

TOPIC AREA 3: HOW DATA CAN BE ACCESSED AND MANAGED ACROSS PLATFORMS 3

3.1 Application Programming Interfaces (API)

3.2 User access controls

3.3 Permissions

TOPIC AREA 1: UNDERSTANDING DATA 1

1.1 Data, information and knowledge

1.2 Big Data

Topic on a Page

for Unit F200: Fundamentals of Data Analytics

*Cambridge Advanced National
(Certificate and Extended Certificate)*

zigzageducation.co.uk

**POD
12845**

Publish your own work... Write to a brief...
Register at **publishmenow.co.uk**

Follow us on Bluesky or X **@ZigZagComputing**

Contents

Product Support from ZigZag Education	ii
Terms and Conditions of Use	iii
Teacher's Introduction.....	iv
Topic Area 1: Understanding data	1
1.1 Data, information and knowledge	1
1.2 Big data.....	1
1.3 Data and file formats.....	2
1.4 Data types and classifications	2
Topic Area 2: Managing data.....	3
2.1 Data lifecycle management (DLM) and data analytics pipeline (DAP).....	3
2.2 Creation and capture: data assurance considerations and data gathering	4
2.3 Storage: data states, data stores, data storage, onsite storage, cloud storage	5
2.4 Data transformation: data wrangling, data maintenance.....	6
2.5 Usage and analysis: data analytics, types of data analytics	6
2.6 Usage and visualisation: presenting data, visualising data	6
2.7 Archival.....	7
2.8 Destruction.....	7
Topic Area 3: How data can be accessed and managed across platforms	8
3.1 Application Programming Interfaces (API)	8
3.2 User access controls	8
3.3 Permissions	8
Topic Area 4: Legal considerations	9
4.1 Legislation and the role of the ICO when using data	9
Topic Area 5: Job roles, skills and attributes in data analytics.....	10
5.1 Job roles related to data analytics	10
5.2 Personal attributes	10
5.3 Communication skills	10

All posters are provided in both A3 and A4 formats

Teacher's Introduction

This resource is intended for use by students studying the **OCR Level 3 AAQ Cambridge Advanced National in IT: Data Analytics, Unit F200 Fundamentals of Data Analytics**, first teaching 2025. This is a mandatory external unit for this qualification and is assessed by an exam.

The intention of this resource is to provide a condensed 'Topic on a Page' which provides an overview of the content of each topic area, which will enable students to review their learning and apply it to the supplied activity sheets.

Remember!

Always check the exam board website for new information, including changes to the specification and sample assessment material.

How to use this resource

The resource consists of:

- 10 A3 posters covering the topics as listed below, labelled: 1 — 10
- 10 A3 activity posters which are partially completed and provide opportunities for students to fill in gaps to show and answer exam-style questions. These are labelled: 1 — 10

Opportunities for use:

- Printed out and displayed on classroom walls
- Individual copies to be given to students as the topic area is delivered
- Activity sheets can be given out at the end of topic delivery to check understanding
- Can be given to students as revision aids

Topic Area 1: Understanding data

1

- 1.1 Data, information and knowledge
- 1.2 Big data

2

- 1.3 Data and file formats
- 1.4 Data types and classifications

Topic Area 2: Managing data

3

- 2.1 Data lifecycle management (DLM) and data analytics pipeline (DAP)

4

- 2.2 Creation and capture: data assurance considerations and data gathering

5

- 2.3 Storage: data states, data stores, data storage, onsite storage, cloud storage

6

- 2.4 Data transformation: data wrangling, data maintenance

7

- 2.5 Usage and analysis: data analytics, types of data analytics
- 2.6 Usage and visualisation: presenting data, visualising data
- 2.7 Archival
- 2.8 Destruction

Topic Area 3: How data can be accessed and managed across platforms

8

- 3.1 Application Programming Interfaces (API)
- 3.2 User access controls
- 3.3 Permissions

Topic Area 4: Legal considerations

9

- 4.1 Legislation and the role of the ICO when using data

Topic Area 5: Job roles, skills and attributes in data analytics

10

- 5.1 Job roles related to data analytics
- 5.2 Personal attributes
- 5.3 Communication skills

TOPIC AREA 1: UNDERSTANDING

1.1 Data, information and knowledge

What is data analytics?

Our daily lives are filled with data. Everything we interact or clickstream with online becomes a data source, from a social media comment, to uploading different types of files to the Internet, such as text, images and video. More and more, we are reliant upon technology and the IoT (Internet of Things), which facilitate our daily lives and have access to our data. This all adds to an ever-increasing explosion of data which is being stored.

Data analytics is about making sense of this data, collecting it, transforming, modelling and interpreting it into useful information, which in turn can be used to make better decisions, draw conclusions, create new products and guide decision-making.



Sources of Data

Data is drawn from data repositories, data warehouses, spreadsheets, databases.

Sources of Information

Data can be drawn from an organisation's internal databases, using queries to extract information.

Sources of Knowledge

Knowledge is gained from a result of analysing data and information as above.

Sources of Data

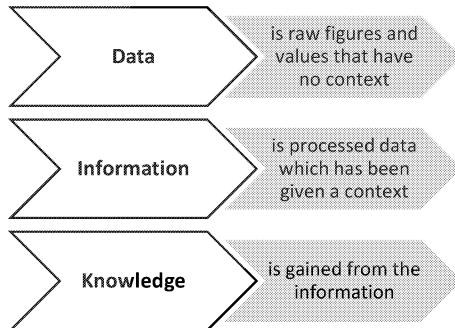
Business Data: Relevant to an organisation – financial, functional, client data

IoT Data: Generated in a real-time format, from industries – health, manufacturing, transportation, specialised software tools for

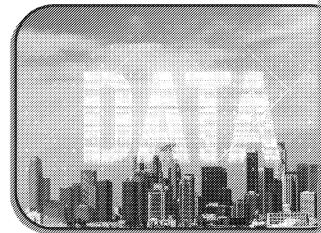
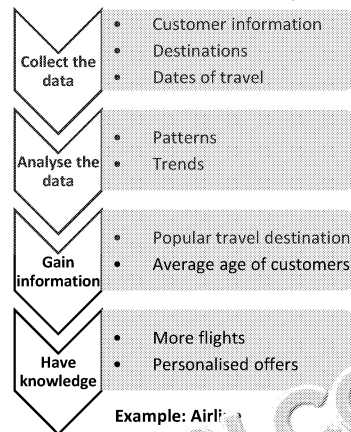
Public Data: Data collected at an academic and government organisation – statistics, census data.

Big Data: Data formed from huge sets, hard to organise and manage, pages, social media.

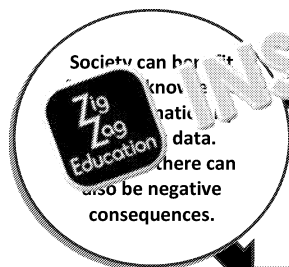
What data, information and knowledge are:



Interaction of data, information and knowledge



Data and information in society



Benefits and limitations of the use of data and information to organisations and individuals

Benefits: healthcare, transport, public services, competitive markets, consumer behaviour

Limitations: cost, data privacy, security, quality of data, lack of skilled professionals

How data and information can have negative consequences

Security issues, cybersecurity, hackers, loss of reputation, data protection issues. Misused in the workplace.

