CompTIA DA0-002 Exam Questions

Total Questions: 100+
Demo Questions: 25
Version: Updated for 2025

Prepared and Verified by Cert Empire – Your Trusted IT Certification Partner

For Access to the full set of Updated Questions – Visit:
CompTIA DA0-002 Exam Questions by Cert Empire

Question: 1

Data Concepts and Environments

Which of the following best enables the retrieval and manipulation of data that is stored in a relational database?

A. XML

B. SQL

C. Excel

D. JavaScript

Answer:

B

Explanation:

SQL (Structured Query Language) is the standard, domain-specific language designed for managing data held in a relational database management system (RDBMS). Its core components include the Data Query Language (DQL) for retrieval (e.g., SELECT statement) and the Data Manipulation Language (DML) for modifying data (e.g., INSERT, UPDATE, DELETE). It is the fundamental and most direct way to interact with relational data structures.

Why Incorrect Options are Wrong:

A. XML: XML (eXtensible Markup Language) is a format for encoding and transporting data, not a query language for manipulating data within a relational database.
C. Excel: Excel is a spreadsheet application. While it can connect to and import data from databases, it is not the underlying language used for retrieval and manipulation.
D. JavaScript: JavaScript is a general-purpose scripting language, primarily for web development, not the specialized language for querying relational databases.

References:

1. Silberschatz, A., Korth, H. F., & Sudarshan, S. (2019). Database System Concepts (7th ed.). McGraw-Hill. In Chapter 3, "Introduction to SQL," the text states, "SQL is the most influential commercially marketed and implemented query language... The data-manipulation language (DML) is the subset of SQL that allows users to query the database and to modify the database." (p. 81, Section 3.1).
2. Garcia-Molina, H., Ullman, J. D., & Widom, J. (2008). Database Systems: The Complete Book (2nd ed.). Pearson. Chapter 6, "SQL: The Query Language," introduces SQL as "the principal language used to describe and manipulate data in a relational database." (p. 221, Chapter 6

Introduction).
3. The PostgreSQL Global Development Group. (2023). PostgreSQL 16 Documentation. In Part I, Chapter 2, "The SQL Language," the introduction states its purpose is to "perform simple operations" on the database, proceeding to demonstrate data retrieval (SELECT) and manipulation (INSERT). (Section 2.1, "Introduction").

Question: 2

Visualization and Reporting

A data analyst is designing a report for the business review team. The team lists the following requirements for the report:

• Specific data points

• Color branding

• Labels and terminology

• Suggested charts and tables

Which of the following components is missing from the requirements?

A. Source validation

B. Design elements

C. Delivery method

D. Report type

Answer:

C

Explanation:

The provided requirements focus on the content and aesthetics of the report, such as which data to show (specific data points), how to present it (charts and tables, labels), and the visual style (color branding). However, the list completely omits the logistical aspects of how the report will be distributed and accessed by the end-users. The delivery method, which includes the format (e.g., static PDF, interactive dashboard), frequency (e.g., daily, weekly, on-demand), and medium (e.g., email, BI portal, embedded in another application), is a critical requirement that fundamentally influences the report's design, tools, and implementation.

Why Incorrect Options are Wrong:

A. Source validation: This is a crucial data governance activity performed by the analyst to ensure data quality, not typically a design requirement specified by the business team.
B. Design elements: This is incorrect as the requirements explicitly list several design elements, including "Color branding" and "Suggested charts and tables."
D. Report type: The context ("business review team") and specified content ("charts and tables")

strongly imply the report type, such as a performance or analytical report.

References:

1. Wexler, S., Shaffer, J., & Cotgreave, A. (2017). The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios. Wiley. In Chapter 2, the authors emphasize understanding the audience and their needs, which includes how they will consume the information. The choice between a static report and an interactive dashboard (a key aspect of the delivery method) is presented as a foundational design decision.

2. Microsoft. (2023). Power BI adoption roadmap: Content delivery. Microsoft Learn. This official vendor documentation for a major BI tool treats content delivery as a distinct and critical stage in the business intelligence lifecycle, separate from content creation. It outlines various strategies for sharing and distributing reports, underscoring its importance as a separate requirement.

3. Turban, E., Sharda, R., & Delen, D. (2014). Business Intelligence and Analytics: Systems for Decision Support (10th ed.). Pearson. In Chapter 3, "Data Warehousing, Business Intelligence, and Data Analytics," the authors describe the components of a BI architecture. The "Business Analytics (BA) user interface" layer explicitly includes methods for report distribution and interaction, identifying it as a key component that must be defined during the requirements phase.

Question: 3

Data Acquisition and Preparation

A data analyst needs to remove all duplicate values between two tables, "Employees" and "Managers," using SQL SELECT statements. Which of the following should the analyst use for this task?

A. SELECT * FROM Employees UNION ALL SELECT * FROM Managers

B. SELECT * FROM Employees UNION SELECT * FROM Managers

C. SELECT * FROM Employees JOIN SELECT * FROM Managers

D. SELECT * FROM Employees CROSS JOIN SELECT * FROM Managers

Answer:

B

Explanation:

The SQL UNION operator is designed to combine the result sets of two or more SELECT statements into a single result set. A key characteristic of the UNION operator is that it automatically removes duplicate rows from the final result. The analyst's task is to merge records from the "Employees" and "Managers" tables and eliminate any duplicates between them. The statement SELECT FROM Employees UNION SELECT FROM Managers correctly performs this operation by vertically appending the two tables and filtering out any identical rows, thus achieving the desired outcome.

Why Incorrect Options are Wrong:

A. UNION ALL combines result sets but includes all rows, including duplicates, which fails to meet the requirement to remove them.
C. The JOIN syntax is invalid, and its purpose is to combine columns from different tables based on a related key, not to combine rows.
D. CROSS JOIN creates a Cartesian product, pairing every row from the first table with every row from the second, which is not the required operation.

References:

1. PostgreSQL 16 Documentation. (n.d.). 7.4. Combining Queries. PostgreSQL. Retrieved from https://www.postgresql.org/docs/16/queries-union.html. In Section 7.4, the documentation states, "The UNION operator computes the set union of the rows returned by the involved SELECT statements. A row is in the set union of two result sets if it appears in at least one of the result sets. ... Duplicate rows are eliminated from the result..."
2. Oracle Database SQL Language Reference, 23c. (n.d.). Set Operators. Oracle. Retrieved from

https://docs.oracle.com/en/database/oracle/oracle-database/23/sqlrf/Set-Operators.html#GUID-B0F24731-B6B5-465F-940E-25D617553A0E. The documentation specifies, "The UNION operator returns the distinct rows from the result sets of the two queries." It contrasts this with UNION ALL, which "does not eliminate duplicate rows."
3. Ullman, J. D., & Widom, J. (n.d.). CS145 Introduction to Databases, Lecture Notes: SQL 2. Stanford University. Retrieved from https://web.stanford.edu/class/cs145/notes/SQL-2.pdf. On page 2, under "Union, Intersection, Difference," the notes explain that (SELECT ... FROM ... WHERE ...) UNION (SELECT ... FROM ... WHERE ...) computes the set union and that "UNION removes duplicates."

