DA0-001 Q&As

CompTIA Data+

Pass CompTIA DA0-001 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.pass4itsure.com/da0-001.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by CompTIA Official Exam Center

QUESTION 1

An analyst has conducted a review of business questions. Which of the following should the analyst do next to conduct an analysis?

A. Determine the data needs and review the observations.

B. Determine the data needs and sources for analysis.

C. Determine the data needs and schedule interviews.

D. Determine the data needs and begin the analysis.

Correct Answer: B

After conducting a review of the business questions, the next step for the analyst is to determine the data needs and sources for analysis. This involves identifying the relevant data elements, variables, and metrics that are required to answer the business questions, as well as the data sources, formats, and quality that are available to access and use. This step will help the analyst to plan the data collection, preparation, and integration processes, as well as to assess the feasibility and limitations of the analysis1.

QUESTION 2

What SQL command is used to delete an entire table from a database?

A. DROP.

B. MODIFY.

C. DELETE.

D. ALTER.

Correct Answer: A

QUESTION 3

Which of the following database schemas features normalized dimension tables?

A. Flat

B. Snowflake

C. Hierarchical

D. Star

Correct Answer: B

Explanation: The correct answer is B. Snowflake.

A snowflake schema is a type of database schema that features normalized dimension tables. A database schema is a way of organizing and structuring the data in a database. A dimension table is a table that contains descriptive attributes or characteristics of the data, such as product name, category, color, etc. A normalized table is a table that follows the rules of normalization, which is a process of reducing data redundancy and improving data integrity by organizing the data into smaller and simpler tables12 A snowflake schema is a variation of the star schema, which is another type of database schema that features denormalized dimension tables. A denormalized table is a table that does not follow the rules of normalization, and may contain redundant or duplicated data. A star schema consists of a central fact table that contains quantitative measures or facts, such as sales amount, order quantity, etc., and several dimension tables that are directly connected to the fact table. A snowflake schema differs from a star schema in that the dimension tables are further split into sub-dimension tables, creating a snowflake-like shape13 A snowflake schema has some advantages and disadvantages over a star schema. Some advantages are: It reduces the storage space required for the dimension tables, as it eliminates the redundant data. It improves the data quality and consistency, as it avoids the update anomalies that may occur in denormalized tables. It allows more detailed analysis and queries, as it provides more levels of dimensions. Some disadvantages are: It increases the complexity and number of joins required to retrieve the data from multiple tables, which may affect the query performance and speed. It reduces the readability and simplicity of the schema, as it has more tables and relationships to understand. It may require more maintenance and administration, as it has more tables to manage and update13

QUESTION 4

Jhon is working on an ELT process that sources data from six different source systems.

Looking at the source data, he finds that data about the sample people exists in two of six systems.

What does he have to make sure he checks for in his ELT process?

Choose the best answer.

