BCS Higher Education Qualification

Diploma

October 2023

EXAMINERS' REPORT

Big Data Management

Questions Report:

A1	
	This question was about management issues for big data.
	Candidates answered this question on management issues for big data very well. Over two thirds of candidates achieved a pass mark.
	The first two parts, a) and b), were answered well by almost all candidates and there were no specific issues evident.
	In part c), a GDPR related question, many candidates were unclear on what precisely personal data referred to. Several candidates included sex, gender and purchase history, which would have been relevant if they were directly linked to a known individual.
	In part d), many candidates struggled to describe unstructured data. Some suggested unstructured data lacked any form of structure. Very few answers indicated why unstructured data was a particular problem in the management of personal data and only a few alluded to the diversity of sources being a key issue.
	Answers in part e) were mixed. Some candidates found it difficult to state government objectives in introducing data privacy regulations. Many cited social reasons in addressing the rights of ownership and not centred on human rights and protection of individuals from corporate exploitation.
A2	
	This question had a very high pass rate as almost all candidates answered the question very well. There were no issues evident.
A3	
	Only a handful of candidates answered this question on Tools and techniques for analysis.
	The question in part d) on neural networks was poorly answered.
	Candidates had difficulty describing three applications of neural networks. Good and acceptable answers would include the notion of mirroring working process of the brain, the use of layered neurons and aspects of deep learning.
B4	
	This question was answered well with more than half of all candidates achieving a pass mark.
	Part a) and part b) had no significant issues.
	Part c) required four characteristics of streamed data and posed a problem for some candidates. Many did not acknowledge the heterogenous nature of streamed data.

		Only a few answers acknowledged the nature of streamed being continuous with no end, and no way to go back in time with the stream to rectify missing items.

In part d) most candidates were able to adequately describe event stream processing but found it difficult to describe complex event streaming. Very few answers provided a full comparison which mentioned the crucial difference in event stream having its continuous nature and time stamped with complex event streams arriving with different time stamps. |
| **B5** | | |
| | | This question was about storage and architectures for big data.

Part a) concerned Kafka. Candidates' answers indicated a reasonable appreciation of what it is and adequately described some of the principal features. However, there was no acknowledgement of the use of fast TCP between devices with the architecture.

Part b) required an explanation of why Kafka is of particular benefit in big data. Candidates' answers indicated a good appreciation of the capabilities in handling huge volumes and can deal with real time streams. However, there was difficulty in providing an example that would particularly emphasise Kafka's expertise, such as LinkedIn. |
| **B6** | | |
| | | In part a), most answers showed a good appreciation of the basic features of a NoSQL database. A few candidates wrote extensively on this part of the question, when the maximum mark possible was 4, and used a significant amount of time elaborating their answer.

In part b), most answers showed a good appreciation of document and column-oriented databases. However, the description of graph based proved difficult for most candidates. Very few offered an answer which acknowledged edges storing relationships between entities and access being through specific edges.
High-speed property being enabled by queries being persisted in the database was not mentioned in any of the answers.

Part c) asked for an explanation as to why MongoDB maintains partition tolerance with consistency but not availability. Very few answers gave an account of default behaviour, setting read/write to the primary node allowing a point of availability failure. Similarly, few answers indicated the issues in possible non-default settings eventually allowing lack of availability.

Part d) was about R. Very few candidates attempted this question. Of those that attempted it, their answers showed no issues, with many citing the use of R in unstructured and semi-structured data handling and the R ecosystem providing packages specifically for big data analytics. |