AS and A LEVEL
Information Technology
9626

Chapter 1

Data Processing and Information

Faisal Chughtai
info.sirfaisal@gmail.com
www.faisalchughtai.com

Data and Information

Data

Data is raw numbers, letters, symbols, sounds or images with no meaning.

Some examples of data are:

- P952BR
- @bbcclick
- 359

The data P952BR could have several meanings. It could be:

- A product code
- A postal/ZIP code
- A car registration number

Because we do not know what the data means, it is meaningless.

Information

When data items are given context and meaning, they become information. Information is much more refined data, that has evolved to the point of being useful for some form of analysis.

Data is given context by identifying what sort of data it is.

Data	Context	Comment
P952BR	A product code	This is a product code, but it is still not known what it is a product code for, so it is still data.
@bbcclick	A Twitter handle	This is an address used for Twitter, but it is not information unless it is known to be a Twitter handle or used within Twitter software. It's also not known whose address it is.
359	Price in Pakistani Rupees	This is a currency value, but it is not known what the price is for, so it is still data.

Table 1.1: Examples of data being given context.

For the data to become information, it needs to be given meaning.

Data	Context	Meaning
P952BR	A product code	A product code for a can of noodles.
@bbcclick	A Twitter handle	The Twitter address for the BBC's weekly technology show, Click, which is worth watching on BBC World News and BBC2 to keep up to date with technology.
359	Price in Pakistani rupees	The price of a mobile phone cover.

Table 1.2: Examples of data being given context and meaning to become information.

Data Sources

Direct data

Data collected from a **direct data source** (primary source) must be used for the same purpose for which it was collected.

It is often the case that the data will have been collected or requested by the person who intends to use the data. The data must not already exist for another purpose though.

When collecting the data, the person collecting should know for what purpose they intend to use the data.

Indirect data

Data collected from an **indirect data source** (secondary source) already existed for another purpose.

Although it can still be collected by the person who intends to use it, it was often collected by a different person or organization.

Examples of direct and indirect data

Example 1: Online shop

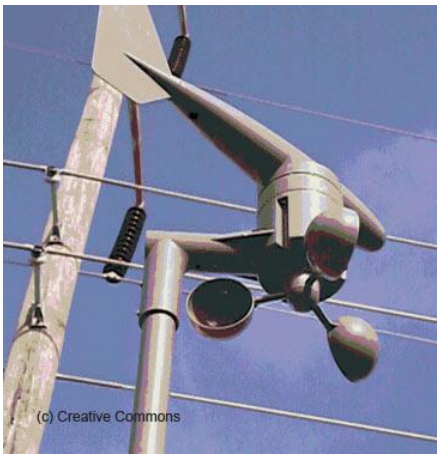
An online shop stores your email address and details of items you have bought. They use this data to help with their stock control and to send you an order confirmation. This is **direct** or **original source data**.

But then they might sell your data (email address) to a similar company (with your permission).

This second company might then email you with a list of related items you might be interested in **(indirect data)**.

For example, you buy a computer game from one online company, then an email arrives from a different company asking if you would like to buy a strategy book for the game.

Example 2: Weather data



Data loggers are set up all over the country to measure local weather conditions.

All this data is gathered together by the 'Met office' to allow weather forecasts to be made. **(Direct or original source data)**

But this 'data set' may also be purchased by a local business who wants to see how sales of their ice-cream relates to the weather.

This information is used to plan ice-cream production ahead of time. **(indirect data)**

Direct data sources

Questionnaires

Questionnaires are often used to collect data from individuals. They can be hard copy or completed online. Questionnaires can make it easier to analyse information because all respondents are asked the same questions.

Interviews

Interviews allow you to collect more data from people as responses can be given in greater depth.

Observation

Data gatherers observe what is happening during an activity, then record and analyse the resulting data.

Data logging
This is the use of sensors to produce data that can be gathered and interpreted.

Uses of direct data

An example of use of direct data could be planning a new high speed train line

The government are investigating the feasibility of introducing a new high speed rail link between the capital and a major industrial city. Before they decide whether to proceed or not, they will need to collect some direct data.

This direct data will include:

- The time it takes to get from the capital to the other city using the existing rail line.
- The number of trains and passengers who use the existing rail line.
- How many passengers would use the new system.
- What people who live on or near the existing route think about the effect it would have on their environment.

Here are some examples of how the data could be collected.

The time it takes to get from the capital to the other city using the existing rail line:

This information can be collected from existing train timetables, however this method would not be using a direct data source. Original data could be collected by actually travelling on trains periodically and timing the journeys, but this might not be practical given the time it would take.

The number of trains and frequency on the existing line:

The suggested method to be used is a data logger. A sensor is placed on the rail line. This sensor is attached to a roadside data logger. As trains pass over the sensor, their speed, time of day, number of carriages and frequency are logged. The advantage of a data logger is that it gathers physical data automatically.

The number of passengers:

The method could be to use infra-red sensors fitted around each door on the train to count the number of passengers getting on and off the train at each station. From these it can be calculated how many passengers are on the train at any point along its route. The data is fed back to a microprocessor.

How many passengers would use the new system:

This could make use of questionnaires: passengers on the existing route and airline passengers in the capital are asked to complete the questionnaires.

The advantage of questionnaires is that they can be collected and analysed reasonably quickly. The disadvantage is that only a proportion are returned, making the sample size quite small.

What local residents think:

Face-to face interviews would be best. The advantage of interviews is that they may gather some unexpected data and obtain personal attitudes that a simple questionnaire would not. However, it takes time for many interviews to take place.

Indirect data sources

Indirect data sources are third-party sources that the data gatherer can obtain data from.

Electoral register

This is a list of adults who are entitled to vote in a local or national election. An edited version of the register can be purchased and used for any purpose.

Businesses collecting personal information

Businesses sell the information that they collect from their customers. For example when someone purchases something online they are often asked to tick a box authorising the business to share this with other organisations. Customers often provide personal information that has a commercial value. Businesses use this information to create mailing lists that can be purchased by any other organisation/individual to send emails or even brochures through the post.

Uses of indirect data sources

- Apart from elections and other government purposes, the electoral register can only be used to select individuals for jury service or by credit reference agencies. These agencies are allowed to buy the full register to help them check the names and addresses of people applying for credit. They are also allowed to use the register to carry out identity checks in an attempt to deal with money laundering.
- Businesses which collect personal information often use it to create mailing lists that they then sell to other organisations, which are then able to send emails or even brochures through the post.
- Any organisation that provides data or information to the general public for use by them can be said to be an indirect source.
- Another scenario could be studying pollution in rivers. Direct data sources could be used, of course; questionnaires could be handed out to local landowners and residents in houses near to the river, asking about the effects on them of the pollution, and they could also be interviewed. Computers with sensors could be used to collect data from the river. However, indirect data sources could also be used; documents may have been published by government departments showing pollution data for the area.

Advantages of direct data sources

- Only as much or as little data is gathered as needed.
- Exactly where the data came from, and therefore how reliable it is, is known.
- There may be an opportunity to sell the data for other purposes.
- Gathering data directly addresses specific issues, as the data gatherer controls the methods of collecting the data to fit their needs.