INSPECTION COPY

Educational

Utilities

Financial

Police

Transport

Healthcare

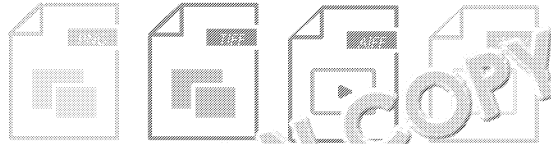
COPYRIGHT
PROTECTED



TOPIC AREA 1: UNDERSTANDING

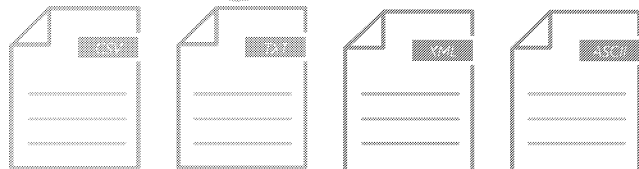
1.3 Data and file formats

To be used and recognised by the different software systems in use, data in the form of encoded and stored files must have a specific file format.



Benefits and limitations of data/file formats

Data	File	Limitations
Encoding data	ASCII	Limited character set, difficult to represent other language, low data security.
Transport of data	XML	Large files, slow, not good with binary data. Compatibility and memory issues.
Data interchange	JSON	Supports a limited set of data types, low data security.
Plain Text data	CSV TXT	No formatting. .txt uses TAB to separate each field. .csv uses a comma.
Numeric data	Integer/decimal/float	Float and decimal slower to process.
Audio	AIFF WAV	Large file formats, high quality.
Image	JPEG PNG TIFF	JPEG compressed, easy to transport, but because of compression, can reduce image quality and become pixelated. PNG does not support CMYK colour mode, and not printable. TIFF large file format, high quality image.



1.4 Data types

Data can be formatted in different ways. It also has different classes.

Benefits and limitations

Data type	Benefits
Boolean	Two values: true or false.
Character	Text, punctuation, fixed length.
Date	dd/mm/yyyy stored as text.
Integer	Whole numbers.
Real/float	Decimal numbers.
String	Mostly text. Holds characters.

Classifications of data

Qualitative
Quantitative
Structured
Unstructured

INSPECTION COPY

COPYRIGHT
PROTECTED



TOPIC AREA 2: MANAGING

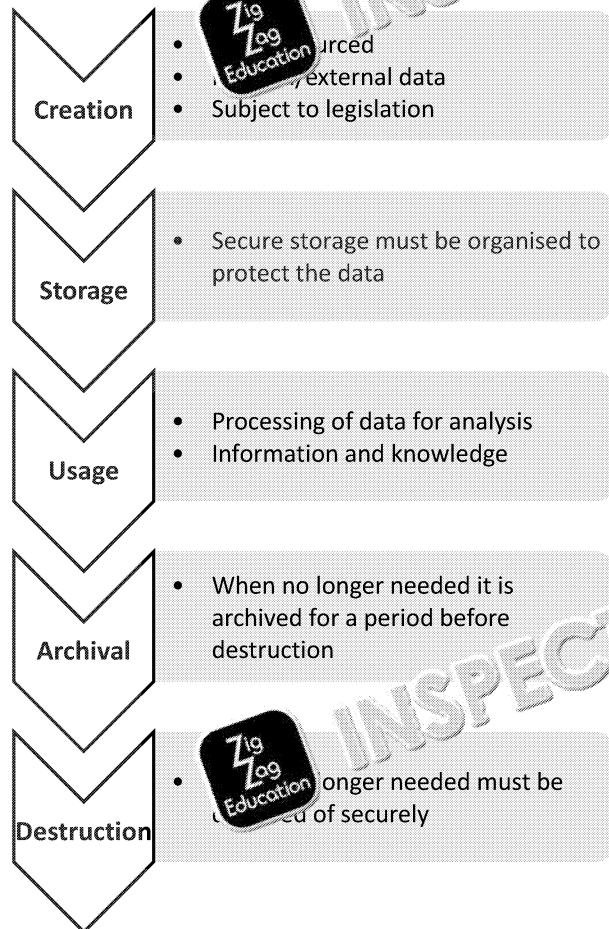
2.1 Data lifecycle management (DLM) and the data analytics pipeline

2.1.1 Data lifecycle management (DLM)

What is data lifecycle management (DLM)?

An approach to managing data with five phases.

The interactions and iterations of the five phases

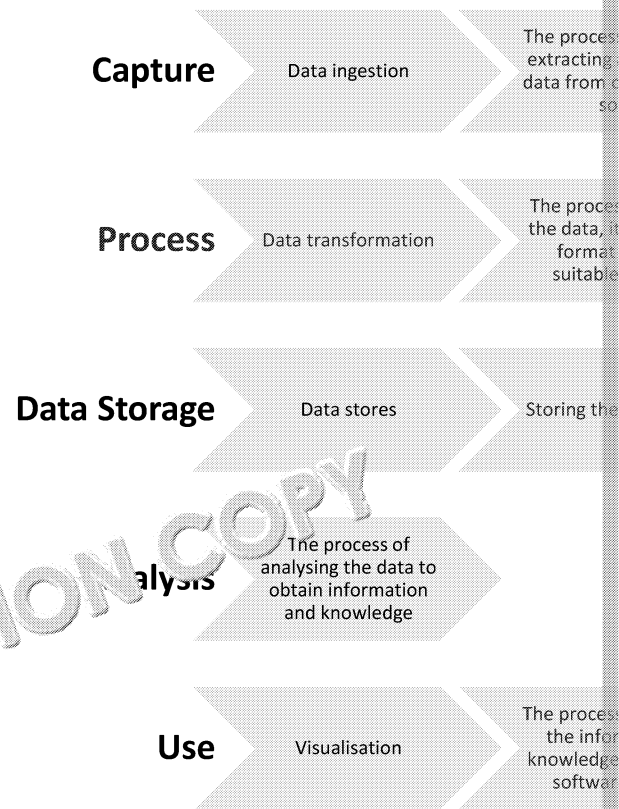


2.1.2 Data analytics pipeline

What is the data analytics pipeline?

An approach to data analysis with five phases.

The interactions and iterations of the five phases



INSPECTION COPY

COPYRIGHT
PROTECTED



TOPIC AREA 2: MANAGING

2.2 Creation and capture

2.2.1 Data assurance considerations

What is data assurance?

When considering data which is to be collected, how do we know that the data is accurate and reliable? There must be a system in place to ensure that data supplied has been quality assured.

The purpose and importance of data assurance considerations

Why do we have to have data assurance considerations? If data does not go through these assurance processes, then the data becomes unreliable.

GIGO Garbage in, Garbage out

Poor data = unreliable information = inaccurate knowledge



How each consideration affects the collection and use of data

Although there are data assurance measures to be considered, it is important that there is an accountability process to follow.

Data profiling:

Analysing structure

Data cleansing:

Identifying redundancy

Data entry:

Validation methods implemented

Data documentation:

Guidance provided about the data

Data audits:

To ensure reliability of the data assurance process



Confidence in data

If organisations cannot be sure that the data has gone through data assurance considerations, they will not have confidence in the data they have been provided with. This can lead to a bad reputation for the data provider.



2.2.2 Data gathering

What is data gathering?

Data gathering is the process of collecting data from various sources, as opposed to data that is already available, which is data quality, which is a measure of the accuracy and reliability of the data.

Methods

Documents and records

Focus groups

Interviews

Observations

Online tracking

Questionnaires and Surveys

Social media monitoring

Transactional tracking

Verbal histories

Factors in effective data gathering

It is important to consider the factors that affect data gathering. There are clear factors that are needed. The meaning of data needs to be clear if possible. It is important to consider the factors that affect data gathering and the form of the data.

INSPECTION COPY

COPYRIGHT
PROTECTED



TOPIC AREA 2: MANAGING

2.3 Storage

2.3.1 Data states

What is a data state?

When data is used by computing equipment, it can be subject to different states, depending upon the stage reached in the DLM. (See 2.1.1)

It has three states:

- Data in transit
- Data at rest
- Data in use

Data in transit

This is data which is being transferred over a network. It may also be in the process of transferring from one local device to another. It will be encrypted for security purposes.

Data at rest

This is data which is stored for later use. It may be stored in the cloud or on a physical backup. It will be encrypted for security purposes.

