Backup Less Archive More

White paper on why organisations should consider data archiving

Author – Ray Quattromini – Managing Director

28th October 2010

Fortuna Power Systems Ltd
1 Woodmere Croft
Basingstoke
Hampshire
RG22 5HB

Table of Contents

BACKUP LESS, ARCHIVE MORE

Time and time again I hear the word Backup and the problems associated with it. Well this document will hopefully convince people and companies that if they archive more they can backup less!

Why do we backup?

1. In the event of a disaster we can restore to a point in time prior to the event.
2. To restore data in the event of malicious or accidental data erasure.
3. To protect against virus's and hackers.
4. Restore data from one system to another.
5. Just in case anything goes wrong with a system upgrade.

What are the issues with backup?

1. Time to backup everything.
2. Ever increasing data volumes.
3. Backup failure.
4. Financial impact of continuously renewing backup software and systems.
5. Managing backup software and agents.
6. To de-dupe or not to de-dupe? This is an interesting one.
7. Hard disk or tape?
8. Encrypt or not?
9. Onsite, offsite.
10. Media collection costs.
11. Data Management.
12. Regulation and compliance.
13. Data retention periods.
14. Wear and tear.

We distribute a large number of data archiving technologies that provide huge long term cost savings by being Energy Efficient. We hope after reviewing the contents of this paper you will have a better understanding of why you should Backup Less and Archive More.

Backup Issues

Time

The world today is generating information faster than ever before. The issue is that we are not going to get more hours in a day; the backup window is ever decreasing and companies are fast running out of idea's to backup their live data within the timeframe they have.

I am now aware of systems that can restore data the second it was lost. This system is a continuous backup solution running 7x24 and we can provide this technology upon request.

Increasing Data Volumes

Data is typically growing faster than our current backup systems can cope. Today some companies are selectively backing up data on a daily basis, in order to at least have 1 copy every 48 hours; this issue is not going to get easier and should be resolved.

Backup Failures

I have lost count as to how many times people have mentioned “the backup failed again last night”. They can fail for so many reasons and it’s not all down to human error. With the ever increasing data sprawl, trying to keep a handle on this problem is always going to end in tears. In order to manage our backups we firstly need to manage our data more efficiently by prioritising the data in to more manageable chunks. Rather than create a single backup job to backup everything, try to prioritise and create clearly defined backup jobs. At least this way it should help to find out the individual causes of backup failure, rather than reading a huge report that lists everything, the reports are easier to read and analyse. Ensure you then resolve the route cause and run the backup again.

Things to consider about backup:

1. Do you need to backup desktops or personal data?
2. Do you have the correct agents to backup your operating systems or platforms?
3. Should you be looking at an alternative supplier or vendor?
4. When did you last replace your backup tapes?
5. When did you last clean your tape drive?
6. Do you need to backup everything?
7. Have you removed or added additional servers or applications?
8. Do you have networking issues?
9. Do you have the correct access permissions to backup the servers?
10. Who has access to the backups and what level can they see?
11. When was the last time you performed a restore?

If you backup everything fantastic, but what if you need to restore everything if disaster strikes? Time is precious and backup is just a matter of housekeeping and continual management, by reducing data volumes will dramatically increase your chances of achieving a 100% success rate.

Financial Impact

Every so often a new faster technology comes along to hopefully alleviate some of the issues we are experiencing with backup.

These technologies only provide a short term fix to a long term problem.

1. LTO-5 – Much greater tape capacity, not much faster than (140MB/s vs 120MB/s) LTO-4
2. VTL – A disk array with software that mimics a tape library
3. De-duplication – Looks for common blocks and creates a hash table
4. Compression – Gives us more for less, providing the data is compressible
5. RAID Array – Cheaper than a VTL and uses the backup software to create the virtual tape library
6. A new version of backup software comes out. Do we upgrade or forklift upgrade to something else.
7. A new SAN and virtualized environment needs additional agents.

