Data Protection – History, Evolution, Best Practices

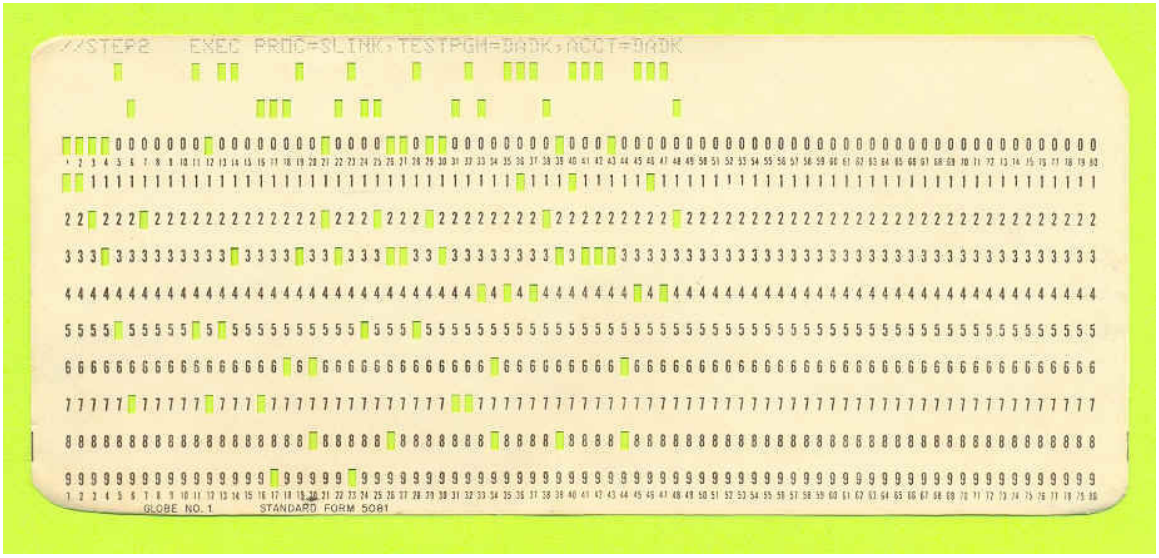
By Global Data Vault

Introduction

As business data processing has evolved, so have the methods and best practices for data protection. This white paper will provide some historical context and, more importantly, address a sea change which now affects best practices for data protection for small businesses.

History

The origins of automated business data processing date back to the U. S. Census of 1890 which employed the Hollerith punched card and unit record equipment also known as tabulating machines. Punched cards were the medium of data storage for many years thereafter. They later became known as IBM cards. The card shown here is actually an instruction telling the computer to execute a program which had just been loaded from similar cards.



If a card containing data was lost or destroyed, it was easy enough to make a new one. Large decks of cards could be copied using duplicating machines which could copy big stacks of cards quickly by punching a second set. This method of data backup was the dawn of data protection.

Open reel magnetic tape was introduced in the 1950's. This photo shows a 10 inch, 9 track tape which was in common use from the 1950's through the last 1970's.

These tapes could store 5MB to 150MB of data and marked an evolutionary step in data storage and data protection. Data was initially stored on tape and later on disk. Data was copied to a "backup tape" for safe storage. Tapes were rotated in a "Son, father, grandfather" retention policy. Tapes were often moved offsite to provide a disaster recovery capability.

If an original was lost, damaged or destroyed, a backup tape could and often did save the day. And it was easy to save the day because backups were easily interchangeable with the originals but more importantly it was the system architecture of the day which enabled easy recovery.

Early computers were very expensive and to recover the manufacturing and development costs, these big "mainframes" were required to be utilized 24 hours per day and to do many different jobs. Each "job" involved an application program and the related data. The application program was loaded into the mainframe via cards or tape and the data was presented in a similar manner. The job was "run" and the results – most often financial accounting information – were rendered.

Everything was separate or discrete. The hardware, the operating system, the application program and the data were all separate components. That enabled easy methods of data protection. Because one only needed a good copy of each element and this was addressed at every level. Most data processing sites had agreements with “backup sites”. If your primary site suffered a disaster, you would pack up your application programs and data – on tape and cards – and head over to your disaster recovery “DR” site. Your DR site had been selected because it had similar or identical hardware and operating systems and hopefully some spare capacity to fit in your work. This was a workable and usually successful strategy.

Technology evolution

In the last 15 years three things have created a new and more challenging data protection environment. Lower cost machines built in distributed networks have changed the way applications access computing power. More sophisticated operating systems have been built to harness and distribute that power. And most importantly, the way in which application programs are installed has dramatically changed the environment.

Key distinctions

The two most important changes impacting data protection are first, today’s method of installing application programs. Instead of “loading” an application every time it is to be used, we now “install” an application and the related data and leave them in place within the computing infrastructure. Second, data is now accessed in real time and no longer processed and then left at rest. There are several important new layers of complexity created.

Now we must identify distinct “restore points”. These are points in time at which data can be restored to a state of “referential integrity”. Referential integrity is said to exist if all elements of a given data set are properly synchronized. For example when a customer makes a purchase on account, and an invoice is created, we must also update revenue, inventory, and the customers balance. Generally these happen quickly in sequence but care must be taken to assure that any backup which was taken included all of these transactions – or none of them.

And, now that data is also installed and permanently available to the application systems, any restore process must take into consideration the additional complexity associated with installing the data. Today backup methods often change the data from the format in use in production to a format more suitable for remote offsite storage.

Current methods

Current systems attempt to address these two additional complexities, referential integrity and data format changes. Many complex systems have been developed to solve these difficulty problems. They are, by their nature, complex, and complexity leads to difficulty.

The fatal flaw in most of these methods is that they attempt to separate applications, also known as the application lawyer, from the related data. Now, with applications and data so tightly interwoven, we need a better solution.

Best practices

Today's best practices require take a different and more holistic view of all of the elements needed to perform computing tasks. Now we must recognize that the ability to recover to a given prior state is dependent on addressing these new realities.

Today's best backup and restore processes no longer depend on distinctions about programs versus data. Today's best processes focus on creating a restorable image of the entire environment sometimes referred to as a workload.

Virtualizing servers, a method and technology pioneered by VMWare, allows us to view the operating system, application programs and all of the data on a server as a discrete unit, or a workload. These units or workloads can be completely encapsulated and captured in one data set called an image. That image is a complete copy of everything in the workload. Today's best practices for data protection work at this level, and they also address referential integrity issues.

It is no longer sufficient to simply make an additional copy of your data, keep multiple versions or generations and store it in a secure offsite backup location.

Technology now exists to make referentially integrated and complete copies of a given workload. The resulting data sets can then be restored onto the original or even different hardware. They can be moved to secure offsite locations and even mounted into virtual environment in remote data center locations, providing a true disaster recovery capability.

Future

Unfortunately today most small and medium sized businesses make simple, old fashion data backups. The capabilities to make workload backups are just now coming into reach as economically viable for small and medium sized businesses. There is a significant opportunity and an important challenge to provide this new higher level of data protection for these businesses.

Conclusion

Data protection requirements and technologies have evolved for over 100 years. In the past 10 years a fundamental shift has occurred in both the required level of solution and in the available solutions to meet that challenge. Full workload copies which are aware of referential integrity and are stored at secure remote offsite data storage locations are now affordable for small and medium sized businesses.

For more information about these advanced data protection methods, visit www.globaldatavault.com.