BIG DATA FRAMEWORK®

The Enterprise Big Data Professional Examination

Rationale

1 CO0102 - Big Data Concepts and Key Drivers

B

Recall the names of the four characteristics of Big Data.

a) Incorrect: See Rationale B.
b) Correct: The characteristics of Big Data are commonly referred to as the four Vs: Volume (size of the data sets), Velocity (speed with which data is generated), Variety (sources and structure) and Veracity (quality of the data). Ref: 1.4
c) Incorrect: See Rationale B.
d) Incorrect: See Rationale B.

2 ST0201 - Big Data Strategy

C

Understand the primary reason organizations struggle to realize a competitive advantage through Big Data.

a) Incorrect: Organizations may be treating information as a strategic asset but with the growth of Big Data and Big Data solutions, the reason why so many companies are struggling to realize their competitive advantage through Big Data is because they have not (adequately) defined a Big Data strategy. Ref: 3.1.
b) Incorrect: Less than half an organization's structured data is actively used in making decision – and less than 1% of its unstructured data is analyzed or used at all. Ref: 3.1
c) Correct: The reason why so many companies are struggling to realize their competitive advantage through Big Data is because they have not (adequately) defined a Big Data strategy. In many organizations, Big Data is still project-based, instead of embedded into the veins of the organization. In order to avoid these pitfalls and realize a long-term competitive advantage, the Big Data Framework starts with defining and formulating a Big Data Strategy. Ref: 3.1
d) Incorrect: Organizations are competing on the preciseness of their data but this not a reason why organizations are struggling to realize a competitive advantage from Big Data. See Rationale C.

3 AL0202.2 - Big Data Algorithms

D

Recall the characteristics of negative skew.

a) Incorrect: See Rationale D.
b) Incorrect: See Rationale D.
c) Incorrect: A distribution is positively skewed when the tail of the curve is longer on the right side or skewed to the right, and the mean is greater than the median and mode. The majority of the values exist on the left side of the curve. Ref: 5.2
d) Correct: A distribution is negatively skewed when the tail of the curve is longer on the left side or skewed to the left, and the mean is less than the median and mode. The majority of the values exist on the right side of the curve. Ref: 5.2

4 CO0103.1 - Big Data Concepts and Key Drivers

C

To recall the techniques commonly associated with supervised machine learning.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: Supervised machine learning is mostly associated with classification and regression techniques. Ref: 1.8
d) Incorrect: See Rationale C.

5 CO0206.1 - Big Data Concepts and Key Drivers

B

Understand the characteristics of structured data.

a) Incorrect. (4) Structured data is data that adheres to a pre-defined data model and is therefore straightforward to analyze. Structured data conforms to a tabular format with relationships between the different rows and columns. Ref. 1.6
b) Correct. (3). Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in structures databases. Ref. 1.6
c) Incorrect. (2). Common examples of structured data are Excel files or SQL databases. Ref: 1.6
d) Incorrect. (1). Structured data is data that adheres to a pre-defined data model and is therefore straightforward to analyze. Structured data conforms to a tabular format with relationship between the different rows and columns. Ref. 1.6

6 FU0203 - Big Data Functions

D

Understand the six organization success factors for Big Data.

a) Incorrect: Establish a vision on how to create value is the first milestone to gain a clear vision of what the organization is trying to accomplish. As several projects have been completed this should already have bee done. Ref: 7.5
b) Incorrect: This is about making it clear from the very beginning who is responsible for what and designing effective data governance and data management processes. As several projects have been completed this should already have been done. Ref: 7.5
c) Incorrect: A centralized BDCoE provides a uniform point where expertise about Big Data practices and technologies is combined. As this organization is at early days with its Big Data implementation and only a few projects have been run it is unlikely to be ready to set a BDCoE up yet. Ref: 7.5
d) Correct: Knowledge and skills are the most important key to success, yet one of the most difficult elements to obtain. Setting up an ongoing Big Data education program will increase the competency of the organization and embeds a culture of continuous learning. Ref: 7.5

7 AI0203.2 - Artificial Intelligence

C

Understand where deep learning is predominantly used.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: Deep Learning breaks down raw data into a number of layers and subsequently compares these layers with each other. Using this technique, it becomes more efficient to break down large data sets into structured pieces of information that can be analyzed. Deep Learning is therefore predominantly used in processing images, video, speech and audio. Ref: 8.5
d) Incorrect: See Rationale C.