These are direct costs associated with backup and always increase yearly due to the following:

1. Purchase new or upgrade existing technology
2. Software and hardware maintenance renewals
3. Purchase additional backup agents and software

4. Purchase additional tapes
5. Increased offsite storage costs

The indirect costs associated with backup are as follows:

1. Purchase / upgrade tier 1 storage to cope with increased data volumes
2. Slower performance due to increased disk fragmentation
3. Greater deployment of servers and management
4. Increased running costs
5. Increased staffing costs

Backup Agents

A backup agent connects your servers and or desktops to allow backups to be performed. These agents might also be application based such as Exchange, SQL, Oracle, VMware etc or Open File agents that can backup databases without the need to shutdown the applications. To ensure a 100% successful backup night after night is nigh impossible. There will always be something that goes wrong, permissions, missing file, volumes offline etc. As the number of applications grows, so does the complexity.

To De-dupe or not De-dupe that is the question

More hype has been placed on this backup saviour called Data De-duplication than any other backup technology. Why has this technology so over hyped and mis-understood? Firstly let's explain how data de-duplication works.

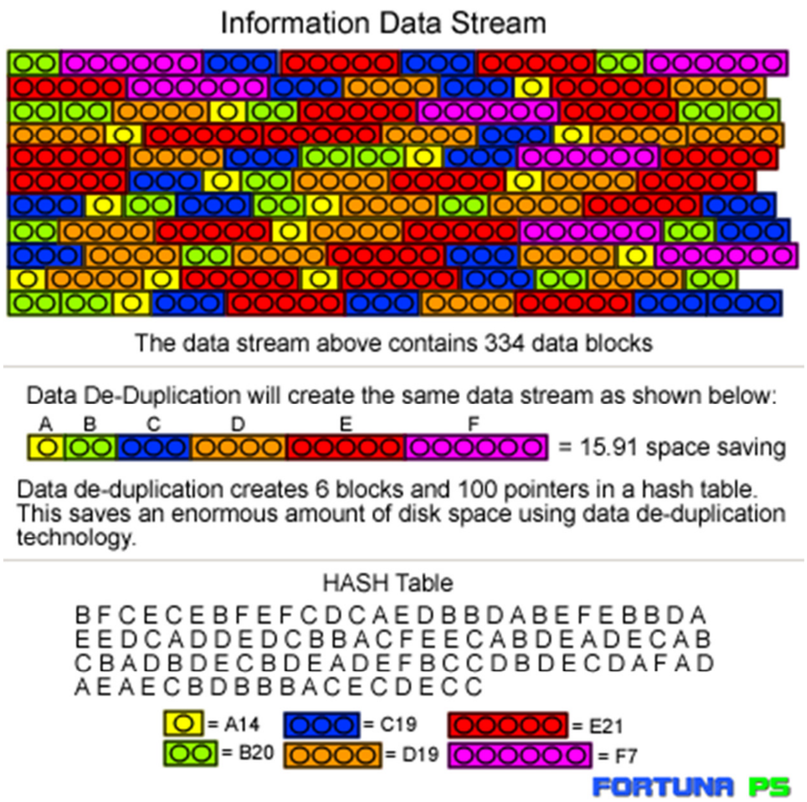
Data De-duplication is a technology that initially creates a full backup of your data, but not as files only as hash tables. These data de-duplication solutions can be pre, post or in-band let me explain.

1. **Pre** – The data is analysed and hashed before landing on the disk. Usually requires a large amount of cache and fast processors.
2. **Post** – A backup is taken and the file is stored on disk, once the backup is finished the hashing takes place. Requires more disk than the Pre process, but typically faster.
3. **In-band** – As the files are backed up they are on the fly hashed and then placed on disk. Requires fast processors and not as fast as the previous two.

