Data security and integrity

Security is the prevention of loss or damage to data.

Data is valuable so must be kept secure. Threats to data can be categorised as follows:

- Natural threats – no human intention:
  - Fires, floods…
- Deliberate human action:
  - Hackers and unauthorised users.
  - Viruses: A malicious program deliberately written to cause damage. Viruses are self-replicating pieces of software.
    - Boot sector viruses: not so common now.
    - Deleting/Corrupting files: viruses 'attach' themselves to particular files at random, or target specific files.
    - Macro viruses: very easy to create using high-level languages such as visual basic. Spread themselves through e-mail systems.
    - DoS attacks: denial of service attacks designed to clog up the Internet by flooding web servers with requests.
  - Terrorists.
- Human error:
  - Faulty software.
  - Operator error.
- Hardware fault.

Various safeguards can be taken to protect against these threats:

- Natural threats – no human intention:
  - Backup locations.
  - Protected locations.
- Deliberate human action:
  - Physically secure location (security guards).
  - Use an effective firewall, a hardware/software system that prevents unauthorised access to system. Passwords may be used to protect the system. However, there are limitations as passwords:
    - Are often forgotten, so people write them down or share them, making the system insecure.
    - Therefore many systems require passwords to be changed every 40 days, and old passwords are remembered so passwords cannot be reused.
    - Passwords are stored in an encrypted form, using a DES (data encryption standard) and a "key" is used to unlock the encrypted data.
  - Scan for viruses using an up-to-date scanner and antivirus software. The software:
    - Looks for a generic pattern, or a pattern for specific viruses in files.
    - Can scan incoming mail/files.
    - Keeps a watch-out on a computer every time you open a file.
    - Can work in a sophisticated manner over a network to update all machines simultaneously.
- Human error:
  - Test system fully.

    - Document system fully.
    - Train users thoroughly so fewer mistakes are made.
- Hardware fault:
    - Use backup hardware (RAID drive mirroring).
    - Protect using anti-surge equipment.

However, if these methods fail then a backup is needed to restore data. A backup is a copy of data kept preferably at a physically separate location. When keeping a backup the following must be considered:

- When?
    - Daily, weekly or monthly depending on how often the data is changed.
- How much?
    - A full backup may take many hours to complete, even to very fast media.
    - Incremental backups (backing up only what has changed since the last backup) on a daily basis and full backups on a weekly or months basis are therefore often more practical.
- Where?
    - Store the backup off site if possible.
- What media?
    - Tape is very common as it is cheap, quick and very easily portable.

Backups are often created using a generation system. The most recent backup created is the son – the 'live' backup. The father and grandfather are the previous backup and the one before that. If the live backup is corrupted it can be recreated using the father and a transaction file, a log of all the changes since the father was created. The grandfather is a final backup If both the father and son have become corrupted.

Integrity is the correctness and accuracy of data.

Validation aims to reduce errors, and does not guarantee correctness. Validation asks questions of the data to see if it *could* be correct.

- Maximum/minimum/required length.
- Range of number.
- List lookup.
- Type.
- Format/Picture – date, for example. In *Microsoft Access* these are called "Input masks." The first character must be alphabetic, the second numeric and the third a hyphen, for example.
- Presence check – required data.
- Uniqueness check – primary keys require no duplication.

Check digits are a form of validation, generated by a function a value is added to the end of a string of data. When inputted into a system the same function is performed on the data, the check digit calculated and compared to the inputted value. This is a simple yet generally effective method of checking data. *Modulus 11* is the most common check digit function.

Batch integrity methods

- Batch total, the number of items of data there should be. For example, the total number of bills or meter readings.
- Control total, the total of a field with meaning. For example the total gas used.

- Has total, again the total of a field but with no meaning. For example the sum of all the housenumbers.

Verification asks if the data *is* correct.

- Tends to involve double checking data:
    - Asking you to enter a password twice.
    - "*Are you sure you want to delete these files*" messages.
    - Data duplicated on a magnetic stripe on a card.
- Data can be entered twice, by two people for example.

However, verification can be very expensive, yet the integrity of data in some systems, for example gas readings, can be very important.