Data in use

This is data which is being used, in other words in the process of being analysed. Because it is being used, it does not need encryption.

2.3.2 Data stores

What is a data store?

The purpose of a data store is to store data.

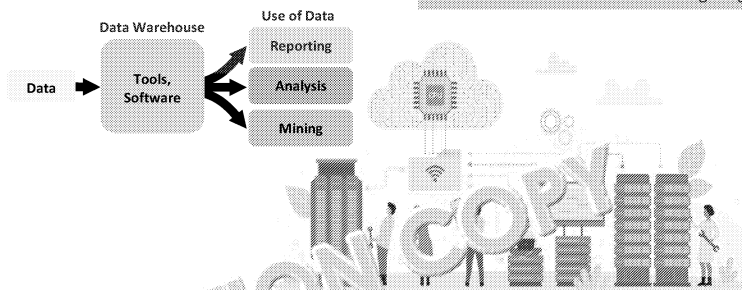
It allows users to access and use the data in a data store.

Data stores are primarily designed to allow users to access and store data. They are designed to store data in a way which enable the user to access and analyse the data.

An organisation may store its data on its own networked storage or in cloud storage.

Big data (see 1.2) – stores which contain vast amounts of data sets are:

- **Data warehouse** – a central repository of data ready to be analysed
- **Data lake** – a central repository of data in a raw format
- **Data mart** – contains subsets of data
- **Data silo** – contains data which has been isolated from other data sets



2.3.3 Data storage

There are three different ways that data can be stored within cloud storage – block, file and object storage.

When considering data storage types, the following must be considered.

- Is the data set large or small?
- Is it structured or unstructured?
- What does the organisation want to do?

Block Storage

Block storage is suitable for smaller data sets. Data is split into uniquely identified, equal-sized blocks which can be accessed quickly.

Limitations: It is not suitable for larger data sets as it becomes more expensive to maintain.

File Storage

Benefits: File storage is suitable for unstructured data, and is simply a hierarchical file and folder system. It has a single path – think of how you access a folder on File Explorer.

Limitations: It can become quite vast as it continues to scale outwards as more folders are added.

Object Storage

Benefits: Object storage is suitable for objects such as video, image and sound. They are given a unique identifier, being the metadata contained in the object. It also contains details such as the age and access rights of the object. It is a cheap system, and as it has a flat structure, large quantities of data can be stored.

Limitations: An object cannot be edited directly within the store.

2.3.4 Onsite storage

Organisations may choose to keep their data onsite. This is storage which is located on the premises. Each has its own security measures.

File server

In a file server organisation, all computers will generally be networked to a central server. It can be used for block and file storage.

Hard drive

A hard drive is contained within a computer or laptop to store data.

- **HDD – Hard Disk Drive** – Inside a computer, it is typically stored in a binary format.
- **SSD – Solid State Drive** – Inside a laptop, it is small. An SSD drive is expensive. Data is stored in NAND flash memory.

Network attached storage - NAS

This is very similar to a file server. The limitation, however, is that it is not as fast. They are typically file-based storage, rather than in block storage.

Portable storage device

A portable storage device can be an external hard drive, USB drive, etc. It is portable, which means it can be used anywhere and anytime.

The limitations are that it can be easily lost, stolen or damaged.

Storage area network - SAN

A storage area network is a network which can have multiple servers connected to a single fibre network and best suited to large organisations and data centres.

2.3.5 Cloud storage

Organisations can store their data in the cloud. All are provided with their own security measures.

Community

Multiple organisations share their data in a dedicated storage area.

Private

Organisations store their own data in a dedicated storage area.

Public

Accessible to anyone with the right credentials.

Hybrid

A mix of private and public storage.

INSPECTION COPY

COPYRIGHT
PROTECTED



TOPIC AREA 2: MANAGING

2.4 Data transformation

2.4.1 Data wrangling

What is data wrangling?

The purpose of data wrangling is to transform raw data into a usable format. This makes it easier to analyse and to make decisions with the data. It is important, as it makes sure that the data is reliable and accurate for use.

2.4.2 Data maintenance

What is data maintenance?

Another important process which ensures that the data is regularly checked to make sure that it continues to be reliable and accurate for use.

2.5 Usage and analytics

2.5.1 Data analytics

What is data analytics?

Data analytics is the processing of data.

The different processes are:

- Collecting data
- Analysing data
- Interpreting data

The overall purpose of data analysis is to find patterns and trends in the data, which will then enable an organisation to make informed decisions, solve problems and ultimately improve business functions.

2.5.2 Types of data analytics

In order to process and analyse data, there needs to be a specific purpose. There are different reasons, and different questions to raise about the data. Therefore, there are different types of data analytics processes which can be used.

All of the different types of data analytics have one purpose, to make informed decisions – which are of course benefits. The limitations of the systems are that interpretations could be biased, especially if there has been an incorrect interpretation.

Cognitive	This method uses machine learning algorithms and AI to take a look at the structure of data. The purpose is to extract further insights and make predictions. This type of data analytics can be used to predict such things as the effectiveness of treatments, potential market trends, anticipating customer behaviour and more. It can be used to identify trends, based on the data collected. e.g. Netflix gathers this type of data to determine what it is going to recommend to its users upon the historical data. If something is popular, it will offer more of it.
Descriptive	This method looks at what happened and why. It looks at descriptive analytics and digs deeper to find the root of the problem. It is used to test a hypothesis, such as if we do this, will this happen? It also looks at correlations between different variables which might explain and examine customer behaviour. Hello Fresh may find that there is a trend in fish dishes being ordered, and, therefore, it will supply more fish recipes.
Diagnostic	This method looks at an outcome and then considers what should or could happen next. It uses historical data and looks at trends to predict future outcomes. Organisations can use this to determine their future financial needs, staffing solutions and marketing trends.
Predictive	This method looks at past decisions and events to estimate the likelihood of different outcomes and to choose the best course of action, e.g. on TikTok the For You feed is based on what you have watched previously and assumes you want to watch more of the same.

2.6 Usage and visualisation

2.6.1 Presenting data

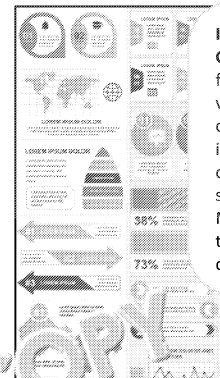
What is data presentation

Once the data has been analysed and information and knowledge has been gained, the next step is to present the information. When presenting data, it is important to have information laid out clearly and also to consider people with accessibility issues.

Information can be presented in the following formats:

- Graphical
- Tabular
- Textual

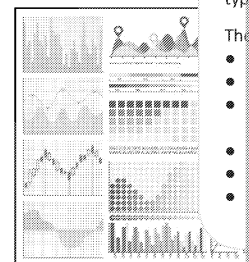
2.6.2 Visualising data



Infographics

Graphical, Textual, Tabular format. Infographics are a visual representation of data and information using images, symbols, charts and diagrams. They are used to simplify information. Minimal text is used and tables can be used for a clear layout.

There are many ways to present data visually. It is important to consider who the audience is and who the data is for to show the intended purpose.



Graphical, Textual, Tabular format.

These are the different types of data presentation.

INSPECTION COPY

COPYRIGHT
PROTECTED



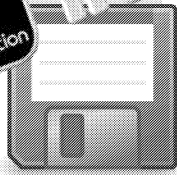
TOPIC AREA 2: MANAGING

2.7 Archival

Why do we archive data?

Once the data has been analysed and knowledge has been presented as information, it then needs to be archived. This means storing it securely, not deleting it, as it may be needed at a later time.

There are different ways to archive data: in cloud storage, networked storage or on-site storage. (See 2.3 Storage)



2.8 Destruction

Why do we have to destroy data?

