

Backup

In information technology, a **backup** or the process of **backing up** refer to making copies of data so that these additional copies may be used to *restore* the original after a data loss event. These additional copies are typically called "backups." The verb is *back up* in two words, whereas the noun is *backup* (often used like an adjective in compound nouns).^[1]

Backups are useful primarily for two purposes. The first is to restore a state following a disaster (called disaster recovery). The second is to restore small numbers of files after they have been accidentally deleted or corrupted. Data loss is also very common. 66% of internet users have suffered from serious data loss.^[2]

Since a backup system contains at least one copy of all data worth saving, the data storage requirements are considerable. Organizing this storage space and managing the backup process is a complicated undertaking. A data repository model can be used to provide structure to the storage. In the modern era of computing there are many different types of data storage devices that are useful for making backups. There are also many different ways in which these devices can be arranged to provide geographic redundancy, data security, and portability.

Contents

- 1 Storage, the base of a backup system
 - 1.1 Data repository models
 - 1.2 Storage media
 - 1.3 Managing the data repository
- 2 Selection, extraction and manipulation of data
 - 2.1 Selection and extraction of file data
 - 2.2 Selection and extraction of live data
 - 2.3 Selection and extraction of metadata
 - 2.4 Manipulation of data and dataset optimisation
- 3 Managing the backup process
 - 3.1 Objectives
 - 3.2 Limitations
 - 3.3 Implementation
 - 3.4 Measuring the process
- 4 Lore
 - 4.1 Confusion
 - 4.2 Advice
 - 4.3 Events

Storage, the base of a backup system

Data repository models

Any backup strategy starts with a concept of a data repository. The backup data needs to be stored somehow and probably should be organized to a degree. It can be as simple as a sheet of paper with a list of all backup tapes and the dates they were written or a more sophisticated setup with a computerized index, catalog, or relational database. Different repository models have different advantages. This is closely related to choosing a backup rotation scheme.

Unstructured

An unstructured repository may simply be a stack of floppy disks or CD-R/DVD-R media with minimal information about what was backed up and when. This is the easiest to implement, but probably the least likely to achieve a high level of recoverability.

Full + Incrementals

A Full + Incremental repository aims to make storing several copies of the source data more feasible. At first, a *full* backup (of all files) is taken. After that, any number of *incremental* backups can be taken. There are many different types of incremental backups, but they all attempt to only backup a small amount of data relative to the full backup. Restoring a whole system to a certain point in time would require locating the full backup taken previous to that time and the incremental backups that cover the period of time between the full backup and the particular point in time to which the system is supposed to be restored.^[3] The scope of an incremental backup is typically defined as a range of time relative to other full or incremental backups. Different implementations of backup systems frequently use specialized or conflicting definitions of these terms.

Differential backup

A differential backup copies files that have been created or changed since the last normal or incremental backup. It does not mark files as having been backed up (in other words, the archive attribute is not cleared). If you are performing a combination of normal and differential backups, restoring files and folders requires that you have the last normal as well as the last differential backup.

Continuous data protection

This model takes it a step further and instead of scheduling periodic backups, the system immediately logs every change on the host system. This is generally done by saving byte or block-level differences rather than file-level

differences.^[4] It differs from simple disk mirroring in that it enables a roll-back of the log and thus restoration of old image of data.

Storage media

Regardless of the repository model that is used, the data has to be stored on some data storage medium somewhere.

Magnetic tape

Magnetic tape has long been the most commonly used medium for bulk data storage, backup, archiving, and interchange. Tape has typically had an order of magnitude better capacity/price ratio when compared to hard disk, but recently the ratios for tape and hard disk have become a lot closer. Tape is a sequential access medium, so even though access times may be poor, the rate of continuously writing or reading data can actually be very fast. Some new tape drives are even faster than modern hard disks. A principal **advantage of tape** is that it has been used for this purpose for decades (much longer than any alternative) and its characteristics are well understood.

Hard disk

The capacity/price ratio of hard disk has been rapidly improving for many years. This is making it more competitive with magnetic tape as a bulk storage medium. **The main advantages of hard disk storage** are low access times, availability, capacity and ease of use. **The main disadvantages of hard disk backups** are that they are easily damaged, especially while being transported (e.g., for off-site backups), and that their stability over periods of years is a relative unknown.

Optical disc

A recordable CD can be used as a backup device. **One advantage of CDs** is that they can in theory be restored on any machine with a CD-ROM drive. (In practice, writable CD-ROMs are not always universally readable.) In addition, recordable CD's are relatively cheap. Another common format is recordable DVD. Many optical disk formats are WORM type, which makes them useful for archival purposes since the data can't be changed. Other rewritable formats can also be utilized such as CD-RW or DVD-RAM. The newer HD-DVDs and Blu-ray Discs dramatically increase the amount of data possible on a single optical storage disk,

Floppy disk

During the 1980s and early 1990s, many personal/home computer users associated backup mostly with copying floppy disks. The low data capacity of a floppy disk makes it an unpopular and obsolete choice today.^[7]

Solid state storage

Also known as flash memory, thumb drives, USB flash drives, CompactFlash, SmartMedia, Memory Stick, Secure Digital cards, etc.,

these devices are relatively costly for their low capacity, but offer excellent portability and ease-of-use.