A. Duplicate Data.

B. Redundant Data.

C. Invalid Data.

D. Missing Data.

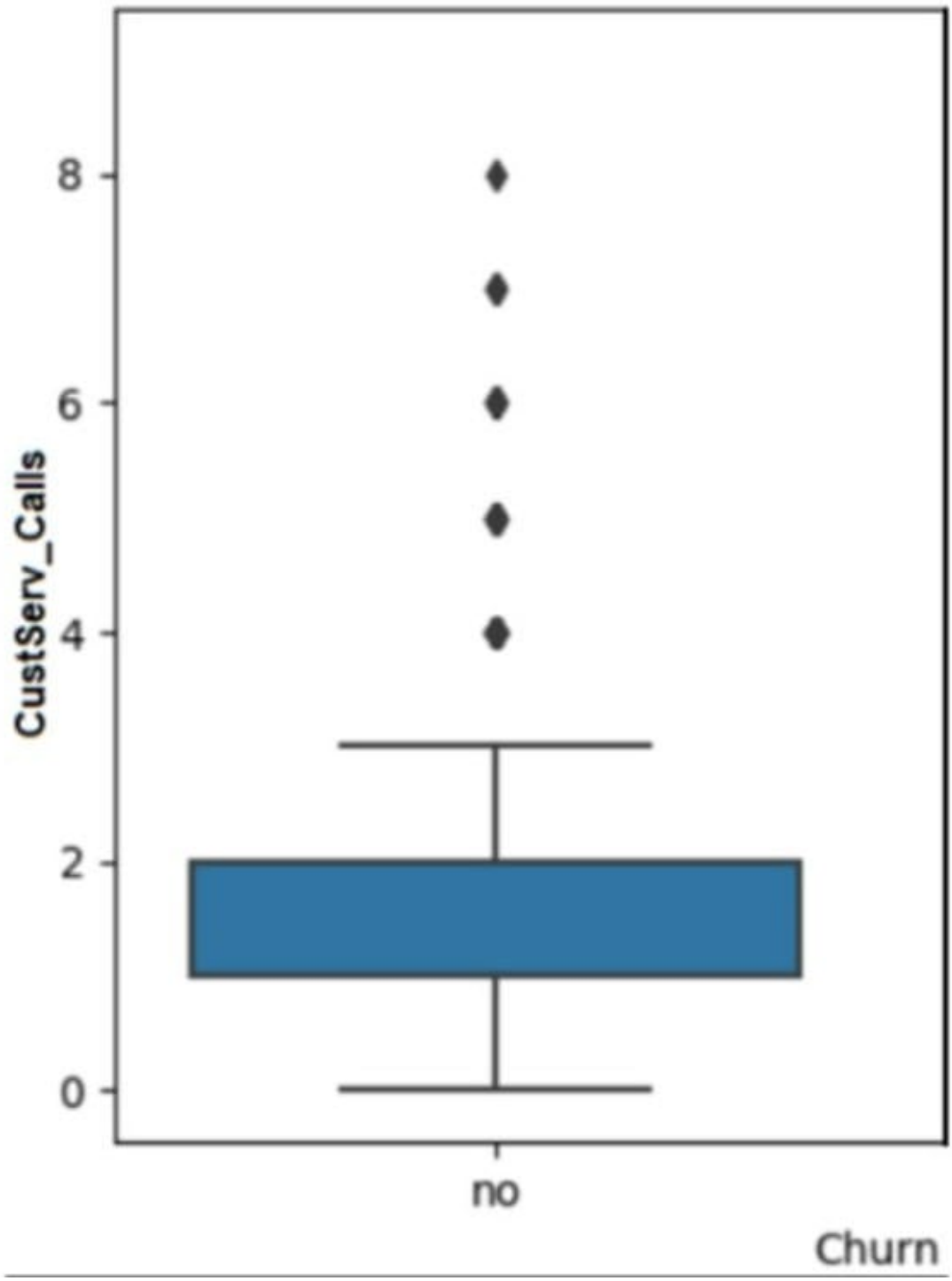
Correct Answer: C

Duplicate Data.

While invalid, redundant, or missing data are all valid concerns, data about people exists in two of the six systems. As such, Jhon needs to account for duplicate data issues.

QUESTION 5

Given the image below:



The data should be cleaned because of the presence of:

A. outlier B. non-parametric data.

C. multicollinearity.

D. invalid data.

Correct Answer: A

Explanation: The answer is A. Outlier.

Short explanation: An outlier is a data point that differs significantly from the rest of the data in a dataset. An outlier can indicate an error, an anomaly, or a rare event in the data. An outlier can affect the statistical analysis and visualization of

the data, such as skewing the mean, variance, or distribution of the data. Therefore, data should be cleaned to identify and remove or correct any outliers.

The image below shows a box plot graph with a vertical axis labeled "Customer Calls" and a horizontal axis labeled "Churn". The box plot is blue in color and the median value is around 2. There are 7 outliers above the box plot, ranging from

4 to 8.

image)

A box plot is a type of graph that can show the distribution of data values using five summary statistics: minimum, maximum, median, first quartile, and third quartile. The box represents the interquartile range (IQR), which is the difference

between the first and third quartiles. The median is shown as a line inside the box. The whiskers extend from the box to the minimum and maximum values, excluding any outliers. Outliers are shown as dots or circles outside the whiskers.

In this graph, we can see that most of the customer calls are between 0 and 4, with a median of 2. However, there are 7 outliers that have more than 4 customer calls, up to 8. These outliers may indicate some customers who have more

issues or complaints than others, or some errors or anomalies in the data collection or recording process. These outliers can affect the analysis and interpretation of the customer calls and churn relationship, such as making it seem that more

customer calls lead to less churn, which may not be true for the majority of the customers. Therefore, data should be cleaned to investigate and handle these outliers appropriately.

QUESTION 6

Which of the following is a control measure for preventing a data breach?