Disadvantages of direct data sources

- Data gathering may be expensive as other companies may have to be hired to get it.
- It may involve having to buy equipment such as data loggers and computers.
- It may not be possible to gather original data due to the time of year e.g. winter snowfall data may be required but it is now the middle of summer.
- Compared to indirect data sources, using direct data sources may be very expensive in preparing and carrying out the gathering of data. Costs can be incurred in, for example, producing the paper for questionnaires, or the equipment for an experiment.
- It takes longer to gather data than to acquire data from an indirect data source.
- By the time the project is complete the data may be out-of-date.
  The sample size may be small.

Advantages of indirect data sources

- Indirect data sources may allow a larger set of data to be examined using less time and money than direct data collection would require.
- The use of indirect data sources allows data to be gathered from subjects (e.g. people) to which the data gatherer does not have physical access.
- A larger sample size can be used. Direct data gathering can have limitations due to the availability of the people being interviewed, but by using indirect data sources, the size of the sample can be increased giving rise to greater confidence in the findings.
- Using indirect data sources can be done at a relatively low cost, although this varies.
- Quite often the data can be in an easily accessible location such as the internet whereas for direct data sources, travelling expenses and time taken to collect data can be great.
- Information can be of a higher quality. Data collected indirectly has already been collated and grouped into meaningful categories and, for example, poorly-written responses to questionnaires or interview transcripts do not have to be read through to create the data source.

Disadvantages of indirect data sources

- The various purposes for which data was collected originally may be quite different to the purpose of the current research and unnecessary data may need to be filtered out.
- There may be no data available – the data required has simply never been recorded.
- There may be sampling bias – data from only one section of the community (whether it is based on educational level, level of income etc.) may have been collected but what is required is data from a representative cross-section of the community.

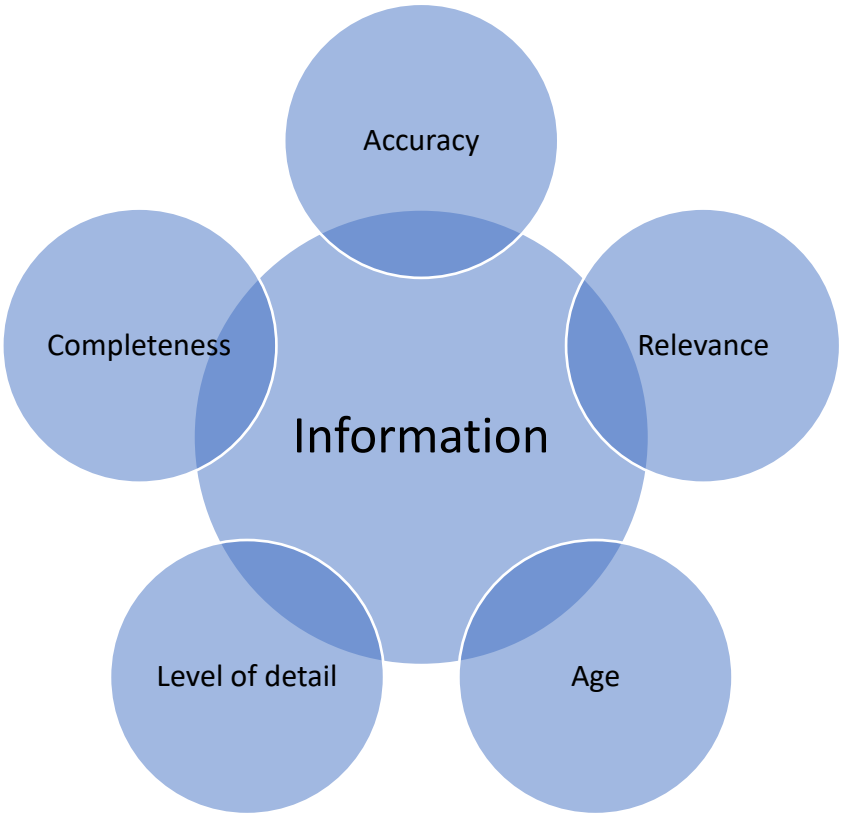
Quality of Information

Measuring the quality of information is sometimes based on the value which the user attaches to the information collected.

Poor quality data can lead to serious consequences. Poor data may give a distorted view of business dealings, which can then lead to a business making poor decisions. Customers can be put off dealing with businesses that give poor service due to inaccurate data, causing the business to get a poor reputation.

With poor quality data it can be difficult for companies to have accurate knowledge of their current performance and sales trends, which makes it hard for them to identify worthwhile future opportunities.

Factors that affect quality of information



Accuracy

As far as possible, information should be free from errors and mistakes. The accuracy of information often depends on the accuracy of the collected data before it is processed. If the original data is inaccurate then the resulting information will also be inaccurate.

Examples

- Consider a simple stock check. If it is carried out manually, a quantity of 62 could easily be copied down as 26 if the digits were accidentally transposed. This information is now inaccurate.
- If people are collecting the data manually e.g. recording answers to questionnaires, writing down instrument measurements, they might make a mistake.
- If data is being collected automatically by sensors or other instruments then the results could be inaccurate if the instruments were not correctly calibrated at the start of the data collection period.

Relevance

Relevance is an important factor because there has to be a good reason why that particular set of data is being collected. Data captured should be relevant to the purposes for which it is to be used. It must meet the requirements of the user.

There are a number of ways in which the data may or may not be relevant to the user's needs. It could be too detailed or concentrate too much on one aspect.

On the other hand, it might be too general, covering more aspects of the task than is necessary.

Examples

- It may relate to geographical areas that are not really part of the study. Where the study is meant to be about pollution in a local area, for example, data from other parts of the country would not be relevant.
- In an academic study, it is important to select academic sources. Business sources or sources which appear to have a vested interest should be ignored.
- You might be feeling unwell and want to make an appointment to see your doctor. You phone up the local surgery to find out when the doctor has a spare appointment time. It wouldn't be very useful or relevant to you if the receptionist told you how many appointment times were available to see the nurse.
- Consider a school situation. You need to study a tremendous amount of information to prepare for your exams. How would you feel if your teachers chose to spend several lessons talking about aspects of the subject that they found really interesting.

Age

How old information is can affect its quality. As well as being accurate and relevant, information needs to be up-to-date. Most information tends to change over time and inaccurate results can arise from information which has not been updated regularly.

The age of information is important, because information that is not up-to-date can lead to people making the wrong decisions. In turn, that costs organizations time, money, and therefore, profits.

Examples

- Choosing a holiday can be great fun. And you would probably go to the travel agent to get the latest brochures. Why? Well, because if you used last year's brochures the holiday may not even be available any more. And certainly the prices would be different. So you need up-to-date information.
- Your parents are still thinking of selling their house. They need an idea of how much their house is worth right now. It would be no use to them being told by the estate agent how much their house was worth five years ago.

Level of detail

For information to be useful, it needs to have the right amount of detail. Sometimes, it is possible for the information to have too much detail, making it difficult to extract the information you really want.

Information should be in a form that is short enough to allow for its examination and use. Information should be in a form that is short enough to allow for its examination and use.