Question: 4

Data Concepts and Environments

Which of the following tables holds relational keys and numeric values?

A. Fact

B. Graph

C. Dimensional

D. Transactional

Answer:

A

Explanation:

A fact table is the central table in a dimensional model, such as a star or snowflake schema. Its design is purpose-built to store two primary types of data: 1) numeric, additive values called measures (or facts) that quantify a business process (e.g., sales amount, units sold), and 2) foreign keys that connect to the primary keys of the surrounding dimension tables. These foreign keys are the "relational keys" that provide the context for the measures. This structure is optimized for data aggregation and analysis.

Why Incorrect Options are Wrong:

B. Graph: This describes a data structure (nodes and edges) used in graph databases, not a table type in a relational or dimensional model.
C. Dimensional: A dimension table contains descriptive, textual attributes (e.g., customer name, product category) and a primary key, not the core numeric measures.
D. Transactional: While a transactional table contains keys and numeric data, the term "Fact table" specifically defines this structure designed for analytics in a data warehouse.

References:

1. Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley. In Chapter 2, "The Core Concepts of Dimensional Modeling," a fact table is defined as containing "the numeric measures of performance" and "a set of foreign keys that point to the primary keys of the associated dimension tables" (p. 21).
2. Microsoft Documentation. (2023). Understand star schema and the importance for Power BI. In the section "Fact and dimension tables," it states, "Fact tables store observational or event data... A fact table contains dimension key columns that relate to dimension tables and numeric measure columns."
3. University of Washington, Paul G. Allen School of Computer Science & Engineering. (n.d.).

CSE 444: Database Systems Internals, Lecture 21: Data Warehousing. Slide 10 defines a Fact Table as containing "1. A multipart key, which is formed by combining the foreign keys of the referenced dimension tables. 2. One or more numerical measures (or 'facts')."

Question: 5

Data Analysis

A data analyst is creating a pivot table for a large dataset for an upcoming board meeting. Which of the following is the purpose of the pivot table?

A. To visualize the data in a dashboard

B. To retrieve and clean data from several sources

C. To summarize and analyze the data

D. To organize the data for reporting

Answer:

C

Explanation:

The primary purpose of a pivot table is to summarize, group, and analyze large datasets. It allows a data analyst to transform rows and columns of data into an interactive summary table. This process enables the calculation of aggregates like sums, counts, and averages across different dimensions of the data, making it easier to identify patterns, trends, and insights. This is essential when preparing concise, high-level information for a board meeting from a complex dataset.

Why Incorrect Options are Wrong:

A. While pivot tables are often the data source for pivot charts (visualizations), the table's fundamental purpose is data summarization, not visualization itself.
B. Retrieving and cleaning data are data preparation steps that occur before a pivot table is created, typically using tools like Power Query or SQL.
D. This statement is too general. While a pivot table does organize data, its specific function is to organize it through summarization and aggregation for analysis, making option C more precise.

References:

1. Microsoft Corporation. (n.d.). Create a PivotTable to analyze worksheet data. Microsoft Support. Retrieved from https://support.microsoft.com/en-us/office/create-a-pivottable-to-analyze-worksheet-data-a9a84538-bfe9-40a9-a8e9-f99134456576.
Reference Detail: The first paragraph states, "A PivotTable is a powerful tool to calculate, summarize, and analyze data that lets you see comparisons, patterns, and trends in your data." This directly supports the "summarize and analyze" function.
2. Google. (n.d.). Create & use pivot tables. Google Docs Editors Help. Retrieved from https://support.google.com/docs/answer/1272900.

Reference Detail: The introduction notes, "You can use pivot tables to narrow down a large data set or see relationships between data points." The section "What you can use pivot tables for" explicitly lists "Summarize data."

3. MIT Libraries. (n.d.). Data Analysis in Excel: Pivot Tables. Data Management Services. Retrieved from https://libguides.mit.edu/c.php?g=176299&p=1159301.

Reference Detail: The guide states, "A pivot table allows you to extract the significance from a large, detailed data set... In the example below, the pivot table is used to summarize the data." This highlights the core function of summarization for analysis.

Question: 6

Data Concepts and Environments

An administrator needs to design a table that will include foreign words. Which of the following is the best option for storing non-native language characters?

A. VARCHAR

B. NVARCHAR

C. CLOB

D. CHAR

Answer:

B

Explanation:

The NVARCHAR data type is the best option for storing non-native language characters. The 'N' prefix stands for "National Language Character Set" and specifies that the column will store data using a Unicode character encoding (e.g., UTF-16 or UTF-8). Unicode is a universal character encoding standard that assigns a unique number to every character, regardless of the platform, program, or language. This allows NVARCHAR to correctly store and retrieve characters from virtually all writing systems in the world, preventing data corruption or loss that can occur when using non-Unicode types like VARCHAR.

Why Incorrect Options are Wrong:

A. VARCHAR: This type typically uses a single-byte character set (like ASCII) based on the database's default collation, which cannot represent the full range of international characters.
C. CLOB: A Character Large Object is designed for storing extremely large blocks of text (megabytes or gigabytes) and is inefficient for storing standard-length words or phrases.
D. CHAR: This is a fixed-length data type that also typically uses a non-Unicode character set by default, and it wastes space by padding shorter strings.

References:

1. Microsoft SQL Server Documentation: In the official documentation for nchar and nvarchar (Transact-SQL), it states, "When you use nchar or nvarchar, we recommend that you: Use nchar when the sizes of the column data entries are likely to be similar. Use nvarchar when the sizes of the column data entries are likely to vary considerably..... Using nchar or nvarchar helps you to avoid having character conversion problems." This confirms NVARCHAR is the standard for variable-length Unicode data. (Source: Microsoft, nchar and nvarchar (Transact-SQL), SQL Server Documentation).

2. Oracle Database Documentation: The Oracle SQL Language Reference specifies the NVARCHAR2 data type for this purpose: "The NVARCHAR2 data type is a variable-length character string... This data type can store Unicode character data." This highlights its role in handling international character sets. (Source: Oracle, Database SQL Language Reference, 19c, Section: "Data Types", Subsection: "NVARCHAR2 Data Type").
3. Stanford University Courseware: Database course materials explain the fundamental difference between character sets. They emphasize that to support internationalization (i18n), a database must use a multi-byte character encoding like Unicode. Data types prefixed with 'N' (like NVARCHAR) are the SQL standard for handling such encodings. (Source: Stanford University, CS245, Notes on Relational Data Models and SQL).

Question: 7

Visualization and Reporting

An analyst needs to create a collection of dashboards for multiple teams within their organization. Which of the following should the analyst do first before starting the project?

