**1. Definition of Terms and Stages of Data Collection**

**Definition of terms and Stages of data collection**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **a) Data**Data is defined as a collection of raw facts (figures, letters, special symbols or a combination of both) that convey little meaning by themselves. example: in a statement like  **James, John and Caro*** what about them?
* why
* how

We actually don't know anything about these three people thus it does not have much meaning to us until its fully processed for us to understand**b) Information**Information is defined as data has undergone processing and is meaningful to the user when he/she needs it. The statement simply indicates that information is not just a by-product of data processing, but it must have certain qualities for it to qualify to be a resource to the organization. Thus the qualities of timeliness, correctness and relevancy must be observed for it to be good information.  example: **James, John and Caro were awarded certificates for being the cleanest students in our school.**this statement now satisfies our curiosity because it has been attached much more meaning useful to us but we need also to when. This statement can therefore be regarded as information**c) Data Processing**  Data processing can be regarded as the act of collecting data, manipulating it to get information through (input, Processing and output) and finally distributing it to the relevant organs for use.Data Processing PlatformData Processing PlatformData processing cycle are the stages (phases) through which data moves before it becomes information. these stages include:1. Data collection
2. Data input
3. Processing
4. Output

data processing cycleData Processing Cycle**Stages of data processing cycle****i) Data Collection**Data collection is the act of collecting data from its point of origin to the computer in a form suitable for processing**Stages of Data Collection**Data collection is procedural, thus it attracts the following stages**A) Data Creation**Data creation is the process of capturing data from the source location using data capturing methods like:***(scanning, digital camera, touch screen, voice input etc)***. **B) Data Transmission**This is a term used to refer to the transfer of data from the point of collection (source) to the point where processing is to be done (data processing department) this is possible through computer-to-computer communication, physically by post office or via courier services if the source is away from the data processing department.**C) Data Preparation**Data preparation is the term used to describe the process by which data is prepared and made ready for entry in the computer for processing.**ii) Data input**This is the process where the collected data is converted from human readable form into machine readable form (binary form) To ensure quick and error free data the following features should be considered:* **Validation** - this is the counter checking whether what is on the input document is exactly what is entered into the computer
* **Media Conversion** - this is the act of converting data from different storage medium e.g from a flash disk to a hard disk
* **Sorting** - Data should be arranged in a predefined order before being processed

**iii) Processing**This is the manipulation of input data into a more meaningful output that can be regarded as being meaningful to the user. Processing involves arithmetic, sorting and logical operations**iv) Output**This is the final activity in data processing cycle where the expected output (information) is displayed. This information is then distributed to places **(information dissemination)** where its needed to influence decision making process. Information dissemination may include: electronic presentations over radio or television, distribution of hard copies, broadcasting messages over the internet or mobile phones.**Questions**1. Define the following terms    a) Data Processing    b) Data processing cycle.    c) Data    d) Information2. using an illustration, describe the four primary stages of the data processing cycle3. Outline the stages of data collection4. List the three feature involved during data input stage to minimize errors5. What is the essence of output stage?**2. Description of errors in data processing** **Description of errors in data processing**

|  |  |
| --- | --- |
|  | **Introduction:**Computers make errors because people program them to make those errors. Thus the acronym GIGO - Garbage in Garbage out- errors can be described as:1) Transcription errors2) Computation errors and3) Algorithm errors**1) Transcription errors**    Transcription errors occur during data entry. These errors include:    i) Misreading errors    ii) Transposition errors**i) Misreading errors**These errors are caused by incorrect reading of the source document by the user and hence        entering wrong values e.g. a user may enter 5 instead of S or 0 instead of o and so forth.**ii) Transposition errors**These are errors that occur due to incorrect arrangement of characters. like putting characters in the wrong order. e.g. the user may enter 369 instead of 396. However these errors can be avoided by using modern data capture devices like the bar code reader, optical character reader, digital cameras, scanners, configuring the right data-types in the database e.t.c.**2) Computation errors** These are errors that occur when an arithmetic operation does not produce the expected results. They include:**overflow, underflow, truncation and rounding errors****a> Overflow**These are errors that occur from a calculation which appear too large to be stored in the allocated memory space. e.g. if a byte is represented by 8 bits, an overflow will occur if the result of a calculation gives a 9-bit number.**b> Underflow****Underflow** (or "floating point underflow", or just "underflow") is a condition in a computer program where the result of a calculation is a smaller number than the computer can actually store in memory.Arithmetic underflow can occur when the true result of a floating point operation is smaller in magnitude (that is, closer to zero) than the smallest value representable as a normal floating point number in the target datatype. Underflow can in part be regarded as negative overflow of the exponent of the floating point value. For example, if the exponent part can represent values from −127 to 127, then a result with absolute value less than 2−127 may cause underflow.**c> Truncation**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. The resulting number is not rounded off.**d> Rounding**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 5. In this case, the successor is 6 therefore 30.666 rounded up to one decimal place is 30.7. If the successor is below 5, e.g. 30.635, we round down the number to 30.6.**3) Algorithm 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. **Data Integrity**Data integrity refers to the correctness and completeness of data entered in a computer or received from the information system. Integrity is measured in **accuracy**, **timeliness** and **relevancy** of data**A> Accuracy**This is how close an approximation is to an actual value. example: in a number like 34.247545, 34.2475 is more accurate than than 34.2 this is because the deviation on the former is much lesser than the latter.**B> Timeliness**This is the relative accuracy of data in respect to the current state of affairs for which it is needed. Information should be available on time for decision making. **For example:** In a data base to prepare worker's salaries, information on the presence of workers over the month should be processed on time before salary arrears are released every month.**C> Relevancy**Data entered into the computer must be relevant in order to get the expected output. it must meet pertinent needs at hand and must meet the requirements of the processing cycle.**Way to minimize data integrity threats**1. Backup data on secondary storage devices or on online storage stores like **Dropbox and Google Drive**
2. Control access to data by enforcing security measures
3. Design user interfaces that minimize chances of invalid data entry
4. Using error detection and correction software when transmitting data
5. Using devices that directly capture data from the source such as bar code readers, digital cameras, optical character reader e.t.c.