Examples

- For example, it is usual to summarize statistical data and produce this information either in the form of a table or using a chart. Most people would consider a chart to be more concise than data in tables, as there is little or no unnecessary information in a chart.
- Suppose a car company director wants to see a summary of the sales figures of all car models for the last year; the information with the correct level of detail would be a graph showing the overall figures for each month. If the director was given figures showing the sales of each model every day of the previous 12 months in the form of a large report, this would be seen as the wrong level of detail because it is not a summary.
- A person orders a pizza. They ask for a large pepperoni to be delivered. They forgot to say what type of base they wanted and where it should be delivered to. The pizza company does not have enough information to fulfil the order.

Completeness

In order for information to be of high quality it needs to be complete. To be complete, information must deal with all the relevant parts of a problem. If it is incomplete, there will be gaps in the information and it will be very difficult to use to solve a particular problem.

Examples

- Consider the car company director mentioned above who wants to see a summary of the sales figures for the last year. If the director was given figures showing the sales for the first six months, this would be incomplete. If the director was shown the figures for only the best-selling models, this would be incomplete.
- Imagine you are feeling ill and you need to make an appointment to see your doctor. How useful would it be if the receptionist just told you that you could have an

appointment at quarter past two? Does she mean today, tomorrow or next week? The information in incomplete.

- You want to plan the family picnic for tomorrow. However, when you check the weather forecast you are only told what the weather in the morning will be like. There is nothing about the afternoon. You can't really make a decision just based upon what the morning weather is likely to be.

Encryption

The need for encryption

When data is transmitted over any public network (wired or wireless), there is always a risk of it being intercepted by, for example, a hacker. Using encryption helps to minimize this risk.

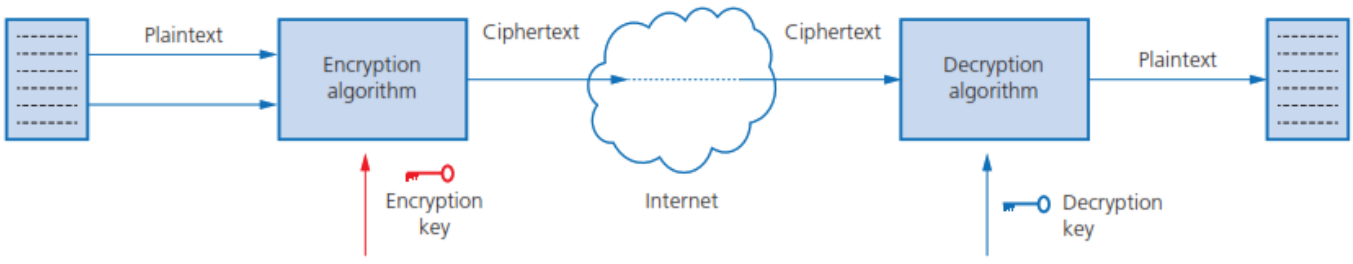
Once it is intercepted the information can be changed or used for purposes such as identity theft, cyber-fraud. If, however, the information is intercepted but it is unreadable or cannot be understood, it becomes useless to the hacker or interceptor.

Encryption is a way of scrambling data so that only authorized people can understand the information. It is the process of converting information into a code which is impossible to understand. This process is used whether the data is being transmitted across the internet or is just being stored. It does not prevent cyber criminals intercepting sensitive information, but it does prevent them from understanding it. This is particularly important if the data is sensitive or confidential for example, credit card/bank details, medical history or legal documents.

Methods of encryption

Encryption is the name given to converting data into a code by scrambling it, with the resulting symbols appearing to be all jumbled up. The algorithms which are used to convert the data are so complex that even the most dedicated hacker would be extremely unlikely to discover the meaning of the data.

Encrypted data is often called **ciphertext**, whereas data before it is encrypted is called **plaintext**.



The way that encryption works is that the computer sending the message uses an encryption key to encode the data. The receiving computer has a corresponding decryption key that can translate it back again.

A key is just a collection of bits, often randomly generated by a computer. The greater the length of the key, the more effective the encryption.

Modern encryption uses 256- bit keys which makes this form of encryption virtually impossible to crack. The key is used in conjunction with an algorithm to create the ciphertext.

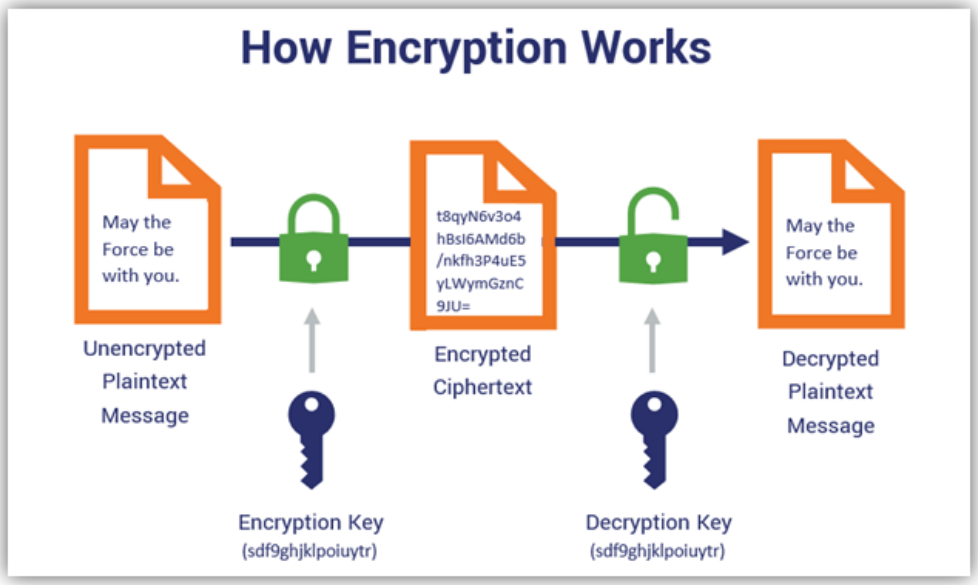
There are two main types of encryption. One is called **Symmetric Encryption** and the other is **Asymmetric Encryption**, which is also referred to as **public-key encryption**.

Symmetric encryption

Symmetric encryption is a type of encryption that uses the same key to encrypt and decrypt data. Both the sender and the recipient have identical copies of the key, which they keep secret and don't share with anyone. This differs from **asymmetric encryption**, which uses two keys, a public key (that anyone can access) to encrypt information and a private key to decrypt information.

How symmetric encryption works

- The sender uses an encryption key (usually a string of letters and numbers) to encrypt their message.
- The encrypted message, called ciphertext, looks like scrambled letters and can't be read by anyone along the way.
- The recipient uses a decryption key to transform the ciphertext back into readable text.



In the example above, we used the same key for encryption and decryption, which means this is symmetric encryption.

Only these two parties (sender and recipient) can read and access the data. This is why it's also sometimes called **secret key encryption**, **secret key cryptography**, **private key cryptography**, **symmetric cryptography** and **symmetric key encryption**.

