CompTIA Data+ Certification Test Bank

Welcome to the Bare Metal Cyber Audio Academy, an audio-first learning library built to help you pass certification exams through clear, practical instruction you can use anywhere. Each course is designed for busy learners who want structured coverage of the exam objectives, high-yield recall practice, and scenario-ready decision skills without needing slides, labs, or extra downloads. The handouts that come with each course are meant to reinforce what you hear, giving you quick reference material like question banks, key definitions, and review prompts you can use before a study session or right before exam day. The goal is simple: help you recognize what the exam is asking, select the best answer confidently, and avoid common traps. You can find the full catalog of courses, books, and resources at https://baremetalcyber.com/, where everything is organized to support steady progress from first listen to test-day readiness.

Find more for free at BareMetalCyber.com

Contents

Bank 1

1. Which scenario most strongly supports choosing a relational database over a non-relational database?
   A. Storing many small image thumbnails with flexible metadata that changes often
   B. Running frequent joins across well-defined tables with consistent keys and constraints
   C. Capturing high-volume clickstream events where the shape of each event varies by source
   D. Caching session tokens in a simple key-value pattern for fast lookups

2. Which description best matches a common strength of many non-relational databases?
   A. They require a fixed schema and reject records that do not match it
   B. They are designed primarily for complex multi-table joins
   C. They handle changing or variable data structures without forcing one strict table shape
   D. They only store flat files like CSV and XLSX

3. Which choice is the clearest example of structured tabular data stored in a plain-text format?
   A. XLSX
   B. JPG
   C. CSV
   D. DAT

4. Which file type most naturally represents nested objects and arrays in a single document-like structure?
   A. JSON
   B. CSV
   C. TXT
   D. JPG

5. Which statement about DAT files is the most accurate for exam decision-making?
   A. DAT always means a spreadsheet file created by Excel
   B. DAT is a generic “data file” label and the contents can vary by application
   C. DAT always stores human-readable text that can be opened safely in any text editor
   D. DAT is an image format similar to JPG

6. Which scenario most strongly supports choosing a non-relational database over a relational database?
   A. A finance ledger requiring strict consistency and multi-step transactions

B. A small inventory list with stable columns and a need for clear referential integrity
C. A dataset made of documents where fields differ by record and evolve frequently
D. A reporting dataset that must be normalized into related tables for long-term use

7. Which format is most likely to require special handling because it is typically binary rather than plain text?
A. TXT
B. XLSX
C. CSV
D. JSON

8. Which choice best explains why "schema" matters in database selection decisions?
A. Schema determines whether the data is encrypted at rest
B. Schema guarantees that all stored data is correct without validation
C. Schema controls whether data can be accessed over an API
D. Schema describes how data is organized, which affects how reliably it can be queried and joined

9. Which file type is most likely to be used for freeform notes or simple logs when no strict structure is required?
A. TXT
B. JSON
C. XLSX
D. CSV

10. Which situation most strongly suggests a relational model because relationships between entities must be enforced?
A. Storing raw photos and thumbnails with tags for later browsing
B. Keeping user profiles where every record has different optional fields
C. Maintaining customers, orders, and order items where keys must match and relationships must be consistent
D. Capturing sensor readings as semi-structured messages with changing attributes

---

1. Correct Answer: B. Running frequent joins across well-defined tables with consistent keys and constraints. Explanation: Relational databases are a strong fit when the work depends on predictable tables and reliable joins across shared keys. Constraints and consistent structure reduce ambiguity in how rows relate, which supports accurate queries.

2. Correct Answer: C. They handle changing or variable data structures without forcing one strict table shape. Explanation: Many non-relational designs accept records that do not all share the same fields, which helps when data evolves quickly. This flexibility can reduce friction when sources produce different shapes of data.

3. Correct Answer: C. CSV. Explanation: CSV is a plain-text format that represents rows and columns using delimiters, which fits structured tabular data. It is commonly used for simple exports and imports where a spreadsheet-like table is needed.

4. Correct Answer: A. JSON. Explanation: JSON naturally represents nested structures using objects and arrays, which fits document-style data. That makes it a common choice when data is hierarchical rather than strictly row-and-column.

5. Correct Answer: B. DAT is a generic “data file” label and the contents can vary by application. Explanation: DAT is often just a generic extension used by different programs, so the structure is not guaranteed. The safest exam choice is to treat DAT as ambiguous until the producing system is known.

6. Correct Answer: C. A dataset made of documents where fields differ by record and evolve frequently. Explanation: Non-relational databases often fit document-like data where each record can have different fields without constant schema redesign. This reduces friction when sources change and the structure is not uniform.

7. Correct Answer: B. XLSX. Explanation: XLSX is typically a binary packaged format rather than a simple text file, so it often needs a tool or library to parse reliably. Plain-text formats like CSV, JSON, and TXT are usually simpler to inspect directly.

8. Correct Answer: D. Schema describes how data is organized, which affects how reliably it can be queried and joined. Explanation: Schema is the blueprint for how data is structured, which influences query patterns and the clarity of relationships. Clear structure supports consistent analysis, especially when joins and keys must behave predictably.

9. Correct Answer: A. TXT. Explanation: TXT is commonly used for simple unstructured content like notes or basic logs. It does not imply a fixed tabular or nested structure, which can be appropriate when structure is not required.

10. Correct Answer: C. Maintaining customers, orders, and order items where keys must match and relationships must be consistent. Explanation: Relational models are well-suited when entities must relate through enforced keys and consistent relationships. This reduces errors such as orphan records and supports accurate multi-table queries.

Bank 2

1. Which example best fits **structured** data?
   A. A free-form customer complaint paragraph in a text file
   B. A JSON event payload with nested fields that vary by event type
   C. A table of orders with fixed columns like OrderID, Date, Amount
   D. A folder of scanned receipts saved as JPG images

2. Which example best fits **unstructured** data?
   A. A fact table with foreign keys to dimension tables
   B. A schema with defined columns and data types
   C. A JSON file with consistent keys across all records
   D. A collection of images and PDFs with no consistent fields

3. Which representation most naturally supports **nested** structures such as objects within objects and arrays of items?
   A. A structured table with fixed columns
   B. A JSON document
   C. A plain TXT file of notes
   D. A JPG image

4. Which scenario most strongly suggests using a **schema** to improve reliability of analysis?
   A. A dataset where every record has different fields and new fields appear weekly
   B. A single long narrative field used only for keyword searching
   C. A reporting dataset where consistent joins and repeatable metrics are required
   D. A set of images that will be manually reviewed one time

5. Which statement best describes a **schema**?
   A. A set of encryption settings applied to a database
   B. A blueprint that defines how data is organized, including tables, fields, and relationships
   C. A dashboard layout standard used by BI tools
   D. A sampling method used to reduce bias

6. Which choice is the best example of a **fact** table in a dimensional model?
   A. A Customers table containing names, addresses, and customer segment
   B. A Products table containing SKU, category, and brand
   C. A SalesTransactions table containing date, product ID, customer ID, quantity, and revenue
   D. A Calendar table containing date, fiscal quarter, and holiday flags

7. Which choice is the best example of a **dimension** table in a dimensional model?
   A. A table that records each payment and its dollar amount
   B. A table that records each website click event

C. A table that stores descriptive attributes like product category and brand
D. A table that stores daily totals already aggregated by month

8. What is the most accurate description of a **slowly changing dimension**?
   A. A fact table that grows slowly because it only stores monthly totals
   B. A dimension table where descriptive attributes can change over time and should be handled intentionally
   C. A table that must be stored in a data lake instead of a warehouse
   D. A schema that prevents any updates after initial load

9. Which example most clearly involves a **data type** decision that can break analysis if handled incorrectly?
   A. Choosing whether to store a chart as a PNG or JPG
   B. Choosing whether “00123” should be treated as a numeric value or an identifier string
   C. Choosing whether to use a bar chart or a map for a report
   D. Choosing whether to publish a report weekly or monthly

10. Which pairing is most likely to cause problems in joins due to **data type mismatch**?
    A. Joining an integer customer ID to a string customer ID that includes leading zeros
    B. Joining a date field to another date field in the same format
    C. Joining two string fields that both store ISO country codes
    D. Joining two numeric fields that both store revenue in dollars

---

1. Correct Answer: C. A table of orders with fixed columns like OrderID, Date, Amount. Explanation: Structured data has a consistent shape with defined fields, which fits rows and columns in a table. That consistency supports reliable filtering, grouping, and joining.

2. Correct Answer: D. A collection of images and PDFs with no consistent fields. Explanation: Unstructured data does not have a predictable field layout that tools can query like a table. Images and PDFs often require extra processing to extract meaning into usable fields.

3. Correct Answer: B. A JSON document. Explanation: JSON is built to represent nested objects and arrays directly. That makes it a natural fit for hierarchical or event-style data with embedded structures.

4. Correct Answer: C. A reporting dataset where consistent joins and repeatable metrics are required. Explanation: A schema creates a reliable structure so the same logic produces consistent results over time. This reduces ambiguity when multiple tables must join and metrics must remain stable.

5. Correct Answer: B. A blueprint that defines how data is organized, including tables, fields, and relationships. Explanation: A schema describes the structure of stored

data and how pieces relate. Clear structure supports consistent querying and reduces confusion about meaning.

6. Correct Answer: C. A SalesTransactions table containing date, product ID, customer ID, quantity, and revenue. Explanation: Fact tables record measurable events and typically include numeric measures plus keys to dimensions. Transactions with quantity and revenue are classic fact data.
7. Correct Answer: C. A table that stores descriptive attributes like product category and brand. Explanation: Dimension tables provide descriptive context used to slice and filter facts. Attributes like category and brand describe the "who or what" behind measures.
8. Correct Answer: B. A dimension table where descriptive attributes can change over time and should be handled intentionally. Explanation: Slowly changing dimensions capture changes like a customer address or product category over time. Handling the change intentionally preserves accurate historical reporting.
9. Correct Answer: B. Choosing whether "00123" should be treated as a numeric value or an identifier string. Explanation: Treating identifiers as numbers can drop leading zeros and change meaning. Correct data typing preserves identity and prevents failed joins or incorrect groupings.
10. Correct Answer: A. Joining an integer customer ID to a string customer ID that includes leading zeros. Explanation: If one side is numeric and the other is a string with leading zeros, values may not match even when they represent the same entity. This mismatch can create missing joins and misleading results.