When a decision is made that the data is no longer needed, it is important that data is securely destroyed. Under GDPR regulations, data must be kept for no longer than it is needed, or used for any other purpose than the one it was intended for, otherwise the regulation is being contravened and there can be legal consequences.

When data is destroyed, the procedure chosen to destroy it must ensure that it is not readable, and that it cannot be used.

The different methods used can contribute to e-waste and it is always best to make an environmentally conscious decision.

Data destruction methods

Degaussing

This method is used for destroying any magnetic storage device, such as a HDD (hard disc drive).

A degaussing machine is used, which uses a strong magnetic field to rearrange the magnetic structure of the hard drive.

✓ **Benefits** – This is a reliable form of data destruction and

✗ **Limitations** – The machine can only be used on one storage organisation. The demagnetising process renders the hard drive unusable. The magnetic force must be very strong, however.

Drive destruction

This method is used for destroying any type of hardware storage device, such as a HDD, SSD, USB, SD card.

It is physically destroyed by either burning, breaking or crushing.

✓ **Benefits** – This is a reliable form of data destruction, as it

✗ **Limitations** – It can be a costly and time-consuming process and has an environmental impact and then be subject to e-waste regulations.

Erasure/Overwriting

This method is used for any type of hardware storage device.

It involves completely erasing the storage device, or overwriting the old data with new data. There is also specific software which can be used to perform these actions.

✓ **Benefits** – This is a reliable form of data destruction, as it is a cost-effective method as many devices can be erased simultaneously.

✗ **Limitations** – It can be a time-consuming process. Software

Shredding

This method is used for destroying any type of hardware storage and basically involves cutting the hardware up into small pieces and then burning them.

✓ **Benefits** – This is a reliable form of data destruction, as it

✗ **Limitations** – This method also can contribute to environmental

INSPECTION COPY

COPYRIGHT
PROTECTED



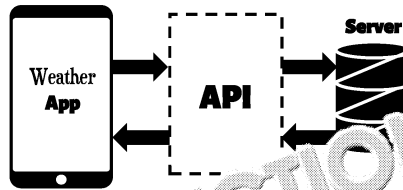
TOPIC AREA 3: HOW DATA CAN BE ACCESSED AND MANAGED

3.1 Application Programming Interfaces (API)

The role of an API

An API provides an interface between two applications. It allows them to communicate with one another.

For example, a weather app (Client) on your mobile phone will collect the data from the source data (Server) using an API.



API certifications

Open:	These are APIs with no authentication methods. The Wikimedia API enables anyone to create an app and link to the vast amount of knowledge it has. It supplies data to your app. They are commonly used with mobile phone apps.
Public:	These are paid for APIs. They will require authentication to connect an app. The Hunter API links an app to professional email addresses, which can be checked for authenticity.
Composite:	This API combines several APIs and can perform more than one job at once. When we purchase something from Amazon, APIs are used to add a product to basket, update the basket, checkout the basket and pay.
Internal/Private:	These are private APIs and restricted to private organisations for their use only. These are the most common type of API and are very quick to develop. These are generally used to improve organisational functions.
Partner:	These are APIs which can be free to or paid for by a limited audience. They are limited to invitation only and have restrictive authentication methods. They are, therefore, very secure. Organisations will use this type to monetise their products.

Benefits: Reduced costs, secure, scalability, greater productivity, data accuracy.

Disadvantages: Vulnerable to security breaches, complex to maintain, compatibility issues.

Types of API design

JSON-RPC	Stands for JavaScript Object Notation - Remote Procedure Call. Most platforms use this type of API to send and receive data. Request and response architecture build. It uses HTTPS request and response.
SOAP	Stands for Simple Object Access Protocol. It is an older system, and, therefore, slower. It uses a message envelope, and messages are usually transmitted via a web service. It is designed to move data around organisations. It is secure and is used in financial transactions.
REST	Stands for Representational State Transfer API. The most commonly used API and used with social media platforms and platforms such as Uber. It generally uses JSON-RPC as well as HTTP requests to interact with data.
XML-RPC	Stands for Extensible Markup Language - Remote Procedure Call and uses HTTP to transport an XML format to encode data. It is known for its information security. Generally used in content management, tasks and password management systems.

3.2 User access controls

What is user access control?

This is how access to computer systems can be controlled. It helps to keep data secure and prevent unauthorized access. This helps to keep data secure.

Attribute-based Access Control (ABAC)

This type of access is dependent upon a set of attributes such as user, their user ID, their age or their location. If the attributes match, access is granted. This is commonly used in financial institutions.

Discretionary Access Control (DAC)

This type of access allows persons who have control of access to grant or deny access to others. An example of this is 'sharing' a Google document. By sharing, you are giving access to others.

Mandatory Access Control (MAC)

This type of access is determined by an agreed policy, set by an administrator. It is generally used in health organisations.

Role-based Access Control (RBAC)

This type of access is allocated by the role of the person using the system. This is a reliable and secure system. Users can be given read, write, or delete access.

Rule-based Access Control (RBAC)

This type of access is determined by an agreed set of rules. It is used to control access to data in the event of a cybersecurity threat and if there is infrastructure overload. It is determined by type of organisation, the sensitivity of the data, and who has access.

3.3 Permissions

User Rights

Once a user has been granted access, they can be given different permissions. These are then granted to the user. What rights the user can be given read, write, edit, delete access depending on their level of access.

Administrative Access

Administrators can control access levels, privileges. They can grant access to everyone or no one.

There are different categories of user privilege:

User level: This is determined by the user rights allocated.

User group level:

- **Administrator:** Full access.
- **Standard user:** General access, controlled by administrator.
- **Guest:** Temporary access, some restrictions in place.

File and folder level:

Again, this is determined by access level rights and different permissions can be granted.

INSPECTION COPY

COPYRIGHT
PROTECTED



TOPIC AREA 4: LEGAL CONSID

4.1 Legislation and the role of the ICO when using data

When working with data, there are several areas of legislation to consider and to adhere to, particularly with personal, identifiable information.

Legislation

	What it is	Risks	Non-compliance
Computer Misuse Act 1990 (CMA)	<ul style="list-style-type: none"> This act provides legislation on crimes connected to the use of computers. Unauthorised access to computer material, including emails, text messages and mobile phones. Unauthorised access to financial data with the intent to commit fraud, introducing malware or viruses. Unauthorised acts causing or creating risk of serious damage. Example is cybercrime, which has a serious risk to human welfare, damage to the economy and national security. 	<ul style="list-style-type: none"> Staff are educated about the associated risks with access to data. Policies and procedures to ensure compliance; failure to comply could result in disciplinary procedures. IT administrators should implement appropriate user access levels and permissions, and put robust security measures in place. Operating systems to be kept up to date with security patches. 	Ranging from fines to prison time , from a few months right up to life imprisonment for the most serious offences under Section 3ZA.
Data Protection Act 2018 (DPA)	<p>This act controls how organisations use our personal information. There are seven principles that organisations must adhere to.</p> <ol style="list-style-type: none"> Lawfulness, fairness and transparency – When collecting information, it must be clear what it is being used for. Purpose limitation – It must be clear what the data is being collected for and it should only be used for that purpose. Data minimisation – You must only collect the data you need for the purpose. Accuracy – Data collected must be checked for accuracy, and updated as necessary. Storage limitation – Data collected must only be kept for the time it is needed and no longer. Integrity and confidentiality – Data collected must be stored securely and kept confidential. It must not be passed on to a third party. Accountability – Individuals have the right to deal with their personal information. Organisations must understand the principles and apply them. 	<ul style="list-style-type: none"> A legal requirement to register with the Information Commissioner's Office (ICO) if an organisation processes personal data. Examples of who might hold your personal data are education, health and public service systems. A person is entitled to ask for access to their information, have it amended and have it deleted. 	<p>An organisation can be penalised at two levels:</p> <ul style="list-style-type: none"> Higher maximum level – up to 4% of the organisation's worldwide income. Standard maximum level – up to 2% of the organisation's worldwide income.
Freedom of Information Act 2000 (FOIA)	<p>This act gives the public the right to access information held by UK public authorities.</p> <p>Some data is exempt, e.g. personal, confidential and sensitive data if it is information that has been recorded and stored by the public authority.</p>	<p>Public authorities must publish certain information about their activities.</p> <p>If there is a request for public information, the organisation must provide the information if it holds it, and respond within 20 days.</p>	If the public authority does not comply with a FOI request, the ICO can issue fines and enforcement notices.