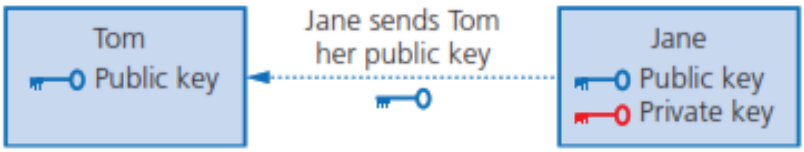
Asymmetric encryption

Asymmetric encryption is also known as public-key cryptography. Asymmetric encryption overcomes the problem of **symmetric encryption** keys being intercepted by using a pair of keys. This will include a public key which is available to anybody wanting to send data, and a private key that is known only to the recipient. The key is the algorithm required to encrypt and decrypt the data.

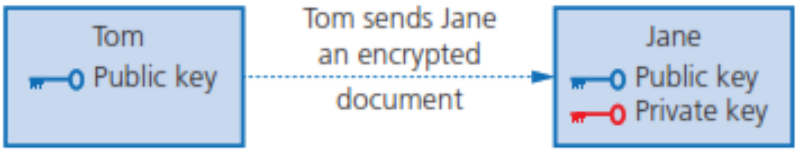
Using an example to explain how this works; suppose Tom and Jane work for the same company and Tom wishes to send a confidential document to Jane:

1. Jane uses an algorithm to generate a matching pair of keys (private and public) that they must keep stored on their computers; the matching pairs of keys are mathematically linked but can't be derived from each other.

2. Jane now sends her public key to Tom.



3. Tom now uses Jane's public key to encrypt the document he wishes to send to her. He then sends his encrypted document (ciphertext) back to Jane.



4. Jane uses her matching private key to unlock Tom's document and decrypt it; this works because the public key used to encrypt the document and the private key used to decrypt it are a matching pair generated on Jane's computer. (Jane can't use the public key to decrypt the message.)

Encryption protocols

An encryption protocol is the set of rules setting out how the algorithms should be used to secure information. There are several encryption protocols.

IPsec (internet protocol security)

is one such protocol suite which allows the authentication of computers and encryption of packets of data in order to provide secure encrypted communication between two computers over an internet protocol (IP) network. It is often used in VPNs (virtual private networks).

SSH (secure shell)

is another encryption protocol used to enable remote logging on to a computer network, securely. SSH is often used to login and perform operations on remote computers, but it can also be used for transferring data from one computer to another.

Transport Layer Security (TLS) and Secure Socket Layer (SSL)

The most popular protocol used when accessing web pages securely is transport layer security (TLS). TLS is an improved version of the secure sockets layer (SSL) protocol and has now, more or less, taken over from it.

The three main purposes of SSL/TLS are to:

- Enable encryption in order to protect data
- Make sure that the people/companies exchanging data are who they say they are (authentication)
- Ensure the integrity of the data to make sure it has not been corrupted or altered.

Many websites use SSL/TLS when encrypting data while it is being sent to and from them. This keeps attackers from accessing that data while it is being transferred. SSL/TLS should be used when storing or sending sensitive data over the internet. The SSL/TLS protocol enables the creation of a secure connection between a web server and a browser. Data that is being transferred to the web server is protected from eavesdroppers.

Uses of encryption

There are many reasons to encrypt data:

- Companies often store confidential data about their employees, which could include medical records, payroll data, as well as personal data.
- An employee in a shared office may not want others to have access to their work which may be stored on a hard disk, so it needs to be encrypted.
- A company's head office may wish to share sensitive business plans with other offices using the internet. If the data is encrypted, they do not have to worry about what would happen if it were intercepted.
- When individuals are emailing each other with information they would want to remain confidential. They need to prevent anybody else from reading and understanding their mail.
- People use websites for online shopping and online banking. When doing so, the debit/credit card and other bank account details should be encrypted to prevent fraudulent activity taking place.

Applications of encryption

Hard disk encryption

Hard-drive encryption is a technology that encrypts the data stored on a hard drive using sophisticated mathematical functions. Data on an encrypted hard drive cannot be read by anyone who does not have access to the appropriate key or password. This can help prevent access to data by unauthorized persons and provides a layer of security against hackers and other online threats.

When a file is written to the disk, it is automatically encrypted by specialised software. When a file is read from the disk, the software automatically decrypts it while leaving all other data on the disk encrypted. The encryption and decryption processes are understood by the most frequently used application software such as spreadsheets, databases and word processors.

The whole disk is encrypted, including data files, the OS and any other software on the disk. Full disk encryption is your protection should the disk be stolen, or just left unattended. So, even if the disk is still in the original computer, or removed and put into another computer, the disk remains encrypted and only the keyholder can make use of its contents.

Email encryption

Email encryption involves encrypting, or disguising, the content of email messages in order to protect potentially sensitive information from being read by anyone other than intended recipients. Email encryption often includes authentication.

Email is a vulnerable medium, particularly when emails are sent over unsecured, or public, Wi-Fi networks. Even emails sent within a secure company network can be intercepted by other users, including your login credentials. Encryption renders the content of your emails unreadable as they travel from origin to destination, so even if someone intercepts your messages, they can't interpret the content.

Email encryption: what to encrypt?

1. The connection from your email provider.
2. Your actual email messages.
3. Your stored, cached, or archived email messages.

Encryption in HTTPS websites

Normal web pages that are not encrypted are fetched and transmitted using Hypertext Transfer Protocol (HTTP). Anybody who intercepts web pages or data being sent over HTTP would be able to read the contents of the web page or the data. This is particularly a problem when sending sensitive data, such as credit card information or usernames and password.

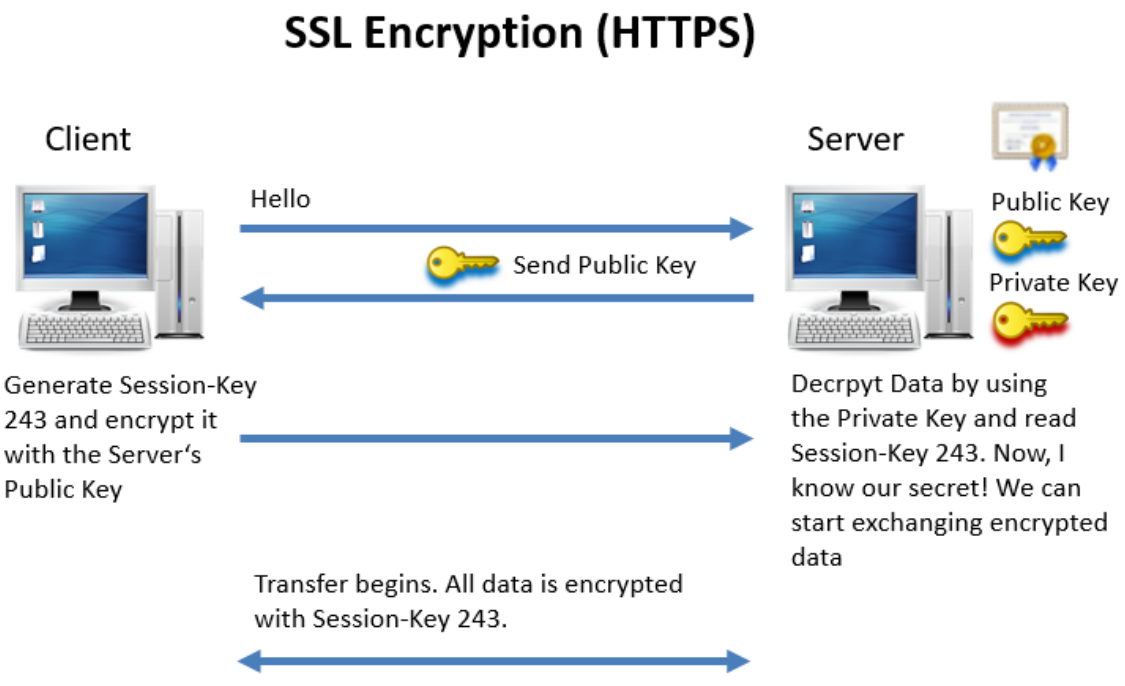
HTTPS, or Hypertext Transfer Protocol Secure, is the secure version of HTTP, which is the only primary protocol browsers use to connect to web servers and display web pages to users. HTTPS uses asymmetric encryption to secure the data in transport between the web server and client.