8 FR0202.4 - The Big Data Framework

C

Understand the characteristics of the five levels of Big Data maturity.

a) Incorrect: At level 2 there are pockets of analytics across the enterprise, however functioning in silos and no overarching data or analytics strategy. Ref: 2.4
b) Incorrect: At level 3 there are expanding siloed functional analytics to shared operational level analytics with support and commitment from the C-suite. Ref: 2.4
c) Correct: At level 4 data and analytics are viewed as an enterprise priority. The organization is developing enterprise wide analytics capabilities across all domains to create meaningful content and ideas. Ref: 2.4
d) Incorrect: At level 5 there is trusted insight created by enterprises with analytics that support strategic decision-making. The enterprise is reaping the benefits and is focused on optimization of analytics. Ref: 2.4

9 AI0102.2 - Artificial Intelligence

A

Recall the two main reasons cognitive analytics differentiates from other forms of analytics.

a) Correct: Cognitive analytics makes decisions based on the perceived environment and personalized characteristics. Ref: 8.3
b) Incorrect: The perceived environment is determined using specific information. It is not a general assessment. Ref: 8.3
c) Incorrect: The perceived environment is determined using specific information. It is not a general assessment. Ref: 8.3
d) Incorrect: Cognitive analytics makes decisions based on the perceived environment and personalized characteristics. Ref: 8.3

10 ST0101 - Big Data Strategy

A

Recall the approach to formulating a Big Data strategy.

a) Correct. Formulating a Big Data strategy can be reviewed as a five-step approach, and effectively consists: 1. Define business objectives; 2. Execute a current state assessment; 3. Identify and prioritize Use Cases; 4. Formulate a Big Data Roadmap; and 5. Embed through Change Management. Ref. 3.3
b) Incorrect. See Rationale A.
c) Incorrect. See Rationale A.
d) Incorrect. See Rationale A.

11 AL0207 - Big Data Algorithms

B

Understand the implications of population size for sampling Big Data.

a) Incorrect: See Rationale B.
b) Correct: With Big Data, it becomes possible to analyze massive quantities of data. The larger the data set become, the closer it is to the actual population, and the less likely it is that the data set becomes biased. Ref: 5.3
c) Incorrect: See Rationale B.
d) Incorrect: See Rationale B.

12 AR0101 - Big Data Architecture

A

Recall what a reference architecture is.

a) Correct: In summary, a reference architecture can be thought of as a resource that documents the learning experiences gained through past projects. Ref: 4.2
b) Incorrect: See Rationale A.
c) Incorrect: See Rationale A.
d) Incorrect: See Rationale A.

13 CO0101 - Big Data Concepts and Key Drivers

D

Recall the definition of Big Data.

a) Incorrect: See Rationale D.
b) Incorrect: See Rationale D.
c) Incorrect: See Rationale D.
d) Correct. This is the definition of Big Data used in the Body of Knowledge. Ref 1.1

14 AR0204.1 - Big Data Architecture

C

Recognize examples of the 3 types of storage systems for massive data.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: Examples of DAS include hard drives, solid-state drives, optical disc drives, and storage of external drives. Ref: 4.4
d) Incorrect: See Rationale C.

15 AR0102.2 - Big Data Architecture

C

Recall the names of the roles in the structure of the NIST Big Data reference architecture.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: The Data Provider role introduces new data or information feeds into the Big Data system for discovery, access, and transformation by the Big Data system. Ref: 4.2
d) Incorrect: See Rationale C.

