Chapter 24 Data Management

Content

Syllabus Learning Outcomes

3.3 Databases and Data Management

Understand, create and use SQL and NoSQL databases, as well as understand techniques to protect the privacy and integrity of data.

3.3.9 Understand the need for privacy and integrity of data.
3.3.10 Describe methods to protect data.
3.3.11 Explain the difference between backup and archive.
3.3.12 Describe the need for version control and naming convention.
3.3.13 Explain how data in Singapore is protected under the Personal Data Protection Act to govern the collection, use and disclosure of personal data.

1 Data Lifecycle

In computer science, data is information coded and structured for subsequent processing. It is any sequence of one or more symbols.

Data requires interpretation to become information. In computers, data is stored digitally, i.e.

data is represented using the binary number system of ones and zeros, instead of analogue representation. This allows us to use the computer to process the data.

The data life cycle refers to the entire period of time that data exists in the system. This life cycle encompasses all the stages that the data goes through, from its creation to its eventual disposal, encompassing the entire lifespan of data within an organisation or system.

2 Data Privacy

Data privacy is the requirement for data to be accessed by, or disclosed to, authorised users only. In other words, it is about keeping data private rather than allowing it to be available in the public domain. The term may be applied to both individuals and organisations.

It is important that unauthorised people do not have access to data they are not supposed to have. Unfortunately, in today's digitised society, it is easier than ever to gather a person's data and use it to obtain valuable insights, track his/her movements, or commit fraud.

For instance, data on which websites you visit often can reveal which products you are more likely to purchase as a shopper. This information can be highly valuable to an advertiser or a company.

As more services become available online, the risk of fraudulent use of data increases. For instance, with a photo of your identity card, a person can impersonate you and register for a new phone line on a telco website. Previously, such a transaction would have required the person to personally register over the counter with a physical identity card.

As technology becomes increasingly powerful, machines can gather information on a person easily, like performing facial recognition on surveillance videos to track down the whereabouts of an individual in a particular area.

Organisations also not only have to restrict access to data that is private to them, such as their management meeting minutes, but they are also responsible under privacy protection legislation to protect the data that they collect from individuals.

3 Data Integrity

Data integrity refers to the accuracy and validity of data. Data integrity covers data in storage, during processing, and while in transit. Data integrity can be compromised in several ways and at different stages during data processing.

- Errors on input – Data that is keyed in may be wrongly transcribed. A batch of transaction data could go astray, or be keyed in twice by mistake.

- Errors in operating procedures – An update program could for example be run twice in error and quantities on a master file would then be updated twice.
- Program errors – These could lead to corruption of files; a new system may have errors in it that will not surface for some time, or errors may be introduced during program maintenance.
- Viruses – Files can be corrupted or deleted if disk becomes infected with a virus. ● Transmission errors – Interference or noise in a communications link may cause bits to be wrongly received.

Each time data is replicated or transferred, it should remain intact and unaltered between updates. Error-checking methods and validation procedures are typically relied on to ensure the integrity of data that is transferred or reproduced without the intention of alteration.

4 Threats to data

- **Lapses in users' behaviour**
  Users may not be careful when dealing with data. This includes making errors in data entry, using weak passwords, or revealing password/OTP to others.

  Case study:
  - https://www.straitstimes.com/singapore/courts-crime/2278m-lost-to-top-10-scams-in-first-half-of-2022-as-overall-crime-rises-by-36

- **Mismanagement by multiple users**
  Multiple users working on the same file may accidentally overwrite each other's data, causing data to be inaccurate or invalid.

- **Natural disasters**
  Natural disasters can destroy the computer system and/or storage device physically.
- **Unauthorised intrusion into the system**
  If the system is vulnerable to hackers, data may be corrupted or lost. Other cyberattacks such as virus, worms and trojan are equally damaging to the system and its data.

  Case studies:
  - https://www.straitstimes.com/singapore/personal-info-of-15m-singhealth-patients including-pm-lee-stolen-in-singapores-most
  - https://en.wikipedia.org/wiki/WannaCry_ransomware_attack
  - https://www.wired.com/story/confessions-marcus-hutchins-hacker-who-saved-the-internet/

- **Malicious software**
  Malicious software entering the computer system will cause harm to a system and the data stored in the system.

5 Methods to protect data

5.1 Data backup

Modern storage equipment usually provides built-in disk clustering and redundancy in the event data in the primary storage is lost or corrupted. Users can also duplicate data or do a data backup.

Data backup involves creating copies of data and storing them separately. It is intended to be used as a safety precaution or prevention against unexpected loss of data due to unintended data corruption or errors such as disk drive failing, files accidentally being deleted, or a data center going offline during a catastrophic event.

A backup is typically a short-term solution and may be performed on a regular basis (e.g., daily, weekly) to ensure that data is up to date and readily available in case of a data loss event.