Bank 3

1. Which choice is the best example of a database data source?
   A. A CRM system exposing tables through SQL
   B. A dashboard PDF exported from a BI tool
   C. A slide deck summarizing quarterly results
   D. A JPG image of a handwritten form

2. Which choice is the best example of an API data source?
   A. A manual copy and paste of values into a spreadsheet
   B. A web service that returns JSON responses to requests
   C. A printed report scanned into a PDF
   D. A folder of archived images

3. Which situation best supports using web scraping as a data source method?
   A. A public web page displays tables you need, but there is no provided API
   B. A secured internal database is available through SQL
   C. A vendor sends daily CSV exports by email
   D. A data warehouse already contains curated tables

4. Which data source is most associated with event-by-event operational records such as authentication attempts or application errors?
   A. Web scraping
   B. Files stored in XLSX format
   C. Logs
   D. Static infographics

5. Which statement best describes a data lake?
   A. A repository designed mainly for highly curated, modeled tables for reporting
   B. A repository that stores raw and varied data types, often before heavy modeling
   C. A repository that only stores structured tables in third normal form
   D. A repository that replaces the need for governance and access control

6. Which statement best describes a data warehouse in exam terms?
   A. A place for raw, unprocessed data that will not be transformed
   B. A repository optimized for curated, structured data used for analytics and reporting
   C. A system used only for unstructured data like images and audio
   D. A short-term cache that is automatically deleted after each query

7. Which choice best describes a data mart?
   A. A full enterprise repository covering every subject area
   B. A subject-focused slice of analytics data built for a specific team or domain
   C. A temporary table created during query optimization
   D. A log repository used only for security events

8. Which scenario most strongly suggests the presence of a silo?
   A. Multiple teams use a shared repository with consistent definitions and access controls
   B. A single source of truth exists for customer definitions across all systems
   C. One department keeps its own dataset with unique definitions that are not shared
   D. Raw and curated data are stored together with a common catalog

9. Which pairing is most accurate for choosing a repository type based on intended use?
   A. Data lake for curated reporting tables; data warehouse for raw ingestion
   B. Data warehouse for curated reporting tables; data lake for raw varied storage
   C. Data mart for raw ingestion at enterprise scale; warehouse for ad hoc personal files
   D. Silo for shared governance; warehouse for disconnected spreadsheets

10. Which option is the best reason to compare repositories during exam decision questions?
   A. Repository type determines chart colors and branding rules
   B. Repository type determines whether statistics can be computed
   C. Repository type affects how data is stored, governed, accessed, and queried for the goal
   D. Repository type eliminates the need to check data types and nulls

---

1. Correct Answer: A. A CRM system exposing tables through SQL. Explanation: Databases are a common source when data is stored in structured tables that can be queried directly. SQL access supports filtering, joining, and aggregating in a repeatable way.

2. Correct Answer: B. A web service that returns JSON responses to requests. Explanation: APIs provide data through request and response patterns, often returning JSON. They are a common way to pull data from SaaS platforms without direct database access.

3. Correct Answer: A. A public web page displays tables you need, but there is no provided API. Explanation: Web scraping is often used when needed data is visible on web pages but not available through an official interface. It is typically a fallback approach when structured access is not provided.

4. Correct Answer: C. Logs. Explanation: Logs capture time-ordered event records like errors, logins, and system actions. They are commonly used as a source for troubleshooting, monitoring, and behavior analysis.

5. Correct Answer: B. A repository that stores raw and varied data types, often before heavy modeling. Explanation: Data lakes commonly hold diverse data, including

semi-structured and unstructured content. They are often used earlier in the pipeline before data is fully modeled for reporting.

6. Correct Answer: B. A repository optimized for curated, structured data used for analytics and reporting. Explanation: Data warehouses typically emphasize cleaned, structured, and modeled data for consistent analytics. They support repeatable reporting and business-facing metrics.
7. Correct Answer: B. A subject-focused slice of analytics data built for a specific team or domain. Explanation: Data marts are designed around a department or subject area such as finance or marketing. They simplify access by focusing on the subset of data that group needs most.
8. Correct Answer: C. One department keeps its own dataset with unique definitions that are not shared. Explanation: Silos form when teams maintain isolated data with inconsistent definitions and limited sharing. This creates conflicting numbers and duplicated effort across the organization.
9. Correct Answer: B. Data warehouse for curated reporting tables; data lake for raw varied storage. Explanation: Warehouses commonly serve structured, modeled analytics use cases, while lakes commonly store raw or mixed-format data. This pairing matches typical exam decision logic about purpose and structure.
10. Correct Answer: C. Repository type affects how data is stored, governed, accessed, and queried for the goal. Explanation: Repository choice changes how data is organized and controlled, which influences what is feasible and efficient. It also shapes how teams find data, apply governance, and produce consistent results.

Bank 4

1. Which environment choice best fits a scenario where an organization keeps some workloads on-prem but also uses a public cloud for scalability?
   A. On-prem only
   B. Hybrid
   C. Single-tenant SaaS only
   D. Air-gapped only

2. Which environment choice best fits a scenario where all compute and storage run in a provider-managed platform rather than in the organization's own data center?
   A. Cloud
   B. On-prem
   C. Hybrid
   D. Offline local

3. Which benefit is most commonly associated with containers in data environments?
   A. They permanently increase internet bandwidth
   B. They package an application and its dependencies for consistent deployment
   C. They eliminate the need for any security controls
   D. They guarantee queries run faster without tuning

4. Which choice is the most appropriate tool for interactive, exploratory analysis and quick iteration in many data workflows?
   A. Notebook environment
   B. Image editor
   C. Ticketing system
   D. Password manager

5. Which choice best matches an IDE's typical role in data work?
   A. A place to run purely visual dashboards with no code
   B. A tool for writing, organizing, and debugging code with developer features
   C. A storage platform that replaces databases
   D. A method for collecting data through surveys

6. Which tool category is most directly associated with building dashboards and sharing reporting artifacts with stakeholders?
   A. BI platform
   B. Package manager
   C. Operating system shell
   D. Compression utility

7. Which environment selection most directly depends on data residency, control requirements, and latency constraints?
   A. Choosing chart type
   B. Choosing a KPI

C. Choosing cloud vs on-prem vs hybrid
D. Choosing a file extension

8. Which statement best describes a lakehouse in exam terms?
A. A repository that must be unstructured only
B. A model that combines elements of a data lake and a data warehouse
C. A dashboard template library
D. A storage device connected directly to a laptop

9. Which AI concept best describes systems that generate new text, images, or code based on patterns learned from training data?
A. Robotic process automation
B. Generative AI
C. Data normalization
D. Referential integrity

10. Which term best matches the idea of automating repetitive rule-based business tasks like copying fields between systems?
A. Deep learning
B. Natural language processing
C. Robotic process automation
D. Data drift

---

1. Correct Answer: B. Hybrid. Explanation: Hybrid environments combine on-prem resources with cloud services to balance control and scalability. This choice matches scenarios where some systems stay in a data center while others run in a public cloud.

2. Correct Answer: A. Cloud. Explanation: Cloud environments use provider-managed infrastructure for compute and storage rather than organization-owned hardware. This matches scenarios where the organization relies on a cloud provider's platform services.

3. Correct Answer: B. They package an application and its dependencies for consistent deployment. Explanation: Containers bundle runtime dependencies so the same workload runs consistently across environments. This supports repeatability across development, testing, and production systems.

4. Correct Answer: A. Notebook environment. Explanation: Notebooks support interactive exploration where code, output, and commentary can be iterated quickly. This fits common exam scenarios focused on rapid analysis and experimentation.

5. Correct Answer: B. A tool for writing, organizing, and debugging code with developer features. Explanation: IDEs provide features like debugging, linting, project

structure, and code navigation. These capabilities support maintaining larger codebases beyond quick exploration.

6. Correct Answer: A. BI platform. Explanation: BI platforms focus on dashboards, reports, and shared analytics artifacts. They are commonly used to publish and distribute reporting outputs to business audiences.
7. Correct Answer: C. Choosing cloud vs on-prem vs hybrid. Explanation: Residency, control, and latency constraints directly influence where workloads should run. Environment choice must align with governance needs and performance realities.
8. Correct Answer: B. A model that combines elements of a data lake and a data warehouse. Explanation: Lakehouse approaches aim to blend lake-style storage flexibility with warehouse-style analytics capabilities. This concept shows up in repository comparison decisions.
9. Correct Answer: B. Generative AI. Explanation: Generative AI refers to models that produce new content such as text, images, or code. The key idea is generation based on learned patterns, not just classification.
10. Correct Answer: C. Robotic process automation. Explanation: Robotic process automation targets repetitive rule-based tasks, often by mimicking user actions in systems. It is distinct from machine learning methods that learn from data to make predictions.