HTTPS is more favorable where privacy is more relevant. These can be situations where we are making online transactions, logging into our bank, or other tasks that would include the usage of sensitive documents. A green padlock, or simply a padlock, is shown, which signifies the usage of HTTPS.

HTTPS uses **Secure Socket Layer (SSL)** or **Transport Layer Security (TLS)** to encrypt and decrypt pages and information sent and received by web users.

How HTTPS works

1. You start your web browser and request a secure page by using the https:// prefix on the URL.
2. Your web browser contacts the web server on the HTTPS port and requests a secure connection.
3. The server responds with a copy of its SSL certificate.
4. Your web browser uses the certificate to verify the identity of the remote server and extract the remote server's public key.



5. Your web browser creates a session key, encrypts it with the server's public key and sends the encrypted key to the server.
6. The server uses its private key to decrypt the session key.
7. The client and server use the session key to encrypt all further communications.

Checking the Accuracy of Data

Data validation

Validation is one way of trying to reduce the number of errors in the data being entered into your system. Validation is performed by the computer at the point when you enter data. It is the process of checking the data against the set of validation rules.

Validation aims to make sure that data is sensible, reasonable, complete and within acceptable boundaries.

Data validation can be performed by using a number of validation checks.

Range Check

A range check is commonly used when you are working with data which consists of numbers, currency or dates/times.

A range check allows you to set suitable boundaries:

Boundary	Description	Validation
Upper limit	The maximum price of any item in a shop is £100	Less than **OR** equal to **100**
Lower limit	In a shop, you cannot sell a negative number of items, however you can sell no items	Greater than **OR** equal to **0**
A range	to achieve a B grade you must score between 75% - 84%	Greater than or equal to **75** **AND** Less than or equal to **84**

Type Check

When you begin to set up your new system you will choose the most appropriate data type for each field.

A type check will ensure that the correct *type* of data is entered into that field. For example, in a clothes shop, dress sizes may range from 8 to 18. A number data type would be a suitable choice for this data. By setting the data type as number, only numbers could be entered e.g. 10, 12, 14 and you would prevent anyone trying to enter text such as 'ten' or 'ten and a half'.

Check Digit

This is used when you want to be sure that a range of numbers has been entered correctly. There are many different schemes (algorithms) for creating check digits.

For example, the ISBN-10 numbering system for books makes use of 'Modulo-11' division. In modulo division, the answer is the remainder of the division. For example

8 Mod 3 = 2 i.e. the remainder of dividing 8 by 3 is 2.

Consider the ISBN number:

ISBN 1 84146 201 2

The check digit is the final number in the sequence, so in this example it is the final '2'.

The computer will perform a complex calculation on all of the numbers and then compare the answer to the check digit. If both match, it means the data was entered correctly.

Length Check

Sometimes you may have a set of data which always has the same number of characters.

For example, a UK landline telephone number has 11 characters.

A length check could be set up to ensure that exactly 11 numbers are entered into the field. This type of validation cannot check that the 11 numbers are correct but it can ensure that 10 or 12 numbers aren't entered.

A length check can also be set up to allow characters to be entered within a certain range.

For example, postcodes can be in the form of:

CV45 2RE (7 without a space or 8 with a space) or

B9 3TF (5 without a space or 6 with a space).

So you could set a length check for postcode to accept data which has a minimum number of 5 characters and a maximum number of 8.

Lookup Check

Where you have a field which only allows a limited list of items to be entered then a lookup list can help to reduce errors.

For instance, the image opposite shows a 'look-up' list being used in a spreadsheet that only allows 8, 10, 11 or 12 to be entered.

For example:

- a shop might put the dress sizes into a lookup list

- a car showroom might put the car models into a lookup list

- a vet might list the most popular types of animals that they deal with

Picture/Format Check

You may see this validation technique referred to as either a picture or a format check, they are the same thing.

Some types of data will always consist of the same pattern.

Example 1

Think about a postcode. The majority of postcodes look something like this:

CV36 7TP

WR14 5WB

Replace either of those examples with L for any letter which appears and N for any number that appears and you will end up with:

LLNN NLL

This means that you can set up a picture/format check for something like a postcode field to ensure that a letter isn't entered where a number should be or a number in place of a letter.

Example 2

A National Insurance number must be in the form of XX 99 99 99 X. The first two and the last characters must be letters. The other six characters are numbers. Any format entered differently to this will be rejected.

Presence Check

There might be an important piece of data that you want to make sure is always stored.

For example, a school will always want to know an emergency contact number, a video rental store might always want to know a customer's address.

A presence check makes sure that a critical field cannot be left blank, it must be filled in. If someone tries to leave the field blank then an error message will appear and you won't be able to progress to another record or save any other data which you have entered.

Consistency check

A consistency check is a type of logical check that confirms the data has been entered in a logically consistent way. It checks that data across two fields is consistent.

An example is checking if the delivery date is after the shipping date for a parcel.

When entering the gender of 'M' or 'F', a consistency check will prevent 'F' from being entered if the title is 'Mr' and will prevent 'M' from being entered if the title is 'Mrs' or 'Miss'.

When entering data about dispatching products, it would not be possible to mark an item as being dispatched until after it has been packaged.

Limit check

A limit check is similar to a range check, but the check is only applied to one boundary.

For example, in the UK you are only allowed to drive from the age of 17, but there is no upper limit. If somebody enters a number lower than 17 when asked to enter their age when applying for a driving license, for example, this will generate an error message.

Data verification

Verification means to check that the data on the original source document is identical to the data that you have entered into the system. Verification can be performed in two ways; double entry method, visual check.

Double entry

Think about when you choose a new password, you often have to type it in twice. This lets the computer check if you have typed it exactly the same both times and not made a mistake. It verifies that the first version is correct by matching it against the second version.
Whilst this can help to identify many mistakes, it is not ideal for large amounts of data.

- It could take a person a lot of time to enter the data twice.
- They could enter the same mistake twice and so it wouldn't get picked up.
- You would end up with two copies of the data.

Visual check

This saves having to enter the data twice. It can help pick up errors where data has been entered incorrectly or transposed.

However, it isn't always that easy to keep moving your eyes back and forth between a monitor and a paper copy. Also, if you are tired or your eyes feel 'blurry' then you might miss errors. An alternative method is to print out the data entered and compare the printout side by side with the source document.