Regulations

UK General Data Protection Regulation 2018 (UK GDPR)

Privacy and Electronic Communications Regulations 2003 (PECR)

The Information Commissioner's Office (ICO)

The ICO is the regulatory body for data protection legislation, as well as freedom of information.

- Data Protection Act 2018
- General Data Protection Regulation (GDPR)
- Freedom of Information Act 2000
- Privacy and Electronic Communications Regulations 2003

INSPECTION COPY

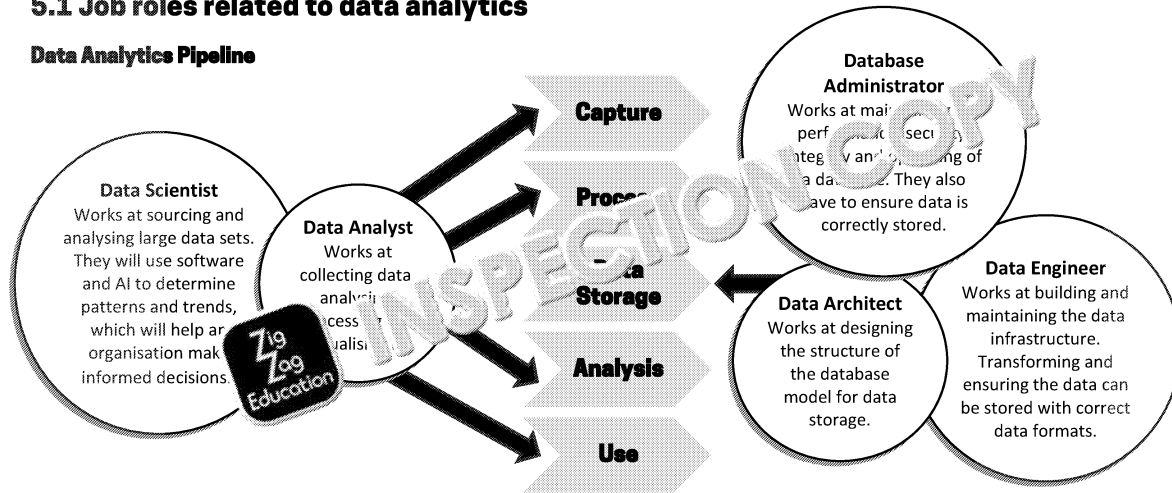
COPYRIGHT
PROTECTED



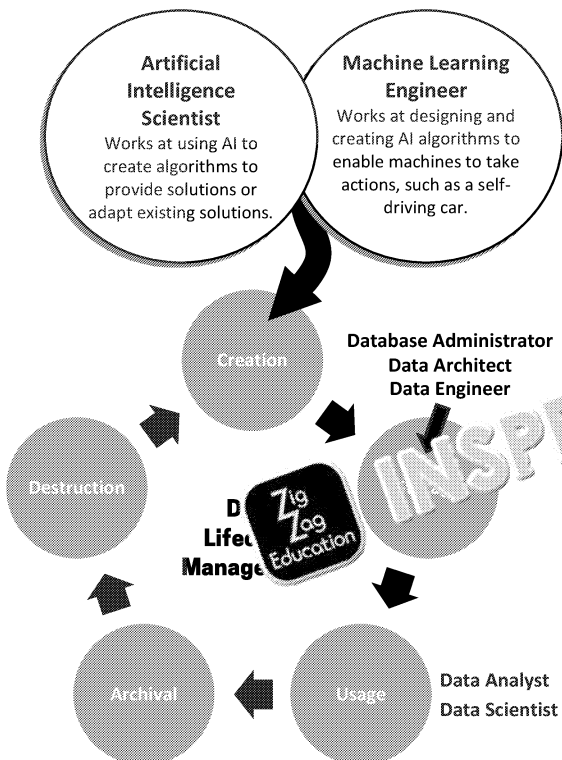
TOPIC AREA 5: JOB ROLES, SKILLS AND ATTRIBUTES

5.1 Job roles related to data analytics

Data Analytics Pipeline



Data Lifecycle Management



Wages – per annum	Entry level	Experienced
Source: Glassdoor Dec 2024		
Database Administrator	£20,000	£27,000
Data Analyst	£25,000	£60,000
Machine Learning Engineer	£30,000	£75,000
Data Scientist	£35,000	£82,500
Artificial Intelligence Scientist	£38,000	£75,000
Data Engineer	£53,000	£75,000
Data Architect	£58,000	£125,000

Appropriate Use of Language

- Provide context and audience.
- Avoid using technical language the audience won't understand.
- Take into account cultural differences.
- Do not use casual language in formal situations.
- The language you use should always be appropriate for the particular situation.

Non-verbal Communication

- Facial expressions – smiling.
- Body language – appropriate gestures such as nodding and pointing.

Questioning Techniques

- Closed questions – 'Yes' or 'No' – easy to analyse.
- Open questions – allow for full answers, which can be useful but harder to analyse.
- Give people time to respond to a question.
- Limit questions to the most useful.

5.2 Personal Attributes

As so much of the work is done on a computer, it is also the person's ability to communicate with others that is important. These personal attributes are:

- Analytical
- Communication
- Independent
- Leadership
- Planning
- Organisation
- Problem-solving
- Self-motivation
- Teamwork
- Time management

INSPECTION COPY

COPYRIGHT
PROTECTED



1.1 Data, information and knowledge

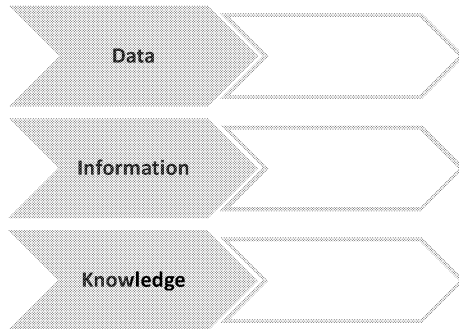
What is data analytics?

Our daily lives are filled with data. Everything we interact or clickstream with online becomes a data source, from a social media comment, to uploading different types of files to the Internet, such as text, images and video. More and more, we are reliant upon technology and the IoT (Internet of Things), which facilitate our daily lives and have access to our data. This all adds to an ever-increasing explosion of data which is being stored.

Data analytics is about making sense of this data, collecting it, transforming, modelling and interpreting it into useful information, which in turn can be used to make better decisions, draw conclusions, create new products and guide decision-making.



data, information and knowledge are:



Describe what each of these are. ▲

Sources of Data

Data is drawn from data repositories, data warehouses, spreadsheets, databases.

Sources of Information

Data can be drawn from an organisation's internal databases, using queries to extract information.

Sources of Knowledge

Knowledge is the result of analysing data and information as above.