Bank 5

1. When combining two sources, what is the most important step to preserve meaning and avoid incorrect matches?
   A. Convert all columns to text so matching is easier
   B. Sort both datasets in the same order before combining
   C. Remove all nulls before any merge occurs
   D. Confirm what the keys represent in each source and how they align

2. Which merge approach is best when you need to add columns from a second table based on a shared key?
   A. Union
   B. Join
   C. Concatenation
   D. Aggregation

3. Which join type returns only rows where matching keys exist in both tables?
   A. Inner join
   B. Left join
   C. Full outer join
   D. Right join

4. Which operation is best when you need to stack two datasets with the same columns into one longer dataset?
   A. Join
   B. Aggregation
   C. Union
   D. Nested query

5. Which action is required to calculate totals by category, such as revenue by region?
   A. Sorting by region then filtering
   B. Grouping by region and applying an aggregate
   C. Joining the table to itself
   D. Converting all numerics to strings

6. What is a common reason to use a nested query or subquery?
   A. To permanently store results in a new database
   B. To convert JSON into a table without parsing
   C. To avoid using any filters in the query
   D. To filter or select records based on an intermediate result, such as an aggregated threshold

7. Which change is most likely to improve performance when filtering or joining repeatedly on the same fields?
   A. Add indexes on the join and filter columns
   B. Use more DISTINCT operations to reduce duplicates

C. Apply functions to the indexed fields to standardize formatting
D. Increase the number of columns returned to avoid lookups

8. Why are parameterized queries valuable in repeatable analytics work?
   A. They force the database to ignore indexes for consistency
   B. They automatically remove nulls and outliers
   C. They allow the same query logic to be reused safely with different inputs
   D. They guarantee correct results without validation

9. Which choice best describes the difference between ETL and ELT?
   A. ETL is only for APIs, while ELT is only for files
   B. ETL always uses SQL, while ELT never uses SQL
   C. ETL loads before extracting, while ELT extracts after transforming
   D. ETL transforms data before loading it into the target, while ELT loads first then transforms in the target

10. Which approach best fits a situation where you need opinions from a target population, not just system-generated events?
    A. Log collection
    B. Surveys
    C. Web scraping
    D. Indexing

---

1. Correct Answer: D. Confirm what the keys represent in each source and how they align. Explanation: Keys can look similar but represent different things, like account ID versus customer ID, which leads to incorrect matches. Confirming key meaning protects accuracy before choosing join type or transformation steps.

2. Correct Answer: B. Join. Explanation: A join adds columns from another table by matching rows on a shared key. This is the standard pattern when one dataset provides additional attributes for the same entities.

3. Correct Answer: A. Inner join. Explanation: An inner join keeps only the records where a matching key exists on both sides. This prevents unmatched rows from appearing in the result set.

4. Correct Answer: C. Union. Explanation: Union stacks rows from datasets that share the same column structure, creating a longer table. It is the appropriate choice when combining similar records from two sources into one dataset.

5. Correct Answer: B. Grouping by region and applying an aggregate. Explanation: Grouping defines the categories, and aggregates compute summaries like totals or averages within each group. Without grouping, the aggregate would apply to the entire dataset rather than each region.

6. Correct Answer: D. To filter or select records based on an intermediate result, such as an aggregated threshold. Explanation: Nested queries are commonly used when one step depends on the output of a prior step, such as finding categories above a certain total. This keeps the logic precise when a simple row-level filter is not enough.

7. Correct Answer: A. Add indexes on the join and filter columns. Explanation: Indexes help the database locate and match rows more efficiently for frequent filters and joins. This often reduces scans and improves runtime when the query pattern aligns with the index.

8. Correct Answer: C. They allow the same query logic to be reused safely with different inputs. Explanation: Parameterization supports repeatability without rewriting the query for each value or date range. It also reduces copy-paste errors and makes query behavior more consistent.

9. Correct Answer: D. ETL transforms data before loading it into the target, while ELT loads first then transforms in the target. Explanation: ETL emphasizes transformation upstream before the destination store, while ELT relies on the target environment to do transforms after loading. The distinction affects tooling choices and where compute work happens.

10. Correct Answer: B. Surveys. Explanation: Surveys collect direct responses from people and are designed to capture opinions, perceptions, or self-reported behaviors. Logs and scraping capture observed events or published content rather than elicited responses.

Bank 6

1. What is the most reliable first step when you suspect missing values are affecting results?
   A. Drop any row with a null immediately to simplify the dataset
   B. Convert all nulls to zero so calculations do not fail
   C. Identify where nulls occur and look for patterns by column and source
   D. Change the chart type so nulls are less visible

2. Which situation most strongly suggests that missing values are not random and need extra caution?
   A. Nulls appear evenly across all columns and sources
   B. Nulls cluster in one field after a specific source system change
   C. Nulls occur only in derived fields created during analysis
   D. Nulls appear only in a text notes column

3. Which action is the best first move when you discover duplicate records in a dataset?
   A. Remove duplicates without checking why they exist
   B. Determine what should be unique and confirm the intended key or grain
   C. Convert duplicates into separate tables automatically
   D. Replace duplicate values with nulls

4. Which example best matches redundancy rather than duplication?
   A. The same transaction row appears twice with identical values
   B. A dimension key is missing from several rows
   C. The dataset contains both "StateName" and "StateAbbrev" that describe the same concept
   D. A numeric field contains negative values unexpectedly

5. Which issue most directly threatens the completeness of a dataset?
   A. A small number of outliers in a numeric field
   B. Missing records for an entire day due to a pipeline failure
   C. A column that uses both uppercase and lowercase text
   D. A report that refreshes every hour instead of daily

6. Which scenario most clearly indicates a validation issue rather than a natural outlier?
   A. A handful of customers spend far more than average during a holiday
   B. A temperature sensor reads higher values during summer months
   C. A revenue field contains a negative value where refunds are not possible in the business rules
   D. Website traffic spikes after a marketing campaign

7. What is the most important reason to separate true outliers from data errors before analysis?

A. Outliers always need to be deleted
B. Outliers guarantee a model will overfit
C. Outliers reduce the need for normalization
D. Errors can mislead conclusions, while true extremes may be meaningful signals

8. Which check best helps detect duplication caused by joining at the wrong grain?
A. Confirm expected row counts before and after the join and look for unexpected multiplication
B. Convert the join keys to text
C. Remove all nulls before joining
D. Sort both tables alphabetically before joining

9. Which situation best represents a completeness issue tied to a source limitation rather than a processing bug?
A. A file import fails for one day because the server was down
B. A survey only includes respondents from one region, leaving others unrepresented
C. A join drops rows because the join keys are mismatched types
D. A report shows stale data due to refresh delays

10. Which step best supports defensible decisions when you detect quality issues like duplicates, outliers, and missing data?
A. Hide the problematic rows so the report looks clean
B. Change the chart colors to reduce attention
C. Document the issue, the checks performed, and the resolution or decision
D. Skip validation if the results look reasonable

---

1. Correct Answer: C. Identify where nulls occur and look for patterns by column and source. Explanation: The safest first step is to measure and locate missingness so you understand scope and potential cause. Patterns often point to a specific field, system, or time window that explains the problem.

2. Correct Answer: B. Nulls cluster in one field after a specific source system change. Explanation: Clustering after a known change suggests a systematic cause, not random missingness. That kind of pattern can bias results and usually requires investigation before imputation or removal.

3. Correct Answer: B. Determine what should be unique and confirm the intended key or grain. Explanation: Duplicate handling depends on what "unique" means for the dataset's purpose, such as one row per transaction or one row per customer-day. Confirming the grain prevents deleting valid repeated events or keeping invalid duplicates.

4. Correct Answer: C. The dataset contains both "StateName" and "StateAbbrev" that describe the same concept. Explanation: Redundancy is about repeated

information across fields, which can cause confusion and inconsistency. Duplication is repeated rows, while redundancy is repeated meaning across columns.

5. Correct Answer: B. Missing records for an entire day due to a pipeline failure. Explanation: Completeness is about having all required records for the period and scope. A missing day creates a gap that can distort trends and totals.

6. Correct Answer: C. A revenue field contains a negative value where refunds are not possible in the business rules. Explanation: When values violate known rules, the issue is likely validation rather than a meaningful extreme. Business-rule checks help separate true signals from data entry or processing errors.

7. Correct Answer: D. Errors can mislead conclusions, while true extremes may be meaningful signals. Explanation: Treating all outliers as errors can delete important real-world behavior, like fraud spikes or rare events. Treating errors as true values can distort averages, trends, and model behavior.

8. Correct Answer: A. Confirm expected row counts before and after the join and look for unexpected multiplication. Explanation: Wrong-grain joins often multiply rows, creating inflated totals and apparent duplicates. Row-count checks and spot checks catch this early before downstream metrics are corrupted.

9. Correct Answer: B. A survey only includes respondents from one region, leaving others unrepresented. Explanation: This is a completeness issue caused by the collection design and coverage, not a processing failure. The dataset is incomplete for the intended population even if the pipeline ran correctly.

10. Correct Answer: C. Document the issue, the checks performed, and the resolution or decision. Explanation: Quality issues require a traceable explanation of what was found and what was done about it. Documentation supports review, repeatability, and trust in the final analysis.

Bank 7

1. Which task best represents using regular expressions in data cleaning?
   A. Calculating an average from a numeric column
   B. Sorting a table by a date field
   C. Extracting a ZIP code pattern from an address string
   D. Adding an index to a join key

2. Which approach is most appropriate when converting a text field into a consistent, analysis-ready format?
   A. Standardization of strings through parsing and conversion
   B. Adding decorative effects to a chart for readability
   C. Replacing all text fields with null values
   D. Using a full outer join to preserve all rows

3. Which example best fits parsing rather than simple standardization?
   A. Converting "texas" and "Texas" to "TEXAS" consistently
   B. Trimming extra spaces from names
   C. Splitting "Last, First" into separate LastName and FirstName fields
   D. Converting a numeric ID to a string to keep leading zeros

4. Which reshape operation is best when you need to turn a single column containing multiple values into multiple rows?
   A. Appending
   B. Exploding
   C. Deleting
   D. Binning

5. Which reshape operation is best when you need to stack two datasets with the same columns into one longer dataset?
   A. Appending
   B. Deleting
   C. Scaling
   D. Imputation

6. Which reshape operation is most associated with combining related tables using keys to create a wider dataset?
   A. Exploding
   B. Augmenting
   C. Merging
   D. Binning

7. Which feature engineering technique best matches grouping continuous values into ranges such as ages 0–9, 10–19, and so on?
   A. Scaling
   B. Binning

C. Parsing
D. Concatenation

8. Which feature engineering technique is most directly aimed at bringing numeric values onto a comparable scale across features?
A. Scaling
B. Deleting
C. Exploding
D. Web scraping

9. Which approach best describes imputation in data preparation?
A. Removing all rows that contain null values without review
B. Converting all null values to zero in every column
C. Filling missing values using a defined method appropriate to the context
D. Encrypting sensitive columns before analysis

10. Which example best represents creating a derived variable during feature creation?
A. Renaming a column from “DOB” to “BirthDate”
B. Converting all text to lowercase
C. Joining two tables on CustomerID
D. Creating “DaysSinceLastPurchase” from date fields

---

1. Correct Answer: C. Extracting a ZIP code pattern from an address string. Explanation: Regular expressions are used to match patterns in text and extract or validate parts of strings. ZIP code extraction is a classic pattern-matching use case.