A. Evaluate the user persona type for the dashboards.

B. Determine the number of team members who need to access the dashboards.

C. Determine the delivery method of the dashboards.

D. Evaluate the KPIs for the dashboards.

Answer:

A

Explanation:

The foundational step in designing effective dashboards is to understand the target audience. Evaluating the user persona involves identifying the users' roles, goals, technical skills, and how they will use the information to make decisions. This understanding directly informs all subsequent design choices, including which Key Performance Indicators (KPIs) are most relevant, the complexity of the visualizations, and the most suitable delivery method. For a project involving multiple teams, this step is critical as each team may have a distinct persona with unique requirements, ensuring the final dashboards are tailored, relevant, and actionable for each group.

Why Incorrect Options are Wrong:

B. Determine the number of team members who need to access the dashboards.
This is a logistical consideration for deployment and licensing, not the primary step in designing the dashboard's content and structure.
C. Determine the delivery method of the dashboards.
The delivery method (e.g., web, mobile, email) should be chosen based on the needs and work habits of the user persona, making it a subsequent decision.
D. Evaluate the KPIs for the dashboards.
Relevant KPIs are selected based on the goals and responsibilities of the target user. Defining the user persona first ensures the chosen KPIs are meaningful and useful.

References:

1. Microsoft Power BI Documentation: In the official learning path for report design, the first step outlined is audience identification. "The first step in planning your report is to identify your audience. Ask yourself who will be using this report? What are their goals for using the report?" Source: Microsoft Learn, "Plan a report in Power BI" module, "Identify the audience" unit.
2. University Courseware: Academic literature on visualization design emphasizes a user-centered approach. The nested model for visualization design and validation starts with understanding the domain problem and the target users' tasks.
Source: Munzner, T. (2014). Visualization Analysis and Design. CRC Press. Chapter 3, "A Nested Model for Visualization Design and Validation," outlines the process starting with characterizing the tasks and data of a specific real-world domain, which is synonymous with understanding the user and their context.
3. Official Vendor Documentation (Tableau): Best practices for visual analytics consistently prioritize understanding the audience before any other step.
Source: Tableau Whitepaper, "Visual Analysis Best Practices: A Guidebook". In the section "Cycle of Visual Analysis," the process begins with "Get Data" followed by defining the task and audience questions. The document states, "A good visualization starts with a good question... The question should be tailored to the audience of your visualization." (p. 6).

Question: 8

Data Concepts and Environments

Which of the following file types separates data using a delimiter?

A. XML

B. HTML

C. JSON

D. CSV

Answer:

D

Explanation:

A Comma-Separated Values (CSV) file is a plain text format used to store tabular data, such as a spreadsheet or database. Each line in the file corresponds to a row in the table, and within each line, fields (or columns) are separated by a specific character known as a delimiter. While the comma is the most common delimiter, other characters like tabs or semicolons can also be used. The fundamental structure of a CSV file is based on this delimiter-separated principle.

Why Incorrect Options are Wrong:

A. XML: XML (Extensible Markup Language) is a markup language that uses a hierarchical structure of tags and attributes to define and organize data, not a simple delimiter.
B. HTML: HTML (HyperText Markup Language) is a markup language for creating web pages. It uses tags to structure content for display in a browser, not for storing delimited data.
C. JSON: JSON (JavaScript Object Notation) structures data using key-value pairs and nested objects/arrays. It is a semi-structured format, not a flat, delimiter-separated one.

References:

1. Internet Engineering Task Force (IETF). RFC 4180: Common Format and MIME Type for Comma-Separated Values (CSV) Files. October 2005. Section 2, "Definition of the CSV Format," states: "Within each record, there may be one or more fields, separated by commas." Available: https://www.rfc-editor.org/rfc/rfc4180.html#section-2
2. World Wide Web Consortium (W3C). Extensible Markup Language (XML) 1.0 (Fifth Edition). November 26, 2008. Section 2.1, "Well-Formed XML Documents," defines the structure based on start-tags, end-tags, and elements, which is fundamentally different from a delimited format. Available: https://www.w3.org/TR/xml/#sec-well-formed
3. Internet Engineering Task Force (IETF). RFC 8259: The JavaScript Object Notation (JSON)

Data Interchange Format. December 2017. Section 2, "JSON Grammar," defines the structure of JSON objects as collections of name/value pairs and arrays as ordered lists of values, not as delimiter-separated fields. Available: https://www.rfc-editor.org/rfc/rfc8259.html#section-2
4. VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Chapter 3, "Data Manipulation with Pandas," discusses various data formats, describing CSV files as "text files with values separated by commas (or another delimiter)." (This is a highly respected academic-level text, often used in university curricula).

Question: 9

Data Concepts and Environments

Which of the following data repositories stores unformatted data in its original, raw form?

A. Data warehouse

B. Data silo

C. Data mart

D. Data lake

Answer:

D

Explanation:

A data lake is a centralized storage repository designed to hold vast amounts of data in its native, raw format. Unlike traditional data warehouses, a data lake ingests structured, semi-structured, and unstructured data without pre-defining a schema or requiring transformation. This "schema-on-read" approach allows data to be stored "as-is" in its original, unformatted state, making it available for various future analytical and machine learning purposes. The primary characteristic of a data lake is its ability to store raw data from diverse sources.

Why Incorrect Options are Wrong:

A. Data warehouse: Stores structured, transformed, and processed data from operational systems, organized by a predefined schema for specific business intelligence tasks.
B. Data silo: Refers to an isolated data repository that is not integrated with other systems; this term describes accessibility, not the format of the data.
C. Data mart: A focused subset of a data warehouse, containing structured and processed data tailored to a specific department or business function.

References:

1. Official Vendor Documentation: Amazon Web Services (AWS). (n.d.). What is a data lake? AWS. Retrieved from https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/. In the section "What is a data lake?", it states, "A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data..."
2. Official Vendor Documentation: Microsoft Azure. (n.d.). What is a Data Lake? Microsoft. Retrieved from https://azure.microsoft.com/en-us/solutions/data-lake/. The documentation states, "A data lake is a centralized repository designed to store, process, and secure large amounts of

structured, semi-structured, and unstructured data. It can store data in its native format..."
3. Peer-Reviewed Academic Publication: Nargesian, F., et al. (2019). Data Lake Governance: A Survey. arXiv preprint arXiv:1905.01912. Section 2, Paragraph 1. "A data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data." https://doi.org/10.48550/arXiv.1905.01912
4. University Courseware: Stanford University. (2021). CS 229: Machine Learning - Unsupervised Learning, Recommender Systems, Reinforcement Learning. Lecture Notes. The course materials often contrast data lakes and warehouses, noting that data lakes store raw data while warehouses store processed data. The concept is fundamental to modern data architecture discussions in such courses.

Question: 10

Data Analysis

A data analyst needs to create a report that anticipates the number of calls received daily. Which of the following is the best statistical method to use?