5.2 Data archive

Archiving data involves storing data that is not actively used but kept for historical references or auditing purposes. It ensures that important records remain available years after they were created. In most cases, the purpose of data archiving is to meet legal and compliance requirements. For example, a doctor's office might be required to keep patient records for a certain period of time. Similarly, a bank may need to retain transaction records.

As far as the fundamental difference between backups and archives is concerned, backup helps with data recovery in the event of recent data loss or corruption or hardware failure. Archives, on the other hand, serve the dual purposes of long-term retention and space management.

Comparison chart between data backup and data archive

	Backup	**Archive**
Data Storage Method	The original data remains in place, while a backup copy is stored in another location	Archived data is moved from its original location to an archive storage location
Data State	Backed-up data is constantly changing	Once you create an archive, you do not modify it
Data Retention Policy	You periodically delete or overwrite data backups that are too old to be useful	Data archives are designed for long-term storage

Storage Type	Hot cloud storage or easily accessible local storage locations	Cold cloud storage or tape archives
Data Scope	All of your data, with the exception of unimportant information like temporary files	Specific files that you must retain for compliance purposes

Note: Hot storage refers to fast, easy-to-access data storage like your local hard drive or a quick-access cloud storage provider like Google Drive. In contrast, cold storage is for archival data that is rarely accessed and usually stored off-site.

5.3 Restricting access

We can also ensure that only authorised users can access our data. This includes using:

- Passwords
  - It is important that users follow general guidelines to ensure the password is strong. Certain organisations may have more stringent rules.
- Two-factor Authentication
  - This can be in the form of security tokens or One-Time-Password (OTP)
- Biometrics
  - This can be in the form of facial recognition and fingerprint scanning.

Such authentication and authorisation controls help us verify credentials and assure that user privileges are applied correctly. These measures are typically used as part of an identity and access management (IAM) solution and in combination with role-based access controls (RBAC).

5.4 Version Control

Version control is a system that records changes to a file or set of files over time so that you can recall specific versions later.

Git is a commonly used version tool. It allows users to rollback to previous working versions of processes to fix mistakes while minimising disruption to other team members. Git has three main states that your files can reside in: modified, staged, and committed:

- Modified means that you have changed the file but have not committed it to your database yet.
- Staged means that you have marked a modified file in its current version to go into your next commit snapshot.
- Committed means that the data is safely stored in your local database.

More details can be found here:
https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

5.4 File Naming Convention (FNC)

File naming convention is essentially a framework for naming files in a way that describes what they contain and how they relate to other files. This helps to minimise the chances of files being misplaced or lost unintentionally due to poor organisation of files. Such a convention also enables users to locate files quickly.

Developing an FNC is done by identifying the key elements of the project, and the important differences and commonalities between your files.

File naming best practices:

- Files should be named consistently
- File names should be short but descriptive (<25 characters) (Briney, 2015)
- Avoid special characters or spaces in a file name
- Use capitals and underscores instead of periods or spaces or slashes
- Use date format ISO 8601: YYYYMMDD
- Include a version number (Creamer et al. 2014)
- Write down naming convention in data management plan

Elements to consider using in a naming convention are:

- Date of creation (putting the date in the front will facilitate computer-aided date sorting)
- Short Description
- Work
- Location
- Project name or number
- Sample
- Analysis
- Version number

Examples of files following an FNC:

- 20160104_ProjectA_Ex1Test1_SmithE_v1.xlsx
- 20160104_ProjectA_MeetingNotes_SmithE_v2.docx

5.5 Validation and Verification

We have learnt how to implement data validation and verification earlier in the Data Validation and Verification lesson. Essentially, data validation and verification are measures to check data when it originally enters the system (i.e. 1st stage of the data life cycle) or when it is transmitted from one system (or process) to another. (subsequent stages)

5.6 Disaster recovery

Disaster recovery is a set of practices and technologies that determine how an organisation deals with a disaster, such as a cyberattack, natural disaster, or large-scale equipment failure. The disaster recovery process typically involves setting up a remote disaster recovery site with copies of protected systems and switching operations to those systems in case of disaster.

5.7 Data encryption

Data encryption alters data content according to an algorithm that can only be reversed with the

right encryption key. Encryption protects your data from unauthorised access even if data is stolen by making it unreadable.

5.8 Data Erasure

Data erasure is more secure than standard data wiping because it uses software to completely overwrite data on any storage device. It verifies that the data is unrecoverable. limits liability by deleting data that is no longer needed. This can be done after data is processed and analysed or periodically when data is no longer relevant. Erasing unnecessary data is a requirement of many compliance regulations.

6 Governance of personal data

6.1 Personal Data

Personal data refers to data about an individual who can be identified from that data, or from that data and other information to which the organisation has or is likely to have access.

Personal data can range from names and contact numbers to other types of data that form part of an individual's record.