In addition to the above they can also be:

1. **File level data de-duplication** typically provides compression rates up to 5:1 and removes repetive files.
2. **Fixed block data de-duplication** provides compression rate up to 10:1. Views the blocks of a given fixed size and removes the repetive blocks.
3. **Variable block data de-duplication** by far the best and most flexible technology providing compression ratios of 50:1 or more, although in real world applications you will get between 20-30:1

Now whatever data de-duplication process we deploy, they all effectively do the same job. We now have a full backup and know it consists of repetitive blocks. The next time we perform the backup, we only need to backup the changed blocks and append the hash tables accordingly. The diagram below will hopefully explain how this works in more detail.


Information Data Stream

The data stream above contains 334 data blocks

Data De-Duplication will create the same data stream as shown below:

Data de-duplication creates 6 blocks and 100 pointers in a hash table. This saves an enormous amount of disk space using data de-duplication technology.

HASH Table

B F C E C E B F E F C D C A E D B B D A B E F E B B D A
E E D C A D D E D C B B A C F E E C A B D E A D E C A B
C B A D B D E C B D E A D E F B C C D B D E C D A F A D
A E A E C B D B B B A C E C D E C C

Data de-duplication is a great technology providing your data isn't video based, audio, images or other media that is already compressed. Yes it will reduce your backup window but not to the extent you were lead to believe.

File or Hash

Now remember earlier we mentioned that data de-duplication systems do not store files only hashes. This is a double edged sword great for reducing backup windows as we only backup the changed blocks, but restores take far longer (see below).

Restore

We need urgently to restore downed systems, with a non de-duped solution just hit restore and the files come back instantly. Data De-duplication doesn't work like this, here are the steps.

Providing your appliances are working and you can get access to them. Firstly re-constitute the de-duped blocks to files and then start the restore. This is fine providing you want to recover a small amount of data but typically it will take 2X longer to perform a restore using data de-duplication than a conventional backup. It also requires extra disk capacity to act as a staging area for the de-duped data to reside, prior to being fully restored.

Let's do some simple arithmetic

Currently the fastest data de-duplication appliances backup at 6.4TB's hour (why not quote restore speeds as well?).

This equates to 1.77GB/s, fantastic if we can actually sustain these backup rates. Equivalent to twelve LTO-5 tape drives!

We need to restore 50TB's of data to 20 servers.

Tape method – The restore speed to each server is 50MB/s and our tape restore speed is 100MB/s. This means we can restore 3.6TB's per hour (50x 3600 x 20). Therefore to restore 50TB's would take just under 14 hours.

De-duplication method – Using the same server speed as above our hashed duplication data firstly needs to be re-constituted to files. So let's assume 1GB/s, once the file exists it needs to be copied to server. 1GB/s x 3600 = 3.6TB's hour to our staged storage first and then copied to our servers 50x 3600 x 20. In effect this would take twice as long as normal to perform a restore. Now in a disaster scenario which option would you choose?

Costs

A 4TB data de-duplication appliance costs $20k and would hold 100TB's @ 25:1 compression. A similar 84TB RAID/MAID system would cost $38.2k. Visit www.big-data-storage.co.uk for more information. For you $20k your backup speed would be up to 400GB/hr, whereas the RAID/MAID would backup 2.5TB's per hour or 6x faster! Again your restore speeds are also going to be affected by the disk performance and clearly de-duplication saves money at the expense of backup and restores windows.

One chunk or Two?

In 99% of instances you will need to purchase two! Why is this, firstly the data de-duplication solutions are disk based and require periodic maintenance or will fail (it's just a matter of when). Some allow backup to tape from the appliance, but let's not confuse the issue. The real reason you need to purchase two, is if something gets corrupted on the first appliance, you have a backup of a backup! Do you do this for tape; of course not, you take the backup tapes offsite or create copies.

Prolong the inevitable

Data De-duplication is not a backup utopia! It should not be viewed as a permanent answer to your backup; it will only delay the inevitable issue of backing up your data by introducing just another complex and expensive backup layer to administer.

Hard disk or tape?

This depends on your corporate compliance and governance legislation.