A. Predictive

B. Diagnostic

C. Inferential

D. Descriptive

Answer:

A

Explanation:

The core requirement of the task is to "anticipate the number of calls received daily," which is a forecasting activity. Predictive analytics is the statistical method specifically designed for this purpose. It utilizes historical data, statistical algorithms, and machine learning techniques to model and predict future outcomes. The analyst would build a model (e.g., a time-series forecast) based on past call data to estimate the number of future calls, directly fulfilling the request.

Why Incorrect Options are Wrong:

B. Diagnostic: This method is used to understand the root cause of a past event (e.g., "Why did call volume spike last Tuesday?"), not to predict future events.
C. Inferential: This method uses a sample of data to make generalizations about a larger population (e.g., estimating the satisfaction of all customers from a survey sample).
D. Descriptive: This method summarizes and describes past data (e.g., calculating the average number of calls per day last month) but does not make future predictions.

References:

1. University Courseware: Bertsimas, D., & Dunn, J. (2017). Lecture 1: Introduction to Analytics. MIT OpenCourseWare, Course 15.071 (The Analytics Edge), Spring 2017. On slide 11, "The Analytics Toolkit," Predictive Analytics is defined as: "Predict future outcomes based on historical data (e.g., predict which customers will 'churn')." This directly aligns with predicting future call volumes.
2. Official Vendor Documentation: Microsoft Azure Documentation. (2023). What is automated machine learning (AutoML)?. In the section "When to use AutoML: classify, regress, & forecast," the task of "Forecasting" is described as a type of regression, a core predictive technique, used to

"predict future values based on historical time-series data," such as predicting future sales. This is analogous to predicting future call volumes.

3. Peer-Reviewed Academic Publication: Shmueli, G. (2010). To Explain or to Predict?. Statistical Science, 25(3), 289-310. In Section 2, "Explanatory and Predictive Modeling," the paper distinguishes between modeling for explanation (understanding causal effects) and modeling for prediction (forecasting new observations). The analyst's task of anticipating calls is a clear case of predictive modeling. DOI: https://doi.org/10.1214/10-STS330

Question: 11

Data Analysis

A product goes viral on social media, creating high demand. Distribution channels are facing supply chain issues because the testing and training models that are used for sales forecasting have not encountered similar demand. Which of the following best describes this situation?

A. Model bias

B. Data drift

C. Incorrect sizing

D. Skewing

Answer:

B

Explanation:

This scenario is a classic example of data drift. Data drift occurs when the statistical properties of the production data, on which the model makes predictions, change or "drift" away from the data the model was originally trained on. The viral event caused a sudden, significant shift in the distribution of sales demand, making the historical training data no longer representative of the current reality. Consequently, the forecasting model's performance degrades because the patterns it learned are now obsolete.

Why Incorrect Options are Wrong:

A. Model bias: This refers to inherent, systematic errors in the model's predictions due to flawed assumptions, not a change in data distribution over time.
C. Incorrect sizing: This describes a potential consequence of the inaccurate forecast (e.g., insufficient inventory), not the underlying statistical problem with the model's data.
D. Skewing: This is a measure of a distribution's asymmetry. While the new data may be skewed, "data drift" is the specific term for the phenomenon of the data distribution changing over time.

References:

1. Microsoft Azure Official Documentation. In the "What is data drift?" section for Azure Machine Learning, it states: "Data drift is one of the top two reasons for model accuracy degradation. It occurs in most machine learning applications... Data drift is the change in model input data that leads to model performance degradation." This directly aligns with the scenario where a change in demand patterns (input) causes the sales forecasting model to fail.
Source: Microsoft Corporation. (2023). Detect data drift (preview) on datasets. Azure Machine

Learning Documentation. Section: "What is data drift?".

2. Google Cloud Official Documentation. The documentation for Vertex AI Model Monitoring defines data drift (also called feature skew) as occurring when the statistical properties of input features change over time. It distinguishes between training-serving skew and prediction drift. The scenario described is a form of prediction drift, where feature data distributions change between training and serving.

Source: Google Cloud. (2023). Introduction to Vertex AI Model Monitoring. Vertex AI Documentation. Section: "Data drift and concept drift".

3. Academic Publication. A foundational survey on the topic defines the problem: "In a changing environment, the data distribution may change over time, which is known as concept drift... This change may hurt the performance of a model trained on old data." The viral event represents a sudden change in the environment, causing the data distribution to shift.

Source: Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys (CSUR), 46(4), 1-37. Section 2: "Concept Drift". DOI: https://doi.org/10.1145/2523813

4. University Courseware. Stanford University's course on Machine Learning Systems Design explicitly covers this issue. Lecture notes describe data drift as a change in the input data distribution over time, which is a common reason for performance degradation in production ML systems.

Source: Stanford University. (2021). CS 329S: Machine Learning Systems Design, Lecture 4: Data. Course materials. Section: "Data Distribution Shifts".

Question: 12

Data Concepts and Environments

Given the following table:

ID Value 1 1.5 2 24.456 3 113

Which of the following data types should an analyst use for the numeric values in the Value column?

A. Double

B. Float

C. Boolean

D. Integer

Answer:

B

Explanation:

The Value column contains numeric data with decimal places (e.g., 1.5, 24.456), which are floating-point numbers. A Float is a single-precision data type suitable for representing such numbers. While a Double (double-precision) could also be used, it consumes more storage space. A fundamental principle in database design and data management is to select the most storage-efficient data type that can accurately represent the data. Since the values shown do not require the high precision of a Double, Float is the most appropriate and optimal choice.

Why Incorrect Options are Wrong:

A. Double: While it can store the data, it uses more memory than necessary for the given low-precision values, making it a less optimal choice.
C. Boolean: This data type is for true/false values and cannot represent the numeric data in the 'Value' column.
D. Integer: This data type is for whole numbers and cannot store values with decimal points like 1.5 or 24.456.

References:

1. PostgreSQL 16 Documentation. (n.d.). Chapter 8. Data Types, 8.1. Numeric Types. Retrieved from https://www.postgresql.org/docs/16/datatype-numeric.html. The documentation specifies real (a 4-byte single-precision floating-point number) and double precision (an 8-byte double-precision floating-point number), highlighting the difference in storage and precision.
2. MySQL 8.0 Reference Manual. (n.d.). 11.1.6 Floating-Point Types (Approximate Value) - FLOAT, DOUBLE. Retrieved from https://dev.mysql.com/doc/refman/8.0/en/floating-point-types.html. This source details that FLOAT uses 4 bytes of storage while DOUBLE uses 8 bytes, confirming that FLOAT is the more storage-efficient option for numbers that do not require double precision.
3. Bryant, R. E., & O'Hallaron, D. R. (2016). Computer Systems: A Programmer's Perspective (3rd ed.). Pearson. In Chapter 2, "Representing and Manipulating Information," Section 2.4 discusses floating-point numbers, detailing the IEEE 754 standard for single-precision (float) and double-precision (double) formats, which forms the basis for how these data types are implemented and their respective precision and range capabilities.