Sources of Knowledge

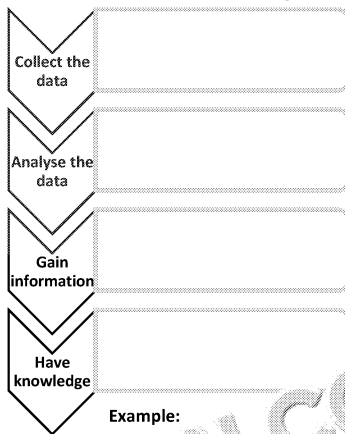
Business Data: Relevant to an organisation – financial, functional, client data

IoT Data: Generated in a real-time format, from industries – health, manufacturing, transportation, specialised software tools for

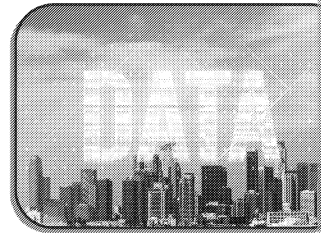
Public Data: Data collected at an academic and government organisation – statistics, census data.

Big Data: Data formed from huge sets, hard to organise and manage, pages, social media.

Interaction of data, information and knowledge



Example:



Add the labels for each of the descriptions. ▶

Data and information in society

Society can benefit from the use of data. However, there can also be negative consequences.

benefits and limitations of the use of data and information to organisations and individuals

▲ What are the benefits and limitations of the use of data and information to organisations and individuals?

How data and information can have negative consequences

Security issues, cybersecurity, hackers, loss of reputation, data protection issues. Misused in the workplace.

COPYRIGHT
PROTECTED





INSPECTION COPY

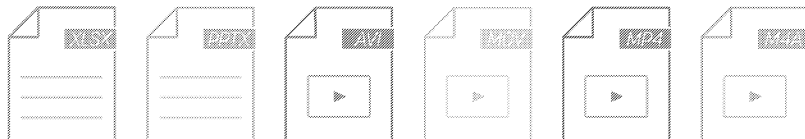
1.3 Data and file formats

To be used and recognised by the different software systems in use, data in the form of encoded and stored files must have a specific file format.



Benefits and limitations of data/file formats

Data	File	Limitations
Encoding data		Industry standard, easy to use, uses less memory space. Limited character set, difficult to represent other language, low data security.
	XML	Used for exchanging data between different systems, stores in plain text format. Large files, slow, not good with binary data. Compatibility and memory issues.
Data interchange		Used for data exchange in web services. Fast and simple syntax. Supports a limited set of data types, low data security.
Plain Text data		No formatting. .txt uses TAB to separate each field. .csv uses a comma. .csv only used in spreadsheet. .txt can be used in any application.
	Integer/decimal/float	Integer can be processed faster. Float and decimal slower to process.
	AIFF WAV	Large file formats, high quality. As they are lossless they need to be compressed to MP4 for transportation.
Image		JPEG compressed, easy to transport, but because of compression, can reduce image quality and become pixelated. PNG does not support CMYK colour mode, and not printable. TIFF large file format, high quality image. TIFF, because of large file size, not good for transportation.



1.4 Data types

Data can be formatted in different ways. It also has different classes.

Benefits and limitations

Data type	
	Two or false, true, false
Character	
Date	dd/mm/yyyy stored
	Whole numbers
Real/float	
	Most characters

Classifications of data

INSPECTION COPY

COPYRIGHT
PROTECTED

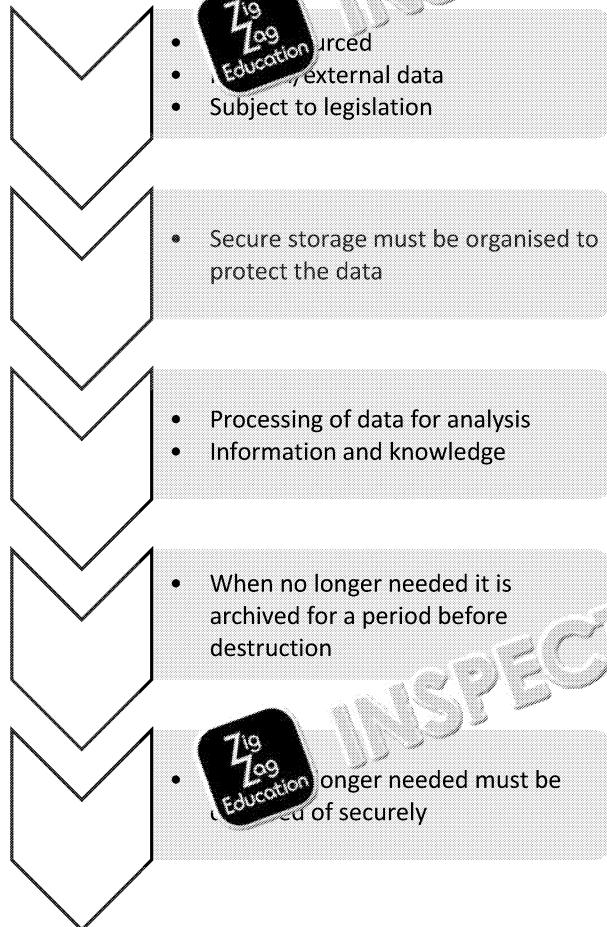
2.1 Data lifecycle management (DLM) and the data analytics pipeline

2.1.1 Data lifecycle management (DLM)

What is data lifecycle management (DLM)?

An approach to managing data with five phases.

▼ Fill in the gaps of the data lifecycle management.

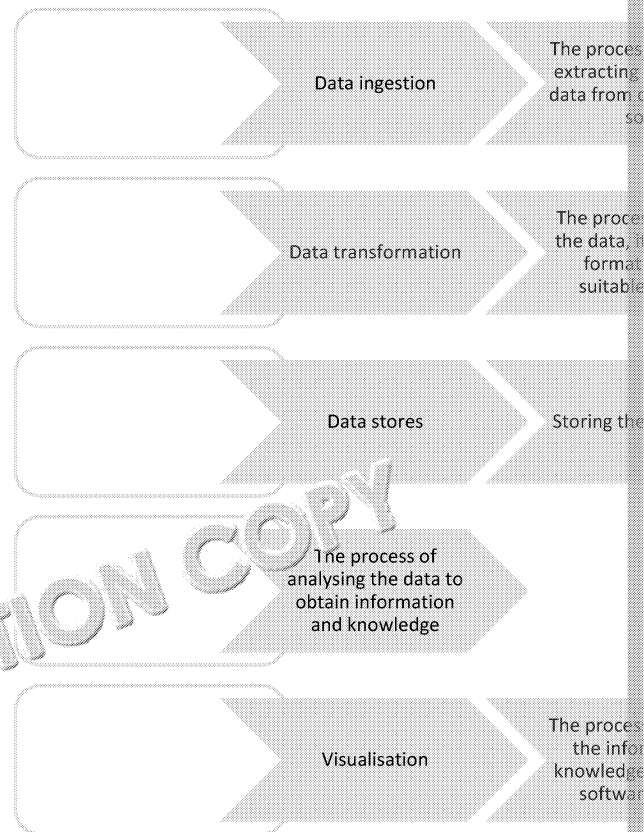
The interactions and iterations of the five phases

2.1.2 Data analytics pipeline

What is the data analytics pipeline?

An approach to data analysis with five phases.

▼ Fill in the gaps in the data analytics pipeline.

The interactions and iterations of the five phases

INSPECTION COPY

COPYRIGHT
PROTECTED

2.2 Creation and capture

2.2.1 Data assurance considerations

What is data assurance?

▲ Explain what data assurance is

What areas must be considered?

Accuracy

The data supplied must be accurate and relevant.

.....
The data supplied needs to have gone through checks to ensure that it is accurate and valid.

Redundancy

Errors in data must be removed; if there are duplicates, such as the same data stored in different places, this provides unreliable data.

Reliability

Organisations rely on data having gone through the data assurance process to be reliable.

.....
That the data supplied is up to date. If it is not, then it will not provide accurate information.

Validation

A defined set of rules can be implemented for data entry, so that the correct data is input. Methods such as drop-down lists, input messages all help to prevent GIGO.

