Temasek Junior College

2024 JC2 H2 Computing

Data Management & Security 1 – Backup and Archival

Section	3	Data & Information
Unit	3.3	Databases and Data Management
Objectives	3.3.11	Explain the difference between backup and archive

1 The Need for Data Security

Data is one of the most valuable assets that an organisation has. Its security is thus critical. There are many different reasons behind data loss, some of which are given below:

Hardware failure

It is common misconception that data can remain safe on secondary storage devices. However, the ugly truth is that all secondary storage devices are prone to failure.

In the normal course of use, these devices may experience wear and tear over time, reducing their reliability and increasing the risk of data loss.

The table below also gives some common causes of failure to common secondary storage devices:

Device	**Possible Cause of Failure**
Magnetic hard disk drives (HDD)	Head crashes Mechanical failure Circuit failure Heat or water damage Power issues
Solid state drives (SSD)	Chip failure Power failure Water damage Bad blocks
Optical discs	Surface damage Exposure to light Extreme temperatures Humidity Chemical degradation
Digital tapes	Jamming Coming off the spool Excessive humidity High temperatures

Human actions

Humans may accidentally delete or overwrite data. They may also format or partition the wrong drive or handle storage media improperly, leading to data loss.

Data may also be deleted or destroyed will ill-intent.

Software corruption

Issues with software, operating system failures, or bugs can cause data corruption or loss. This can occur during updates, installations, or due to malware or viruses.

Power failures
Sudden power outages or voltage fluctuations can interrupt data transfers or damage storage devices, leading to data loss.

Natural disasters
Fires, floods, earthquakes, or other natural disasters can physically damage storage devices and result in data loss.

Theft or loss
If a device containing important data is stolen or lost, the data may be permanently inaccessible.

Malware and ransomware
Malicious software can encrypt, delete, or compromise data, making it inaccessible unless a ransom is paid.

Viruses and cyber-attacks
Viruses and Cyber Attacks: Viruses, worms, Trojans, or other cyber-attacks can corrupt or delete data, compromise system integrity, or cause data loss.

Software or hardware incompatibility
Using incompatible software or hardware components can lead to data corruption or loss.

To mitigate the risk of data loss, it is important to regularly backup important data, implement robust security measures, and use reliable hardware and software.

2 Backup

Backing up data means making a copy of the data and storing it on a different storage device.

2.1 Frequency of Backup

A backup allows data to be restored to the state it was at the time the backup was made.

Hence the timing of the backup is a critical factor. The key question is "How much data can you afford to lose?".

A small office with no online systems may choose to backup its data **overnight**. This would mean that any files created or changed during the day would not be backed up. If there was a failure, they would lose a day's worth of work, and this may be considered an acceptable risk by the business.

A business that sold event tickets online 24-hours a day could not afford to lose any data. Any lost bookings would result in angry customers, and thousands of lost bookings could result in business failure. This business needs a **real-time backup** system that makes a copy of every transaction as soon as it is made.

2.2 Location of Backup

It would be illogical to store a backup on the same storage device as the original data. If the device failed, both the original and backup copies would be lost.

In addition, the backups should also be stored in a separate location in case the original location was affected by fire, flood, earthquake, or other disasters. The original location may also have its integrity or security compromised, putting the assets there at risk.

Increasingly, cloud storage is used to back up data. The responsibility for keeping this data safe is down to the cloud storage provider.

2.3 Lifespan of Backup

Most businesses run a **backup cycle** that keeps multiple versions of backup files.

For example, there may be a rolling seven-day cycle where each day's backup overwrites the previous week's version. This allows data to be recovered to its original state on any of the previous seven days and is a useful approach if a problem isn't immediately detected.

Some schools back up students' work at the end of each year so that when the next academic year starts and there is lost data, their work can still be recovered.

These cycles are sometimes referred to as generations, because a common approach was to keep three versions, often known as "grandfather", "father" and "son". (This naming approach may sound quite old-fashioned today.)

2.4 What to Backup?

It is tempting to back up everything 'just in case', but this is rarely needed.

Some files do not change and some can be recovered by other means. For example, software files can be downloaded again or reinstalled from disk. Some files may not change often, so they only need to be backed up as and when they have changed.

2.5 Incremental Backups

An **incremental backup** targets only files that have been changed.

Files that have not been updated are not backed up. This can save a lot of time, but it is more complex to recover as a single backed up version will not contain all of the files.

2.6 Backup Policies

All responsible organisations should have a **backup policy** that lays out what is backed up and how frequently it is backed up. Organisations should also have tried and tested recovery procedures, so that systems can be up and running again in the shortest possible time.