16 AL0208 - Big Data Algorithms

C

Understand why correlations are useful in Big Data Science.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: Correlation is any of a broad class of statistical relationships involving dependence, although it is mostly used to indicate whether two variables have a linear relationship. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. Ref: 5.4
d) Incorrect: Regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Ref: 5.5

17 CO0204.4 - Big Data Concepts and Key Drivers

A

Understand the characteristic Veracity and how they distinguish Big Data from traditional data analysis.

a) Correct. Veracity refers to the quality of the data that is being analyzed. High veracity has many records that are valuable to analyze and that contribute in a meaningful way to the overall results. Ref: 1.4
b) Incorrect: Variety makes Big Data really big. Big Data comes from a great variety of sources and generally has in three types: structured, semi structured and unstructured (as discussed in the next section). Ref: 1.4
c) Incorrect: Variety makes Big Data really big. Big Data comes from a great variety of sources and generally has in three types: structured, semi structured and unstructured (as discussed in the next section). Ref: 1.4
d) Incorrect: The volume of data refers to the size of the data sets that need to be analyzed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. Ref: 1.4

18 FU0101.2 - Big Data Functions

A

Recall the key characteristics of the five major pillars of a Big Data Centre of Excellence.

a) Correct: An important requirement for the Big Data labs is to have the hardware compatible for Big Data processing. In general, Big Data labs require hardware with sufficiently larger RAM than usual for Big Data processing. Ref: 7.2
b) Incorrect: the environment should be a creative space to experiment and run test data analyses in order to achieve the desired results. Ref: 7.2
c) Incorrect: A well-designed Big Data lab contains isolated work possibilities where data analysts can ‘crunch the number’ without distractions. Ref: 7.2
d) Incorrect: A well-designed Big Data lab contains open work-spaces that allow for communication and collaboration. Ref. 7.2

19 PR0201.3 - Big Data Processes

B

Understand the characteristics of the six types of problems that shape the business objectives of Big Data Projects (3. Inferential)

a) Incorrect: A descriptive business objective aims to summarize the characteristics of different datasets, originating from either inside or outside the enterprise. The business objective is to collect and summarize data in order to make decisions. Ref: 6.2
b) Correct: An inferential business objectives aim to find characteristics about a population by studying a sample of the data. Inferential business objectives are prevalent in targeting (potential) customers in marketing and sales organizations within the enterprise. Ref: 6.2
c) Incorrect: A predictive business objective aims to predict future behaviors by analyzing and extrapolating data sets, such as predicting which products customers are likely to buy. With predictive objectives, the outcome is uncertain and Big Data is used in order to find the best possible answer. Ref: 6.2
d) Incorrect: A mechanistic business objective aims to find how variables influence outcomes of data sets. It requires a deeper understanding of the underlying relationships and patterns within data sets. Ref: 6.2

20 AR0207 - Big Data Architecture

D

Understand how Hadoop overcomes the risk of losing data in Big Data environments.

a) Incorrect: Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. Ref: 4.6
b) Incorrect: Parallel programming models have been developed that increase the performance of NoSQL databases, they do not overcome the risk of losing data. Ref: 4.4
c) Incorrect: Whilst SANs are primarily used to enhance storage devices this is not to overcome the risk of losing data. Ref: 4.4 and 4.6
d) Correct: One of the core properties of the Hadoop is the Hadoop Distributed Filing System where each of the data parts is replicated multiple times and distributed across multiple nodes within the cluster. If one node fails, another node has a copy of that specific data package that can be used for processing. Because of this property, data can still be processed and analyzed even when one of the nodes fails because of a hardware failure. Ref: 4.6

21 AL0212 - Big Data Algorithms

A

Understand how outlier detection is used in the context of Big Data.

a) Correct: Outliers are generally data points that appear to be unexpected in comparison with the rest of the data – they do not fit into the pattern of the other data points. Ref: 5.8
b) Incorrect: See Rationale A.
c) Incorrect: See Rationale A.
d) Incorrect: See Rationale A.

22 FR0201 - The Big Data Framework

A

Understand the relevance of Big Data Strategy in establishing a Big Data organization.

a) Correct: In order to achieve tangible results from investments in Big Data, enterprise organizations need a sound Big Data strategy. How can return on investments be realized, and where to focus effort in Big Data analysis and analytics? (Ref: 2.2)
b) Incorrect: See Rationale A.
c) Incorrect: See Rationale A.
d) Incorrect: See Rationale A.

23 CO0207 - Big Data Concepts and Key Drivers

B

Understand the role of Hadoop in distributed storage and processing.

a) Incorrect. Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. Ref. 1.7.
b) Correct. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework. Ref. 1.7.
c) Incorrect. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework. Ref. 1.7.
d) Incorrect. It is important to note that most Big Data solutions make use of the Hadoop framework as their underlying software framework. The term ‘Hadoop’ has therefore also become known as the eco-system that connects different Big Data solutions (and commercial vendors) together. Ref. 1.7.