.....
Data must have a proper structure and be fit for its intended purpose, e.g. data for a specific system.

.....
Data that is input must be checked to ensure that it is correct. This can be done by checking against an up-to-date source.

▲ Provide the missing headings.

The purpose and importance of data assurance considerations

Why do we have to have data assurance considerations?
If data does not go through these assurance processes, then the data becomes unreliable

GIGO Garbage in, Garbage out

Poor data = unreliable information = inaccurate knowledge



How each consideration affects the collection and use of data

Although there are data assurance measures to be considered, it is important that there is an accountability process to follow.

Data profiling:

Analysing structure

Data cleansing:

Identifying redundancy

Data entry:

Validation methods implemented

Data documentation:

Guidance provided about the data

Data audits:

To ensure reliability of the data assurance process



Confidence in data

If organisations cannot be sure that the data has gone through data assurance considerations, they will not have confidence in the data they have been provided with. This can lead to a bad reputation for the data provider.

2.2.2 Data governance

What is data governance?

Methods

Documents and records

Interviews

Observations

Online tracking

Social media monitoring

Transactional tracking

▲ Fill in the table

Factors for effective data governance

It is important that there are clear policies in place. It is needed. The meaning depends on the need to be transparent if possible. Consider the context and the form

INSPECTION COPY

COPYRIGHT
PROTECTED



2.3 Storage

2.3.1 Data states

What is a data state?

When data is used by computing equipment, it can be subject to different states, depending upon the stage reached in the DLM. (See 2.1.1)

What are the three data states?

Fill in the missing headings.

.....
This is data which is being transferred over a network. It may also be in the process of transferring from one local device to another. It will be encrypted for security purposes.

.....
This is data which is stored for later use. It may be stored in the cloud or on a physical backup. It will be encrypted for security purposes.

.....
This is data which is being used, in other words in the process of being analysed. Because it is being used, it does not need encryption.

2.3.2 Data stores

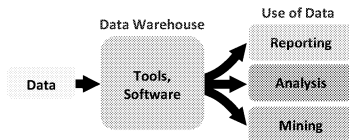
What is a data store?

Explain what a data store is. ▼

Name the different types of big data and data stores. ▼

Big data (see 1.2) – stores which contain vast amounts of data sets are:

- – a central repository of data ready to be analysed
- – a central repository of data in a raw format
- – contains subsets of data
- – contains data which has been isolated from other data sets



2.3.4 Onsite storage

File server

In an organisation, all computers will generally be networked to a central file server. It can be used for block and file storage.

Hard drive

A hard drive is contained within a computer or laptop to store data. There are two types:

-
-

Network attached storage - NAS

Portable storage device

A portable storage device can be an external hard drive, USB drive, etc. It is portable, which means it can be used anywhere and anytime.

The limitations are that it can be easily lost, stolen or damaged.

Storage area network - SAN

2.3.5 Cloud storage

Organisations can store data in the cloud. Internet-based storage is available. All are provided by third parties.

.....
Multiple organisations can share their data on dedicated servers.

.....
Organisations can store their own data in the cloud.

.....
Accessible from anywhere.

.....
A mix of private and public cloud storage.

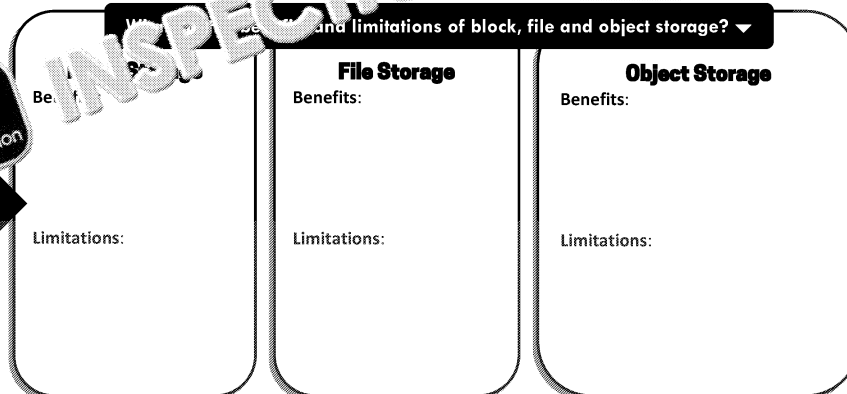
.....
What are the advantages of cloud storage?

2.3.3 Data storage

There are three different ways that data can be stored with cloud storage – block, file and object storage.

When considering data storage types, the following must be considered.

- Is the data set large or small?
- Is it structured or unstructured?
- What does the organisation want to do?



INSPECTION COPY

COPYRIGHT
PROTECTED

2.4 Data transformation

2.4.1 Data wrangling

What is data wrangling?

▼ Explain what data wrangling is.

2.4.2 Data maintenance

What is data maintenance?

Another important process which ensures that the data is regularly checked to make sure that it continues to be reliable and accurate for use.

2.5 Usage and analysis

2.5.1 Data analytics

What is data analytics?

▼ Describe what data analytics is, its purpose and the three different processes involved.

2.5.2 Types of data analytics

In order to process and analyse data, there needs to be a specific purpose. There are different reasons, and different questions to raise about the data. Therefore, there are different types of data analytics processes which can be used.

All of the different types of data analytics have one purpose, to make informed decisions – which are of course benefits. The limitations of the systems are that interpretations could be biased, especially if there has been an incorrect interpretation.

▼ Add the names of the five different types of data analytics.

	This method uses machine learning algorithms and AI to take specific types of structured data. The purpose is to extract further insights and make predictions which the data analytics might miss. This type of data analytics can be used to predict things such as health treatments, potential market trends, anticipating customer behaviour.
	This method uses past data to identify patterns in the past behaviours and events. It can be used to identify trends on the basis of past data, e.g. Netflix gathers this type of data to determine what it is most popular to watch. It is dependent upon the historical data. If something is popular, it will offer more of the same.
	This method looks at what happened and why. It looks at descriptive analytics and digs deeper to find the root of the problem. It is used to test a hypothesis, such as if we do this, will this happen? It also looks at correlation of different variables which might explain and examine customer behaviour. Hello Fresh may find that there is a trend in fish dishes being ordered, and, therefore, it will supply more fish recipes.
	This method looks at an outcome and then considers what should or could happen next. It uses historical data and looks at trends to predict future outcomes. Organisations can use this to determine their future financial needs, staffing solutions and marketing trends.
	This method looks at past decisions and events to estimate the likelihood of different outcomes and to choose the best course of action, e.g. on TikTok the For You feed is based on what you have watched previously and assumes you want to watch more of the same.

2.6 Usage and visualisation

2.6.1 Presenting data

What are the three ways that data can be presented? ▶

Methods of data presentation

Once the data has been analysed and information and knowledge has been gained, the next step is to present the information. When presenting data, it is important to have information laid out clearly and also to consider people with accessibility issues.

2.6.2 Visualising data

Infographics

Describe what an infographic is, the visual format it uses and the benefits of using one. ▼

There are many ways to visualise data. It is important to choose the right one and who the audience is to show the data in the intended way.

Graph types
These are the different types of graphs used to visualise data.

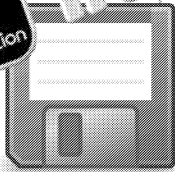
INSPECTION COPY

COPYRIGHT
PROTECTED

2.7 Archival

Why do we archive data?

▲ Explain the purpose of archiving data.



2.8 Destruction

Why do we have to destroy data?

▼ the purpose of destroying data. ▼

Data destruction methods

▼ Identify and explain how each data destruction method works.

Benefits – This is a reliable form of data destruction and a

Limitations – The machine can only be used on one storage device.
The demagnetising process renders the hard drive storage device unusable.
The magnetic force must be very strong, however, and technical

Benefits – This is a reliable form of data destruction, as the

Limitations – It can be a costly and time-consuming process.
It can have a negative environmental impact and then be subject to e-waste regulations.

