The Enterprise Big Data Professional Examination

Multiple Choice

90 Minute Paper

Instructions

1. All 60 questions should be attempted.
2. All answers are to be marked on the answer grid provided.
3. Please use a pencil and NOT ink to mark your answers in the Answer sheet provided.
4. There is only one correct answer per question.
5. You have 90 minutes for this paper.
6. You must get 39 or more correct to pass.

Candidate Number: .....…

1 Which is one of the four characteristics of Big Data?

a) Validity
b) Volume
c) Value
d) Variability

2 Which is one of the primary reasons that so many organizations find it difficult to realize a competitive advantage through Big Data?

a) They treat information as a strategic asset
b) All of their Big Data is actively being used
c) They have not adequately defined a Big Data strategy
d) They compete on the preciseness of their data

3 What type of skew is indicated when most values in a data set are greater than the mean?

a) Major
b) Minor
c) Positive
d) Negative

4 What techniques are commonly associated with supervised machine learning?

a) Classification and correlation
b) Correlation and regression
c) Classification and regression
d) Regression and clustering

5 Which of the following statements apply to structured data?

1. Is organized in a pre-defined format or table
2. Is typically stored in Excel files or SQL databases
3. Is typically text heavy, but may also contain dates, numbers and facts
4. Shows where relationships are established between fields

a) 1, 2, 3
b) 1, 2, 4
c) 1, 3, 4
d) 2, 3, 4

6 Having completed a few simple Big Data projects, an organization now wants to ensure their Big Data Team retain their Big Data knowledge and adopt a culture of continuous learning.

Achieving which organization success factor would **MOST** support this?

a) A clear view of what the organization is aiming to achieve with Big Data
b) Effective data governance and data management processes
c) A centralized Big Data Centre of Excellence
d) An ongoing Big Data professional training program

7 For the analysis of which kind of data are Deep Learning techniques predominantly used for?

a) Categorical and random
b) Alpha and numeric
c) Images and audio
d) Machine and integer

8 What level on the Big Data maturity scale is demonstrated by an organization that is managing data and analytic capabilities across all domains of the organization but has yet to fully optimize all its Big Data activities?

a) Level 2 – Localized Analytics
b) Level 3 – Analytical Operation
c) Level 4 – Analytical Enterprise
d) Level 5 – Data Driven Enterprise

9 What differentiates cognitive analytics from other forms of analytics?

a) Decisions are made based on the perceived environment and personalized characteristics
b) General awareness and emotional intelligence are factored into the decision process
c) Rationale perception and a generalized set of conditions are factored into the decision process
d) Decisions are automated and irrelevant of any individual preferences

10 When formulating a Big Data strategy, in what sequence should the following occur?

1. Define business objectives
2. Identify and prioritise Use Cases
3. Execute a current state assessment
4. Formulate a Big Data Roadmap

a) 1, 3, 2, 4
b) 1, 3, 4, 2
c) 3, 1, 2, 4
d) 3, 2, 4, 1

11 What effect does Big Data have on data samples used to make predictions?

a) Smaller subsets of data are required for results
b) Sample sizes are closer to the entire population
c) Relationships are created between all data samples
d) Samples become independent of populations

12 Which statement **BEST** describes a reference architecture?

a) A set of documents to which a project manager or other interested party can refer for best practices
b) A source of information configured using a common structure within a computer
c) A piece of shared advice that builds upon common knowledge gained by others
d) A recommendation based on a standard design principle that others have experience of

13 What is the definition of Big Data?

a) The grouping of bulky information that cannot be processed manually or through automated computer software
b) The creation of computer programs that are able to process large quantities of data
c) The consolidation of internal, external, structured, semi-structured, unstructured and enterprise data
d) The exploration of techniques, skills and technology to deduct valuable insights out of massive quantities of data

14 What are solid stated drives (SSD), optical disk devices or external hard drives that are connected to a computer examples of?