A. Data transmission

B. Data attribution

C. Data retention

D. Data encryption

Correct Answer: D

Explanation: This is because data encryption is a type of control measure that prevents a data breach, which is an unauthorized or illegal access or use of data by an external or internal party. Data encryption can prevent a data breach by protecting and securing the data using a code or a key that scrambles or transforms the data into an unreadable or incomprehensible format, which can only be decoded or restored by authorized users who have the correct code or key. For example, data encryption can prevent a data breach by encrypting the data in transit or at rest, such as when the data is sent over a network or stored in a device. The other control measures are not used for preventing a data breach. Here is why:

Data transmission is a type of process that transfers and exchanges data between different sources or systems, such as databases, cloud services, or web applications. Data transmission does not prevent a data breach, but rather exposes the data to potential risks or threats during the transfer or exchange. However, data transmission can be made more secure and less vulnerable to a data breach by using encryption or other methods, such as authentication or authorization. Data attribution is a type of feature or function that assigns and tracks the ownership and origin of the data, such as the creator, modifier, or source of the data. Data attribution does not prevent a data breach but rather provides information and evidence about the data provenance and history. However, data attribution can be useful for detecting and responding to a data breach by using audit logs or metadata to identify and trace any unauthorized or illegal access or use of the data. Data retention is a type of policy or standard that specifies and regulates the storage and preservation of the data, such as the duration, location, or format of the data. Data retention does not prevent a data breach, but rather affects the availability and accessibility of the data for future use or reference. However, data retention can be optimized and aligned with the legal and ethical requirements and standards of the industry or the organization to reduce the risk or impact of a data breach.

QUESTION 7

Given the following data: Which of the following BEST describes the data set?

Name	Gender	Age	Annual income
Ralph	M	27	$75,000
Jessie	F	3	$75,000
Monica	F	31	$125,000
Carlos	M	53	$75
Sara	F	43	$0

A. There is data bias.

B. The data is incomplete.

C. The data is inconsistent.

D. The data is outliers.

Correct Answer: C

Explanation: This is because inconsistency is a type of data quality issue that occurs when the data does not follow a common format, structure, or rule across different sources or systems, which can affect the efficiency and performance of the analysis or process. Inconsistency can be caused by having different spellings, punctuations, capitalizations, or abbreviations for the same or similar values in a data set, such as "M", "m", "Male", or "male" for gender in this case. Inconsistency can be eliminated or reduced by using data cleansing techniques, such as standardizing or normalizing the data values. The other options are not correct descriptions of the data set. Here is why:

Data bias is a type of data quality issue that occurs when the data is not representative or proportional of the population or the parameter, which can affect the validity and reliability of the analysis or process. Data bias can be caused by having a sample that is too small, too large, or too skewed for the population or the parameter, such as having only male customers for a product that targets both genders in this case. Data bias can be eliminated or reduced by using sampling techniques, such as stratified or cluster sampling. The data is incomplete is a type of data quality issue that

occurs when the data is absent or missing in a data set, which can affect the accuracy and reliability of the analysis or process. The data is incomplete can be caused by various factors, such as human error, system error, or non-response. The data is incomplete can be addressed by using various methods, such as replacing or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression. The data is outliers is a type of data quality issue that occurs when the data has values that are unusually high or low compared to the rest of the data set, which can affect the quality and validity of the analysis or process. The data is outliers can be caused by various factors, such as measurement error, natural variation, or extreme events. The data is outliers can be addressed by using various methods, such as removing or filtering out the outliers, or using robust statistics that are less sensitive to outliers, such as median, interquartile range, or box plot.

QUESTION 8

Which of the following are reasons to create and maintain a data dictionary? (Choose two.)

A. To improve data acquisition

B. To remember specifics about data fields

C. To specify user groups for databases

D. To provide continuity through personnel turnover

E. To confine breaches of PHI data

F. To reduce processing power requirements

Correct Answer: AB

Explanation: The reasons to create and maintain a data dictionary are to improve data acquisition and to remember specifics about data fields. A data dictionary is a document or a database that describes the structure, meaning, and usage of the data elements in a data source or a database. A data dictionary can help to improve data acquisition by providing clear and consistent definitions, rules, and standards for the data collection process. A data dictionary can also help to remember specifics about data fields by providing information such as data type, format, length, range, default value, constraints, relationships, etc. The other options are not reasons to create and maintain a data dictionary, as they are related to other aspects of data management or security. A data dictionary does not specify user groups for databases, as this is a function of access control or authorization. A data dictionary does not provide continuity through personnel turnover, as this is a function of documentation or knowledge transfer. A data dictionary does not confine breaches of PHI data, as this is a function of encryption or anonymization. A data dictionary does not reduce processing power requirements, as this is a function of optimization or compression. Reference: [What is a Data Dictionary? - DataCamp]

QUESTION 9

A military commander would like to see the health scorecards of the troops daily and filter them based on gender and rank. Considering this data is PHI, which of the following would be the best way for the commander to view the information?

