INDEX

Introduction

- **Data** refers to the raw facts that do not have much meaning to the user and may include numbers, letters, symbols, sound or images.
- **Information** refers to the meaningful output obtained after processing the data.
- **Data processing** therefore refers to the process of transforming raw data into meaningful output i.e. information.

Data processing cycle

- It refers to the sequence of activities involved in data transformation from its row form to information. It is often referred to as cycle because the output obtained can be stored after processing and may be used in future as input.
- The four main stages of data processing cycle are:
  - Data collection
  - Data input
  - Data processing
  - Data output

1. **Data collection**

- Also referred to as data gathering or fact finding, it involves looking for crucial facts needed for processing.

2. **Data input:**

- **Input validation**: data entered into the computer is subjected to validity checks by a computer program before being processed to reduce errors as the input.

3. **Processing**

- This is the transformation of the input data by the software to a more meaningful output (information).Some of the operations performed on the data include calculations, comparing values and sorting.

4. **Output**

- The final activity in the data processing cycle is producing the desired output also referred to as information. This information can be distributed to the target group or stored for future use.

Description of some errors in data processing

1. **Computational errors**

- Occurs when an arithmetic operation does not produce the expected results. The most common computation errors include **overflow**, **truncation** and **rounding**
    - **Overflow errors:** Occurs if the result from a calculation is too large to be stored in the allocated memory space. For example if a byte is represented using 8 bits, an overflow will occur if the result of a calculation gives a 9-bit number.
    - **Truncation errors:** Result from having real numbers that have a long fractional part which cannot fit in the allocated memory space. The computer would truncate or cut off the extra characters from the fractional part. For example, a number like 0.784969 can be truncated to four digits to become 0.784
    - **Rounding errors:** Results from raising or lowering a digit in a real number to the required rounded number. for example, to round off 30.666 to one decimal place we raise the first digit after the decimal point if its successor is more than or equal to five. In this case the successor is 6 therefore 30.666 rounded up to one decimal place is 30.7.if the successor is below five,e.g.30.635,we round down the number to 30.6

2. **Transcription errors**

- Occurs during data entry. Such errors include **misreading** and **transposition** errors
    - Misreading errors (Format Errors)
        - Are brought about by the incorrect reading of the source by the user and hence entering wrong values. For example a user may misread a handwritten figure such as 589 and type S89 instead i.e. confusing 5 for S.
    - Transposition errors

        - Results from incorrect arrangement of characters i.e. putting characters in the wrong order. For example the user might enter 396 instead of 369.
        - These errors may be avoided by using modern capture devices such as bar code readers, digital cameras etc which enter data with the minimum user intervention.

3. **Algorithm or logical errors**

    - An algorithm is a set of procedural steps followed to solve a given problem. Algorithms are used as design tools when writing programs. Wrongly designed programs would result in a program that runs but gives erroneous output. Such errors that result from wrong algorithm design are referred to as algorithm or logical errors.

Student task

- Discuss about what is the difference, when validating data, between “formal validation” and “business validation”, in other words, “input validation versus business rules validation”.
    - Look for examples

Data integrity

- Data integrity refers to the accuracy and completeness of data entered in a computer or received from the information system. Integrity is measured in terms of **accuracy**, **timeliness** and **relevance** of data.

    - Accuracy: It refers to how close an approximation is to an actual value. As long as the correct instructions and data are entered, computers produce accurate results efficiently. In numbers, the accuracy of a real number depends on the number. For example, 72.1264 is more accurate than 72.13.

- Timeliness: This is the relative accuracy of data in respect to the current state of affairs for which it is needed. This is important because data and information have a time value attached to them. If received late, the information may have become useless to the user. For example, information in the newspaper that is meant to invite people for a meeting or occasion must be printed prior to the event and not later.

- Relevance: Data entered into the computer must be relevant so as to get the expected output. In this case, relevance means that the data entered must be pertinent to the processing needs at hand and must meet the requirements of the processing cycle. The user also needs relevant information for daily operations or decision making.

Threat to data integrity

- **Threats to data integrity can be minimized through the following ways:**
    - Backup data preferably on external storage media.
    - Control access to data by enforcing security measures.
    - Design user interfaces that minimize chances of invalid data entry.
    - Using error detection and correction software when transmitting data
    - Using devices that directly capture data from the source such as bar code readers, digital cameras, and optical scanners.

Student task:

- Think about examples of data that is not consistent or lacks of integrity according to your criteria.
- Optional: What does ACID means when talking about Relational Databases?

1.1 Atomicity o atomicidad
1.2 Consistency o consistencia
1.3 Isolation o aislamiento
1.4 Durability o durabilidad

Computer files

- **A file** can be defined as a collection of related records that give a complete set of information about a certain item or entity.
- **Elements of computer file:** A computer file is made up of **fields** and **records**.
    - **Field:** A field is a single character or collection of characters that represents a single piece of data. For example, the student's admission number is an example of a field.
    - **Records:** A record is a collection of related fields that Represents a single entities, e.g. in a class score sheet, detail of each student in a row such as admission number, name, total marks and position make up a record.

1st Classification: Logical and physical files

- Computer files are classified as either **physical** or **logical**

- **Logical files:** A computer file is referred to as logical file if it is viewed in terms of what data item it contains and details of what processing operations may be performed on the data items. It does not have implementation specific information like field, data types, size and file type.

- **Physical files:** As opposed to a logical file, a physical file is viewed in terms of how data is stored on a storage media and how the processing operations are made possible. Physical files have implementation specific details such as characters per field and data type for each field.

2nd Classification: Binary and text files

All files are saved in one of two file formats - binary or text. The two file types may look the same on the surface, but their internal structures are different.