2. Correct Answer: A. Standardization of strings through parsing and conversion. Explanation: Converting and standardizing text ensures consistent representation, which improves grouping, matching, and downstream analysis. Parsing and conversion are common steps to make strings analysis-ready.

3. Correct Answer: C. Splitting “Last, First” into separate LastName and FirstName fields. Explanation: Parsing breaks a field into meaningful components based on separators or rules. Standardization changes representation but does not necessarily separate components into new fields.

4. Correct Answer: B. Exploding. Explanation: Exploding turns one record with a multi-valued field into multiple records, one per value. This makes the data easier to analyze in row-based tools.

5. Correct Answer: A. Appending. Explanation: Appending stacks rows from datasets that share the same schema into a single dataset. This is the typical reshape choice when combining like-for-like records.

6. Correct Answer: C. Merging. Explanation: Merging combines datasets using keys to produce a wider result with more columns. This is the common reshape pattern when enriching records with related attributes.

7. Correct Answer: B. Binning. Explanation: Binning converts continuous values into discrete ranges, which can simplify analysis and modeling. Age ranges are a common example of binning.

8. Correct Answer: A. Scaling. Explanation: Scaling adjusts numeric features so they are comparable in magnitude, which can help certain analytical methods. It reduces dominance of large-scale features over small-scale ones.

9. Correct Answer: C. Filling missing values using a defined method appropriate to the context. Explanation: Imputation replaces missing values using a chosen rule such as mean, median, or a model-based approach. The method should match the meaning of the field and the reason the value is missing.

10. Correct Answer: D. Creating “DaysSinceLastPurchase” from date fields. Explanation: A derived variable is computed from existing fields to capture useful information in a new form. Time-since measures are common derived features that improve analysis and decision-making.

Bank 8

1. Which artifact is most appropriate for aligning expectations with a requester before building a final dashboard?
   A. A mock-up showing layout and key elements
   B. A binary DAT export from the source system
   C. A refreshed snapshot feed at one-minute intervals
   D. A normalization script that removes duplicates

2. Which accessibility choice most directly improves comprehension for viewers with color-vision differences?
   A. Use subtle brand colors with low contrast to look clean
   B. Rely only on color and remove labels to reduce clutter
   C. Ensure sufficient contrast and avoid encoding meaning with color alone
   D. Use gradients and 3D effects to make values pop

3. Which situation most strongly suggests tailoring the communication style for a non-technical audience?
   A. Sharing query execution plans and index statistics
   B. Explaining findings using plain language and focusing on impact
   C. Publishing raw log lines for transparency
   D. Presenting a schema diagram with primary and foreign keys

4. Which approach best fits communicating results to an external audience compared to an internal technical team?
   A. Provide only SQL code and a data dictionary
   B. Provide detailed stack traces and error logs
   C. Provide context, definitions, and careful wording with only necessary technical detail
   D. Provide no explanations because the numbers speak for themselves

5. Which choice best represents selecting the right level of detail using a persona?
   A. Using the same dense technical write-up for every reader
   B. Adjusting detail based on a defined reader role such as executive, analyst, or operator
   C. Removing all numbers so nothing can be challenged
   D. Switching every chart to a table to avoid visuals

6. Which scenario most directly involves sensitivity considerations in reporting?
   A. Deciding whether to use mean or median for skewed data
   B. Choosing a bar chart rather than a pie chart
   C. Sharing customer-level results that include personal identifiers
   D. Joining two tables on a shared key

7. Which KPI design best matches the idea that metrics should answer the business question?

A. A metric chosen because it is easy to compute, even if it does not map to the goal
B. A metric that directly measures progress toward the stated outcome
C. A metric that changes every week so it stays interesting
D. A metric that is only a technical system count without context

8. Which example best fits a KPI framing choice rather than a chart formatting choice?
A. Deciding whether to track “on-time delivery rate” to measure fulfillment performance
B. Choosing a font size for chart titles
C. Selecting a color palette for accessibility
D. Removing gridlines to reduce clutter

9. Which statement best describes a common mistake when choosing the “right detail” for an audience?
A. Providing the exact same depth of detail to all audiences regardless of need
B. Using clear labels rather than legends
C. Including definitions for important terms
D. Checking accessibility contrast

10. Which action best supports clear communication when results could be misread or taken out of context?
A. Remove caveats and assumptions so the message is shorter
B. Add context, define terms, and state limits in plain language
C. Show only the most extreme values for impact
D. Hide uncertainty by rounding aggressively

---

1. Correct Answer: A. A mock-up showing layout and key elements. Explanation: Mock-ups help confirm requirements and expectations before investing in a full build. They reduce rework by making the intended layout and message explicit early.

2. Correct Answer: C. Ensure sufficient contrast and avoid encoding meaning with color alone. Explanation: Accessibility improves when viewers can distinguish elements without relying solely on color. Good contrast and redundant cues like labels reduce misinterpretation.

3. Correct Answer: B. Explaining findings using plain language and focusing on impact. Explanation: Non-technical audiences usually need outcomes and implications more than implementation detail. Plain language supports comprehension and decision-making.

4. Correct Answer: C. Provide context, definitions, and careful wording with only necessary technical detail. Explanation: External audiences often need more context and less internal jargon because they lack inside knowledge. Careful wording reduces misinterpretation and protects credibility.

5. Correct Answer: B. Adjusting detail based on a defined reader role such as executive, analyst, or operator. Explanation: Personas help decide what details matter for a reader's decisions. This prevents both overloading and oversimplifying the message.
6. Correct Answer: C. Sharing customer-level results that include personal identifiers. Explanation: Sensitivity concerns increase when information can identify individuals or reveal private details. Handling identifiers requires careful scope, access control, and appropriate aggregation.
7. Correct Answer: B. A metric that directly measures progress toward the stated outcome. Explanation: KPIs are most useful when they map cleanly to the business question and desired outcome. This keeps reporting focused on decisions rather than vanity metrics.
8. Correct Answer: A. Deciding whether to track "on-time delivery rate" to measure fulfillment performance. Explanation: KPI framing is about selecting what to measure to reflect success. Chart formatting is about how to display a chosen measure, not which measure matters.
9. Correct Answer: A. Providing the exact same depth of detail to all audiences regardless of need. Explanation: Different audiences need different levels of detail to make decisions effectively. Overly technical detail can confuse, while overly shallow detail can mislead.
10. Correct Answer: B. Add context, define terms, and state limits in plain language. Explanation: Context and definitions prevent readers from making incorrect assumptions about what the numbers represent. Stating limits builds trust and reduces the chance of overclaiming.

Bank 9

1. Which approach focuses on summarizing what happened in the data using summaries and basic patterns rather than predicting future outcomes?
   A. Descriptive
   B. Predictive
   C. Prescriptive
   D. Inferential

2. Which approach focuses on forecasting outcomes such as demand next month or churn risk?
   A. Inferential
   B. Descriptive
   C. Predictive
   D. Prescriptive

3. Which approach focuses on recommending actions, such as which option should be chosen to reach a goal?
   A. Prescriptive
   B. Predictive
   C. Descriptive
   D. Inferential

4. Which approach most directly addresses drawing conclusions about a population from a sample using statistical reasoning?
   A. Predictive
   B. Inferential
   C. Descriptive
   D. Prescriptive

5. Which measure of central tendency is most sensitive to extreme outliers?
   A. Mode
   B. Median
   C. Mean
   D. Standard deviation

6. Which measure of central tendency is typically best for a highly skewed distribution, such as income?
   A. Mean
   B. Median
   C. Mode
   D. Variance

7. Which measure identifies the most frequently occurring value in a dataset?
   A. Mean
   B. Standard deviation

C. Median
D. Mode

8. Which measure describes spread by using the average squared distance from the mean?
A. Standard deviation
B. Variance
C. Median
D. Mode

9. Which measure is the square root of variance and is commonly used to describe typical spread in the same units as the data?
A. Mean
B. Mode
C. Standard deviation
D. Range

10. Which statement best describes why date and string functions appear in many exam scenarios?
A. They replace the need for data cleaning
B. They help transform fields for grouping, filtering, and consistent reporting
C. They guarantee correct KPIs without validation
D. They eliminate the need for joins and keys

---

1. Correct Answer: A. Descriptive. Explanation: Descriptive approaches summarize the data as it is, using counts, averages, and basic patterns. The goal is understanding what happened, not predicting or prescribing actions.

2. Correct Answer: C. Predictive. Explanation: Predictive approaches use historical data to forecast likely future outcomes. The focus is estimating what will happen, such as risk, demand, or probability.

3. Correct Answer: A. Prescriptive. Explanation: Prescriptive approaches recommend what to do based on goals and constraints. The output is a suggested action or decision, not just a forecast.

4. Correct Answer: B. Inferential. Explanation: Inferential statistics use samples to draw conclusions about a broader population. This includes estimating parameters and testing whether observed differences are likely real.

5. Correct Answer: C. Mean. Explanation: The mean is pulled toward extreme values because it uses every value directly in the calculation. A single very large or small outlier can shift it noticeably.

6. Correct Answer: B. Median. Explanation: The median depends on rank order rather than magnitude, so outliers have less influence. That makes it a better “typical” value for skewed distributions.

7. Correct Answer: D. Mode. Explanation: The mode is the value that occurs most often. It is useful for categorical data and for spotting common repeated values in numeric sets.

8. Correct Answer: B. Variance. Explanation: Variance measures spread by averaging squared deviations from the mean. Squaring emphasizes larger deviations and is foundational for standard deviation.

9. Correct Answer: C. Standard deviation. Explanation: Standard deviation is the square root of variance, so it returns to the original units. It is commonly used to describe typical spread around the mean.

