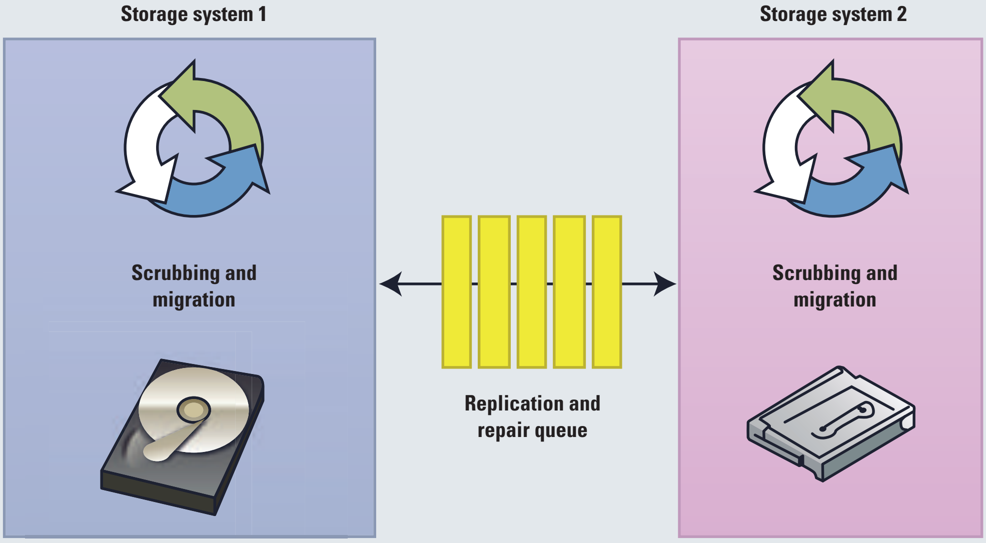
The most reliable and cost-effective format that is best for long-term data storage is Linear Tape-Open (LTO) data tape. Power requirements are low in this method and data densities in modern libraries are very high, while backwards compatibility is good, with a well-defined roadmap going forward that guards against obsolescence. Adoption is also high across almost all industries, so good availability and support is assured.

Linear tape file systems make data tape storage open and interoperable, so LTO is ideal for offline escrow copies, in addition to supporting high-speed rapid access which is possible through automated tape libraries. LTO is also more cost-effective for archiving than running spinning disk drives, with research from the Clipper Group finding that the total cost for the long-term storage of data over a 12 year period on a hard disk is about 15 times more expensive than data tape (2).

A service-oriented approach can be used and delivered by the enterprise, but with applications such as digital data preservation, outsourcing to a

> “Most organisations underestimate the long-term total cost of ownership of archiving, especially where the stringent data retention requirements of GxP compliance need to be met”

Figure 2: Active archiving



Source: Arkivum

specialist provider can deliver significant benefits. An alternative to this is to outsource archiving.

Regulators recognise the benefits that can be derived from contracting specialist services from external suppliers. For example, GCP guidelines specifically addressing the contracting out of archive facilities state that:

- It is fine to use a commercial vendor for storage
- The responsibility lies with the sponsor/investigator
- Integrity, confidentiality, quality and retrieval must be subject to satisfactory service level agreements
- The suitability of facilities must be assessed in advance
- It is critical to use a formal contract with an archive company
- The location of the documents and records must be known at all times

Cloud Storage

Cloud storage can be attractive as it replaces large in-house cap-ex costs with much more manageable ongoing op-ex costs. However, most cloud providers offer no absolute guarantee of long-term data integrity, because typically their focus is on storage for back-up or for hosting live systems – the requirements for which are different to archive – as information is continually being accessed and the ability to achieve high availability and rapid recovery from failures is needed. As a result, cloud storage services do not usually have all the checks and measures in place associated with long-term data retention where integrity needs to be guaranteed for more than a lifetime.

Ongoing Security

In order to mitigate the risk of data loss, organisations should keep several copies of the data in different locations – both online and offline – secured and actively managed. Active management is needed to ensure media is migrated regularly, the condition of multiple data copies is monitored, and all actions are tightly controlled to guarantee proper processes are followed and only by those suitably trained. Security should be applied through user access control, authentication and ISO 27001 hosting environments. Above all, every aspect of the system needs constant and thorough monitoring for a full and up-to-date view of who has done what and when, both inside and out of the organisation.

Additionally, if archive hosting is outsourced, a succession plan must be in place in case the service provider ceases to operate. This could include the requirement that at least one copy of the data is kept in a separate escrow account.

Life sciences companies are facing massive growth in data volumes compounded with ever-increasing pressure on IT budgets and resources. GxP compliance adds an extra dimension of complexity and when document and data retention periods can run to multiple decades, simply keeping data on enterprise storage is untenable.

References

1. McKinsey and Company, Big data: The next frontier for innovation, competition, and productivity, 2011
2. Reine D and Kahn M, In search of the long-term archiving solution – tape delivers significant TCO advantage over disk, Clipper Group, 2010

About the author

Jim Cook is CEO and Co-Founder of Arkivum, a company operating in the rapidly growing cloud archive sector. During a career that has spanned more than three decades, he has been instrumental in helping organisations to achieve their IT and business ambitions. He is a Chartered Engineer and gained a BSc in Electrical and Electronic Engineering from the University of Bath, Avon. He is also a Fellow of the Royal Society for the Encouragement of Arts, Manufactures and Commerce and a Member of the Institution of Engineering and Technology. **Email: jim.cook@arkivum.com**