Question: 13

Visualization and Reporting

The human resources department wants to understand the relationship between the ages and incomes of all employees. Which of the following graphics is the most appropriate to present the analysis?

A. Scatter plot

B. Area plot

C. Bar chart

D. Pie chart

Answer:

A

Explanation:

A scatter plot is the most suitable visualization for examining the relationship or correlation between two continuous numerical variables. In this scenario, 'age' and 'income' are both continuous variables. Each employee can be represented as a point on the graph, with their age plotted on the x-axis and their income on the y-axis. This visual representation allows the human resources department to easily identify patterns, trends (e.g., does income increase with age?), and outliers within the dataset.

Why Incorrect Options are Wrong:

B. Area plot: An area plot is typically used to show how a quantitative value changes over a continuous dimension, such as time, and is not suitable for correlating two distinct variables.
C. Bar chart: A bar chart is designed to compare values across discrete categories, not to illustrate the relationship between two continuous variables like age and income.
D. Pie chart: A pie chart is used to display the proportional composition of a single categorical variable, representing parts of a whole, and cannot show a relationship between two variables.

References:

1. Wilke, C. O. (2019). Fundamentals of Data Visualization. O'Reilly Media, Inc. In Chapter 12, "Visualizing associations among two or more quantitative variables," the text states, "The standard visualization for an association between two quantitative variables is the scatterplot." The chapter proceeds to provide examples using this chart type.
2. University of Virginia Library, Research Data Services. (n.d.). Data Visualization: The Scatterplot. Retrieved from the University of Virginia Library website. The guide specifies, "A scatterplot is a type of data visualization that shows the relationship between two numerical

variables... Scatterplots are used to observe and show relationships between two numeric variables."
3. Healy, K. (2018). Data Visualization: A Practical Introduction. Princeton University Press. In Chapter 3, "Make a Plot," the section on scatterplots explains, "Scatterplots are the default choice for visualizing the relationship between two continuous variables" (p. 56).

Question: 14

Data Concepts and Environments

Which of the following best describes the function of a data type?

A. To provide a generic identifier for files used in analysis

B. To identify the program needed to open a file

C. To differentiate the real value of the field in its context

D. To make the addition of individual records simpler

Answer:

C

Explanation:

A data type is a fundamental classification that specifies the nature of a value a variable or database field can hold. Its primary function is to inform the computer system how to interpret the underlying binary data, thereby defining the value's meaning, the set of valid values, and the operations that can be performed on it. For instance, the binary sequence 01000001 is interpreted as the integer 65 if the data type is integer, but as the character 'A' if the data type is char. The data type provides the essential context to differentiate the value.

Why Incorrect Options are Wrong:

A. This describes a file naming convention or metadata tag, which is distinct from the data type of the values contained within the file.
B. This is the function of a file extension (e.g., .csv, .docx) or MIME type, which associates a file with an application at the operating system level.
D. While using appropriate data types facilitates efficient data manipulation, this is a benefit derived from the primary function, not the function itself.

References:

1. Guttag, J. V. (2016). Introduction to Computation and Programming Using Python, With Application to Understanding Data (2nd ed.). MIT Press. In Chapter 2, Section 2.1, "Objects, Expressions, and Numerical Types," it is explained that "Each object has a type that defines the kinds of things that programs can do with objects of that type." This establishes that the type provides the context for interpretation and operation.
2. PostgreSQL 16 Documentation. (2023). Chapter 8: Data Types. The introduction states, "PostgreSQL has a rich set of native data types... Every data type has an external representation... a set of input and output functions... These functions determine how the type appears in strings... and how the type is interpreted by the system." This directly supports that the

data type's function is to define interpretation.
3. Pierce, B. C. (2002). Types and Programming Languages. The MIT Press. In Chapter 1, Section 1.1, a type system is described as a method for "classifying phrases according to the kinds of values they compute," which aligns with differentiating the real value of a field in its context.

Question: 15

Visualization and Reporting

A project manager requests an unscheduled report that provides a list of clients. Which of the following frequencies is best for this report?

A. Annual

B. Daily

C. Weekly

D. Ad hoc

Answer:

D

Explanation:

The key term in the scenario is "unscheduled report." This indicates a one-time, on-demand request that falls outside of any regular reporting cycle. Ad hoc reporting is the process of creating reports as needed to answer specific, often unforeseen, business questions. Since the project manager's request is not part of a pre-planned daily, weekly, or annual schedule, it is best classified as an ad hoc report.

Why Incorrect Options are Wrong:

A. Annual: This is a long-term, scheduled frequency and is inappropriate for an immediate, unscheduled request.
B. Daily: This is a high-frequency, scheduled report, which contradicts the "unscheduled" nature of the manager's request.
C. Weekly: This is a regular, recurring reporting schedule, not a one-off request made on an as-needed basis.

References:

1. Turban, E., Sharda, R., & Delen, D. (2014). Business Intelligence and Analytics: Systems for Decision Support (10th ed.). Pearson. In Chapter 3, Section 3.4, "Business Reporting, Visual Analytics, and Business Performance Management," the text defines ad hoc queries as "queries that answer unplanned information requests to support a non-routine decision." This directly aligns with the "unscheduled" nature of the project manager's request.
2. University of Illinois at Urbana-Champaign. (n.d.). Enterprise Data Warehouse - Reporting. Technology Services. Retrieved from https://techservices.illinois.edu/services/enterprise-data-warehouse/reporting. In the "Ad Hoc Reporting" section, it is stated, "Ad hoc reports are created on the fly to answer a specific

business question... Ad hoc reporting empowers end-users by enabling them to ask their own questions of the data, without having to rely on IT for a canned report."
3. Popovic, A., Hackney, R., Coelho, P. S., & Jaklic, J. (2012). Towards business intelligence systems success: Effects of maturity and culture on analytical decision making. Decision Support Systems, 54(1), 729-739. https://doi.org/10.1016/j.dss.2012.08.017. The paper discusses different BI functionalities, implicitly distinguishing between standardized reporting and ad hoc analysis (p. 731), which supports the concept of ad hoc as a distinct, non-scheduled activity.

Question: 16

Data Governance

Which of the following supports capabilities such as automatic versioning, corruption checks, KPIs, and user authentication?

A. Notebook

B. REST API

C. Pipeline

D. Source control

Answer:

D

Explanation:

Source-control systems (e.g., Git, Subversion) are specifically designed to 1) record every change as a new commit, giving automatic versioning, 2) store objects by cryptographic hash so any bit-level corruption is detected, 3) enforce user authentication/authorization for every push, pull, or commit, and 4) expose quantitative metrics (commit counts, merge rates, contributor activity) that organizations use as KPIs. Notebooks, REST APIs, and data-processing pipelines do not inherently provide this full combination of functions.

