Exam Snapshot

Series code: DA0-002 (V2)   **Launch:** Oct 14, 2025
Questions: max 90 (multiple-choice + performance-based)
Time: 90 minutes   **Passing score:** 720 (100–900 scale)

Domain Weights (study by %)

Domain	Weight
Data concepts & environments	20%
Data acquisition & preparation	22%
Data analysis	24%
Visualization & reporting	20%
Data governance	14%

Core Workflow (what questions really test)

- Start with the business question, audience, and success metric (KPI).
- Profile the data: types, grain (row meaning), keys, and null patterns.
- Acquire + combine safely: filters first, then joins; validate row counts.
- Clean + transform: fix missing values, duplicates, outliers, and formats.
- Analyze: pick the simplest method that answers the question and explain limits.
- Report: choose the right visual and state the takeaway in one sentence.
- Govern: protect sensitive data and document lineage, retention, and access.

High-Yield Concepts

- Structured vs semi-structured vs unstructured data; common file types (CSV, JSON, Parquet).
- Keys: primary key, foreign key; cardinality (1:1, 1:many, many:many).
- OLTP vs OLAP; data lake vs warehouse; star schema (facts vs dimensions).
- Data quality dimensions: accuracy, completeness, consistency, timeliness, validity, uniqueness.

Joins & Set Ops (quick rules)

- **INNER**: only matching keys. **LEFT**: keep all left rows (watch nulls on right).
- **FULL**: keep all rows; expect nulls on both sides.
- **UNION** removes duplicates; **UNION ALL** keeps duplicates (faster).
- Trap: many-to-many joins can multiply rows; validate with row counts and distinct keys.

Cleaning & Transformation

- Missing data: drop, fill (mean/median/mode), or flag; choose based on meaning.
- Outliers: confirm if real; use IQR/z-score only after understanding context.
- Text + dates: trim, split, parse; watch time zones and inconsistent formats.
- Leakage: never use future information when training/testing a model.

Stats & Analysis Cheat Sheet

- Mean vs median (median is safer with skew/outliers).
- Standard deviation: spread around the mean; variance = SD squared.
- Correlation ≠ causation; check confounders and time ordering.
- Sampling bias and survivorship bias: who is missing from the data?
- Classification basics: confusion matrix; precision, recall, F1; pick based on risk.

Visualization & Reporting

- Bar = compare categories; line = trend over time; scatter = relationship; histogram = distribution.
- Start axes at zero for bars; label units; keep scales consistent.
- Avoid 3D and decoration that changes perception; prefer direct labels over legends.
- Accessibility: readable fonts, contrast, and do not rely on color alone.

Governance & Risk

- Least privilege access; masking vs encryption; tokenization where needed.
- Lineage: where data came from, what changed, and when; document owners and definitions.
- Retention + disposal: keep only what you need, for as long as required.

Test-Day Tactics (fast mental model)

- Read the last line first, then the stem: identify the task and constraints.
- Watch qualifiers: BEST, MOST LIKELY, FIRST, NOT; eliminate 2 distractors quickly.
- If a PBQ looks long, flag it and come back after faster questions.
- Pace: ~1 minute per question, then use the remaining time for flagged items.
- Final pass: verify units, date ranges, and that the answer matches the question asked.