Personal data may include the following:

- Full name
- NRIC or passport number
- Photograph or video image of an individual
- Mobile telephone number
- Personal email address
- Thumbprint
- Name and residential address

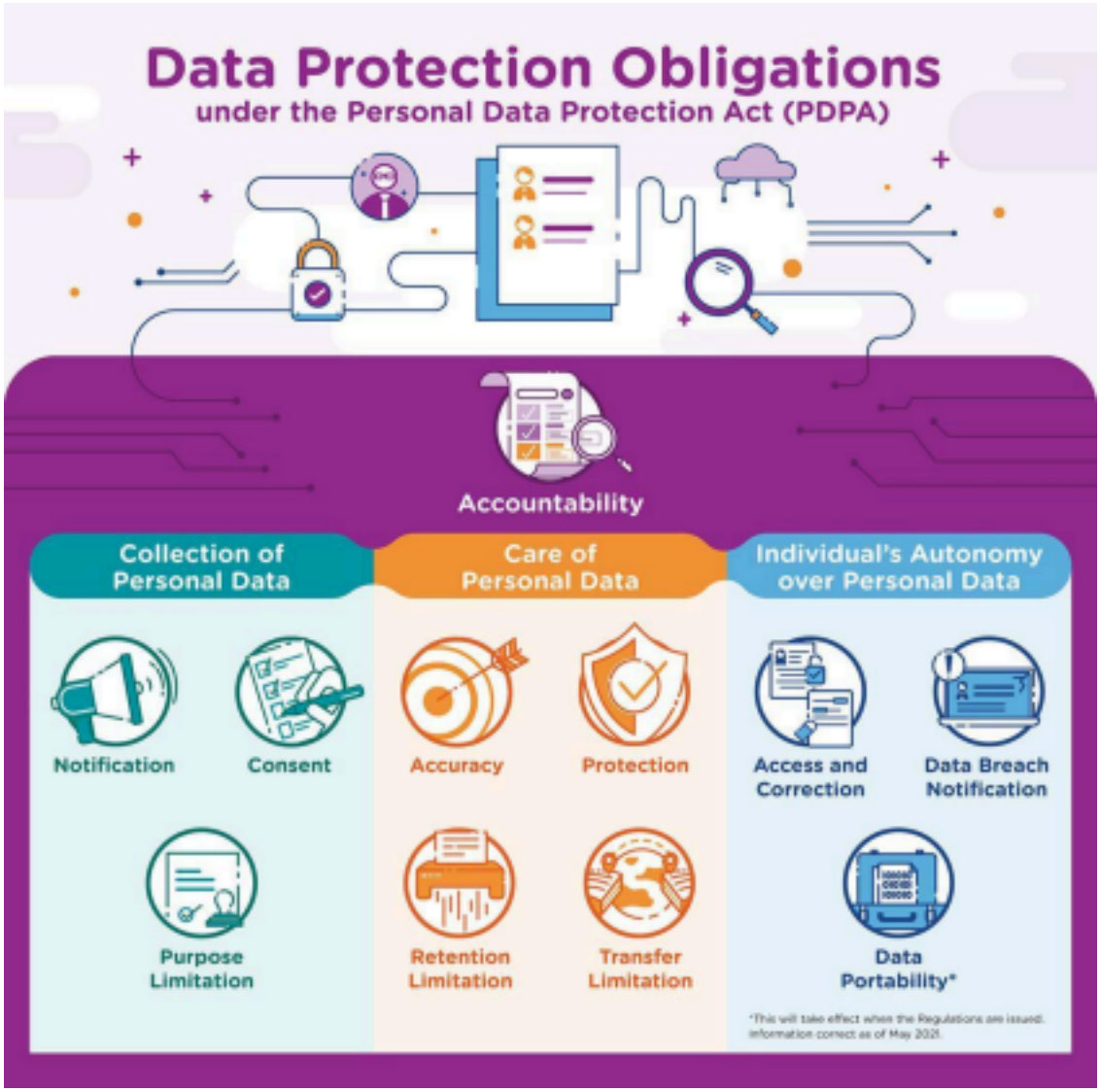
6.2 PDPA

The Personal Data Protection Act (PDPA) is a data protection law that protects against the misuse of personal data by governing the collection, use, disclosure, and care of personal data in Singapore. It recognises both the rights of individuals to protect their personal data, including rights of access and correction, as well as the needs of organisations to collect, use or disclose personal data for legitimate and reasonable purposes. PDPA helps to safeguard personal data from misuse and to maintain individuals' trust in organisations that manage their data.

By regulating the flow of personal data among organisations, the PDPA also aims to strengthen Singapore's position as a trusted hub for businesses.

6.3 Data Protection Obligations

Organisations are required to comply with the various data protection obligations if they undertake activities relating to the collection, use or disclosure of personal data.