- Visual checking can be rather time-consuming and possibly costly as a result.
- Another problem is that the person who is checking that the data has been entered correctly may be the same person who entered it. It is very easy for them to overlook their own mistakes. A possible way around this is to get somebody else to do the check.

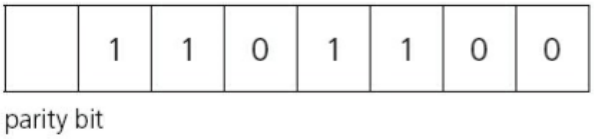
Parity check

A parity bit is a check bit, which is added to a block of data for error detection purposes. It is used to validate the integrity of the data. The value of the parity bit is assigned either 0 or 1 that makes the number of 1s in the message block either even or odd depending upon the type of parity. Parity check is suitable for single bit error detection only.

The two types of parity checking are:

- **Even Parity** − Here the total number of bits in the message is made even.
- **Odd Parity** − Here the total number of bits in the message is made odd.

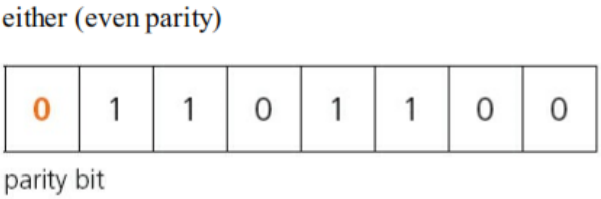
Consider the following byte of data;



If this byte is using even parity, then the parity bit needs to be 0 since there is already an even number of 1-bits (in this case, 4).

If odd parity is being used, then the parity bit needs to be 1 to make the number of 1-bits odd.

Therefore, the byte just before transmission would be;



or (odd parity)

1	1	1	0	1	1	0	0

parity bit

If a byte has been transmitted from 'A' to 'B', and even parity is used, an error would be flagged if the byte now had an odd number of 1-bits at the receiver's end.

sender's byte:

0	1	0	1	1	1	0	0

receiver's byte:

0	1	0	0	1	1	0	0

In this case, the receiver's byte has three 1-bits, which means it now has odd parity whilst the byte from the sender had even parity (four 1-bits). This clearly means an error has occurred during the transmission of the data.

Parity bits only check to see if an error occurred during data transmission. They do not correct the error. If an error occurs, then the data must be sent again.

Parity checks can find an error when a single bit is transmitted incorrectly, but there are occasions when a parity check would not find an error if more than one bit is transmitted incorrectly.

Checksum

A checksum is a value used to verify the integrity of a file or a data transfer. In other words, it is a sum that checks the validity of data. Checksums are typically used to compare two sets of data to make sure they are the same. Some common applications include verifying a disk image or checking the integrity of a downloaded file. If the checksums don't match those of the original files, the data may have been altered or corrupted.

A checksum is also sometimes called a *hash sum* and less often a *hash value*, *hash code*, or simply a *hash*.

A checksum can be calculated in many different ways, using different algorithms, for example a simple checksum could simply be the number of bytes in a file. Just as we saw with the problem with transposition of bits deceiving a parity check, this type of checksum would not be able to notice if two or more bytes were swapped; the data would be different, but the checksum would be the same.

The common protocols used to determine checksum numbers are the transmission control protocol (TCP) and the user diagram protocol (UDP). While checksum values that do not match can signal something went wrong during transmission, a few factors can cause this to happen, such as;

- An interruption in the internet or network connection.
- Storage or space issues including problems with the hard drive.
- A corrupted disk or corrupted file.
- A third party interfering with the transfer of data.

Different algorithms can be used to generate the checksum. Popular algorithms include SHA-256, SHA-1 and MD5.

Hash total

A method for ensuring that data in a file have not been altered. A hash total is the numerical sum of one or more fields in the file, including data not normally used in calculations, such as account number. When necessary, the hash total is recalculated and compared with the original. If data are lost or changed, a mismatch occurs which signals an error.

Let's consider a simple example. Sometimes, school examinations staff are asked to do a statistical analysis of exam results. Here we have a small extract from the data that might have been collected.

Student ID	Number of exam passes
4762	6
0153	8
2539	7
4651	3

Normally, the Student ID would be stored as an alphanumeric type, so for the purpose of a hash check, it would be converted to a number. The hash check involves adding all the Student IDs together. In this example it would perform the calculation 4762 + 153 + 2539 + 4651 giving us a hash total of 12105.

The data would be transmitted along with the hash total and then the hash total would be recalculated and compared with the original to make sure it was the same and that the data had been transmitted correctly.

Control total

A control total is calculated in exactly the same way as a hash total, but is only carried out on numeric fields. There is no need to convert alphanumeric data to numeric. The value produced is a meaningful one which has a use.

In the example above, we can see that it would be useful for the head teacher to know what the average pass rate was each year. The control total can be used to calculate this average by dividing it by the number of students. The calculation is 6 + 8 + 7 + 3 giving us a control total of 24. If that is divided by 4, the number of students, we find that the average number of passes per student is 6.

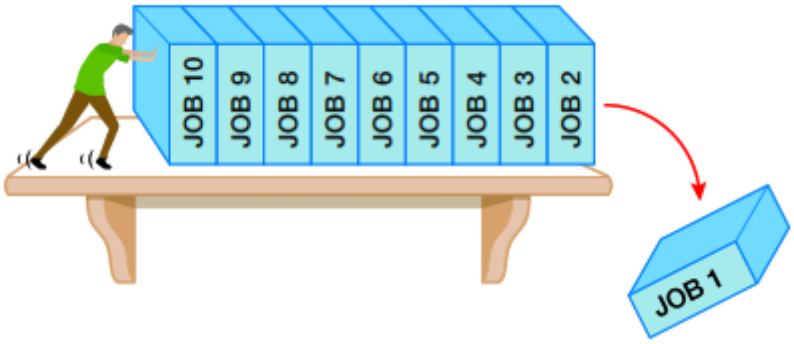
Data Processing

Data processing occurs when data is collected and translated into usable information. Data can include personal data, transaction data, sensor data and much more. Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

There are several different methods of data processing, but the three most popular ones are batch, online, and real-time.

Batch processing

Batch Processing system is an efficient way of processing large volumes of data. where a group of transactions is collected over a period of time. Data is collected, entered, processed and then the batch results are produced. The main function of a batch processing system is to automatically keep executing the jobs in a batch. These batches of data are stored until a set time when they will be processed and an output produced.



Master and transaction files

In this type of system, the important data that the computer stores all of the time is kept in a file called the master file. The data in the **master file** is **sorted** into order using one of the fields in the records in the file, known as the **primary key field**. The primary key field must uniquely identify each record in the file.

Each piece of input data (which will update the contents of the master file) is known as a **transaction**. All the input data is put together into a batch in a file called the **transaction file**. There are three different types of transaction that any processing system will have to cope with. They are;

- Add a new record to the master file
- Delete a record from the master file
- Amend / update an existing record in the master file.

At some predetermined time (e.g. the end of the day or week) the computer system will **process** the data stored in the transaction file and make any changes that are necessary to the master file as a result of the transactions. This will produce an **updated master file** and an **error report** detailing any transactions that could not be processed for some reason. Generating a single updated master file from the old master file and the transaction file is known as **merging** files.

