Unit 4: Exchanging Data
(4a. Compression, Encryption and Hashing, A Level Only Content)

Marks: /32

Answer **all** the questions.

1. A software company decides to release a duplicate file finder which it has named “De-Duplicator”. Duplicate files are files that are exactly the same (bit for bit identical). Space is often wasted on computers by having multiple versions of the same file. Duplicate file finders are programs that find and identify duplicate files on a hard drive so that they can be removed.

   Every time the program encounters a file it takes a hash of the file and checks it against a list. If the hash exists in the list, the file is marked to be deleted. If the hash does not exist it is added to the list.

   (i) Explain **two** characteristics you would look for in a hashing algorithm for this purpose.

   1 _____

   _____

   _____

   _____

   2 _____

   _____

   _____

   _____ **[4]**

   (ii) After running the program a user finds that they still have apparent duplicates of some of their images. Explain why these apparent duplicates might still be present.

   _____

   _____

   _____

   _____

   _____ **[2]**

2(a). A charitable organisation is trying to make the works of William Shakespeare available to more people.

The organisation decides to make a copy of Shakespeare's entire works available as a downloadable text file from its website. It further decides to compress the file before making it available to download.

(i) State an advantage to the website's visitors of the file being compressed.

---[1]

(ii) Explain why the company should use lossless and not lossy compression.

---[3]

(b).

The organisation looks at using either run length encoding or dictionary encoding to compress the file described in **part (a)**.

Discuss the **two** compression methods and justify which you would recommend. You may refer to the extract of text below to illustrate your argument.

What's in a name? that which we call a rose
By any other name would smell as sweet;
So Romeo would, were he not Romeo call'd,

[12]

3(a). The XOR operator can be used to encrypt data.

Show the effect of applying XOR on Text and Key, by completing the last row of the table below.

Text	O								C								R							
Value	0	1	0	0	1	1	1	1	0	1	0	0	0	0	1	1	0	1	0	1	0	0	1	0
Key	A								B								C							
Value	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	1
XOR																								

[2]

(b). Show the effect of applying XOR on your answer to part (a) and Key, by completing the first and last rows of the table below.

(a)																								
Key	A								B								C							
Value	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	0	1	1
XOR																								

[2]

(c). Explain whether the type of encryption described above is symmetric or asymmetric.

---

[2]

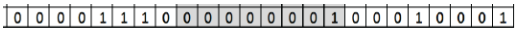
(d). Explain why asymmetric encryption is more suited to transactions over the internet than symmetric encryption.

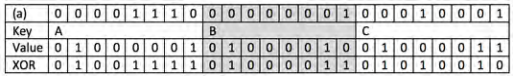
---

[4]

Question			Answer/Indicative content	Marks	Guidance
1		i	• Low chance of collision (i.e. different inputs giving same output) (1 – AO1.2) to reduce risk of different files being marked as the same (1 – AO2.1). • Quick to calculate (1 – AO1.2) as lots of files need to be hashed / needs to be quicker than a bitwise comparison to make it worthwhile (1 – AO2.1). • Provides a smaller output than input (1 – AO1.2) so quicker to compare hashes than original data (1 – AO2.1).	4	1 mark for each correct identification (AO1.2) up to a maximum of two identifications 1 mark for each valid explanation (AO2.1) up to a maximum of two explanations. No credit for function being one way as this serves no benefit in this scenario.
		ii	• Hashing works on the data / bits (1) and so two images may appear the same but not be identical at a bit level (1). This could be because they are different file types (1) / different sizes (1). Even the change of a single bit may result in a completely different hash (1).	2	Up to 2 marks for a valid explanation. Accept any other sensible examples of changes to images that might not be immediately apparent to someone viewing the image.
			Total	6	
2	a	i	Downloads quicker. (1) Saves user money by using less bandwidth / on data usage. (1) (Max 1)	1 (AO1.2)	Do not accept 'saves the user space on their device'. Examiner's Comments This question was well received by most candidates, invariably scoring most marks.
		ii	Lossy takes away some of the information from the original. (1) Lossless preserves all the information from the original. (1) With text the loss of small amounts of information will make it unreadable. (1)	3 (AO1.1 – 2 marks AO2.1 – 1mark)	
	b		Mark Band 3–High Level (9–12 marks) The candidate demonstrates a thorough knowledge and understanding of dictionary and run length encoding for compression. The material is generally accurate and detailed. The candidate is able to apply their knowledge and understanding directly and consistently to the context provided.	AO1.1 (2) AO1.2 (2) AO2.1	Points may include but aren't limited to: AO1 Knowledge and Understanding Run length encoding relies on consecutive pieces of data / characters being the same. Each set of consecutive symbols can be represented by the symbol and its number of occurrences