For example, a software company may adopt the following backup policy

- The main database is backed up every night by having it 'dumped' to a file which is encrypted on the live server and then copied to a company-owned offsite data centre where it is kept for up to a month.

- After each backup, the technical staff carry out a basic check that the database is larger than the previous version.

- On a regular basis, the technical staff will decrypt and load the backup of the database to a test server to check that the backup is working.

- All source code and content is stored on a server hosted by a third party vendor and all other important documents are stored on a cloud drive provided by the same vendor. Here backup and version control is managed by the vendor.

3 Archive

Data archival refers to the process of identifying, organizing, and storing data in a secure and accessible manner. The purpose is to **preserve** valuable information for legal, regulatory, historical, or business purposes while ensuring efficient use of storage resources.

An archive copy does not need to be available immediately online but should be accessible when needed.

3.1 Importance of Data Archival

Regulatory Compliance
Archiving data helps organizations meet legal and regulatory requirements, such as data retention periods mandated by industry-specific regulations.

Litigation and e-Discovery
Archived data can be crucial in legal proceedings, enabling organizations to retrieve and present evidence when required.

Historical Analysis
Archiving data allows for retrospective analysis and trend identification, aiding in strategic decision-making and historical research.

Disaster Recovery
While the archived copy of the data may not be the latest backup copy of the data, it can still serve as a backup in the event of data loss or system failures, ensuring business continuity.

3.2 Key Considerations for Data Archival

When archiving data, the following criteria needs to be considered:

Data Classification
Data should be prioritised based on its value, sensitivity, and legal/regulatory requirements to determine the appropriate archival strategy.

Retention Policies
The period of retention for the archived data must be based upon legal, industry, or organizational requirements, taking into consideration factors like data sensitivity and business needs.

Storage Infrastructure
To ensure cost-efficiency, it is important to adopt a scalable and reliable storage system capable of accommodating large volumes of data over extended periods.

Metadata and Indexing
Descriptive metadata and indexing mechanisms can be implemented to facilitate efficient search and retrieval of archived data.

Data Security
Archived data should be protected against unauthorized access, tampering, and data breaches by implementing appropriate security controls.

Data Integrity and Preservation
Data integrity checks, periodic validation, and migration strategies should be present to preserve the integrity and usability of archived data over time.

3.3 Data Archival Mediums

Tapes
Tape is a traditional medium for data archival due to their durability, cost-effectiveness, and offline nature. Tape archival involves tape storage libraries for long-term data retention

Disks
Disk-based systems allows for faster access to archived data and is suitable for frequently accessed or dynamic data.

Cloud
Storing data in cloud-based platforms or services brings about the advantage of scalability, accessibility, and off-site redundancy, with the potential for cost savings.

Hierarchical Storage Management (HSM)
Automated approach that dynamically moves data between different storage tiers based on usage patterns and access frequency. Can be implemented using NoSQL database storage systems.

3.4 Data Archival and Retention Policies

All responsible organisations should develop a comprehensive **data archival and retention policy** that is aligned with business goals, regulatory requirements and industry best practices.

The data archival and retention policy should:

- be regularly reviewed and updated to ensure compliance with evolving regulations and organizational needs.

- Include robust **data governance** practices, including data classification, access controls, and audit trails for archived data.

- implement periodic data integrity checks, system validations, and migration tests to ensure the long-term viability and accessibility of archived data.

- have provisions for documenting and maintaining an inventory of archived data, including metadata and indexing information, to facilitate efficient retrieval and management.

3.5 Challenges to & Trends in Data Archival

Balancing Cost and Performance
Striking a balance between storage costs, data accessibility, and retrieval performance remains a challenge, with organizations seeking more cost-effective archival solutions.

Big Data and Unstructured Data
Archiving large volumes of unstructured data, such as social media content or multimedia files, poses unique challenges due to their size, complexity, and diverse formats.

Data Privacy and Protection
Compliance with data privacy regulations, such as the PDPA, necessitates careful consideration of data archival processes to protect privacy rights.

Artificial Intelligence and Machine Learning
Integration of such technologies can enhance data archival by automating classification, indexing, and intelligent search capabilities, improving efficiency and data discovery.

4 Backup vs. Archive

The table below gives the key differences between a backup and an archive:

Backup	**Archive**
Enables rapid recovery of live, changing data	Stores unchanging data no longer in use but must still be retained.
One of multiple copies of data	Usually the only remaining copy of the data
Access to data must be quick to allow rapid restoration of data	Speed of access to the data is usually not crucial
Short term retention of data only for the period when the data is in use	Long term retention of data for the required period or indefinitely
Duplicate copies are periodically overwritten	Data should not be altered or deleted