

WHY ARCHIVE IS ARCHIVE, BACKUP IS BACKUP AND BACKUP AIN'T ARCHIVE

From my favorite Dictionary¹:

Archive

Pronunciation: 'är-"klv

Function: noun

Etymology: French & Latin; French, from Latin archivum, from Greek archeion government house (in plural, official documents), from archE rule, government -- more at ARCH.

Date: 1603

: a place in which public records or historical documents are preserved; also : the material preserved -- often used in plural, ex: archives.

Backup

Pronunciation: -"&p

Function: noun

Usage: often attributive

Date: 1951

1 a : one that serves as a substitute or support <a backup plan> b : musical accompaniment

2 : an accumulation caused by a stoppage in the flow <traffic backup>

3 : the act or an instance of backing up a computer's hard disk

If Merriam-Webster had meant for these two things to be the same, she wouldn't have created two words!

Unfortunately, the definitions, except for backup, don't directly address our industry. If we analyze these definitions, however, we can begin to understand the intent of each. Archive means to save in one place, not two, but backup indicates there are two instances of the object. In our case, archive would mean to remove the object to another place, while backup means to copy an object to another place.

To take this further:

- Archive means to remove from the on-line system those objects no longer in day-to-day use, and place them into long term, retrievable, storage.
- Backup means to make a copy of an object in the event the original becomes lost or damaged.

Finally, backup contemplates two instances:

1. catastrophic loss of data, due to hardware failure, or
2. the inadvertent loss of data, a user accidentally deleting a file.

In both cases, we are usually most interested in restoring the most recent copy of the file, not some version from the far distant past. In fact, in most instances, what we really want is the state of the object right before it was destroyed. Not the way it looked yesterday. So, backup is a snapshot of everything.

Archive contemplates a single scenario: the retrieval of specific information for historical purposes. For instance, the state labor department is interested in seeing payroll records for a specific period of time. In this case, the labor department isn't interested in recent backup data, but rather older data in archive form. In addition, the labor department is probably not interested in the electronic mail or other files surrounding the payroll data. So, archive is snapshot of specific things.

Definition: Backup is short term and archive is long term.

¹ WWWebster Dictionary Copyright© 1996 by Merriam-Webster, Incorporated

Too often, we in the computer industry use the words and the techniques interchangeably. Mostly because the tools available only allowed us to do one thing: backup.

How many times have we heard or been told that we must archive the state of the entire system for legal reasons? How many times have we questioned the requirement? Why is there such a huge difference between the answers to these two questions?

As computer environments continue to grow, our ability to “archive the entire system” is becoming more difficult, if not impossible. Obviously, some rational thinking is required. Unfortunately, this thinking is hard and will require a significant amount of time.

In a recent ADSM course that I taught, I observed a dawning recognition come over the face of one of the participants. “So, you mean archive functions differently than backup? I was having a hard time understanding how I was going to ‘archive’ 600 Gbytes, or all of my data over my 57 hour weekend. I know I can’t move that much data in that little time.”

Bingo!

The discussion then centered around what archive really is. For years, this customer had been doing weekly full and daily incremental backups. In addition, they never reused or discarded a tape. This meant that with a lot of finagling, they could restore any system to any state in the past. They had never actually done this, but they could have.

If they had ever tried to restore the state of a complete system, from the far distant past, would it have worked?

No. Reason #1: The hardware available now would not be like the hardware then.

Never mind restoring the Operating System, let’s just restore the application.

No. Reason #2: The hardware and operating system software available now won’t run the old application.

OK, let’s just restore the data from the application.

No. Reason #3: The application that is available now won’t be able to read the data from long ago. In fact, the entire database has been changed. You remember, when we moved from OpenVMS to UNIX we changed from Ingres to Oracle.

If it isn’t going to work anyway, why keep it around?

Bingo.

So, what does archive really mean? What are the “real” legal requirements? The answers vary, but are similar in the following respects:

- some business data should and must be saved for long periods of time; and,
- most businesses have some legal requirements to maintain certain types of information for regulatory purposes.

Of all of the data in your environment, it is likely that less than 10% falls into this second category. Using our friend’s example from above, only 60 GBytes would require archival on a regular basis. 60 GBytes! That’s a lot of stuff. OK, maybe only 10 GBytes is really required. The problem with this thought process is it begs the question, “If only 10% or less of the storage space is consumed with important business data, what’s the rest filled with?” It’s probably best not to think too much about this.

Now we’ve pared down our data to a manageable amount for archival. How should we archive it? Given Reason 3 from earlier, it would make sense to consider that applications do change over time. Perhaps we should consider some simple format for our archived data. I advocate a simple textual representation of the data. Along with the data one should also save a schematic of the data. In the worst case scenario, it would be possible to import this text data into a spreadsheet and manipulate it from there. Or, a simple import program can be written for the current application to read the archived data. In fact, this functionality can be mandated during the development and maintenance of your critical business applications: “Each application shall have the ability to read a textual representation of the previous, or other, application’s data...”

To design an effective archival system for your computer environment:

1. Pick a tool that offers archival services as a separate function from backup.
2. Determine which data actually requires archival. Use your legal department along with your business users of the data to hone in on this data.
3. Determine the archival retention periods for the various types of data in your environment. Not all data has the same requirements.
4. Determine the format of the archived data. Remember, systems, applications and tools change, so pick a format that will work with many tools now and in the future.
5. Design a solution, using the archival tool, which meets your business needs.
6. Archive your business critical data using the tool.
7. Perform a test restore of information into a second application to verify the efficacy of your solution.

The differences between backup and archive are really quite simple. Often, the stumbling block to understanding the difference is the tools we've chosen to use. Up until recently, it has been possible to use backup software for archival purposes. But now, with the explosion of data in many environments, it is becoming necessary to draw distinctions between the two functions. We can't possibly archive all of the data, and we don't need to. Use your unique business requirements and the proper tools to solve your backup and archive issues.

COPYRIGHT 1999 © STORAGE SOLUTIONS SPECIALISTS, INC. ALL RIGHTS RESERVED
WRITTEN BY KELLY J. LIPP



Storage Solutions Specialists, Inc.
PO Box 51313
Colorado Springs, Co 80949-1313
ssi@storsol.com
1-888-Storsol