10. Correct Answer: B. They help transform fields for grouping, filtering, and consistent reporting. Explanation: Date and string functions are often needed to normalize formats, extract parts of dates, and standardize text values. These transformations support accurate grouping, filtering, and reporting logic.

Bank 10

1. What is the best first check when a dashboard suddenly shows blank visuals due to a suspected connectivity issue?
   A. Immediately rebuild the entire dashboard from scratch
   B. Confirm the data source connection details and credentials still work
   C. Change chart colors to force a refresh
   D. Delete calculated fields to reduce complexity

2. What is the most appropriate first step when you suspect the data source itself is producing corrupted or incomplete records?
   A. Validate the source data at the point of origin and compare it to the expected structure
   B. Increase the refresh frequency so errors are overwritten faster
   C. Replace missing values with zeros so outputs look complete
   D. Disable logging so the system runs faster

3. Which evidence most directly supports diagnosing a user-reported problem as a data refresh issue rather than a calculation bug?
   A. A chart uses a different font than last week
   B. The displayed timestamps are older than the expected refresh interval
   C. The dashboard contains multiple visuals on one page
   D. A KPI uses a percentage sign

4. Which action best supports isolating whether a failure is caused by a SQL syntax issue versus a data issue?
   A. Rewrite the query in a different language to see if it works
   B. Remove all WHERE clauses so the query always returns results
   C. Check the query logs and error messages and reproduce the query against the source
   D. Publish the report to a different workspace

5. Which scenario most strongly suggests a permissions problem rather than a broken query?
   A. The same query works for an analyst but fails for a viewer account
   B. The query returns too many rows for the chart to render
   C. The dataset contains outliers in a numeric field
   D. The report colors changed after a theme update

6. When users report “numbers changed,” which check best confirms whether the change came from the source system rather than the report logic?
   A. Compare current source extracts to prior extracts and check source change history
   B. Replace the KPI with a different metric that looks stable

   C. Remove labels and legends to reduce confusion

   D. Use a different chart type to see if the numbers look better

7. Which artifact is most useful for tracing the root cause of a recurring data issue across refresh cycles?

   A. A brand style guide for reports

   B. Execution logs and refresh history with timestamps

   C. A list of preferred chart types

   D. A set of dashboard mock-ups

8. Which step best supports solving a reported issue by validating assumptions before changing code?

   A. Apply a quick fix in production to satisfy the user

   B. Ask users to stop using the report

   C. Reproduce the issue with the same filters, account, and time window the user used

   D. Add more visuals so the report feels complete

9. Which response best fits using communities or documentation in troubleshooting without guessing?

   A. Copy the first suggested fix from a forum without checking sources

   B. Use official documentation and reputable community threads to interpret a specific error message

   C. Disable constraints so the data loads regardless of validity

   D. Ignore error messages if the report sometimes works

10. Which approach best supports preventing repeat issues after resolving a SQL or user-reported problem?

    A. Remove monitoring alerts so noise is reduced

    B. Stop refreshing the dataset to avoid failures

    C. Hide the failing visual so stakeholders do not notice

    D. Document the root cause and add checks or monitoring tied to the failure mode

---

1. Correct Answer: B. Confirm the data source connection details and credentials still work. Explanation: Connectivity failures often come from expired credentials, changed endpoints, or network access issues. Verifying the connection first avoids wasted effort debugging visuals that are fine but cannot reach data.

2. Correct Answer: A. Validate the source data at the point of origin and compare it to the expected structure. Explanation: If corruption is upstream, fixing dashboards will not correct the underlying bad input. Checking the source structure and expected fields helps identify where the pipeline broke.

3. Correct Answer: B. The displayed timestamps are older than the expected refresh interval. Explanation: Stale timestamps indicate that the data likely did not refresh

or the refresh failed. That points to refresh and connectivity before recalculation logic.

4. Correct Answer: C. Check the query logs and error messages and reproduce the query against the source. Explanation: Error messages and logs often distinguish syntax problems from missing objects or type mismatches. Reproducing against the source confirms whether the issue is query text, permissions, or underlying data.
5. Correct Answer: A. The same query works for an analyst but fails for a viewer account. Explanation: Different outcomes by account commonly indicate permissions or role-based access limits. If the query logic were broken, it would typically fail consistently for all users.
6. Correct Answer: A. Compare current source extracts to prior extracts and check source change history. Explanation: Source systems can change data due to late-arriving records, corrections, or upstream logic updates. Comparing extracts and change history helps separate source change from report calculation issues.
7. Correct Answer: B. Execution logs and refresh history with timestamps. Explanation: Logs show when jobs ran, what failed, and what changed across cycles. Timestamps create a timeline that supports root-cause analysis instead of guesswork.
8. Correct Answer: C. Reproduce the issue with the same filters, account, and time window the user used. Explanation: Reproduction controls variables and confirms the problem is real and repeatable. It also helps isolate whether the cause is filtering, permissions, caching, or data timing.
9. Correct Answer: B. Use official documentation and reputable community threads to interpret a specific error message. Explanation: Documentation and trusted community sources help decode error messages accurately. This reduces trial-and-error changes that can introduce new problems.
10. Correct Answer: D. Document the root cause and add checks or monitoring tied to the failure mode. Explanation: Documentation preserves what was learned and helps future responders fix issues faster. Monitoring and checks catch the same failure earlier and reduce repeat incidents.

Bank 11

1. Which visual is generally best for comparing values across categories, such as sales by product line?
   A. Bar chart
   B. Histogram
   C. Box plot
   D. Scatter plot

2. Which visual is typically best for showing geographic patterns, such as incidents by state?
   A. Pivot table
   B. Map
   C. Infographic
   D. Gauge

3. Which output is most appropriate when a reader needs to inspect exact values and sort or filter them quickly?
   A. Infographic
   B. Map
   C. Pivot table
   D. Word cloud

4. Which design choice most directly reduces clutter and improves readability in a visualization?
   A. Add more gridlines to emphasize precision
   B. Increase the number of colors to separate categories
   C. Use 3D effects to make bars stand out
   D. Remove unnecessary gridlines and visual noise

5. Which practice best supports accessibility when using color in charts?
   A. Encode categories using color only and remove labels
   B. Use sufficient contrast and include non-color cues like labels or patterns
   C. Use subtle low-contrast colors so the design looks modern
   D. Use as many colors as possible to avoid legends

6. Which labeling approach typically improves clarity when it can be done cleanly?
   A. Rely on a legend even when direct labels fit
   B. Remove labels to reduce distractions
   C. Use direct labeling instead of forcing readers to look back and forth to a legend
   D. Use abbreviations everywhere to save space

7. Which encoding is most likely to mislead readers by exaggerating differences between values?
   A. Using a shared zero baseline for bar charts
   B. Truncating the y-axis so small differences look large

C. Using clear axis labels with units
D. Keeping category names consistent across visuals

8. Which choice best represents matching the visual to the message rather than picking a visual based on preference?
A. Using a pie chart for every chart because it is familiar
B. Choosing a map only when the question is about location-based patterns
C. Adding branding elements so stakeholders recognize the dashboard
D. Using gradients and shadows to make visuals more engaging

9. Which action best tests whether a chart's message is clear without extra explanation?
A. Add additional decorative icons to guide attention
B. Remove titles so the viewer interprets freely
C. Describe the chart in words and confirm the description matches the intended takeaway
D. Add more categories so the chart feels complete

10. Which combination of design practices best reduces misinterpretation in business reporting visuals?
A. Decorative effects, multiple axes, and heavy branding
B. Clear labels, consistent units, and uncluttered design
C. No labels, minimal context, and high color variety
D. Truncated axes, gradients, and dense legends

---

1. Correct Answer: A. Bar chart. Explanation: Bar charts are well-suited for comparing categories because values share a common baseline. This makes rankings and differences easy to see quickly.

2. Correct Answer: B. Map. Explanation: Maps are designed to show geographic distribution and location-based patterns. They help viewers connect values to places more naturally than non-spatial charts.

3. Correct Answer: C. Pivot table. Explanation: Pivot tables support quick inspection of exact values along with sorting and filtering. They are useful when the reader needs detail rather than a visual pattern summary.

4. Correct Answer: D. Remove unnecessary gridlines and visual noise. Explanation: Extra marks compete with the data and make charts harder to scan. Removing clutter keeps attention on the values and comparisons that matter.

5. Correct Answer: B. Use sufficient contrast and include non-color cues like labels or patterns. Explanation: Accessibility improves when meaning is not carried by color alone. Contrast and redundant cues reduce the chance of misreading for color-vision differences.

6. Correct Answer: C. Use direct labeling instead of forcing readers to look back and forth to a legend. Explanation: Direct labels reduce cognitive load because the reader does not need to map colors or symbols back to a legend. This typically improves speed and accuracy of interpretation.

7. Correct Answer: B. Truncating the y-axis so small differences look large. Explanation: A truncated axis can exaggerate changes and distort perceived magnitude. This can lead to incorrect conclusions about how big differences really are.

8. Correct Answer: B. Choosing a map only when the question is about location-based patterns. Explanation: Matching the visual to the question ensures the encoding supports the message. A map is appropriate only when geography is relevant to the insight.

9. Correct Answer: C. Describe the chart in words and confirm the description matches the intended takeaway. Explanation: If the chart can be summarized plainly and the summary matches the intended message, comprehension is likely strong. This catches confusing encodings or missing context before sharing.

10. Correct Answer: B. Clear labels, consistent units, and uncluttered design. Explanation: Clear labels and consistent units reduce ambiguity in what the numbers represent. Uncluttered design reduces distractions that can cause misreads or arguments about the meaning.