Advantages of tape:

- Tapes are portable
- Tapes are low cost
- Tapes are energy efficient
- Tapes can backup huge amounts of information

Disadvantages of tape:

- Not particularly suited to individual file restores
- Require the correct tape drive and backup software to read them
- They degrade over time
- Not suitable for archiving

Data Encryption

Encrypting data will introduce a backup and restore overhead. There are many stories of tapes being lost or stolen, but there are also stories about organisations that did encrypt their data and can't recover it.

This could be lost encryption keys, faulty systems etc. Police Forces do not like encryption as it adds a layer of complexity that causes more headaches than it solves. If you are sure you have procedures in place to protect your encryption systems, then this is a good idea. If however you change systems, software or lose the encryption keys what will you do with all your historical backup data?

WAN Bandwidth

When synchronising data across wide area links it is an important consideration to find out the available bandwidth for synchronisation purposes. Below is a table of typical transfer speeds. Another thing to consider is when the synchronisation will be taking place, day/night/weekends, quality of the link and the actual available bandwidth that can is available, excluding other normal network traffic. All of this can have a dramatic effect on our ability to store data offsite.

Mbits/s	Mb/s	MB/min	GB/hr
2	.125	7.5	.45
10	1.25	75	4.5
20	2.5	150	9.0
30	3.75	225	13.5
50	6.25	375	22.5
100	12.5	750	45.0

Onsite / Offsite

Everyone should create a second copy of their data which should be stored in a "DATA SAFE" (not Fire Safe). This can be synchronising from Site A to B using hard disk or physically backing up two separate streams to two sites using tape. If you can't afford the cost of a DR site make two copies and store one offsite.

Collection & Storage Costs

Whilst many organisations have 3rd parties collecting their media, the cost of this service rises with fuel costs, increased number of tapes etc. The only way to reduce these costs are to upgrade your tape technology to backup more data on less tapes or setup your own DR site and have the data sent across the LAN or WAN providing you have enough bandwidth. This is where data de-duplication comes in (see above).

Data Management

No matter what backup software a company runs, daily management reports are run to ensure the backups ran successfully. The more data, servers and applications we add, the more time we spend caring for our backups. Today companies employ staff to manage their backups full-time why? It's all down to paranoia, we add more of everything and our worries grow that something somewhere is going to fail.

When I setup Fortuna Power Systems 16 years ago staff would spend 1-2 hours a day on backups, so what went wrong, surely today systems and technology are more reliable than they were 10-15 years ago? Yes, systems and software are more reliable and advanced, but the problem is we have far more data and servers performing tasks. Organisations that virtualise their infrastructures now have 10-20 servers running on their much bigger and more powerful server, so if someone pulls the plug on this we lose twenty servers and not just one! So the worry here is how to make the virtualized infrastructure highly available against such an event occurring.

Regulation & Compliance

Today more and more organisations are facing legislation to secure their data for regulatory, compliance or corporate governance. Trying to identify what should be kept or deleted becomes more difficult every day and for how long do we need to keep it? The problem is backup software isn't designed to identify aged, changed or modified data. All it is designed to do is reset the archive bit, for the next backup. The challenge companies are going to face is where and how to store this legal information.

Data Retention

Data retention is defined by corporate or legal requirements. How long we have to archive and retain data is defined by these rules and policies.

How long do we need to keep our backups? Typically an organisation would create a backup schedule as follows:

1. Daily backups – Kept for 2 weeks
2. Weekly backup – Kept for 1 month
3. Monthly backups – Kept for 1 year
4. Yearly backup – Kept indefinitely

Restore

A huge emphasis is placed on backup with little thought for the primary focus of backup and this is "RESTORE". You have vaults full of backup tapes, but it's no good if you changed your backup software 18 months ago! How many companies keep their backup tapes and never bother to see if they can perform a restore on their legacy data, if they implement new hardware/software? The answer is not that many and who is going to admit to this oversight.

A restore of your information should be performed at least quarterly on a variety of different data types backed up from various servers and not the same server every time. At least this way should something serious occur you can be sure at least your restores will work as planned.