Question	Answer/Indicative content	Marks	Guidance
	Evidence / examples will be explicitly relevant to the explanation. The candidate is able to weigh up both forms of compression and justify dictionary encoding being the better choice. There is a well-developed line of reasoning which is clear and logically structured. The information presented is relevant and substantiated. **Mark Band 2-Mid Level (5–8 marks)** The candidate demonstrates reasonable knowledge and understanding of dictionary and run length encoding for compression; the material is generally accurate but at times underdeveloped. The candidate is able to apply their knowledge and understanding directly to the context provided although one or two opportunities are missed. Evidence / examples are for the most part implicitly relevant to the explanation. The candidate makes a reasonable attempt to come to a conclusion as to which form of compression is better suited. There is a line of reasoning presented with some structure. The information presented is in the most part relevant and supported by some evidence. **Mark Band 1-Low Level (1–4 marks)** The candidate demonstrates a basic knowledge of dictionary and run length encoding for compression; the material is basic and contains some inaccuracies. The candidate makes a limited attempt to apply acquired knowledge and understanding to the context provided. The candidate provides nothing more than	(3) AO3.3 (5) 12	e.g. AAAABBBBBCCC could be represented as 4A5B3C (or A4B5C3 or any sensible RLE encoding) In dictionary encoding frequently occurring pieces of data / groups of characters are replaced by symbols / tokens / smaller groups of characters / indexes. A dictionary is then used to say which symbols / tokens / characters / indexes match which groups of characters. When decompressed the dictionary is used to replace the tokens with the original text. **AO2.1 Application** Run Length Encoding is very unsuitable for the example text. There are very few consecutive repeating symbols in the text. only instances being ll and ee these still require 2 characters to represent them 2l and 2e Dictionary encoding is well suited. There are lots of repeating groups of characters For example ‘call’ ‘name’ ‘[SPACE]we’ ‘Romeo’ We could for example have: What’s in53? that which2 15 rose By5ny other3 would smell5s sweet; So4would,2re he not41’d 1:call 2:[space]we 3:[space]name 4:[space]Romeo[space] 5:[space]a (NB candidates are unlikely to show full compression, just a demonstration of the principle is sufficient. The best candidates are likely to show an awareness that space is a character that can be used in

Question			Answer/Indicative content	Marks	Guidance
			an unsupported assertion. **0 marks** No attempt to answer the question or response is not worthy of credit.		compression and that upper and lowercase letters are different. Demonstrating this is indicative of but not a requisite of the band.) **AO3.3: Evaluation** Run length encoding is not suited to natural language (more likely to be used in simple images). Applying it to the example the resulting text would be the same size as the original / worse than the original (if we use 1s to represent every individual instance of a character). Dictionary encoding works well. We can already see benefit on small piece of text. Would fare substantially better on full works. Dictionary encoding is the best compression method for this scenario. **Examiner's Comments** Candidates were assessed on the quality of their extended response in this question. Most candidates could describe each of the given types of compression appropriately, with many applying them to the scenario. Many candidates correctly concluded that dictionary encoding was the most appropriate in this case, but few then went on to give clear and appropriate justification for their assertion. In general, most candidates scored well on this question.
			Total	**16**	
3	a		 One byte correct (1) all three bytes correct. (1)	2	

Question			Answer/Indicative content	Marks	Guidance
	b		 One byte correct (1) all three bytes correct. (1)	2	Allow FT if (a) is incorrect but bottom row must match XOR with top row and key.
	c		Symmetric (1) …. as the same key is used to decrypt it as encrypt it (1)	2	Allow FT for asymmetric if (b) indicates asymmetric encryption used
	d		Any four from: Symmetric encryption would require both parties to have copy of the key (1) this couldn't be transmitted over the internet or an eavesdropper monitoring the message may see it (1) Asymmetric gets round this requirement as there are two different keys (1) One key encrypts the data (1) which can be publically distributed (1) and a different key to decrypt it (1) which is kept private (1)	4	
			Total	10	