1. Accountability Obligation

Undertake measures to ensure that organisations meet their obligations under the PDPA such as making information about your data protection policies, practices and complaints process available upon request and designating a data protection officer (DPO) and making the business contact information available to the public.

2. Notification Obligation

Notify individuals of the purposes for which your organisation is intending to collect, use or disclose their personal data.

3. Consent Obligation

Only collect, use or disclose personal data for purposes which an individual has given his/her consent to.

Allow the individual to withdraw consent, with reasonable notice, and inform him/her of the likely consequences of withdrawal. Once consent is withdrawn, make sure that you cease to collect, use or disclose the individual's personal data.

4. Purpose Limitation Obligation

Only collect, use or disclose personal data for the purposes that a reasonable person would consider appropriate under the given circumstances and for which the individual has given consent.

An organisation may not, as a condition of providing a product or service, require the individual to consent to the collection, use or disclosure of his or her personal data beyond what is reasonable to provide that product or service.

5. Accuracy Obligation

Make reasonable effort to ensure that the personal data collected is accurate and complete, especially if it is likely to be used to make a decision that affects the individual or to be disclosed to another organisation.

6. Protection Obligation

Reasonable security arrangements have to be made to protect the personal data in your organisation's possession to prevent unauthorised access, collection, use, disclosure or similar risks.

7. Retention Limitation Obligation

Cease retention of personal data or dispose of it in a proper manner when it is no longer needed for any business or legal purpose.

8. Transfer Limitation Obligation

Transfer personal data to another country only according to the requirements prescribed under the regulations, to ensure that the standard of protection is comparable to the protection under the PDPA, unless exempted by the PDPC.

9. Access and Correction Obligation

Upon request, organisations have to provide individuals with access to their personal data as

well as information about how the data was used or disclosed within a year before the request.

Organisations are also required to correct any error or omission in an individual's personal data as soon as practicable and send the corrected data to other organisations to which the personal data was disclosed (or to selected organisations that the individual has consented to), within a year before the correction is made.

10. Data Breach Notification Obligation

In the event of a data breach, organisations must take steps to assess if it is notifiable. If the data breach likely results in significant harm to individuals, and/or are of significant scale, organisations are required to notify the PDPC and the affected individuals as soon as practicable.

11. Data Portability Obligation*

At the request of the individual, organisations are required to transmit the individual's data that is in the organisation's possession or under its control, to another organisation in a commonly used machine-readable format.

**This will take effect when the Regulations are issued.*

Annex 1 - Do Not Call Registry

The Do Not Call (DNC) Registry lets you opt out of marketing messages addressed to your Singapore telephone number, such as those which promote or advertise a good or service, allowing you to have more control over the kind of messages you receive on your telephone, mobile phone or fax machine.

No Voice Call Register	No Text Message Register	No Fax Message Register
For phone calls	For texts including Short Messaging Service (SMS) and Multimedia Messaging Service (MMS)	For faxes

Do note that the DNC Registry does not cover:

- Marketing messages or calls from organisations with whom you have an ongoing relationship, if the messages are about products or services that are relevant to you. However, you may opt out from receiving such telemarketing messages and the organisation must stop sending such messages to your Singapore telephone number after 21 calendar days;
- Messages that are part of a market survey or research;

- Messages related to charitable or religious causes;
- Personal messages sent by individuals;
- Public messages sent by government agencies;
- Political messages; or
- Telemarketing calls or messages sent to a business (i.e. B2B).

References

Burdett, A. (2016). BCS glossary of computing. BCS Learning and Development Limited.

LANGFIELD, S. D. (2015). Cambridge International AS and A Level Computer Science Coursebook. Place of publication not identified: Cambridge University Press.

Watson, D., & Williams, H. (2019, May 31). Cambridge International AS & A Level Computer Science. London: Hodder Education

Personal Data Protection Commission Singapore. Retrieved from https://www.pdpc.gov.sg/Overview-of-PDPA/The-Legislation/Personal-Data-Protection-Act

https://www.pdpc.gov.sg/overview-of-pdpa/the-legislation/personal-data-protection-act/data-protection-obligations

https://www.pdpc.gov.sg/overview-of-pdpa/do-not-call-registry/individual/do-not-call-registry-and-you

Personal Data Protection Commission Singapore. (2019, September 26). Guide to Notification. Retrieved from https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Notification-260919.pdf.

https://www.fortinet.com/resources/cyberglossary/data-integrity

https://cloudian.com/guides/data-protection/data-protection-and-privacy-7-ways-to-protect-user-data/#protection-v-privacy

https://www.msp360.com/resources/blog/backup-vs-archive/#:~:text=As%20far%20as%20the%20fundamental,term%20retention%20and%20space%20management.

https://libguides.princeton.edu/c.php?g=102546&p=930626

MOE CPDD PDPA Notes