Data Archiving

Over the centuries the human race has tried to keep a history of events, discoveries and tragedies. Initially man carved etchings into stone or wood, then man discovered paper for almost 2000 years this was the way knowledge was passed on for future generations. Now in the 21st century a need to keep our digital history is becoming a huge headache for many organisations.

"The latest figures indicate that during the last 3 years we have created more data than in the previous 2000 years!"

In the world we live in today we are constantly generating information in the form of video, sound, e-mails, texts, music, databases, images, office documents, PDF's, and many other varieties of data are created daily. Some of this information is kept for a day whilst much of this information needs to be kept for a month a year or more than 50 years. We keep this information for a number of reasons; it could be legislation, regulation, insurance, assurance, research, intellectual property, personal, historical, medical or many others.

Storage Technologies

I have been involved in the data storage industry for almost 30 years, during that time I have seen numerous technologies succeed and then fail because their development roadmap was to ambitious or someone designs, develops, invents something that appears better, cheaper, faster, easier to use and more reliable than what is currently available.

We are human and because of this we are seduced into purchasing these new revolutionary technologies. Unfortunately we do not always consider the long-term downsides of adopting these new products and it isn't until we come to upgrade, maintain and enhance this product that we realise it is not as revolutionary as first thought and we are locked in to a solution that cannot evolve with our day to day business demands.

Why archive?

As mentioned throughout this white paper the issues of backup are becoming increasingly problematic. A failure to address the fundamental problem that data volumes are overwhelming the primary storage systems will inevitably result in data loss.

80% of stored data is inactive after 60 days

Numerous organisations have a huge on-going investment in purchasing data storage systems and this investment is increasing year on year as the demands to store more information increases. After 3-5 years this equipment is then replaced.

With the implementation of a Data Archive solution we can actually extend the life of this investment by moving the data to a secure active archive, thereby freeing up valuable disk space on high performing storage solutions and slowing down the necessary and ongoing investment of more storage space giving a huge ROI benefit. An additional benefit with a tiered data archive is that you might be able to utilise your existing older storage systems to archive data.

As a consequence the size of online data volumes are spiralling out of control and storage management has become an ever-increasing challenge.

- Server performance and data access are diminishing
- Business legislation and user demands are requiring companies to increase disk space to alleviate the problem
- Data management puts high overheads on networks
- DR policies for recovery take longer as all data needs to be restored rather than the most recent and essential
- The annual cost of managing this data over its lifetime is more than 5x the initial purchase price of the equipment
- Backup windows are diminishing

Buying bigger and faster storage systems is not solving the problem of data identification. How we classify our data is going to become more of an issue than people first thought. Now we know tools exist to move data by date, last modified, file size, extension, user, file type etc, but how many companies are using this technology?

Archive Storage

Once we have an archive, whether this is static, near line or offline you need to consider whether the archive will remain on the same site as the primary data or relocate the archive to another office or even replicate the archive regularly to the DR site. In the event of a disaster your primary concern should be to restore your most recent information first and then restore the archive if required.

Data Cost

All information that we generate has a cost, companies often mention that the latest data residing on tier 1 storage has the most value, I do not agree with this statement for simple reason that important architectural /engineering drawings, films, health or legal documents etc are important long after they were initially created and if anything are more valuable than when first stored.

Archiving will achieve the following:

1. Save tier 1 disk space
2. Reduce management costs
3. Free up valuable system resources
4. Increase the ROI on existing storage
5. Reduce capital expenditure by deferring upfront purchases, energy and prolong equipment life
6. Reduces the number of backup tapes or disk space required
7. Enable you to backup less
8. Perform faster restores
9. Allow you to achieve greater RPO (Recovery Point Objectives)
10. Create a bigger backup window
11. Enforce and distinguish business data from personal data
12. Protect the data that is most important to the business

Current storage technologies

Below is a list of differing storage technologies which we could utilise to provide an archive.
Storage technologies and archive life of media.