**Quiz**1. Define the following terms a) data processing b) data processing cycle
2. Using an illustration, describe the four primary stages of the data processing cycle
3. outline the stages of data collection
4. you may have come across the term garbage in garbage out (GIGO). what is its relevance to errors in data processing.
5. explain the two types of transcription errors
6. state three types of computational errors

        a) define the term data integrity        b) give three factors that determine the integrity of data        C) state at least five ways of minimizing threats to data integrity. |

**3. Data Processing Methods** posted Jun 26, 2014, 7:08 AM by Maurice Nyamoti   [ updated Feb 8, 2016, 4:52 PM ] **Data Processing Methods**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Introduction:**There are **three** types of data processing methods namely:1. Manual data processing
2. Mechanical data processing
3. Electronic Data Processing

**Manual data processing**A paper and pen are used to process in manual data processing **example:** Most schools in Kenya process their data manually especially while admitting form 1s. As the students report, they are given a form to fill, these forms are filed by the school secretary and stored in the file cabinet- that is, if the school has 450 students, the school secretary will have those 450 files to handle thus retrieve, store, manipulate and update information which is a tedious job. therefore manual data processing can also be defined as the physical entering and organization of data**Mechanical data processing**In **mechanical data processing**, data is processed with different mechanical devices. This can include typewriters and calculators**electronic****Electronic Data Processing****Electronic Data Processing (EDP)** can refer to the use of automated methods to process commercial data. This means that, data processing, retrieval, storage, manipulation, sorting and part of analysis is done through automated methods using devices like computers.**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. A **computer file** is a resource for storing information, which is available to a **computer** program and is usually based on some kind of durable storage. **Advantages of computerized/electronic filing system**1. Ease of access.
2. Speed of access.
3. Increased search facilities.
4. Capability to store huge amounts of information in a small space.
5. Portability - a pen drive or disc enables you to transport enormous amounts of data easily and discreetly.
6. Access to data can be monitored and/or restricted. thus enhancing data security
7. Its easier to update and modify information
8. Enhances data integrity and reduces duplication

**Disadvantages of computerized/electronic filing system**1. Electrical/Electronic failure can cause temporary suspension of business or permanent loss of data.
2. Portability - copied information can be easily hidden on a disc or pen drive.
3. Data is not always secured by correct monitoring of access.
4. Information can be permanently deleted in error.
5. Hacking can release information that is data protected.
6. Cost of installing system.

**Elements of a Computer File**A computer file is made up of three elements namely:1. Characters
2. Fields
3. Records

**Characters**A character refers to a letter, number or symbol that can be entered, stored and output by a computer. This is the smallest element in a computer file.**Fields**A field is a single character or a collection of characters that represent a single entity**Example:**fields and records**Records**A record is a collection of related fields that represent a single entity.**Classification of Computer files**Computer files are classified as:1. Logical files
2. Physical files

**Logical Files**This is computer file viewed in terms of what data items it contains and details of what processing operations may be performed on the data items. It does not have implementation specific information like field and data types, size and file types.**Physical Files**This is a computer file viewed in terms of how data is stored on a storage media and how the processing  operations are made possible. They have specific details such as characters per field and data type for each field.**Questions**1. State the three types of data processing methods
2. Define a computer file and state three advantages and three disadvantages of computer files
3. Define: (i) Character  (ii) field  (iii) record
4. Explain the two classifications of computer files
5. Distinguish between a manual file and a computer file
6. List at least three equipments used during mechanical data processing methods for processing
7. Why is a computer file described as 'durable'?
8. **Types of Computer Processing Files**