24 AL0213.5 - Big Data Algorithms

C

Understand why box plots are useful in Big Data.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: A box plot is very useful in Big Data because it immediately shows the mean, median, mode, Q1 and Q3 values, and any potential outliers. It captures and communicates key information very quickly. Ref: 5.9
d) Incorrect: See Rationale C.

25 AR0201 - Big Data Architecture

D

Understand the benefits of using a Big Data reference architecture.

a) Incorrect: A reference architecture is a document or set of documents to which a project manager or other interested party can refer to for best practices. It does not guarantee fast and accurate restuls. Ref: 4.2
b) Incorrect: It provides consistent methods for implementation of technology to solve similar problem sets. Ref: 4.2
c) Incorrect: It encourages adherence to common standards, specifications and patterns. Ref: 4.2
d) Correct: It provides a common language for the various stakeholders. Ref: 4.2

26 PR0204.3 - Big Data Processes

A

Understand the importance of the data improvement and validation step within the data management process and what occurs within it.

a) Correct: (1) The data improvement and validation activity concerns itself with 'cleaning' up the datasets in order to improve the metrics and performance indicators. This may be as a result of an alert from the previous activity: Monitor and management enterprise data. It is in this process activity that alerts are generated. Ref: 6.4
b) Incorrect: (2) Data sets are monitored against the performance indicators in the previous activity: Monitor and management enterprise data. It is in this process activity that alerts are generated. Ref: 6.4
c) Incorrect: See Rationales A and B.
d) Incorrect: See Rationales A and B.

27 AL0101 - Big Data Algorithms

A

Recall what descriptive statistics are.

a) Correct: Descriptive statistics are summary statistics that quantitatively describe or summarize features of a collection of information. Descriptive statistics provide key values that quickly summarize datasets, and are understandable for everyone that is working with the data. Ref: 5.2
b) Incorrect: See Rationale A.
c) Incorrect: See Rationale A.
d) Incorrect: See Rationale A.

28 PR0203.3 - Big Data Processes

C

Understand the importance of the data governance policies activity within the data governance process.

a) Incorrect: This sets the strategic objectives for the management and direction of all data quality activities. The actual polices are set when developing governance policies. Ref: 6.3
b) Incorrect: This activity looks at the data privacy laws of countries involved and regulatory requirements for data transfer. Ref: 6.3
c) Correct: Big Data Governance policies are publicly available for everyone in the organization. These policy documents contain the decisions the enterprise has taken with regards to their data quality organization and explain the requirements for everyone. Ref: 6.3
d) Incorrect: In this activity role assignment has to ensure accountability, authority and supervision as well as the involvement of senior executives and business management and encourage desirable behaviour in the use of data. Ref: 6.3

29 AR0103 - Big Data Architecture

D

Recall the names of the core components in the Hadoop Architecture.

a) Incorrect: The NameNode acts as a facilitator that communicates where data parts are stored and if they are available. Ref: 4.6
b) Incorrect: The MapReduce framework ensures that tasks are completed by enabling the parallel distributed processing of the data parts across the multiple nodes in the cluster. Ref: 4.6
c) Incorrect: Slave Nodes are the nodes in the cluster that follow directions from the Job Tacker. Ref: 4.6
d) Correct: The Job Tracker- is the node in the cluster that initiates and coordinates processing jobs. Ref: 4.6

30 AL0201.2 - Big Data Algorithms

C

Understand what a high standard deviation indicates.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: The standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. Ref: 5.2
d) Incorrect: See Rationale C.

31 PR0202.5 - Big Data Processes

D

Understand the importance of the data cleansing step within the data analysis process.

a) Incorrect: Data identification determines which data sets need to be processed. It is in the data cleansing step where look-up tables would be examined to correct errors. Ref: 6.2
b) Incorrect: In this step data is obtained for processing. It is in the data cleansing step where look-up tables would be examined to correct errors. Ref: 6.2
c) Incorrect: In this step data is examined to determine whether data sets have been corrupted or there are missing values or conflicting data. It is the data cleansing step where look-up tables would be examined to correct errors. Ref: 6.2
d) Correct: Data cleansing is the process of amending or removing data in a database that is incorrect, incomplete, improperly formatted or duplicated. Typically, a data-cleansing tool includes programs that are capable of correcting a number of specific type of mistakes, such as adding missing zip codes or finding duplicate records. Ref: 6.2

32 AL0206.4 - Big Data Algorithms

A

Understand why skew is important for Big Data.

a) Correct: Skewness in distributions is important in data science because the skewness can indicate potential bias (i.e. not an adequate representation of the actual data) in data sets. Ref: 5.2
b) Incorrect: See Rationale A.
c) Incorrect: See Rationale A.
d) Incorrect: See Rationale A.