a) Storage Area Network (SAN)
b) Network Attached Storage (NAS)
c) Direct Attached Storage (DAS)
d) Direct Area Network (DAN)

15 Which of the following is one of the five main roles within the NIST Big Data Reference Architecture?

a) Data Manufacturer
b) Digital Analyst
c) Data Provider
d) Dialogue Provider

16 What type of analysis could be used to explore whether a relationship exists between the amount of vitamins a subject group take and their life span?

a) Clustering
b) Sampling
c) Correlation
d) Regression

17 Which characteristic applies to data sets in Big Data?

a) Contains a percentage of meaningless details
b) Are generated from a single source
c) Are presented in one common structure
d) Exploit traditional storage capabilities

18 Which of the following are key characteristics of a well-designed Big Data Lab?

1. Open work space for collaboration
2. Isolated work space without distractions
3. Creative environment to experiment
4. Networked computers with 4GB RAM

a) 1, 2, 3
b) 1, 2, 4
c) 1, 3, 4
d) 2, 3, 4

19 An insurance company wants to use the character profile of a sample of its existing customers to determine the most effective advertising media to use for a new insurance product.

What type of business objective would address this need?

a) Descriptive business objective
b) Inferential business objective
c) Predictive business objective
d) Mechanistic business objective

20 How does Hadoop overcome the risk of losing data that is being stored or processed?

a) Includes a system for backing up entire machines to a central database multiple times each day
b) Uses parallel programming models to reduce the risk of losing data during processing
c) Uses Storage Area Networks (SANs) to enhance storage devices
d) Replicates data packages and stores them on a number of different machines

21 An analysis of 200,000 candidate's test results shows that 99.9% of candidates scored between 76% and 89%. There were 18 candidates who scored less than 20%.

What statistical term is used to describe these anomalies?

a) Outliers
b) Random data points
c) Deviations
d) Multivariate observations

22 Identify the missing word in the following sentence.

A Big Data [?] is required to help focus an organization's investment in Big Data analysis and analytics.

a) Strategy
b) Architecture
c) Team
d) Function

23 Which statement does **NOT** apply to Hadoop?

a) Is an open source software framework used for processing datasets of Big Data
b) Consists of modules designed on the assumption that hardware failures rarely occur
c) Assumes that hardware failures should be automatically handled by the framework
d) Has become known as the eco-system that connects different Big Data solutions

24 4,000 students completed an examination paper across the country. These results need to be communicated to all examination centres.

What type of graphical representation would show the highest and lowest scores, the average score, the upper and lower quartile, and any individual scores that do not fit within the general pattern of others?

a) Scatter plot
b) Biplot
c) Box plot
d) Q-Q plot

25 Which of the following is a benefit of using an 'open' Big Data reference architecture?

a) Guarantees fast and accurate results analysis of raw data
b) Supports the use of ever changing solutions to solve similar problems
c) Non-restrictive and open to interpretation in terms of application
d) Provides a common language for stakeholders

26 Within the data management process, which of the following statements about the data improvement and validation activity of the data management process is/are true?

1. The objective is to reduce errors in data sets.
2. It triggers an alert if any corrupt data is detected.

a) Only 1 is true
b) Only 2 is true
c) Both 1 and 2 are true
d) Neither 1 or 2 is true

27 What type of statistic quantitatively describes or summarizes the features of a collection of information?

a) Descriptive
b) Summary
c) Expressive
d) Explanatory

28 What data governance activity determines the policies for how an organization's external contractors should manage data access, retrieval, storage, destruction and back up, to ensure management and protection of data is maintained?

a) Develop quality strategy
b) Review regulatory and privacy requirements
c) Develop data governance policies
d) Assign roles and responsibilities

29 Which of the following is **NOT** a core component in the Hadoop Architecture?

a) NameNode
b) MapReduce
c) Slave Node
d) Job Seeker

30 In dispersion statistics, what is indicated by a high standard deviation for a set of data values?

a) Data entries are all very good quality
b) Values are close to those expected
c) Data is spread out over a wide range
d) Values are clustered near to the average