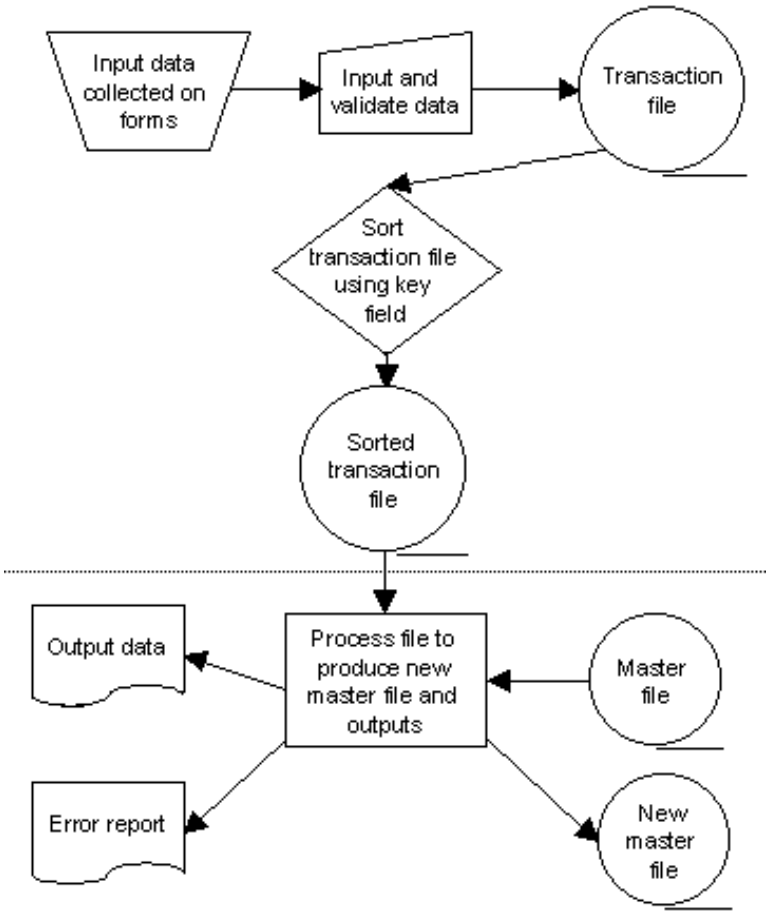
Before this processing can take place, the transactions must be prepared for processing. To do these two operations must be carried out;

1. **Validating**: Each transaction that is entered is **validated** to check that it is unlikely to halt processing by causing an error. It is very important that processing is not halted in a batch processing system because processing often takes place overnight without a human operator being present. If an invalid transaction caused the processing to stop then a whole night's processing time could be lost. Transactions are usually validated as they are keyed in by the operator. Any errors that are found can be reported and corrected straight away.
2. **Sorting**: The data in the transaction file is sorted into order using the same primary key field as the master file. Sorting can not take place until the whole batch of transactions entered and stored in a file. The transaction and master files need to be put into the same order before processing occurs because the two files are processed sequentially. The files are usually stored on a **serial access** medium such as **magnetic tapes**. If the records on the tapes were not sorted into order then whenever a transaction record was read from the transaction file the computer system would have to waste a lot of time searching for the matching record in the master file.

Advantages and disadvantages of batch processing.

Advantages	Disadvantages
It is a single, automated process requiring little human participation which can reduce cost.	There is a delay as data is not processed until the specific time period.
It can be scheduled to occur when there is little demand for computer resources, for example, at night.	Only data of the same type can be processed since an identical, automated process is being applied to all the data.
As it is an automated process there will be none of the transcription and update errors that human operators would produce.	Errors cannot be corrected until the batch process is complete.
There are fewer repetitive tasks for the human operators.	

This system flowchart shows what happens in a batch processing system.




```
1  First record in the transaction file is read
2  First record in the old master file is read
3  REPEAT
4     IDs are compared
5     IF IDs do not match, old master file record is
      written to new master file
6     IF IDs match transaction is carried out
7        IF transaction is D, old master file record
         is not written to new master file
8        IF transaction is C, data in transaction file
         is written to new master file
9     IF IDs match, next record from transaction file
      is read
10    Next record from master file is read
11 UNTIL end of old master file
12 Data in transaction file record is written to new
   master file
13 Any remaining records of the transaction file are
   written to the master file
```

Use of batch processing in payroll

Payroll systems are often implemented using **batch processing**.

- The master file would contain the records for all of a company's employees, including their employee number, rates of pay and how much they have been paid so far this year. The records in the master file would be sorted using the employee number as the primary key field.
- The input data put into the transaction file would consist of records showing how many hours each employee had worked in the current week. Sometimes transactions would be used to add a new record if a new employee started or delete a record when an employee left the company.

Here is an example record for an employee called Anna Jones in a master file together with an entry in the transaction file which will update her record with the number of hours she has worked this week.

Record In Master File

Employee No.	12A004
Surname	Jones
Forenames	Anna Jane
Pay Rate	£7.50
Department	Accounts
Pay This Year	£5575.00

Record In Transaction File

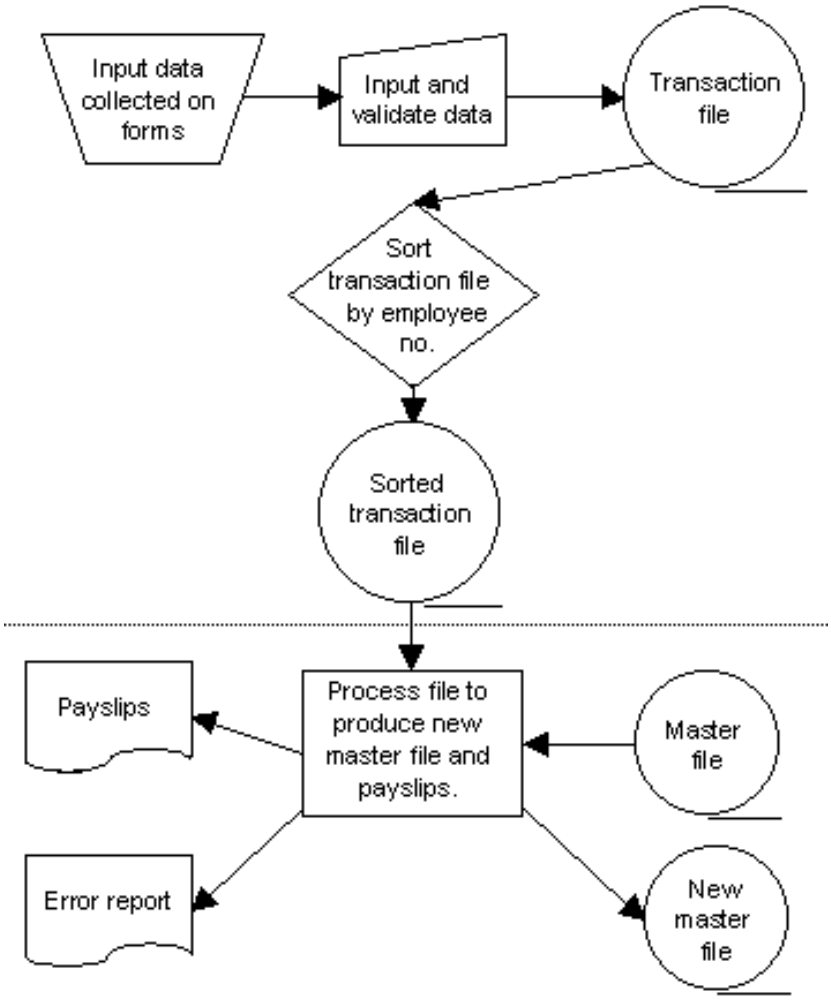
Employee No.	12A004
Transaction Type	Hours Worked
Value	25

After the transaction file was processed, Anna's record in the master file would show that her pay so far this year was Ł5762.50. This would include the Ł187.50 (25 hours at Ł7.50) she had just earned.