33 CO0204.2 - Big Data Concepts and Key Drivers

B

Be able to identify the correct type of analytics to use for a data modelling situation.

a) Incorrect: Descriptive analytics is concerned with monitoring current state performance or results. Ref: 1.5
b) Correct: Diagnostic analytics would be appropriate when seeking to understand (quantify) drivers of performance. Ref: 1.5
c) Incorrect: Predictive analysis is used to forecast likely outcomes Ref: 1.5
d) Incorrect: Prescriptive analytics are used for recommending actions for future decisions. Ref: 1.5

34 PR0102.1 - Big Data Processes

D

Recall how the data identification graph is used in the data analysis process.

a) Incorrect: See Rationale D.
b) Incorrect: See Rationale D.
c) Incorrect: See Rationale D.
d) Correct: A data identification graph first identifies the desired (processed) data and then works backwards to identify where the raw data might be obtained. Ref: 6.2

35 CO0201 - Big Data Concepts and Key Drivers

C

Understand the sequence in which Big Data content has evolved.

a) Incorrect. See Rationale C.
b) Incorrect. See Rationale C.
c) Correct. Big Data Phase 1 was DBMS-based, Phase 2 was web-based, Phase 3 is now evolving to analyze mobile and sensor-based content. Ref. 1.3 and Fig. 1
d) Incorrect. See Rationale C.

36 AR0206.2 - Big Data Architecture

D

Understand the Big Data offline analysis architecture.

a) Incorrect: (1) Offline analysis is used for applications that are less time sensitive and for which the real-time value of data is less urgent. Offline processing (also known as batch processing) imports data on set times and subsequently processes it at time intervals. Most enterprises utilize the offline analysis architecture based on Hadoop in order to reduce costs and improve efficiency of data processing. (Ref: 4.5)
b) Incorrect: (2) The main existing architectures of real-time analysis include parallel processing clusters using traditional relational databases and memory-based computing platforms. Offline processing (also known as batch processing) imports data on set times and subsequently processes it at time intervals. (Ref: 4.5)
c) Incorrect: See Rationales A and B.
d) Correct: See Rationales A and B.

37 ST0202.2 - Big Data Strategy

D

Understand the structure of the Prioritization Matrix.

a) Incorrect: See Rationale D.
b) Incorrect: See Rationale D.
c) Incorrect: See Rationale D.
d) Correct: Prioritization Matrix shows use cases with both meaningful business value (from the business stakeholders' perspectives) and reasonable feasibility of successful implementation. Ref: 3.3

38 AL0204 - Big Data Algorithms

C

Recognize examples of descriptive statistics.

a) Incorrect: The range is the difference between the highest and the lowest value in a set of data (In the example: 22 –17 = 5). Ref: 5.2
b) Incorrect: The interquartile range (IQR), also called the midspread or middle 50%, or is a measure of dispersion, being equal to the difference between 75th and25th percentiles (QR = Q3 - Q1). In other words, the IQR is a statistic that indicates where the middle 50% of values are located (In the example: Q2 is 18, Q1 is 17, Q3 is 21, IQR: 21 – 17 = 4). Ref: 5.2
c) Correct: Variance is the expectation of the squared deviation of a random variable from its mean. The closer the variance is to zero, the more closely the data points are clustered together. (In the example: First, subtract 19o from each value. Next, square each result and sum the values = 24 (Row 3). The variance is the sum of the squared values / population, 24/7 = 3.4 o. Ref: 5.2
d) Incorrect: The standard deviation is a measure that is used to quantify the amount of variation or dispersion of a set of data value. The standard deviation is the square root of the variance. (In example: Sqrt of the variance [3.4] = 1.8439). Ref: 5.2