31 What step in the data analysis process would use a look-up table to cross-reference and correct any area codes that have been entered in the wrong format?

a) Data identification
b) Data collection and sourcing
c) Data review
d) Data cleansing

32 Why is skewness important in data science?

a) Highlights possible distortion of results
b) Establishes limits for deviations from the mean
c) Allows multiple data sets to be combined
d) Presents results from several different angles

33 What type of analytics should be used by a manufacturing company to understand why a certain product performs well in the South East Asia market?

a) Descriptive analytics
b) Diagnostic analytics
c) Predictive analytics
d) Prescriptive analytics

34 How is the data identification graph used in the data analysis process?

a) To collect data from various sources and determine the value of the data
b) To determine whether there are any problems or issues in the data set
c) To determine if the data set contains missing values
d) To identify where the raw data might be obtained

35 In what sequence did the following capabilities of Big Data evolve over time?

1. Location-aware and person-centred analysis
2. Data mining and statistical analysis
3. Web analytics and web intelligence

a) 1, 2, 3
b) 2, 1, 3
c) 2, 3, 1
d) 3, 2, 1

36 Which of the following statements about Big Data analysis architecture is/are true?

1. Offline analysis is more expensive than real-time analysis.
2. Offline analysis enables multiple batches of data to be processed concurrently.

a) Only 1 is true
b) Only 2 is true
c) Both 1 and 2 are true
d) Neither 1 or 2 is true

37 What measures are used to prioritize Use Cases when formulating a Big Data strategy?

a) Level of business involvement and stakeholder buy in
b) Current performance against predefined measures for success
c) Number of user groups affected, and the data sources required
d) Anticipated benefit to users and capability to implement

38 Temperatures (0°c) recorded in London over a week in June are shown in the table below. The mean temperature during this period was 19°c.

	Sun	Mon	Tue	Wed	Thu	Fri	Sat	**Sum**
Temp °c	17	17	18	18	21	20	22	**133**
- 19°c	-2	-2	-1	-1	2	1	3	
(Temp -19°c)2	4	4	1	1	4	1	9	**24**

Result: 24°c / 7 Days = 3.4°c

What measure is represented by the value 3.4?

a) Range
b) Interquartile Range
c) Variance
d) Standard Deviation

39 What is the purpose of a Job Tracker within the Hadoop framework?

a) To follow directions for processing jobs
b) To initiate and coordinate processing jobs
c) To keep track of where data is located
d) To load and analyze the final result

40 A bank has decided to formulate a Big Data strategy following the steps outlined in the Big Data Framework. They have identified several potential Use Cases where Big Data might provide long-term strategic value to the organization. 27 Use Cases have been identified, which is more than the organization can accommodate in terms of budget and resources.

What would be the **BEST** next step to take?

a) Review and prioritize every Use Case based on business impact, budget and resource requirements
b) Formulate a Big Data Roadmap to identify which projects will be executed first
c) Execute a current state assessment to help align the Use Cases to current business goals
d) Find correlations across multiple data sources to identify which Use Cases are most feasible to execute

41 What distribution shape would show how likely it is a customer will buy a new product after an email campaign?

a) Frequency
b) Probability
c) Sampling
d) Normal

42 A high resolution X-ray of a patient's broken bone also includes data on the patient's name, date of birth and patient number. What is the date of birth an example of?

a) Data bias
b) Data visualization
c) Metadata
d) Unstructured data

43 Identify the missing words in the following sentence.

The statistical process of standardization is important in Big Data because it makes it possible to [?] within a data set. eliminate bias identify relationships compare mulitple variables identify outliers

a) eliminate bias
b) identify relationships
c) compare multiple variables
d) identify outliers

44 Which of the following statements is/are true?

1. Correlation estimates the effect of independent variables on the dependent variable in order to make predictions.
2. Regression explores the reason why one variable causes a change in another variable.