Benefits – This is a reliable form of data destruction, as the
It is a cost-effective method as many devices can be erased simultaneously.

Limitations – It can be a time-consuming process. Software

Benefits – This is a reliable form of data destruction, as the

Limitations – This method also can contribute to environmental

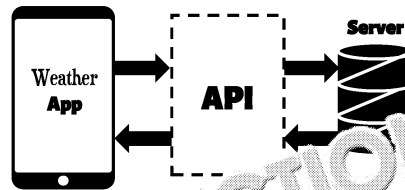


INSPECTION COPY

COPYRIGHT
PROTECTED

3.1 Application Programming Interfaces (API)

The role of an API



▲ Explain what an API is.

API certifications

There are two types of API certifications. The Wikimedia API is a public API that allows developers to create an app and link to the vast amount of knowledge it has. The Wikimedia API supplies data to your app. They are commonly used with mobile phone apps.

Public: These are paid for APIs. They will require authentication to connect an app. The Hunter API links an app to professional email addresses, which can be checked for authenticity.

This API combines several APIs and can perform more than one job at once. When we purchase something from Amazon, APIs are used to add a product to basket, update the basket, checkout the basket and pay.

Internal/Private: These are private APIs and restricted to private organisations for their use only. These are the most common type of API and are very quick to develop. These are generally used to improve organisational functions.

These are APIs which can be free to or paid for by a limited audience. They are limited to invitation only and have restrictive authentication methods. They are, therefore, very secure. Organisations will use this type to monetise their products.

▲ Add the missing certifications.

Describe the benefits and disadvantages of using APIs. ▼

Benefits:

Disadvantages:

Types of API design

What does JSON stand for?

JSON-RPC Stands for JSON Remote Procedure Call. Social media platforms use this type of API to send and request data. Request and response architecture is built into the API. It is used to interact with data.

SOAP Stands for Simple Object Access Protocol. This is an older system, and, therefore, slower. It is a messaging format, and messages are usually transmitted via a web request. It was initially designed to move data around organisations. It is secure and is generally used in financial transactions.

REST Stands for Representational State Transfer API.

◀ REST – what is the most common use of this API?

XML-RPC Stands for Extensible Markup Language - Remote Procedure Call and uses HTTP to transport an XML format to encode data. It is known for its information security. Generally used in content management, tasks and password management systems.

3.2 User access controls

▼ Explain

What is user access control?

What are the five types of user access control? A

This type of access is dependent upon a set of attributes we user, their user ID, their age or their location. If the attribute is in financial institutions.

This type of access allows persons who have control of access. An example of this is 'sharing' a Google document. By sharing

This type of access is determined by an agreed policy, set by person being given access. It is generally used in health organisations.

This type of access is allocated by the role of the person it is. This is a reliable and secure system. Users can be given read

This type of access is determined by an agreed set of rules. cybersecurity threat and if there is infrastructure overload. type of organisation, the sensitivity of the data, and who is

3.3 Permissions

Administr

User Rights

Once a user has been granted permissions, then permissions are then granted to the user. This is to ensure the user has.

What are the four user rights permission? Name them and add what they allow.

There are different categories of user privilege:

User level: This is determined by the user rights allocated.

User group level:

-
-
-

File and folder level: Again, this is determined by access level rights and different permissions can be granted.

► What are the three of account at user group level, and what type access do they allow?

INSPECTION COPY

COPYRIGHT
PROTECTED



4.1 Legislation and the role of the ICO when using data

When working with data, there are several areas of legislation to consider and to adhere to, particularly with personal, identifiable information.

Legislation

Name the three acts. ▼	What it is	Risks	Non-compliance
	<ul style="list-style-type: none"> This act provides legislation on crimes connected to the use of computers. Unauthorised access to computer material, including mobile phones. Unauthorised use of financial data with the intent to commit fraud, introducing malware or viruses. Unauthorised acts causing or creating risk of serious damage. Example is cybercrime, which has a serious risk to human welfare, damage to the economy and national security. 	<ul style="list-style-type: none"> Staff are educated about the associated risks with access to data. Policies and procedures to ensure compliance; failure to comply could result in disciplinary procedures. IT administrators should implement appropriate user access levels and permissions, and put robust security measures in place. Operating systems to be kept up to date with security patches. 	Ranging from fin es to prison time , from a few months right up to life imprisonment for the most serious offences under Section 3ZA.
	<p>This act controls how organisations use our personal information. There are seven principles that organisations must adhere to.</p> <ol style="list-style-type: none"> Lawfulness, fairness and transparency – When collecting information, it must be clear what it is being used for. Purpose limitation – It must be clear what the data is being collected for and it should only be used for that purpose. Data minimisation – You must only collect the data you need for the purpose. Accuracy – Data collected must be checked for accuracy, and updated as necessary. Storage limitation – Data collected must only be kept for the time it is needed and no longer. Integrity and confidentiality – Data collected must be stored securely and kept confidential. It must not be passed on to a third party. Accountability – If an organisation has a legal obligation to deal with people's information, it must understand the principles and follow them. 	<ul style="list-style-type: none"> A legal requirement to register with the Information Commissioner's Office (ICO) if an organisation processes personal data. Examples of who might hold your personal data are education, health and public service systems. A person is entitled to ask for access to their information, have it amended and have it deleted. 	<p>An organisation can be penalised at two levels:</p> <ul style="list-style-type: none"> Higher maximum level – up to 4% of the organisation's worldwide income. Standard maximum level – up to 2% of the organisation's worldwide income.
	<p>This act gives the public the right to access information held by UK public authorities.</p> <p>Some data is exempt, e.g. personal, confidential and sensitive data if it is information that has been recorded and stored by the public authority.</p>	<p>Public authorities must publish certain information about their activities.</p> <p>If there is a request for public information, the organisation must provide the information if it holds it, and respond within 20 days.</p>	If the public authority does not comply with a FOI request, the ICO can issue fines and enforcement notices.

Regulations

Name the two regulations. ▼

The Information Commissioner's Office

What is the ICO's role?

Who are they responsible to?

▲ Explain what the ICO does, who it oversees and its role. ▶

INSPECTION COPY

COPYRIGHT PROTECTED



5.1 Job roles related to data analytics

Data Analytics Pipeline

Add the five job roles involved in the DAP. ▲

Works at sourcing and analysing large data sets. They will use software and AI to determine patterns and trends, which will help an organisation make informed decisions.

Works at collecting data and analysing it to identify patterns and trends.

Capture

Process

Storage

Analysis

Use

Works at maintaining the performance of the database. They ensure the integrity and operation of a database. They also have to ensure data is correctly stored.

Add the missing personal attribute labels. ▶

Works at designing the structure of the database model for data storage.

Works at building and maintaining the data infrastructure. Transforming and ensuring the data can be stored with correct data formats.

Link the job roles to their part in the DAP.

Data Lifecycle Management

Artificial Intelligence Scientist

Machine Learning Engineer

Explain what an artificial intelligence scientist and a machine learning engineer do.



List the five things you should consider in your use of appropriate language.

Appropriate Use of Language

-
-
-
-
-

Non-verbal Communication

- Facial expressions – smiling.
- Body language – appropriate gestures such as nodding and pointing.

Questioning Techniques

- Closed questions – 'Yes' or 'No' – easy to analyse.
- Open questions – allow for full answers, which can be useful but harder to analyse.
- Give people time to respond to a question.
- Limit questions to the most useful.

5.2 Personal Attributes

As so much of the work is done by these people, it is important to have the right personal attributes.

Analytical

Communication

Planning

Organisation

Self-motivation

Teamwork

INSPECTION COPY

COPYRIGHT
PROTECTED