39 AR0208.4 - Big Data Architecture

B

Understand the purpose of a Job Tracker.

a) Incorrect. This is a Slave Node. Ref: 4.6
b) Correct: The Job Tracker is the node in the cluster that initiates and coordinates processing jobs. Additionally, the Job Tracker invokes the Map procedure and the Reduce method. Ref: 4.6
c) Incorrect: This is a Name Node. Ref: 4.6
d) Incorrect: The Job Tracker located the local results and aggregates these into a final result. The final result is loaded into a single node and can be loaded and analyzed. Ref: 4.6

40 ST0203.4 - Big Data Strategy

A

Understand the purpose of Step 4 - Prioritize Data Sources.

a) Correct: After the Use Cases have been developed, the next step is to prioritize all of the Use Cases based on their business impact, budget and resource requirements. By conducting this exercise, enterprises can identify which Big Data initiatives provide most business value. Ref: 3.3
b) Incorrect. Based on the current capability state assessment (step 2) and the identified and prioritized Big Data Use Cases (step 3), the Roadmap can be developed. In the scenario Step 3 prioritization of the Use Cases is still to be completed. Ref: 3.3
c) Incorrect. Executing a current state assessment is step 2 and is completed before prioritization of the Use Cases. Ref: 3.3
d) Incorrect. A correlation is a statistical technique that aims to find relationships between variables. A correlation does not indicate any priority. Ref 3.3

41 AL0205.2 - Big Data Algorithms

B

Understand the characteristics of the probability distribution shapes.

a) Incorrect: A frequency distribution is a table or graph that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample. Ref: 5.2
b) Correct: A probability distribution is a summary graph that depicts the likelihood of all potential outcomes. It is a mathematical function that can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. Ref: 5.2
c) Incorrect: A sampling distribution is the probability distribution of a given statistic based ona random sample. Ref: 5.2
d) Incorrect: A normal distribution represents data that occurs commonly where most values are the same as the average value and only few values are found at the extremities. In a normal distribution, approximately 99% of the values are within three standard deviations of the mean, and the area under the curve is equal to one. Ref: 5.2

42 CO0205 - Big Data Concepts and Key Drivers

C

Understand the function of metadata in Big Data environments.

a) Incorrect: Bias is where the sample size will result in inadequate or wrong predictions about the future, because in inferential statistics assumptions are made about the entire population based on the sample. Ref: 5.3
b) Incorrect: Data visualization condenses large data sets to summary graphs that are easy to understand and easy to discuss. Ref: 5.9
c) Correct. Metadata is data about data. It provides additional information about a specific set of data. In a set of photographs, for example, metadata could describe when and where the photos were taken. The metadata than provides fields for dates and locations which, by themselves, can be considered structured data. Ref: 1.6
d) Incorrect: Metadata can additional information about a specific set of data. But by itself can be considered structured data. Ref: 1.6

43 AL0203 - Big Data Algorithms

C

Understand why standardization is important in Big Data.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: Standardization is one of the most important processes when analyzing Big Data, because standardized score or values allow different variables to be combined. Ref: 5.2
d) Incorrect: See Rationale C.

44 AL0209 - Big Data Algorithms

D

Understand the differences between correlation and regression.

a) Incorrect: Correlation only indicates if a relationship exists between variables. Ref: 5.5
b) Incorrect: Regression is a set of statistical processes for estimating the relationships among variables. Neither simple linear regression nor correlation answer questions of causality. Ref: 5.5
c) Incorrect: See Rationale A and B.
d) Correct: See Rationale A and B.

45 AL0201 - Big Data Algorithms

D

Recognize examples of a classification algorithm.

a) Incorrect: See Rationale D.
b) Incorrect: See Rationale D.
c) Incorrect: See Rationale D.
d) Correct: Classification is the problem of identifying to which of a set of categories a new observation belongs, based on a training set of data containing observations whose category membership is known. Ref: 5.6

46 AR0203 - Big Data Architecture

D

Understanding the differences between local and distributed storage and processing.

a) Incorrect: Because the data in the data warehouse (local storage) is neatly structured, a business intelligence analysis tool (examples are SAP Business Objects or IBM Cognos) can subsequently run queries and provide reports that provide the requested information and insights. Ref: 4.3
b) Incorrect: Analytics and visualization tools are used to display the results of distributed processing and are one of the four design principles of Big Data Architectures. Ref: 4.3
c) Incorrect: See Rationale A and B.
d) Correct: See Rationale A and B.