The transaction file would be processed at the end of each week as workers are paid weekly. Before processing it would have to be sorted into the same order as the master file, i.e. the records would be sorted into order by employee number.

The computer would then process the transactions, using the information about how many hours each employee has worked this week (from the transaction file) and their rates of pay (from the master file) to calculate the employee's wages for the week. Payslips can then be printed and the master file can be updated to increase the amount paid so far this year by the wages paid this week. An error report will also be produced.

This system flowchart shows the operation of the payroll system.



Online/transaction processing

An online processing system is a type of processing system that deals with data in transactions. A certain amount of data is input as a transaction. This amount of data is usually small. Once the data for the transaction is collected it is processed and the next transaction can occur. Transaction Processing is also known as **Interactive processing.**

For some applications the **master file** needs to be kept up to date all of the time. For example, in a travel agency whenever a seat is booked on a flight the number of seats that remain available on the flight must be reduced by one immediately. If this update was not done until the end of a day (as might happen in a **batch processing** system) then the flight could become overbooked with the same seat being booked more than once.

Transaction processing systems are used whenever the **master file must be kept up to date**. A transaction processing system operates like this;

- When a transaction is entered it is placed in a **queue** of transactions waiting to be carried out. The transactions are processed in the order that they are placed into the queue. If there are many people using the system at the same time then there could be lots of transactions being made from different computer terminals.

- The computer system will process one transaction at a time. Once the system starts processing a transaction it will not process any other transactions until the current transaction is finished. When a transaction is processed the master file is updated immediately. Therefore, the master file is always kept up to date.

Transaction processing systems need to use **direct access files**. In a direct access file any record in the file can be updated directly, without having to read through all of the records that come before it in the file. **Serial access media** such as **magnetic tape** cannot be used. The time required to find the record in the master file that a transaction related to would make processing incredibly slow.

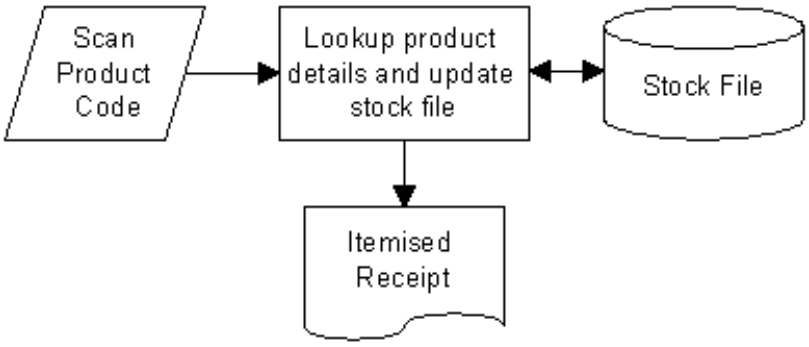
Transaction processing is **interactive**, i.e. processing takes place as a "conversation" between the user and the computer. The computer responds to the user's input by outputting some data before the user can input any more data. This means that the user's input can depend on the computer's previous output.

Use of online/transaction processing in POS automatic stock control system

A supermarket uses a Point of Sale (POS) terminal to keep track of the goods that it has in stock and to produce bills for customers. Whenever an item is sold the product code of the item is read into a till by a **barcode** scanner. Sometimes the barcode is not scanned properly. If this happens the product code must be rescanned or entered using a **keyboard**.

Once the product code has been entered the till looks up the price and name of the product in the supermarket's stock database. This information is printed on the customer's receipt. It also updates the product's record in the stock file to indicate that the product has been sold.

This system can be depicted by this system flowchart.



The system has to be a transaction processing system because a customer at a till would not be prepared to wait for a batch processing system to look up the price and name of the product.

The product file on the computer contains the record of each product that is sold. Each record consists of different fields containing data, for example;

- **Barcode number:** the number which identifies each different product; this is the key field because it is different for each product
- **Product details:** a description, such as tin of beans, packet of teabags and so on price of the product
- **Size:** weight or volume of the product
- **Number in stock:** the current total of that product in stock; this changes every time a product is sold or new stock arrives
- **Re-order level:** the number which the computer will use to see if more of that product needs re-ordering. If the number in stock falls to this level, the supermarket or store must re-order
- **Re-order quantity:** when the product needs re-ordering, this is the number of products which are automatically reordered
- **Supplier number:** the identification number of the supplier which will be used to look for the details on the supplier file.

The processing involved in automatic stock control is as follows:

```
1  The product's barcode is input from the barcode
   reader
2  The computer searches for this barcode number in
   the product file and finds it using direct access
3  The number in stock is reduced by one
4  The computer then compares the number in stock
   with the re-order level
5  If the number in stock is not equal to the
   re-order level then go back to step 1 and repeat
6  If the number in stock is equal to the re-order
   level then the computer creates an automatic order
7    It looks up the re-order quantity of that product
8    It looks up the supplier number of that product
9    It searches the supplier file for the record
     corresponding to the supplier number found in the
     product file
10   It sends the order automatically to the supplier
     using the supplier's contact details
11 Go back to step 1 and repeat
```

Real-time processing

For some tasks a computer must process information to meet some real-world time deadline. The time deadline must be met regardless of how much work the computer has to do. Usually real time computer systems are required to process information very quickly. Most real time systems appear to process data instantly.

A computer system which must **process data extremely quickly** to meet a **real-world deadline** is known as a **real time system.**

Use of real-time processing in a missile guidance system

For example, consider a **control system** that is used to fly a guided missile to hit a target. Once the missile is launched the control system must guide it to its target. The missile's guidance system identifies its position by examining the contours / features of the ground it is flying over. This input data is used to ensure that the missile follows the path it has been programmed with. Using the inputs, the system will decide how to move ailerons/fins to change direction and control the speed of the rocket motor.

Because a missile flies very fast the inputs must be processed very quickly to affect the outputs. If it took two seconds for the missile to change its direction to avoid an unexpected object then the missile would probably hit the object before the change of direction could take place. The data must be processed in thousandths of a second to avoid a disaster. There is a real-world time limit on how long the computer system can take to process data without having terrible consequences. Therefore, a missile control system has to be a real time system.

Most real time systems are **control systems**. Other examples include the cooling system in a nuclear reactor which prevents the reactor overheating or any system which must control a dangerous chemical reaction.