A. An emailed report

B. A password-protected dashboard

C. A daily printout of a report

D. A cloud-hosted spreadsheet

Correct Answer: B

A password-protected dashboard is a type of web-based application that can display the health scorecards of the troops in a secure and interactive way. A password-protected dashboard can provide the following benefits for the commander: It can protect the PHI data from unauthorized access or disclosure by requiring a valid username and password to log in. This can ensure that only the commander and other authorized personnel can view the information12 It can allow the commander to filter the data based on gender and rank by using drop-down menus, sliders, checkboxes, or other controls. This can enable the commander to customize the view and focus on the relevant data13 It can update the data daily by connecting to a data source that refreshes automatically or on demand. This can ensure that the commander always sees the latest and most accurate information14 It can present the data in a visual and intuitive way by using charts, graphs, tables, or other elements. This can help the commander to understand and analyze the data more easily and effectively1

QUESTION 10

A data analyst is developing a data dictionary that aligns with a company\\'s data management processes and policies. Which of the following best describes what should be included in the data dictionary?

A. Information containing the links to business data

B. Information explaining the business methodologies

C. Information containing definitions of the business data

D. Information describing the data analysis phases

Correct Answer: C

QUESTION 11

Which one of the following values will appear first if they are sorted in descending order?

A. Aaron.

B. Molly.

C. Xavier.

D. Adam.

Correct Answer: C

Explanation: The value that will appear first if they are sorted in descending order is Xavier. Descending order means arranging values from the largest to the smallest, or from the last to the first in alphabetical order. In this case, Xavier is the last name in alphabetical order, so it will appear first when sorted in descending order. The other names will appear in the following order: Molly, Adam, Aaron. Reference: Sorting Data - W3Schools

QUESTION 12

A data analyst has been asked to create a sales report that calculates the rolling 12-month average for sales. If the report will be published on November 1, 2020, which of the following months shouts the report cover?

A. October 1, 2019 to October 31, 2020

B. October 31, 2020 to November 1, 2021

C. November 1, 2019 to October 31, 2020

D. October 31, 2019 to October 31, 2020

Correct Answer: A

The report should cover the months from October 1, 2019 to October 31, 2020. A rolling 12-month average is a type of moving average that calculates the average of the last 12 months of data for each month. It is useful for smoothing out seasonal fluctuations and identifying long-term trends in the data. To calculate the rolling 12-month average for sales for November 1, 2020, the analyst needs to use the sales data from the previous 12 months, starting from November 1, 2019 and ending on October 31, 2020. The other options are either too short or too long to cover the required period.

QUESTION 13

What analytics suite is offered by Microsoft and directly integrates with SQL Server Databases?

A. Qlik.

B. Power BI.

C. Domo.

D. Dataroma.

Correct Answer: B

Power BI is a collection of software services, apps, and connectors that work together to turn your unrelated sources of data into coherent, visually immersive, and interactive insights. Your data may be an Excel spreadsheet or a collection of cloud-based and on- premises hybrid data warehouses.

QUESTION 14

Which of the following are reasons to create and maintain a data dictionary? (Choose two.)

A. To improve data acquisition

B. To remember specifics about data fields

C. To specify user groups for databases

D. To provide continuity through personnel turnover

E. To confine breaches of PHI data

F. To reduce processing power requirements

Correct Answer: BD

A data dictionary is a collection of metadata that describes the data elements in a database or dataset. It can help improve data acquisition by providing information about the data sources, formats, quality, and usage. It can also help remember specifics about data fields, such as their names, definitions, types, sizes, and relationships. Therefore, options B and D are correct.

Option A is incorrect because it is not a reason to create and maintain a data dictionary, but a benefit of doing so.

Option C is incorrect because specifying user groups for databases is not a function of a data dictionary, but a function of a database management system or a security policy.

Option E is incorrect because confining breaches of PHI data is not a function of a data dictionary, but a function of a data protection or encryption system. Option F is incorrect because reducing processing power requirements is not a function of a data dictionary, but a function of a data compression or optimization system.

QUESTION 15

Joseph is interpreting a left skewed distribution of test scores. Joe scored at the mean, Alfonso scored at the median, and gaby scored and the end of the tail.

Who had the highest score?

A. Joseph

B. Joe

C. Alfonso

D. Gaby

Correct Answer: C

Alfonso had the highest score. A left skewed distribution is a distribution where the tail is longer on the left side than on the right side, meaning that most of the values are clustered on the right side and there are some outliers on the left side.

In a left skewed distribution, the mean is less than the median, which is less than the mode. Therefore, Joseph, who scored at the mean, had the lowest score, Gaby, who scored at the end of the tail, had the second lowest score, and

Alfonso, who scored at the median, had the highest score.

Reference: Skewness - Statistics How To