47 AL0103.1 - Big Data Algorithms

C

Recall what classification is.

a) Incorrect: This is correlation. See Rationale C.
b) Incorrect: This is clustering. See Rationale C.
c) Correct: Classification is the problem of identifying to which of a set of categories a new observation belongs, based on a training set of data containing observations whose category membership is known. Because the computer is ‘fed’ sample data, classification is a form of supervised machine learning. Ref: 5.6
d) Incorrect: This is data visualization. See Rationale C.

48 AL0211.2 - Big Data Algorithms

B

What form of machine learning is associated with clustering.

a) Incorrect: This is correlation. See Rationale B.
b) Correct: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). Ref: 5.7
c) Incorrect: This is classification. See Rationale B.
d) Incorrect: This is data visualization. See Rationale B.

49 AR0205.3 - Big Data Architecture

C

Understand the storage mechanisms for Big Data.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: Big Data are generally stored on hundreds or thousands of commercial servers. In order to access the data stored on these servers, parallel programming models have been developed that increase the performance of NoSQL databases. Ref: 4.4
d) Incorrect: See Rationale C.

50 CO0203 - Big Data Concepts and Key Drivers

C

Understand the purpose of data analysis and analytics.

a) Incorrect. See Rationale C.
b) Incorrect. See Rationale C.
c) Correct. Both 1 and 2 are true. The primary purpose of data analysis is to review existing data in order to describe patterns that have happened in the past. Whereas data analysis aims to support decision-making by reviewing past data, analytics is primarily concerned with optimizing the future. For this purpose, analytics makes use of algorithms to find patterns in data in order to provide advice on the best possible course of action. Ref. 1.5
d) Incorrect. See Rationale C.

51 FR0101 - The Big Data Framework

C

Recall the six capabilities of the Big Data Framework.

a) Incorrect. Big Data Architecture. In order to work with massive data sets, organizations should have the capabilities to store and process large quantities of data. In order to achieve this, the enterprise should have the underlying IT infrastructure to facilitate Big Data. Ref: 2.2
b) Incorrect: Big Data Algorithms. A fundamental capability of working with data is to have a thorough understanding of statistics and algorithms. Ref: 2.2
c) Correct: The six main elements are Big Data Strategy, Big Data Architecture, Big Data Algorithms, Big Data Processes, Big Data Functions and Artificial Intelligence. Ref: 2.2
d) Incorrect: Big Data Strategy. In order to achieve tangible results from investments in Big Data, enterprise organizations need a sound Big Data strategy. Ref: 2.2

52 CO0208.1 - Big Data Concepts and Key Drivers

B

To be able to recognize example of supervised machine learning.

a) Incorrect: Structured is not a type of machine learning. Ref: 1.8
b) Correct: In supervised machine learning a computer learns a certain task because it is fed labelled training data. In other words, the computer is first confronted with a number of ‘sample cases’, from which it learns what decision to make. In this case it ‘learns’ what recommendation to make based on data about other customer purchases Ref: 1.8
c) Incorrect: Unsupervised – a computer is fed data and needs to infer relationships in the data without any prior knowledge about the data set. Ref: 1.8
d) Incorrect: Unstructured is not a type of machine learning. Ref: 1.8

53 AR0202.5 - Big Data Architecture

B

Understand the functions and activities associated with the logical role System Orchestrator

a) Incorrect. This is the function of the Data Provider. See Rationale B.
b) Correct: Orchestration ensures that the different applications, data and infrastructure components of Big Data environments all work together. In order to accomplish this, the System Orchestrator makes use of workflows, automation and change management processes. Ref: 4.2
c) Incorrect: The Big Data Application Provider is the architecture component that contains the business logic and functionality that is necessary to transform the data into the desired results. See Rationale B.
d) Incorrect: This is the function of the Big Data Framework Provider. See Rationale B.