Bank 11

1. Which deliverable is most appropriate when leaders need a short, decision-focused summary rather than full interactive exploration?
   A. Executive summary
   B. Data lake
   C. Raw log export
   D. SQL stored procedure

2. Which deliverable is most appropriate when users need ongoing interactive access to metrics and the ability to explore trends over time?
   A. Dashboard
   B. Static infographic only
   C. Screenshot collage of charts
   D. Plain TXT notes

3. Which deliverable is most appropriate when an organization wants a single entry point that hosts multiple reports and dashboards for different audiences?
   A. Portal
   B. Pivot table
   C. Schema diagram
   D. Temporary table

4. Which dashboard behavior best matches a scenario where the same content is delivered on a fixed schedule, such as every Monday morning?
   A. Ad hoc
   B. Self-service only
   C. Recurring
   D. Static one-time

5. Which dashboard behavior best matches a scenario where a user changes filters, slices data, and explores questions without requesting a new report each time?
   A. Static
   B. Self-service
   C. Recurring
   D. Ad hoc

6. Which dashboard behavior best matches a scenario where a specific question is answered once and the output is not expected to change?
   A. Static
   B. Dynamic
   C. Self-service
   D. Recurring

7. Which behavior best matches a scenario where the dashboard content changes based on selections and updates as the user interacts?

A. Static
B. Dynamic
C. Ad hoc
D. Snapshot only

8. Which versioning approach best supports comparing “what the report said” last month versus what it says today?
A. Real-time feed only with no stored history
B. Snapshots taken at defined points in time
C. Removing old data to keep storage small
D. Editing past reports so they match the latest results

9. Which choice best describes a real-time feed?
A. Data that updates continuously or near-continuously as new events arrive
B. Data captured once and never refreshed
C. A versioned copy taken at month-end only
D. A dataset that can only be accessed through a manual export

10. Why do refresh intervals matter for reporting accuracy and trust?
A. They guarantee the data is correct without validation
B. They eliminate the need for documentation
C. They determine how current the numbers are and set expectations for users
D. They prevent duplicate records from existing in the source system

---

1. Correct Answer: A. Executive summary. Explanation: Executive summaries are designed to present key findings and decisions in a short format. They reduce detail so leaders can act without needing deep navigation.

2. Correct Answer: A. Dashboard. Explanation: Dashboards provide interactive views of metrics and trends over time. They support ongoing monitoring and exploration without rebuilding the artifact repeatedly.

3. Correct Answer: A. Portal. Explanation: Portals serve as a centralized place to access multiple analytics artifacts. They help organize content by audience and simplify discovery.

4. Correct Answer: C. Recurring. Explanation: Recurring behavior matches scheduled delivery, such as a weekly or monthly cadence. The key idea is predictable timing tied to a refresh or publish routine.

5. Correct Answer: B. Self-service. Explanation: Self-service supports user-driven exploration through filters and slices without requesting new custom builds. It reduces bottlenecks while still using governed content.

6. Correct Answer: A. Static. Explanation: Static outputs are intended to answer a question once without ongoing updates. They are commonly used for one-time analyses or fixed point-in-time presentations.

7. Correct Answer: B. Dynamic. Explanation: Dynamic dashboards respond to user selections and update displayed results as filters and parameters change. This interactivity is central to the “dynamic” behavior description.

8. Correct Answer: B. Snapshots taken at defined points in time. Explanation: Snapshots preserve a point-in-time view so historical comparisons are possible later. Without snapshots, real-time updates can overwrite what was previously reported.

9. Correct Answer: A. Data that updates continuously or near-continuously as new events arrive. Explanation: Real-time feeds reflect new data as it comes in, often with minimal delay. They are used when timeliness is critical for decisions or monitoring.

10. Correct Answer: C. They determine how current the numbers are and set expectations for users. Explanation: Refresh intervals define the lag between reality and what the report shows. Clear expectations prevent confusion when numbers differ from other sources or from “live” systems.

Bank 12

1. Which symptom most directly indicates a report performance problem related to load time rather than data accuracy?
   A. A KPI shows a value that is one percent higher than expected
   B. A report takes a long time to open or visuals render slowly
   C. A date field displays in a different format
   D. A chart title is missing

2. Which factor is most likely to increase refresh time when the dataset becomes very large?
   A. Using fewer filters in the report
   B. Returning fewer rows from the source
   C. Processing and transferring a larger volume of records each refresh
   D. Using consistent units and labels

3. Which scenario most strongly suggests stale data caused by a refresh delay rather than a calculation bug?
   A. The report uses a different color palette than last week
   B. The "last refreshed" timestamp is older than the expected interval
   C. A bar chart is used instead of a line chart
   D. A label contains a typo

4. Which check best confirms whether broken filters are caused by source structure changes?
   A. Compare the report's expected fields to the current source schema and field names
   B. Increase the number of filters to see if one starts working
   C. Replace filters with manual sorting
   D. Switch to a different chart type

5. Which situation most strongly indicates a structure change in the source is impacting the report?
   A. The number formatting shows commas instead of spaces
   B. A column used in a filter was renamed or removed in the source
   C. A chart legend moved to the bottom
   D. A KPI is displayed as a percentage

6. Which action best supports validating calculations and code in a way that reduces preventable mistakes?
   A. Ship the change quickly and wait for user complaints
   B. Disable alerts so the team is not distracted
   C. Rely on one person's memory of the logic
   D. Use review or peer checks before publishing changes

7. Which approach best represents monitoring alerts as a control for report correctness?
   A. Alerts that notify when refresh fails or metrics deviate beyond a threshold
   B. Alerts that change the dashboard theme automatically
   C. Alerts that hide visuals when performance is slow
   D. Alerts that prevent any filtering to reduce load

8. Which action is most appropriate when corrupt data is suspected in a report output?
   A. Delete the entire dataset immediately to prevent exposure
   B. Filter or reprocess affected data and verify outputs against a trusted reference
   C. Increase decorative effects so users do not notice the issue
   D. Change labels to make the numbers look more reasonable

9. Which step best supports distinguishing corrupted source data from a report-layer transformation bug?
   A. Compare report results to a direct extract from the source for the same time window
   B. Remove all filters and publish anyway
   C. Replace nulls with zeros and re-run calculations
   D. Reduce the number of visuals to make the report faster

10. Which response best fits handling corrupt data without spreading incorrect results?
    A. Continue publishing so trends remain uninterrupted
    B. Hide the issue by removing error indicators
    C. Pause distribution, document the issue, and verify corrected results before re-release
    D. Change the metric definition so it matches the corrupted values

---

1. Correct Answer: B. A report takes a long time to open or visuals render slowly. Explanation: Load-time issues show up as slow opening and delayed rendering even if the numbers are correct. This points to performance constraints rather than calculation accuracy.

2. Correct Answer: C. Processing and transferring a larger volume of records each refresh. Explanation: Larger datasets require more time to extract, transform, load, and compute refresh results. More records usually means more work for both the source and reporting layers.

3. Correct Answer: B. The “last refreshed” timestamp is older than the expected interval. Explanation: A stale timestamp indicates the refresh did not run on schedule or failed. That suggests delayed updates rather than incorrect formula logic.

4. Correct Answer: A. Compare the report’s expected fields to the current source schema and field names. Explanation: Filters break when the field names, types, or structures they rely on change upstream. Comparing schema expectations to the current source identifies mismatches quickly.

5. Correct Answer: B. A column used in a filter was renamed or removed in the source. Explanation: Renames and removals directly break references used by filters and calculations. This is a classic structure-change failure mode.

6. Correct Answer: D. Use review or peer checks before publishing changes. Explanation: Peer checks catch logic errors and edge cases that a single author can miss. Review is a practical control to improve correctness before users depend on outputs.

7. Correct Answer: A. Alerts that notify when refresh fails or metrics deviate beyond a threshold. Explanation: Monitoring alerts create early warning when data pipelines fail or metrics change unexpectedly. This supports faster detection and reduces the time incorrect data is visible.

8. Correct Answer: B. Filter or reprocess affected data and verify outputs against a trusted reference. Explanation: When corruption is suspected, the priority is to prevent incorrect outputs while restoring accurate results. Verification against a trusted reference helps confirm the fix is real.

9. Correct Answer: A. Compare report results to a direct extract from the source for the same time window. Explanation: A direct extract provides a baseline to determine whether the source data is already wrong. If the extract is correct but the report is wrong, the issue is likely in transformations or calculations.

10. Correct Answer: C. Pause distribution, document the issue, and verify corrected results before re-release. Explanation: Publishing known-bad data spreads misinformation and damages trust. Documenting and verifying corrections supports accountability and prevents repeat mistakes.

Bank 12

1. Which signal most strongly points to a **load-time** performance problem rather than a refresh or calculation issue?
   A. The report opens slowly even before filters are applied
   B. A single KPI is off by one decimal place
   C. A category label has a spelling mistake
   D. A column contains unexpected nulls

2. Which change is most likely to improve performance when a report struggles because it is pulling far more data than needed?
   A. Add more visuals to spread the work across pages
   B. Switch all charts to tables
   C. Increase the refresh frequency
   D. Limit the result set to essential rows and columns before rendering

3. Which evidence best supports the conclusion that users are seeing **stale data**?
   A. The chart uses a different font than yesterday
   B. The report has fewer filters than last week
   C. The "last refreshed" time is older than the expected schedule
   D. The y-axis starts at zero

4. Which situation most strongly suggests the report filters broke due to a **source structure change**?
   A. The report is slower at noon than in the morning
   B. A field used in a filter was renamed, removed, or had its type changed upstream
   C. A KPI uses a different rounding rule
   D. A chart title was edited

5. Which action is the most direct form of **source validation** when a report suddenly shows unexpected values?
   A. Compare the report's output to a direct extract from the source for the same time window
   B. Change the chart type to see if the pattern looks better
   C. Hide the outliers so the graph looks cleaner
   D. Remove labels to reduce clutter

6. Which practice best reduces preventable errors when updating calculations or code used in reports?
   A. Make changes only during off-hours
   B. Disable alerts to avoid noise
   C. Use review or peer checks before publishing changes
   D. Avoid documenting changes so the team stays flexible

7. Which monitoring approach best helps detect report issues early without waiting for user complaints?

A. Wait for quarterly reviews to compare numbers

B. Refresh manually once a week

C. Remove all filters so fewer things can break

D. Set alerts for refresh failures and for unexpected metric shifts

8. When corrupt data is suspected, which response best supports accurate recovery without spreading incorrect results?

   A. Keep publishing so trends remain continuous

   B. Filter or reprocess the affected data, then verify against a trusted reference

   C. Delete the entire dataset immediately with no checks

   D. Change the metric definition so it matches what is already shown

9. Which check best supports diagnosing a broken filter when the source still has the field name but results look wrong?

   A. Increase the number of filters to see if one works

   B. Remove all labels and legends

   C. Confirm the field's data type and expected values still match what the filter logic assumes

   D. Add decorative effects to make differences stand out

10. Which outcome is the clearest risk when a wrong-grain join or duplicated rows inflate a report's dataset size?

    A. Longer load and refresh times, plus misleading totals

    B. Better accessibility for color-vision differences

    C. More accurate KPIs because there are more rows

    D. Reduced need for validation because the dataset is larger

---

1. Correct Answer: A. The report opens slowly even before filters are applied. Explanation: Slow opening and delayed rendering point to load-time constraints rather than incorrect math. A calculation problem can exist, but it does not usually explain slow rendering on its own.