Why Incorrect Options are Wrong:

A. Notebook - Captures code and narrative but lacks built-in integrity hashes, enterprise auth, and KPI reporting.
B. REST API - Is an interface protocol; by itself it offers no intrinsic versioning or corruption checking.
C. Pipeline - Orchestrates data or CI tasks; may consume version-controlled artifacts but does not supply native VCS functions.

References:

1. Loeliger, J. & McCullough, M. "Version Control with Git", 2nd ed., O'Reilly, 2012, pp. 7-10 (automatic history), pp. 55-57 (SHA-1 integrity).
2. Chacon, S. & Straub, B. "Pro Git", Apress, 2014, ch. 15 "Plumbing & Porcelain", SectionIntegrity (hash checks) and Section15.6 (auth hooks).
3. Apache Software Foundation, "Subversion Authentication & Authorization", Official SVN Book v1.7, ch. 6, Section"Authentication", Section"Repository Integrity".
4. Jiang, Y. et al., "Understanding project health in GitHub", Empirical Software Engineering, 2020, pp. 2423-2425 (KPIs derived from commits). DOI:10.1007/s10664-019-09735-5

5. MIT OpenCourseWare, 6.831 "Software Engineering for Web Apps" (Spring 2020), Lec 7 slides, p. 12 ("Version control systems provide history, auth, integrity, project metrics").

Question: 17

Data Governance

A company gives users adequate data access permissions to allow them to fulfill their duties but nothing more. Which of the following concepts best describes this practice?

A. Active Directory

B. Hierarchical access

C. Zero Trust

D. Least privilege

Answer:

D

Explanation:

The scenario describes the principle of least privilege (PoLP). This is a fundamental information security concept which dictates that a user, program, or process should only have the bare minimum permissions required to perform its specific, authorized function. The phrase "adequate data access permissions to allow them to fulfill their duties but nothing more" is the textbook definition of this principle. It aims to limit the potential damage from errors, malware, or a compromised user account by restricting access rights to only what is strictly necessary.

Why Incorrect Options are Wrong:

A. Active Directory: This is a Microsoft directory service product used to manage users and resources. It is a tool that can implement least privilege, not the principle itself.
B. Hierarchical access: This describes an access control structure based on an organizational hierarchy (e.g., managers have more access than their reports), which does not inherently guarantee least privilege.
C. Zero Trust: This is a broad security framework built on the principle of "never trust, always verify." While least privilege is a core component of Zero Trust, it is not the specific concept described.

References:

1. National Institute of Standards and Technology (NIST) SP 800-53 Rev. 5, Security and Privacy Controls for Federal Information Systems and Organizations, December 2020. In Control AC-6, "Least Privilege," the discussion states: "The principle of least privilege is also applied to non-privileged users to ensure that the users are only granted access to the information and information system resources that the users require to perform their official duties." (Page 101).
2. National Institute of Standards and Technology (NIST) SP 800-207, Zero Trust Architecture,

August 2020. Section 2.1, "Tenets of Zero Trust," lists as a tenet: "Access to individual enterprise resources is granted on a per-session basis... Access should also be granted with the least privileges needed to complete the task." This shows that least privilege is a component of the broader Zero Trust model. (Page 6).
3. Microsoft Documentation, What is Active Directory Domain Services?, October 2021. The documentation defines it as a service that "stores directory data and manages communication between users and domains, including user logon processes, authentication, and directory searches." This confirms it is a technology, not a security principle.
4. Saltzer, J. H., & Schroeder, M. D. (1975). The Protection of Information in Computer Systems. Proceedings of the IEEE, 63(9), 1278-1308. This foundational academic paper on computer security introduces the "Principle of Least Privilege" as a key design principle: "Every program and every user of the system should operate using the least set of privileges necessary to complete the job." (Section I.A.3, Page 1281). https://doi.org/10.1109/PROC.1975.9939

Question: 18

Data Analysis

Software end users are happy with the quality of product support provided. However, they frequently raise concerns about the long wait time for resolutions. An IT manager wants to improve the current support process. Which of the following should the manager use for this review?

A. Infographic

B. KPI

C. Survey

D. UAT

Answer:

B

Explanation:

The IT manager's goal is to improve the support process by addressing the "long wait time for resolutions." Key Performance Indicators (KPIs) are the most appropriate tool for this task. KPIs are specific, measurable metrics used to evaluate the performance and efficiency of a process. By analyzing KPIs such as 'Average Time to Resolution,' 'First Response Time,' and 'Ticket Backlog,' the manager can objectively identify bottlenecks, set performance targets, and track the impact of process changes over time. This data-driven approach is fundamental to process improvement.

Why Incorrect Options are Wrong:

A. Infographic: An infographic is a tool for visually presenting data and findings after an analysis is complete, not for conducting the initial review itself.
C. Survey: The organization already has user feedback indicating the problem (long wait times). Another survey would gather more user opinions but would not analyze the internal process causing the delays.
D. UAT: User Acceptance Testing (UAT) is a phase in the software development lifecycle to validate that a system meets user requirements; it is not used to review an operational support process.

References:

1. University Courseware: In materials for business process management, KPIs are defined as essential tools for process analysis. "Process performance measures (or Key Performance Indicators - KPIs) are a prerequisite for process improvement. They are used to determine how well a process is performing with respect to its goals."
Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). Fundamentals of Business Process Management. Springer. (This is a widely used textbook in university courses). Chapter 8, "Process Monitoring and Mining," extensively covers the use of KPIs for process evaluation.
2. Official Vendor Documentation: Documentation for IT Service Management (ITSM) and data analytics platforms consistently highlights the use of KPIs for service desk improvement. For example, Microsoft's documentation for its customer service analytics tools emphasizes tracking core KPIs to enhance efficiency.
Microsoft. (n.d.). Customer Service historical analytics report. Microsoft Learn. In the "Summary report" and "Agent report" sections, the documentation details key metrics like "Average time to resolution" as critical for evaluating and improving support performance.
3. Peer-Reviewed Academic Publication: Research in information systems and service management confirms that monitoring KPIs is central to improving IT support functions. Studies show that metrics directly related to time and efficiency are crucial for identifying and resolving process bottlenecks.
Iden, J., & Eikebrokk, T. R. (2013). Implementing IT Service Management: A systematic literature review. International Journal of Information Management, 33(3), 512-523. https://doi.org/10.1016/j.ijinfomgt.2013.01.004. This review discusses how ITSM frameworks like ITIL rely on metrics and KPIs to measure, control, and improve service delivery processes.

Question: 19

Data Concepts and Environments

Which of the following is the best reason for a company to use a CSV file to share data instead of an Excel file?

A. CSV files can store different types of encoding.

B. CSV files are not vendor-specific.

C. CSV files are smaller in size.

D. CSV files are easier to change in text editors.

Answer:

B

Explanation:

The best reason for using a CSV (Comma-Separated Values) file for data sharing instead of an Excel file is its vendor-neutral nature. CSV is a simple, text-based, open format, making it universally compatible with a vast range of applications, programming languages, and database systems without requiring specific proprietary software like Microsoft Excel. This high degree of interoperability ensures that the recipient can access and process the data regardless of their software environment, which is a primary concern when exchanging data between different organizations.