**Types of Computer Processing Files**

|  |  |
| --- | --- |
|  | There are various types of files used for storing data needed for processing, reference or backup. However the main common files include:Master files, Transaction, Reference, Backup, Report and sort file.**Master File**A master file refers to a collection of records pertaining to one of the main subjects of an information system, such as customers, employees, products and vendors.**Transaction (movement) file**This is a collection of **transaction** records. The data in **transaction** files is used to update the master files, which contain the data about the subjects of the organization (customers, employees, vendors, etc.).**Reference File**This file is normally used for reference or look-up purposes. Look up information is that information which is stored in a separate file but is required during processing.**Backup File**A backup file is used to hold copies (backups) of data or information from the computers fixed storage (hard disk) **Report File**A report file is used to store relatively permanent records extracted from the master file or generated after processing.**Sort File**A sort file stores data which is arranged in a particular order**Questions**A. State uses of the following files:1. Sort file
2. master file
3. reference file
4. report file
5. transaction file
6. backup file

B. Of the above files, which file can be used:    i.    In a supermarket to fetch the price of a commodity from the master file    ii.    In a supermarket to update the master file    iii.    In a bank to extract and store some important records from the master file    iv.    In a bank to arrange records of customers according to their bank deposits    v.    By the government to a keep copy of listed employees in the master file who are               about to retire in one year time |

 |

5. File Organization Methods **File Organization Methods**

|  |  |
| --- | --- |
|  | **Introduction**File organization refers to the way data is stored in a file. File organization is very important because it determines the method of access, efficiency, flexibility and storage devices to be used. There are four methods of organizing files on a storage media namely: Sequential, random, serial and indexed- sequential **Sequential*** A sequentially organized file consists of records arranged in the sequence in which they are written to the file (the first record written is the first record in the file, the second record written is the second record in the file, and so on). As a result, records can be added only at the end of the file. Attempting to add records at some place other than the end of the file will result in the file begin truncated at the end of the record just written.
* Sequential files are usually read sequentially, starting with the first record in the file. Sequential files with a fixed-length record type that are stored on disk can also be accessed by relative record number (direct access).
* Records in sequential files can be read or written only sequentially.
* After you have placed a record into a sequential file, you cannot shorten, lengthen, or delete the record. However, you can update (REWRITE) a record if the length does not change. New records are added at the end of the file.
* If the order in which you keep records in a file is not important, sequential organization is a good choice whether there are many records or only a few. Sequential output is also useful for printing reports.

**Advantages**1. Sorting makes it easier to access records

**Disadvantages**1. Sorting does not remove the need to access other records as the search looks for a particular record
2. Sequential records cannot support modern technologies that require fast access to stored records
3. The requirement that all records be of the same size is sometimes difficult to enforce

**Random*** In random file organisation, records are stored in random order within the file.  Though there is no sequencing to the placement of the records, there is however, a pre-defined relationship between the key of the record and its location within the file.  In other words, the value of the record key is mapped by an established function to the address within the file where it resides.  Therefore, any record within the file can be directly accessed through the mapping function in roughly the same amount of time.  The location of the record within the file therefore is not a factor in the access time of the record.  As such, random files are also known in some literature as direct access files.
* This method is normally used by optical disks like compact disks

**Advantages**1. Quick retrieval of records
2. The records can be of different sizes

**Serial*** Serial file organization is the simplest file organization method.  In serial files, records are entered in the order of their creation.  As such, the file is unordered, and is at best in chronological order. Serial files are primarily used as transaction files in which the transactions are recorded in the order that they occur.
* This type of access is normally used by magnetic tapes

**Advantages**1. It is simple

**Disadvantages**1. It is cumbersome to access because you have to access all preceding records before retrieving the one being searched
2. Wastage of space on medium in form of inter- record gap
3. It cannot support modern high speed requirements for quick records access