2. Correct Answer: D. Limit the result set to essential rows and columns before rendering. Explanation: Pulling less data reduces transfer, processing, and rendering work. This targets a common root cause when large datasets make reports sluggish.

3. Correct Answer: C. The "last refreshed" time is older than the expected schedule. Explanation: A stale refresh timestamp is direct evidence the report is behind the expected update cadence. That supports a refresh delay diagnosis more than a calculation bug.

4. Correct Answer: B. A field used in a filter was renamed, removed, or had its type changed upstream. Explanation: Filters depend on stable field names and types. Structure changes break those references or change how filter logic behaves.

5. Correct Answer: A. Compare the report’s output to a direct extract from the source for the same time window. Explanation: A direct source extract provides a baseline to see whether the source is already wrong. If the source extract is correct, the issue is more likely in report transformations or calculations.

6. Correct Answer: C. Use review or peer checks before publishing changes. Explanation: Peer checks catch logic errors and edge cases that a single author may miss. This reduces preventable mistakes before users rely on the output.

7. Correct Answer: D. Set alerts for refresh failures and for unexpected metric shifts. Explanation: Alerts create early warning when the pipeline fails or numbers drift unexpectedly. This shortens the time incorrect or missing data is visible.

8. Correct Answer: B. Filter or reprocess the affected data, then verify against a trusted reference. Explanation: Reprocessing addresses the corruption while verification confirms the fix is real. This approach reduces the chance of distributing incorrect results.

9. Correct Answer: C. Confirm the field’s data type and expected values still match what the filter logic assumes. Explanation: Filters can behave incorrectly when a type changes or values no longer match expected formats. Validating types and value shapes helps isolate why filtering results look wrong.

10. Correct Answer: A. Longer load and refresh times, plus misleading totals. Explanation: Duplicated rows increase processing cost and can inflate sums and counts. That creates both performance problems and incorrect business conclusions.

Bank 13

1. Which artifact most directly supports identifying where a dataset came from and how it moved through systems over time?
   A. Lineage documentation
   B. Decorative branding guidelines
   C. A chart color palette
   D. A list of favorite visual types

2. Which concept best matches the idea of establishing a single agreed-upon dataset or definition for a metric across teams?
   A. Source of truth
   B. Outlier handling
   C. Binning
   D. Refresh rate

3. Which documentation artifact is most focused on defining each field's meaning, type, and allowed values?
   A. Data dictionary
   B. Executive summary
   C. Infographic
   D. Container image

4. Which documentation artifact most directly helps explain how data moves from one system to another through steps and transformations?
   A. Flow diagram
   B. Pivot table
   C. Bar chart
   D. Survey form

5. Which statement best describes why metadata matters for governance?
   A. Metadata is only used to speed up queries
   B. Metadata provides context about data such as definitions, owners, and usage
   C. Metadata replaces the need for access control
   D. Metadata guarantees perfect data quality automatically

6. Which question most directly reflects a governance "documentation" need rather than an analysis need?
   A. What is the mean order value this month?
   B. Which chart type best shows rankings?
   C. What does the "CustomerStatus" field mean and who owns its definition?
   D. Which join type keeps unmatched rows?

7. Which artifact is most appropriate when explaining why a model produced a certain output to a non-technical reviewer?
   A. Explainability report

B. Compressed CSV export
C. Raw log dump
D. Unlabeled scatter plot

8. Which practice best supports traceability when a dataset changes over time?
A. Editing past reports so they always match current numbers
B. Using versioning so changes can be tracked and compared
C. Removing older data so storage stays small
D. Disabling refresh intervals

9. Which versioning choice best supports answering "what did we know at the time" during an investigation?
A. Snapshots taken at defined points in time
B. Only real-time feeds with no stored history
C. A single rolling table that is overwritten each refresh
D. Manual copy and paste into a spreadsheet with no dates

10. Which statement best explains why governance artifacts are important for passing scenario-based exam questions?
A. They make charts look more professional
B. They reduce the need to understand joins and filters
C. They provide evidence of definitions, ownership, and traceability when decisions must be defended
D. They guarantee the data is clean without checks

---

1. Correct Answer: A. Lineage documentation. Explanation: Lineage documents where data originated and how it flowed and changed through systems. This supports investigations and helps teams trust how a dataset was produced.

2. Correct Answer: A. Source of truth. Explanation: A source of truth is the agreed authoritative reference for a dataset or metric definition. It reduces conflicting numbers across teams and reports.

3. Correct Answer: A. Data dictionary. Explanation: Data dictionaries define fields, meanings, types, and acceptable values. This prevents misinterpretation and improves consistent use across analysts and tools.

4. Correct Answer: A. Flow diagram. Explanation: Flow diagrams show movement and transformation steps across systems. They make pipelines easier to review and troubleshoot.

5. Correct Answer: B. Metadata provides context about data such as definitions, owners, and usage. Explanation: Metadata adds the "who, what, and why" around the raw values, which supports governance. It helps people interpret data consistently and manage it responsibly.

6. Correct Answer: C. What does the “CustomerStatus” field mean and who owns its definition? Explanation: Governance documentation focuses on definitions, ownership, and consistent meaning. This question targets interpretability and accountability rather than computation.
7. Correct Answer: A. Explainability report. Explanation: Explainability reports communicate how outputs were produced in a way reviewers can understand and evaluate. They support transparency and responsible use of model results.
8. Correct Answer: B. Using versioning so changes can be tracked and compared. Explanation: Versioning preserves a record of changes, which supports auditing and rollback. It prevents confusion when results shift after updates.
9. Correct Answer: A. Snapshots taken at defined points in time. Explanation: Snapshots preserve what the data looked like at a specific moment. This supports investigation questions about historical state and prior reporting.
10. Correct Answer: C. They provide evidence of definitions, ownership, and traceability when decisions must be defended. Explanation: Scenario questions often test whether choices can be justified with documentation and traceability. Governance artifacts support defensible decisions and reduce ambiguity.

Bank 14

1. Which policy topic most directly determines how long a dataset must be kept before it can be deleted?
   A. Retention rules
   B. Chart labeling standards
   C. Join strategy
   D. Binning strategy

2. Which statement best describes why replication rules matter in compliance contexts?
   A. Replication always improves data quality automatically
   B. Replication can create extra copies that must follow the same retention, access, and jurisdiction rules
   C. Replication eliminates the need for backups
   D. Replication guarantees faster dashboards in all cases

3. Which scenario most directly reflects a jurisdiction requirement such as GDPR affecting data handling?
   A. Choosing a bar chart instead of a pie chart
   B. Deciding whether to use mean or median
   C. Storing personal data in a region with strict legal limits on transfer and processing
   D. Increasing refresh intervals to reduce load

4. Which decision best reflects “storage rules” rather than analysis technique?
   A. Choosing a repository for long-term archival versus short-term operational use
   B. Choosing a chart encoding for ranking
   C. Choosing a join type to preserve unmatched rows
   D. Choosing the mode as the central tendency measure

5. Which statement best reflects the safest exam stance when dealing with GDPR or jurisdiction topics in a scenario?
   A. Assume GDPR applies to every dataset in every country
   B. Ignore jurisdiction rules unless the prompt uses the word “privacy”
   C. Treat jurisdiction as a constraint and avoid overclaiming specifics not provided in the scenario
   D. Copy whatever the source system currently does without review

6. Which practice best supports compliance with retention requirements when multiple copies of data exist?
   A. Delete only the primary copy and ignore replicas
   B. Track where copies exist and apply the same retention and deletion logic consistently
   C. Increase the number of replicas so data loss is impossible
   D. Convert all fields to text so retention is easier

7. Which item is most likely to be considered a sensitive data category that requires extra care in sharing and access?
   A. A product catalog list with public SKUs
   B. A weather dataset with city temperatures
   C. A dataset containing personally identifiable information about customers
   D. A chart showing sales by region

8. Which control most directly limits who can view or change sensitive data in a system?
   A. Role-based access control
   B. Decorative branding standards
   C. Binning
   D. A pivot table

9. Which response best fits preparing for an audit in a data governance scenario?
   A. Rely on informal team knowledge because it is faster
   B. Provide traceable documentation, evidence of controls, and clear incident reporting paths
   C. Remove all logs to reduce storage costs
   D. Avoid writing definitions so teams can stay flexible

10. Which statement best reflects why incident reporting appears in audit and compliance topics?
    A. It is only a technical logging feature with no governance impact
    B. It helps show accountability and a documented response when issues occur
    C. It reduces the need for access control
    D. It guarantees no incidents will happen

---

1. Correct Answer: A. Retention rules. Explanation: Retention rules define how long data must be kept before it can be deleted or archived. They are a compliance and governance requirement, not an analysis technique.

2. Correct Answer: B. Replication can create extra copies that must follow the same retention, access, and jurisdiction rules. Explanation: Replication increases the number of stored copies, which expands the compliance surface area. Those copies still need consistent controls and lifecycle handling.

3. Correct Answer: C. Storing personal data in a region with strict legal limits on transfer and processing. Explanation: Jurisdiction requirements constrain where data can be stored and how it can be transferred. GDPR and similar rules can affect processing, access, and cross-border movement.