Why Incorrect Options are Wrong:

A. CSV files can store different types of encoding.
Modern Excel files (.xlsx) also support various encodings, such as UTF-8, so this is not a unique advantage for CSV.
C. CSV files are smaller in size.
While often true, this is a secondary benefit. The primary goal of data sharing is compatibility, and file size differences can be mitigated with compression.
D. CSV files are easier to change in text editors.
This is a technical convenience for developers but not the main business reason for choosing a format for inter-company data exchange.

References:

1. Internet Engineering Task Force (IETF) RFC 4180. Shafranovich, Y. (2005). Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180. The abstract states, "This RFC documents the format known as Comma-Separated Values (CSV). It is used to exchange data between applications that use incompatible data formats." This highlights its primary purpose

as a tool for interoperability.

2. University of California, Berkeley, Foundations of Data Science Courseware. Adhikari, A., DeNero, J., & Wagner, D. (2021). Computational and Inferential Thinking: The Foundations of Data Science. Chapter 6: Tables. The text describes CSV as a "simple, universal format for tabular data" and notes that "most spreadsheet and data analysis programs can import and export data in this format." This emphasizes its universal applicability over vendor-specific formats.

3. Principles of Data Management: Facilitating Information Sharing. Lemahieu, W., van den Broucke, S., & Baesens, B. (2018). Cambridge University Press. Chapter 1, Section 1.3, "Data Exchange and Syntactic Interoperability," discusses the importance of standardized, simple formats for data exchange. It explains that formats like CSV and XML are designed to be application-independent to facilitate information sharing between disparate systems, a core principle that favors CSV over proprietary formats like Excel's .xlsx.

Question: 20

Visualization and Reporting Which of the following is business intelligence software?

A. SAS

B. Python

C. Notepad++

D. Tableau

Answer:

D

Explanation:

Tableau is a leading software platform specializing in data visualization and business intelligence (BI). Its core function is to enable users to connect to various data sources, create interactive and shareable dashboards, and perform self-service analytics to derive business insights. The platform is designed specifically to transform raw data into actionable information through visual analysis, which is the primary purpose of BI software.

Why Incorrect Options are Wrong:

A. SAS is a comprehensive statistical analysis and advanced analytics platform, which is much broader than being exclusively BI software.
B. Python is a general-purpose programming language, not a standalone BI application, although it has libraries used for BI tasks.
C. Notepad++ is a source code and text editor with no inherent data analysis or visualization capabilities.

References:

1. Official Vendor Documentation: Tableau Software, LLC. "What is Tableau?". Tableau.com. Accessed 2023. The official website defines Tableau as a "visual analytics platform transforming the way we use data to solve problems-empowering people and organizations to make the most of their data." This aligns directly with the definition of business intelligence software.
2. University Courseware: Duke University, Fuqua School of Business. "IDS 703: Data Visualization in Tableau". Course Catalog. The existence and description of this course at a reputable institution categorize Tableau as a primary tool for data visualization, a key component of business intelligence.
3. Peer-Reviewed Academic Publication: Cernezel, J., & Vuga, V. (2021). A Comparative Study of Business Intelligence and Data Visualization Tools: A Case Study of Tableau and Microsoft Power BI. International Journal of Engineering and Advanced Technology (IJEAT), 10(3), 118-123. https://doi.org/10.35940/ijeat.C2289.0210321. This paper explicitly identifies and compares

Tableau as a leading business intelligence and data visualization tool (Section I, Introduction, Paragraph 2).

Question: 21

Data Governance

Which of the following best describes an assessment a data analyst would use to validate that the number of records in a dataset matches the expected results?

A. Source control

B. Unit test

C. Stress test

D. Health check

Answer:

B

Explanation:

A unit test is a type of automated test that validates a small, isolated piece of a system. In the context of data analytics, a unit test is used to verify a specific aspect of a data transformation or dataset. Checking if the number of records in a target dataset matches the number of records from a source or a predefined expected value is a classic example of a data-centric unit test. This test ensures data completeness and helps identify issues like data loss during an ETL (Extract, Transform, Load) process.

Why Incorrect Options are Wrong:

A. Source control: Manages and tracks changes to code and files (e.g., Git), but it does not perform validation on the data content itself.
C. Stress test: Evaluates a system's performance and stability under extreme load, not the correctness or completeness of the data records.
D. Health check: A high-level, general assessment to confirm that a system or service is operational and available, not a granular data validation.

References:

1. dbt Labs Documentation, "About dbt tests": "Tests are assertions you make about your data... A singular test is a specific query that you write to assert something about your data. These are highly flexible, and can be used to test for virtually any condition you can write a SQL query for." This includes custom tests for row counts. (Source: dbt Labs, official vendor documentation).
2. Fowler, M. (2014). "Unit Test": "Unit tests are low-level, focusing on a small part of the software system... In object-oriented design, this is often a single class, but it can be as small as a single method." This foundational software engineering principle is directly applied to data pipelines,

where a "unit" can be a single transformation step or a resulting dataset. (Source: martinfowler.com, a highly respected resource in software engineering, often cited in academic contexts).

3. Microsoft Azure Documentation, "Develop and test data mapping flows": "Before you publish your changes, you want to make sure they work as expected. In mapping data flows, you can run your logic on a live Spark cluster... For a more formal test approach, you can use a unit test framework to run a series of tests against your data flow." (Source: Microsoft Azure, official vendor documentation).

4. Zuckerman, D., & Miller, D. (2016). 6.031: Software Construction, Spring 2016. Massachusetts Institute of Technology: MIT OpenCourseWare. In Reading 3: Testing, it is stated, "A unit test is a test for a single module (a unit) of your program... A good unit test is automated, fast, and repeatable." This academic definition supports using the term for an automated, specific check like record count validation. (Source: MIT OpenCourseWare, University Courseware).

Question: 22

Visualization and Reporting

A data company needs a visualization that shows the availability zones from the last ten years and any future availability zones that the company will be using in the next five years. Which of thefollowing is the most appropriate visualization to display this information?

A. Bar chart

B. Mosaic plot

C. Map

D. Pie chart

Answer:

C

Explanation:

The core data element in the question is "availability zones," which are distinct physical, geographical locations. A map is the most appropriate and effective visualization for displaying geospatial data. It allows the data company to plot the exact locations of past and future availability zones, clearly showing their spatial distribution, proximity, and regional coverage. Different colors or symbols can be used on the map to distinguish between historical and planned zones, fulfilling all requirements of the request.

Why Incorrect Options are Wrong:

A. Bar chart: A bar chart is best for comparing discrete quantities across categories, not for illustrating the geographical distribution of locations.
B. Mosaic plot: A mosaic plot is used to visualize the relationship between two or more categorical variables, which is not suitable for displaying geographic points.
D. Pie chart: A pie chart is designed to show the proportional composition of a whole and cannot effectively represent multiple, distinct geographical locations.