54 PR0101 - Big Data Processes

C

Recall the key characteristics of the Big Data Governance process.

a) Incorrect: The data analysis process contains the sequential steps enterprises take in order to process Big Data. Ref: 6.2
b) Incorrect: The data management process safeguards the quality of the data on a day-to-day operational level. Ref: 6.2
c) Correct: The data governance process must define clear roles and responsibilities across divisional boundaries in the enterprise. The role assignment has to ensure accountability, authority and supervision as well as the involvement of senior executives and business management and encourage desirable behavior in the use of data. Ref: 6.3
d) Incorrect: This is not one of the three sub-processes. See Rationale C.

55 FU0202.2 - Big Data Functions

B

Understand the typical responsibilities and skill sets of the Data Scientist.

a) Incorrect: See Rationale B.
b) Correct: The role of the data scientist requires creative thinking and problem solving skills that are necessary to design, develop and deploy algorithms that can retrieve value from Big Data. Ref: 7.3
c) Incorrect: See Rationale B.
d) Incorrect: See Rationale B.

56 AI0202.2 - Artificial Intelligence

B

Understand the knowledge representation capability in artificial intelligence and the key challenges involved.

a) Incorrect: Automated reasoning in Artificial Intelligence is the knowledge capability that concerns itself with understanding reasoning capabilities in computer systems. The goal of automated reasoning is to design computer systems that can reason completely automatically (without human involvement). Ref: 8.4
b) Correct: Knowledge representation incorporates findings from psychology about how humans solve problems and represent knowledge in order to design logical statements that make complex systems easier to design and build. As such, it heavily relies on the application of logic in order to model reasoning. Ref: 8.4
c) Incorrect: The objective of machine learning is to design a system that improves and gets better over time. Just like humans memorize information or relationships when they are presented to them, so can computer systems learn from previous interactions. Ref: 8.4
d) Incorrect: Natural Language Processing (NLP) is the domain that defines the interactions between computers and (natural) human languages, so that people can interact with the computer. NLP needs to detect the language of the person, detect the sequences of word and potential detect emotions in the way the message is communicated. Ref: 8.4

57 AL0102.1 - Big Data Algorithms

C

Recall key facts about correlation.

a) Incorrect: The presence of a correlation is not sufficient to infer the presence of a causal relationship. Also correlation is about the relation between two variables, not the relationship between data sets. Ref: 5.4
b) Incorrect: See Rationale C.
c) Correct: any statistical relationship, whether causal or not, between two random variables or bivariate data. Correlation is any of a broad class of statistical relationships involving dependence, although it is mostly used to indicate whether two variables have a linear relationship. Ref: 5.4
d) Incorrect: See Rationale C.

58 AI0101 - Artificial Intelligence

D

Understand the role of rational agents in cognitive analytics.

a) Incorrect: See Rationale D.
b) Incorrect: See Rationale D.
c) Incorrect: See Rationale D.
d) Correct: An agent perceives data from a specific environment through one of more sensors. The agent subsequently processes this data and subsequently takes a specific action. The decision is autonomous, and similar to the decision a human would take in similar circumstances. Ref: 8.3

59 AI0101.1 - Artificial Intelligence

C

Recall the operational definition of intelligence according to the Turing test.

a) Incorrect: See Rationale C.
b) Incorrect: See Rationale C.
c) Correct: In the Turing test, the interrogator is limited to using written questions. Ref: 8.1
d) Incorrect: See Rationale C.

60 FU0201 - Big Data Functions

B

Understand the benefits of a Big Data Centre of Excellence.

a) Incorrect: Hiring experienced data scientists will help but for obtaining long-term value from big data and become a truly ‘data driven’ organization, it is crucial to set up a Big Data Centre of Excellence. Ref: 7.2
b) Correct: To obtain long-term value from big data and become a truly ‘data driven’ organization, it is crucial to set up a Big Data Centre of Excellence. Ref: 7.2
c) Incorrect: Designing Big Data processes should be something that is done from the start of a Big Data initiative. To obtain long-term value from big data and become a truly ‘data driven’ organization, it is crucial to set up a Big Data Centre of Excellence. Refs: 7.2 & 7.5
d) Incorrect: Procuring appropriate Big Data tools will be part of the Big Data Architecture but are not the best approach for building up capability. To obtain long-term value from big data and become a truly ‘data driven’ organization, it is crucial to set up a Big Data Centre of Excellence. Ref: 7.2