4. Correct Answer: A. Choosing a repository for long-term archival versus short-term operational use. Explanation: Storage rules focus on where data lives, how it is kept,

and how long it persists. These decisions affect retention, access, and compliance obligations.

5. Correct Answer: C. Treat jurisdiction as a constraint and avoid overclaiming specifics not provided in the scenario. Explanation: Exam scenarios often require recognizing legal constraints without inventing details. The safe choice is to note jurisdiction as a requirement and limit claims to what the prompt supports.

6. Correct Answer: B. Track where copies exist and apply the same retention and deletion logic consistently. Explanation: Compliance requires consistent handling across all copies, including replicas and backups when in scope. Tracking locations prevents forgotten copies from violating retention or deletion requirements.

7. Correct Answer: C. A dataset containing personally identifiable information about customers. Explanation: Personally identifiable information can identify individuals and often triggers stricter handling requirements. Access, sharing, retention, and auditability become higher stakes with such data.

8. Correct Answer: A. Role-based access control. Explanation: Role-based access control limits access based on assigned roles and permissions. This directly restricts who can view or modify sensitive data.

9. Correct Answer: B. Provide traceable documentation, evidence of controls, and clear incident reporting paths. Explanation: Audits rely on evidence, not informal understanding. Documentation and incident reporting show accountability and repeatable control operation.

10. Correct Answer: B. It helps show accountability and a documented response when issues occur. Explanation: Incident reporting demonstrates that the organization can detect, respond, and document issues. This supports governance expectations and audit readiness.

Bank 15

1. Which control most directly restricts data access based on job role so only approved users can view sensitive fields?
   A. Data binning
   B. Role-based access control
   C. Descriptive statistics
   D. Web scraping

2. Which practice best protects sensitive data while it is being transmitted over a network?
   A. Encryption in transit
   B. Snapshots
   C. Data mart design
   D. Binning

3. Which practice best protects sensitive data stored on disk or in a database when the storage is compromised?
   A. Encryption at rest
   B. Labels and legends
   C. Inner joins
   D. Parsing

4. Which scenario most directly represents reducing exposure by limiting the amount of sensitive data shared?
   A. Sharing only aggregated results instead of customer-level records
   B. Adding more columns so analysis is easier
   C. Increasing refresh frequency so the report stays current
   D. Converting all identifiers to integers

5. Which data category most strongly fits personally identifiable information?
   A. A public list of product SKUs and prices
   B. A chart of weekly sales totals by region
   C. A patient’s diagnosis code and treatment notes
   D. A customer’s name and email address

6. Which data category most strongly fits protected health information?
   A. A customer’s name and email address
   B. A patient record containing identity plus health details
   C. A public dataset of city temperatures
   D. A product catalog with categories and brands

7. Which technique best supports sharing data for analysis while reducing the chance a person can be identified?
   A. Anonymization
   B. Adding more detailed timestamps

C. Removing labels and legends
D. Increasing the number of joins

8. Which technique best supports allowing realistic testing or analytics while hiding sensitive values from most users?
A. Masking
B. Adding indexes
C. Scaling
D. Concatenation

9. Which approach is most appropriate when a dataset includes sensitive health information and must be shared with a broader internal audience?
A. Share the full dataset as long as it is inside the company network
B. Apply access controls and use masking or anonymization where appropriate
C. Convert the file to CSV so it is easier to review
D. Remove all constraints so work moves faster

10. Which statement best explains why encryption and access control are both needed for sensitive data?
A. They are interchangeable, so using either one is enough
B. Encryption replaces the need for governance documentation
C. Access control limits who can see data, while encryption protects it if storage or transmission is intercepted
D. Access control is only for performance, while encryption is only for formatting

---

1. Correct Answer: B. Role-based access control. Explanation: Role-based access control restricts access based on role assignments, which limits who can view or change sensitive fields. This helps enforce least privilege in shared environments.

2. Correct Answer: A. Encryption in transit. Explanation: Encryption in transit protects data as it moves between systems, such as from a client to a server. It reduces the risk of interception exposing sensitive values.

3. Correct Answer: A. Encryption at rest. Explanation: Encryption at rest protects stored data on disks or databases if the storage media or backups are accessed without authorization. It reduces exposure even when access controls are bypassed through theft or misconfiguration.

4. Correct Answer: A. Sharing only aggregated results instead of customer-level records. Explanation: Aggregation reduces exposure by removing individual-level detail that could identify people. This supports safer sharing while still enabling trend and performance analysis.

5. Correct Answer: D. A customer’s name and email address. Explanation: Personally identifiable information can identify a specific individual, such as a name paired

with contact details. These fields often require stricter access and handling controls.

6. Correct Answer: B. A patient record containing identity plus health details. Explanation: Protected health information combines identity with health-related information. This category typically requires heightened safeguards and limited sharing.

7. Correct Answer: A. Anonymization. Explanation: Anonymization aims to remove or alter identifiers so individuals cannot reasonably be re-identified. This supports broader sharing when individual identity is not needed.

8. Correct Answer: A. Masking. Explanation: Masking hides sensitive values while keeping data usable for many internal purposes like testing or reporting. It limits exposure by showing only partial or substituted values.

9. Correct Answer: B. Apply access controls and use masking or anonymization where appropriate. Explanation: Sensitive health information requires tighter access plus techniques that reduce exposure for broader audiences. Controls and protection methods together reduce risk while still enabling legitimate use.

10. Correct Answer: C. Access control limits who can see data, while encryption protects it if storage or transmission is intercepted. Explanation: Access control reduces exposure by preventing unauthorized viewing. Encryption protects confidentiality even if data is captured from storage or intercepted in transit.

Bank 16

1. Which control best supports catching data issues early by automatically checking expected conditions during processing?
   A. Automated tests for data quality rules
   B. Decorative dashboard branding
   C. Switching chart types
   D. Increasing the number of joins

2. Which practice best supports preventing unreviewed changes to transformation logic from silently breaking outputs?
   A. Source control for code and configuration
   B. Removing peer review to move faster
   C. Copying scripts between folders by hand
   D. Disabling alerts to reduce noise

3. Which step best represents user acceptance testing in a data workflow?
   A. A team validates the output meets agreed business requirements before release
   B. A developer writes a new join without checking row counts
   C. An analyst changes labels to match branding
   D. A system refreshes data more frequently

4. Which activity best matches requirement validation rather than statistical analysis?
   A. Checking whether the delivered KPI matches the defined business question and rules
   B. Computing variance and standard deviation
   C. Selecting a bar chart for rankings
   D. Binning numeric values into ranges

5. Which metric is most appropriate for monitoring data health over time?
   A. A one-time screenshot of a dashboard page
   B. A consistent completeness rate and null percentage tracked each refresh
   C. A list of favorite report colors
   D. A single pivot table saved once a year

6. Which practice best supports detecting subtle changes in incoming data patterns that can degrade analysis even when pipelines still run?
   A. Ignoring drift because it is normal
   B. Data drift monitoring
   C. Removing all outliers automatically
   D. Using more decorative chart effects

7. Which activity best fits data profiling as part of monitoring data health?
   A. Computing distributions, distinct counts, and null rates to understand current shape
   B. Publishing the report to a new portal

C. Creating a new dashboard theme
D. Increasing refresh intervals to reduce load

8. Which scenario most directly indicates a need for quality tests tied to business rules?
A. A chart legend overlaps a label
B. A numeric field sometimes contains negative values that should never occur
C. The dashboard background color changed
D. The report contains too many gridlines

9. Which control most directly helps prevent “it worked on my machine” differences across environments when releasing data transformations?
A. Ad hoc manual edits in production
B. Untracked script copies emailed between team members
C. Versioned, reviewed changes in source control
D. Changing the KPI definitions weekly

10. Which statement best explains why monitoring alerts are valuable for data quality and reporting reliability?
A. Alerts guarantee data will always be accurate
B. Alerts detect failures or unusual shifts early so issues can be addressed quickly
C. Alerts remove the need for documentation
D. Alerts replace access control and encryption

---

1. Correct Answer: A. Automated tests for data quality rules. Explanation: Automated tests check expected conditions, such as valid ranges or required fields, as data moves through the workflow. This catches failures early before bad data reaches reports.

2. Correct Answer: A. Source control for code and configuration. Explanation: Source control tracks changes, supports reviews, and allows rollback when something breaks. It prevents silent drift in transformation logic and improves traceability.

3. Correct Answer: A. A team validates the output meets agreed business requirements before release. Explanation: User acceptance testing confirms the deliverable meets the intended needs from the user or business perspective. It reduces the chance of releasing outputs that are technically correct but operationally wrong.

4. Correct Answer: A. Checking whether the delivered KPI matches the defined business question and rules. Explanation: Requirement validation confirms the output aligns to the agreed definition and intent. This is different from computing statistics, which only describes the data mathematically.

5. Correct Answer: B. A consistent completeness rate and null percentage tracked each refresh. Explanation: Health metrics are meaningful when they are measurable, repeatable, and tracked over time. Completeness and null rates are common indicators that reveal degradation early.

6. Correct Answer: B. Data drift monitoring. Explanation: Drift monitoring detects changes in distributions or patterns that can degrade models and analytics even when jobs succeed. This helps identify when assumptions no longer hold.

7. Correct Answer: A. Computing distributions, distinct counts, and null rates to understand current shape. Explanation: Profiling measures the current characteristics of data to find anomalies and shifts. It provides a baseline for monitoring and quality checks.

8. Correct Answer: B. A numeric field sometimes contains negative values that should never occur. Explanation: Values that violate business rules indicate a need for validation tests tied to those rules. Automated checks can flag the issue immediately during ingestion or transformation.

9. Correct Answer: C. Versioned, reviewed changes in source control. Explanation: Source control plus review reduces environment-specific differences and ensures changes are reproducible. It also provides a record of what changed and when.

10. Correct Answer: B. Alerts detect failures or unusual shifts early so issues can be addressed quickly. Explanation: Alerts shorten the time between a problem and detection, reducing how long bad data is visible. Early detection supports reliability and maintains trust in reporting outputs.