Remote backup service

As broadband internet access becomes more widespread, remote backup services are gaining in popularity. Backing up via the internet to a remote location can protect against some worst-case scenarios such as fires, floods, or earthquakes which would destroy any backups in the immediate vicinity along with everything else. There are, however, a number of **drawbacks** to remote backup services. **First**, internet connections (particularly domestic broadband connections) are generally substantially slower than the speed of local data storage devices, which can be a problem for people who generate or modify large amounts of data. **Secondly**, users need to trust a third party service provider with both privacy and integrity of backed up data. The risk associated with putting control of personal or sensitive data in the hands of a third party can be managed by encrypting sensitive data so that its contents cannot be viewed without access to the secret key. Ultimately the backup service must itself be using one of the above methods, **so this could be seen as a more complex way of doing traditional backups.**

Managing the data repository

Regardless of the data repository model or data storage media used for backups, a balance needs to be struck between accessibility, security and cost. These media management methods are not mutually exclusive and are frequently combined to meet the needs of the situation. Using on-line disks for staging data before it is sent to a near-line tape library is a common example.

On-line

On-line backup storage is typically the most accessible type of data storage, which can begin restore in milliseconds time. A good. This type of storage is very convenient and speedy, but is relatively expensive. On-line storage is quite vulnerable to being deleted or overwritten, either by accident, by intentional malevolent action, or in the wake of a data-deleting virus payload.

Near-line

Near-line storage is typically less accessible and less expensive than on-line storage, but still useful for backup data storage.. A mechanical device is usually involved in moving media units from storage into a drive where the data can be read or written. Generally it has safety properties similar to on-line storage.

Off-line

Off-line storage requires some direct human action in order to make access to the storage media physically possible. This action is typically inserting a tape into a tape drive or plugging in a cable that allows a device to be accessed.

Because the data is not accessible via any computer except during limited periods in which it is written or read back, it is largely immune to a whole class of on-line backup failure modes. Access time will vary depending on whether the media is on-site or off-site.

Off-site data protection

To protect against a disaster or other site-specific problem, many people choose to send backup media to an off-site vault. The vault can be as simple as a system administrator's home office or as sophisticated as a disaster hardened, temperature controlled, high security bunker that has facilities for backup media storage. Importantly a data replica *can* be off-site but also *on-line* (e.g., an off-site RAID mirror). Such a replica has fairly limited value as a backup, and should not be confused with an off-line backup.

Backup site or disaster recovery center (DR center)

In the event of a disaster, the data on backup media will not be sufficient to recover. Computer systems onto which the data can be restored and properly configured networks are necessary too. Some organizations have their own data recovery centers that are equipped for this scenario. Other organizations contract this out to a third-party recovery center. Because a DR site is itself a huge investment, backup is very rarely considered the preferred method of moving data to DR site. More typical way would be remote disk mirroring, which keeps the DR data as up-to-date as possible.

Selection, extraction and manipulation of data

Selection and extraction of file data

Deciding what to back up at any given time is a harder process than it seems. By backing up too much redundant data, the data repository will fill up too quickly. Backing up an insufficient amount of data can eventually lead to the loss of critical information.

Copying files

Making copies of files is the simplest and most common way to perform a backup. A means to perform this basic function is included in all backup software and all operating systems.

Partial file copying

Instead of copying whole files, one can limit the backup to only the blocks or bytes within a file that have changed in a given period of time. This technique can use substantially less storage space on the backup medium, but requires a high level of sophistication to reconstruct files in a restore situation. Some implementations require integration with the source filesystem.

Identification of changes

Some filesystems have an archive bit for each file that says it was recently changed. Some backup software looks at the date of the file and compares it with the last backup, to determine whether the file was changed.

Managing the backup process

It is important to understand that backup is a process. As long as new data is being created and changes are being made, backups will need to be updated. Individuals and organizations with anything from one computer to thousands (or even millions) of computer systems all have requirements for protecting data. While the scale is different, the objectives and limitations are essentially the same. Likewise, those who perform backups need to know to what extent they were successful, regardless of scale.

Objectives

Recovery point objective (RPO)

The point in time that the restarted infrastructure will reflect. Essentially, this is the roll-back that will be experienced as a result of the recovery. The most desirable RPO would be the point just prior to the data loss event. Making a more recent recovery point achievable requires increasing the frequency of synchronization between the source data and the backup repository.

Recovery time objective (RTO)

The amount of time elapsed between disaster and restoration of business functions.

Data security

In addition to preserving access to data for its owners, data must be restricted from unauthorized access. Backups must be performed in a manner that does not compromise the original owner's undertaking. This can be achieved with data encryption and proper media handling policies.

Limitations

An effective backup scheme will take into consideration the limitations of the situation.