**Indexed- Sequential*** indexed file contains records ordered by a **record key**. Each record contains a field that contains the record key. The record key uniquely identifies the record and determines the sequence in which it is accessed with respect to other records. A record key for a record might be, for example, an employee number or an invoice number.
* An indexed file can also use alternate indexes, that is, record keys that let you access the file using a different logical arrangement of the records. For example, you could access the file through employee department rather than through employee number.
 |

a) this isting raw data into a meaningful format called informationn6. Electronic Data Processing Modes Electronic Data Processing Modes

|  |  |
| --- | --- |
|  | Computers are programmed to process data in different ways. Just like humans they can process shortest job first-SJF, First Come First Serve-FCFS or they can just Round Robin giving a time span, important/emergency jobs first-real time. Computers can also multitask or multi-programming etc. Examples of processing modes are:1.   Online Processing2.   Real-Time processing3.   Distributed Processing4.   Time-Sharing5.   Batch Processing6.   Multitasking7.   Interactive ProcessingOnline Processing * In this processing mode, data is processed immediately it is received.
* **Example:** When booking a seat on an airline, the seat is booked immediately.  This is to avoid problems of double-booking.  **other example** of online processing is when playing computer games online

Real-Time Processing* In real time processing the computer processes the incoming data as soon as it occurs, up-dates the transaction file and gives an immediate response that would affect the events as they happen.
* There is no much difference with online processing however with real-time processing, the outcome of the processing is required immediately to influence decision making and not necessarily through online means but can be within a machine.
* **Example:** in nuclear power stations, a certain level of temperature is required to be maintained for effective disintegration of atoms. If the temperatures are not controlled, the system  may cause an emergency situation- like in increase or decrease in slight temperature fluctuations, therefore computers are used to control the air conditioning systems at the plant because they are programmed to provide instant answers upon a slight change in temperature has occurred- that is what we call **real-time processing**
* **Another example**: is the use of humidifiers and dehumidifiers in the computer lab, when there is a lot of humidity, computers turn on the dehumidifiers and vice versa

Distributed data processing* These points to dividing (distributing) processing tasks to two or more computers that are located on physically separate sites but connected by data transmission media; there may be a central computer that receives input from the remote computers (terminals), processes the data and updates the master file. If required, the output can be communicated back to the remote terminals.
* **Example:** When you withdraw money from a bank, your records are transacted and updated in the main server computer which will also update client computers across the bank branches divide.

Time-Sharing* Just as the name suggests, time-sharing refers to many terminals connected to a central computer and given access to the central processing unit apparently at the same time this sounds like **‘round robin’**. Each user is allocated a time slice of the CPU in sequence.
* The amount of time allocated to each user is controlled by a multi-user operating system. If a user’s task is not completed during the allocated time slice, he/she is allocated another time slice later in a round robin manner.

Batch Processing* In batch processing, data is accumulated as a group (batch) over a specified period of time e.g. daily, weekly or monthly.
* The batch is processed at once. **E.g.** in a payroll processing system, employees details concerning number of hours worked, rate of pay, and other details are collected for a period of time, say one month. These details are then used to process the payment for the duration worked.

other areas where batch processing can be applied include:* Processing bank cheques
* Printing of bank statements
* Updating of a stock database

Multiprocessing* Multiprocessing refers to the processing of more than one task apparently at the same time. This is possible in computers like mainframes and network servers.
* A computer may contain more than one independent central processing unit which works together in a coordinated way. At a given time, the processors may execute instructions from two or more programs or from different parts of one program simultaneously.

Multi-programming* Also referred to as *multitasking*- refers to a type of processing where more than one program are processed apparently at the same time by *a single central processing unit*.
* Unlike multiprocessing, in multitasking, the computer has only one CPU. The computer allocates each program a time slice and decides what order they will be executed.

Interactive Processing* There is a continuous dialogue between the user and the computer.
* As the program executes, it keeps on prompting the use to provide input or respond to prompts displayed on the screen

Advantages of electronic processing1. Quick processing especially where all required data is available2. Availability of data and information digitally3. Distance between entities that are processing data is made non-significant4. Support for information sharing and collaboration on a wider scale.Disadvantages of electronic information processing1. Security of data can be compromised during storage or while in transit on networks if appropriate measures are not taken2. Lack of legal frameworks in many countries that should support electronic processing activities3. Lack of ICT skills among many knowledge workers to support electronic data processing.Topical Questions1.   What similarity is there between:i>           Online processing and real-time processingii>          Multiprocessing and multi-programming2.   What differences are there between:                      i.        Online processing and real-time processing                     ii.        Multiprocessing and multi-programming3.   SHULE Secondary school Board of Governors has asked you to help them find a computer processing mode that they will introduce to enroll form one students each year that will allow them to confirm admissions, pay school fees without necessarily having to travel physically to the school. Which processing mode would you recommend and why 4.   Of the processing modes, which mode is effective for filling questionnaires online and why? |

            b) this are a set of stages that data passes through in order to be transformed into                         information            c) These are raw facts which do not have much meaning to the user            d) Processed data which has much meaning and can be used in decision making        2.         3. i) data collection            ii) input            iii) processing            iv) output        4. a) media conversion            b) input validation            c) sorting        5. Display of expected results (information) that can be used in decision making |