a) Only 1 is true
b) Only 2 is true
c) Both 1 and 2 are true
d) Neither 1 or 2 is true

45 What sort of algorithm could be used to give a diagnosis to a patient based on observed symptoms?

a) Calculation
b) Correlation
c) Clustering
d) Classification

46 Which of the following statements about the differences between local and distributed storage solutions is/are true?

1. Business Intelligence analysis tools are best suited to distributed storage solutions.
2. For data held in local data storage solutions it is more appropriate to use analytics and visualization tools.

a) Only 1 is true
b) Only 2 is true
c) Both 1 and 2 are true
d) Neither 1 or 2 is true

47 What do classification algorithms do?

a) Indicate whether a relationship exists between variables
b) Group sets of objects based on similarities in their characteristics
c) Identify which known category a new observation belongs to
d) Condense large data sets into summary graphs

48 What do clustering algorithms do?

a) Indicate whether a relationship exists between variables
b) Group sets of objects based on similarities in their characteristics
c) Identify which known category a new observation belongs to
d) Condense large data sets into summary graphs

49 In a distributed storage system, what has been developed to speed up access to data stored across thousands of servers?

a) File systems
b) NoSQL Databases
c) Programming models
d) Direct Attached Storage (DAS)

50 Which of the following statements is/are true?

1. The primary purpose of data analysis is to review existing data in order to support decision-making.
2. The primary purpose of analytics is to analyze data sets in order to optimize the future.

a) Only 1 is true
b) Only 2 is true
c) Both 1 and 2 are true
d) Neither 1 or 2 is true

51 Which of the following is **NOT** one of the six main elements within the Big Data Framework?

a) Big Data Architecture
b) Big Data Algorithms
c) Big Data Methodologies
d) Big Data Strategy

52 What class of machine learning is used by a website recommendation engine to suggest to an online customer additional products for purchase as they move to the payment screen? The suggestions are based on profiles of other customers who have made similar purchases.

a) Structured
b) Supervised
c) Unsupervised
d) Unstructured

53 What is the function of the System Orchestrator within the NIST Big Data Reference Architecture?

a) Introduce new data or information feeds into the Big Data system
b) Ensure that the components of Big Data environments work together
c) Use business logic and functionality to transform data into the desired results
d) Store and process data based on designs that are optimized for Big Data

54 What Big Data process covers the assignment of roles and responsibilities for maintaining a consistent and proper handling of data across the business?

a) Data analysis
b) Data management
c) Data governance
d) Data organization

55 Which Big Data Team role needs the ability to come up with creative ideas to help them design and develop algorithms?

a) Big Data Analyst
b) Big Data Scientist
c) Big Data Engineer
d) Big Data Architect

56 In Artificial Intelligence, which essential capability exploits the outcome from studies about how humans deal with problems in order to model reasoning?

a) Automated reasoning
b) Knowledge representation
c) Machine learning
d) Natural language processing

57 What is correlation?

a) A causal relationship between two data sets
b) The conversion of data points to a standardized value
c) A statistical relationship between two random variables
d) The difference between the largest and smallest values in a data set

58 In cognitive analytics, what is it that makes decisions and takes specific actions, based on the perceived environment and patterns of specific users?

a) Intelligent algorithm
b) Active mediator
c) Human operator
d) Rational agent

59 According to the Turing test, what is used to assess a machines ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human?

a) Mathematical equations
b) Scientific algorithms
c) Written questions
d) Verbal communication

60 The CIO of a fast-growing company in the telco industry has been tasked to build up Big Data capabilities for the company. She has done some research into possible approaches that she can take to accomplish this.

What is the **BEST** approach that she can take?

a) Hire experienced data scientists from other organizations
b) Initiate the formation of a Big Data Centre of Excellence
c) Design processes for Big Data Analysis, Data Management and Data Governance
d) Identify the most suitable Big Data tool for her organization