While both binary and text files contain data stored as a series of bits, the bits in text files represent characters, while the bits in binary files represent custom data.

Distinguishing between the two is important as different OSs treat text files differently. For example in Unix you end your lines with just \n while in MS OSs you use \r\n and in Macs you use \n\r. Software such as FTP clients try to change the line endings on text files to match the destination OS by adding/removing the characters. This is to make sure that the text file will look properly on the destination OS.

Student task:

- Bits or bytes?
- Search for examples of text and binary files in your computer
- ASCII, UTF-8… What's this?

Types of Computer Processing Files

- There are numerous types of files used for storing data needed for processing, reference or back up. The main common types of processing files include

- Master files,
- Transaction,
- …

1. **Master file:** A master file is the main that contains relatively permanent records about particular items or entries. For example a customer file will contain details of a customer such as customer ID, name and contact address.
2. **Transaction (movement) file:** A transaction file is used to hold data during transaction processing. The file is later used to update the master file and audit daily, weekly or monthly transactions. For example in a busy supermarket, daily sales are recorded on a transaction file and later used to update the stock file. The file is also used by the management to check on the daily or periodic transactions.

Student task:

- Identify examples of mater files and transaction files in your daily life or in a business context which is familiar to you.

File organization methods

File organization refers to the way data is stored in a file. File organization is very important because it determines the methods of access, efficiency, flexibility and storage devices to use. There are four methods of organizing files on a storage media. This include, among others:

- **serial**
- **random,**
- **indexed-sequential**

1. **Serial file organization**
   - Records in a file are stored and accessed one after another.
   - The records are not stored in any way on the storage medium this type of organization is mainly used on magnetic tapes.

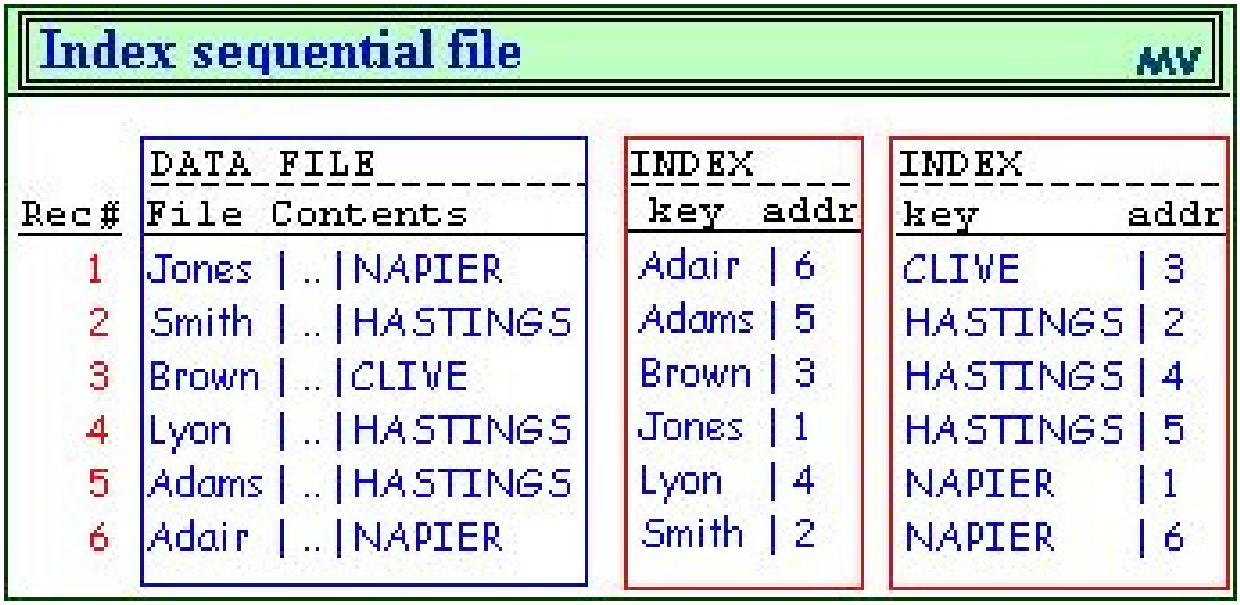
Advantages	Disadvantages
• It is simple • It is cheap	• It is cumbersome to access because you have to access all proceeding records before retrieving the one being searched. • Wastage of space on medium in form of inter-record gap. • It cannot support modern high speed requirements for quick record access.

2. **Random or direct file organization**
   - Records are stored randomly but accessed directly.
   - To access a file stored randomly, a record key is used to determine where a record is stored on the storage media.
   - Magnetic and optical disks allow data to be stored and accessed randomly.

Advantages	Disadvantages
• Quick retrieval of records. • The records can be of different sizes.	• Relatively complex when programming. • System design based on random file organization is complex and costly. • It cannot support modern high speed requirements for quick record access.

3. Indexed-sequential file organization method

- Almost similar to sequential/serial method only that, an index is used to enable the computer to locate individual records on the storage media. For example, on a **magnetic drum**, records are stored sequential/serial on the tracks. However, each record is assigned an index that can be used to access it directly.
- An **index** is an additional file that contains information about where records are stored.
- In the simplest case, an index would contain the **key values** of the records and the **address** where the record is stored.
- The **index file** would be a **sequential file**, with the index records stored **in key value order**. This makes the index easy to search.
- An **indexed sequential file** is a sequential file that also has an index. This would allow sequential access of records as well as indexed access.



Student task:

- Search for advantages and disadvantages of indexed sequential files

Student Task

A practical exercise will be given to the students for the Analysis and Design of the structure of different files that will complement the concepts worked in this unit.

In the next unit the files designed will be constructed using a RDBMS (Relational Database Management System)