Blu-ray	30 - 50 years	No special archival requirements
UDO	30 – 50 years	No special archival requirements
RAID	5 years	Will require checks to ensure no drive or controller failures
Tape	5 – 10 years	Air conditioned and humidity controlled
CD/DVD	10 – 20 years	Kept out or sunlight or direct light sources

Access speeds and energy consumption of the devices.

Blu-ray	Up to 20 secs	Lowest
UDO	Up to 20 secs	Lowest
RAID	Micro seconds	High
Tape	Minutes	Low
CD/DVD	Up to 20 secs	Lowest

Reliance on other technologies to read the data

Blu-ray	None if written in UDF format (Universal Data Format), any device with Blu-ray reader should be capable of reading the information
UDO	Requires UDO reader and correct writing software
RAID	Dependant on drives, controller, operating system
Tape	Requires tape drive and correct backup software, unless using LTFS
CD/DVD	None if written in UDF format, any device with a CD/DVD reader should be capable of reading the information

Archive Media

No matter what type of archiving technology you choose it is important to remember that the media is more important than the device. If you do need to keep media for a considerable time, always buy branded media from a known entity and ask for a certificate of conformance. These are sometimes difficult to obtain but should state that the media is certified for "x" years archive life. It might cost you a few pence more in the beginning, but it could save you thousands over the lifetime of the archive.

To ensure a long-term archive you should ideally make three copies of the media and store it in three separate locations, if possible on three differing types of technology in the hope that in 50 years one of the technologies is still around to read the stored information.

Archival Procedures

When purchasing any media for backup or archive it is important to remember that the technology is only as good as the amount of money you have available to protect the data. So for example a backup tape might have a manufacturer's life of 30 year these claims are made because the tape has been tested in perfect conditions and accelerated tests to achieve the results. After 29 years and 364 days the tape fails you will be sent a new tape from the manufacturer, the manufacturer's warranty only covers the tape and not the contents i.e. the data!

Unless you are prepared to keep your tape in perfectly controlled conditions, you will never get 30 year life out of any tape.

Many years ago I worked for a company that sold tape drives and media to the Middle East and for some unknown reason these drives started to fail dramatically. The request was made to return 50 of drives back to the UK for investigation and another 50 promptly despatched. When we received the drives back and looked in to the problem we saw the issue was down to rust! The tape drives and media were being used in an environment that had high humidity and no air conditioning.

The cost of storing information increases yearly, this could be storage space, media costs, energy consumption, replacement costs etc. Clearly the archive media we choose needs to provide greatest long term costs savings for preserving the information and this is the challenge.

Technology refresh

With any technology you should perform the following:

- Periodically do a sample check on your archived media to ensure that the integrity of the media and data is still intact.
- Perform an analysis of current technologies and decide whether or not to keep or transfer your archive to a newer technology
- Ensure that the equipment is fully maintained and that you continue to purchase branded archive grade media
- Should you decide to change operating systems, always check compatibility with your archived data
- Ensure that the technology you are purchasing has a long track record and is not in the process of selling off products to please investors
- Do not de-dupe or encrypt archived data as ideally if you do need to restore the information in 20 years you want it in native format.
- If you are required to store data for 5+ years put it on optical it's far safer and will save you £'s in the long run.

Archive More

There is no Silver Bullet to the issue of Backup or Archiving. By careful planning a happy medium should be achieved to ensure your most recent data is backed up daily and archived information retained for "X" days, months, years.

An archive needn't be permanent; it could be a SATA disk array that only needs to be backed up once a month. Just deploying this affordable storage would free up your tier 1 disk space, removing none corporate data from the network would free up disk space, reduce network traffic and reduce your backup window. These simple strategies can be easily achieved with a little time, thought and money. We are able to provide a solution that, really will Backup Less and Archive More.

Call us on +44 (0)1256 782030 or email: sales@data-storage.co.uk and we will assist in any way we can.

Websites

Below are some of our websites that provide data storage solutions.

www.data-storage.co.uk
www.big-data-storage.co.uk
www.lto-tape.co.uk
www.blurayjukebox.com
www.raidstoragesystem.com