Backup window

The period of time when backups are permitted to run on a system is called the backup window. This is typically the time when the system sees the least usage and the backup process will have the least amount of interference with normal operations. The backup window is usually planned with users' convenience in mind. If a backup extends past the defined backup window, a

decision is made whether it is more beneficial to abort the backup or to lengthen the backup window.

Performance impact

All backup schemes have some performance impact on the system being backed up. For example, for the period of time that a computer system is being backed up, the hard drive is busy reading files for the purposes of the backup, and its full bandwidth is no longer available for other tasks. Such impacts should be analyzed.

Costs of hardware, software, labor

All types of storage media have a finite capacity with a real cost. Matching the correct amount of storage capacity (over time) with the backup needs is an important part of the design of a backup scheme. Any backup scheme has some labor requirement, but complicated schemes have considerably higher labor requirements. The cost of commercial backup software can also be considerable.

Network Bandwidth

Distributed backup systems can be affected by limited network bandwidth.

Implementation

Meeting the defined objectives in the face of the above limitations can be a difficult task. The tools and concepts below can make that task more achievable.

Scheduling

Using a job scheduler can greatly improve the reliability and consistency of backups by removing part of the human element. Many backup software packages include this functionality.

Authentication

Over the course of regular operations, the user accounts and/or system agents that perform the backups need to be authenticated at some level. The power to copy all data off of or onto a system requires unrestricted access. Using an authentication mechanism is a good way to prevent the backup scheme from being used for unauthorized activity.

Chain of trust

Removable storage media are physical items and must only be handled by trusted individuals. Establishing a chain of trusted individuals (and vendors) is critical to defining the security of the data

Lore

Confusion

Due to a considerable overlap in technology, backups and backup systems are frequently confused with archives and fault-tolerant systems. Backups differ from archives in the sense that archives are the *primary copy* of data, usually put away for future use, while backups are a *secondary copy* of data, kept on hand to replace the original item. Backup systems differ from fault-tolerant systems in the sense that backup systems assume that a fault *will* cause a data loss event and fault-tolerant systems assume a fault *will not*.

Advice

- The more important the data is that is stored on the computer, the greater is the need for backing up this data.
- A backup is only as useful as its associated restore strategy.
- Storing the copy near the original is unwise, since many disasters such as fire, flood, and electrical surges are likely to cause damage to the backup at the same time.
- Automated backup and scheduling should be considered, as manual backups can be affected by human error.
- Backups will fail for a wide variety of reasons. A verification or monitoring strategy is an important part of a successful backup plan.
- It is good to store backed up archives in open/standard formats. This helps with recovery in the future when the software used to make the backup is obsolete. It also allows different software to be used.

Backup software

Backup software is a computer program used to perform a complete backup of a file, data, database, system or server. The backup software enables a user to make an exact duplicate of everything contained on the original source. This software must also be used to perform a recovery of the data or system in the event of a disaster.

Contents

- 1 Key features
 - 1.1 Volumes
 - 1.2 Data compression
 - 1.3 Remote backup
 - 1.4 Access to open files
 - 1.5 Incremental backups

- [1.6 Schedules](#)
- [1.7 Encryption](#)
- [1.8 Transaction mechanism](#)

Key features

There are several features of backup software that make it more effective in backing up data.

Volumes

Voluming allows the ability to compress and split backup data into separate parts for storage on smaller, removable media such as CDs. It was often used because CDs were easy to transport off-site and inexpensive compared to hard drives or servers.

However, the recent increase in hard drive capacity and decrease in drive cost has made voluming a far less popular solution. The introduction of small, portable, durable USB drives, and the increase in broadband capacity has provided easier and more secure methods of transporting backup data off-site.

Data compression

Since hard drive space has cost, compressing the data will reduce the size allowing for less drive space to be used to save money.

Remote backup

Several factors have contributed to a surge in the use of remote or offsite backup of data to geographically distant sites.

1. The rapid growth of data and its importance to business.
2. The rapid adoption of high-speed broadband internet.
3. The falling price of disk drive technology.
4. The rise of risks such as hackers, hurricanes, viruses, hardware failure.

These structural changes present opportunities for young startups, which are serving this growing market with next-generation backup technologies that automatically back up data to offsite data centers (sometimes called vaults) via the Internet. Many banks, stock exchanges, and other large institutions often do this to ensure data integrity.

Access to open files

Many backup solutions offer a plug-in for access to exclusive, in use, and locked files.

Incremental backups

Backup solutions generally support incremental backups in addition to full backups, so only material that is newer or changed compared to the backed up data is actually backed up, in order to dramatically increase the speed of the backup process.

Schedules

Backup schedules are usually supported to reduce maintenance of the backup tool and increase the reliability of the backups.

Encryption

To prevent data theft, some backup software offers cryptography features to protect the backup.

Transaction mechanism

To prevent loss of previously backed up data during a backup, some backup software (e.g. Areca Backup) offer Transaction mechanism (with commit / rollback management) for all critical processes (such as backups or merges) to guarantee the backups' integrity.