References:

1. Wilke, C. O. (2019). Fundamentals of Data Visualization. O'Reilly Media. In Chapter 15, "Visualizing geospatial data," the author states, "Whenever we are working with data that has an intrinsic geographic component... we should consider drawing a map." This directly supports using a map for availability zones.
2. Healy, K. (2018). Data Visualization: A Practical Introduction. Princeton University Press. In Chapter 7, "Draw Maps," it is noted that "Maps are the right tool for the job when your goal is to show the spatial distribution of some quantity or to compare it across different areas."

3. Tableau Software, LLC. (2024). Build a Map View. Tableau Help Documentation. The official documentation states, "When you have geographic data, you can plot it on a map... Maps are effective for showing trends or patterns in your data that have a geographic context." This aligns with plotting the locations of availability zones.

Question: 23

Data Concepts and Environments

Which of the following best describes the method used to combine files, software, and libraries for use on various operating systems and environments?

A. Package manager

B. Code repository

C. Virtual machine

D. Containerization

Answer:

D

Explanation:

Containerization is a method of operating-system-level virtualization used to deploy and run applications in isolated spaces called containers. A container packages an application's code along with all its necessary dependencies, such as libraries, system tools, and runtime environments. This creates a single, lightweight, executable software package that is portable and can run consistently and reliably across various computing environments, from a local machine to a cloud infrastructure.

Why Incorrect Options are Wrong:

A. Package manager: A package manager installs and manages software packages and their dependencies for a specific operating system or programming language, not for cross-environment portability.
B. Code repository: A code repository is a storage location for managing and tracking changes to source code (e.g., Git); it does not package the application for execution.
C. Virtual machine: A virtual machine emulates an entire computer system, including a full guest operating system, making it much more resource-intensive than just packaging an application and its libraries.

References:

1. Docker Inc. (Official Vendor Documentation). "What is a Container?". Docker Resources. Accessed October 2023. The documentation states, "A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another."
2. Bernstein, D. (2014). "Containers and Cloud: From LXC to Docker to Kubernetes." IEEE Cloud Computing, 1(3), pp. 81-84. DOI: 10.1109/MCC.2014.51. This paper (Section: "What Is a

Container?") describes containers as a technology that packages an application with its runtime dependencies, enabling it to run on any infrastructure.
3. University of California, Berkeley. "CS 162: Operating Systems and Systems Programming," Lecture 22: Virtual Machines. The course materials often contrast VMs with containers, defining containers as a lightweight alternative that bundles the application and its dependencies to run on a shared host OS kernel, ensuring environmental consistency.

Question: 24

Data Governance

A recent server migration applied an update to dataset naming conventions. Multiple users are now reporting stale information in an existing dashboard. The date in the dataset confirms a successful data refresh. Which of the following should a data analyst do first?

A. Confirm the dashboard is pointed to the newest dataset.

B. Filter the data in the dashboard.

C. Escalate user permissions on the server.

D. Verify that the dashboard subscription is not expired.

Answer:

A

Explanation:

The problem states that a server migration led to an update in dataset naming conventions, and users are now seeing stale data despite a successful data refresh. This strongly indicates that the dashboard's connection to its data source is broken or pointing to the old, now-un-updated dataset. The most direct and logical first step in troubleshooting is to verify that the dashboard's data source configuration has been updated to point to the new dataset name. This action directly addresses the most probable root cause introduced by the recent changes.

Why Incorrect Options are Wrong:

B. Filter the data in the dashboard.
Filtering manipulates data already loaded into the dashboard; it cannot resolve an issue where the dashboard is connected to the wrong, stale data source.
C. Escalate user permissions on the server.
The issue is stale data, not an access-denied error. Users can view the dashboard, indicating that permissions are not the primary problem.
D. Verify that the dashboard subscription is not expired.
Subscriptions relate to the automated delivery of reports (e.g., via email), not the data freshness of the live, interactive dashboard itself.

References:

1. Microsoft Power BI Documentation: In the official documentation for managing data sources, the process for modifying a connection is detailed. When a data source's location or name changes, the connection must be updated in the "Data source settings." This directly supports verifying the dashboard's pointer to the newest dataset.
Source: Microsoft Learn, "Manage data source settings," Section: "Edit data source or change data source."
2. Tableau Official Documentation: Tableau's documentation explains that workbooks contain a connection to a data source. If the underlying data source (e.g., a file or database table) is renamed or moved, the connection in the workbook must be edited to point to the new source to receive updates.
Source: Tableau Help, "Edit a Data Source," Section: "Replace a data source."
3. University Courseware on Data Governance: Data governance principles emphasize maintaining data lineage, which is the lifecycle of data, including its origins and transformations. A change in a dataset's name is a critical event in its lineage that must be propagated to all dependent systems, such as dashboards, to ensure data integrity and accuracy. Failing to update the connection breaks this lineage.
Source: DAMA International, "The DAMA Guide to the Data Management Body of Knowledge (DMBOK2)," Chapter 3: Data Governance, Section on Metadata Management. (This is a foundational text often used in university data management courses).

Question: 25

Data Governance

A developer builds an online survey that requires all questions to have an answer. Which of the following inconsistencies does this setting prevent?

A. Missing values

B. Duplication

C. Data corruption

D. Completeness

Answer:

A

Explanation:

The setting described, which requires all survey questions to have an answer, is a form of input validation. Its direct purpose is to ensure that no data field is left empty or null upon submission. In data analysis and management, an empty or null entry is referred to as a "missing value." By making all fields mandatory, the developer is proactively preventing the occurrence of missing values in the resulting dataset, which is a fundamental step in maintaining data quality at the point of collection.

Why Incorrect Options are Wrong:

B. Duplication: This setting does not prevent a user from submitting the identical survey multiple times, which is the cause of data duplication.
C. Data corruption: Data corruption involves unintended changes to data during storage or transmission (e.g., bit rot) and is unrelated to whether a field is mandatory.
D. Completeness: While preventing missing values improves data completeness, "missing values" is the specific inconsistency being prevented. Completeness is the broader data quality dimension that is the result of this action.

References:

1. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211-218. In Table 1, "Completeness" is defined as a data quality dimension, and its assessment often involves checking for missing values. The mandatory field setting is a direct mechanism to prevent these missing values. (DOI: https://doi.org/10.1145/505248.506010)
2. Batini, C., & Scannapieco, M. (2016). Data and Information Quality: Dimensions, Principles and

Techniques. Springer. Section 2.2.1, "Completeness," discusses how this dimension is violated by the presence of null or missing values in database tuples. The survey setting directly addresses this cause of incompleteness.
3. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. Chapter 3, "Data Preprocessing," Section 3.2.1, "Missing Values," explicitly details this common data quality problem and discusses methods to handle it. Requiring an answer at the point of entry is a preventative measure against this issue.