CompTIA Data Plus Exam Glossary

Find more at BareMetalCyber.com

1. **Aggregation**
   Aggregation is the process of rolling up detailed data into summaries, like totals, averages, counts, mins, or maxes. On the exam, it often shows up when you're asked to produce metrics (like monthly revenue) from transaction-level rows.
2. **Categorical data**
   Categorical data represents labels or groups rather than numeric measurements, such as department, region, or product category. It's commonly used for grouping, filtering, and comparisons across segments.
3. **Continuous data**
   Continuous data can take any value within a range, such as temperature, time, or weight. It's often analyzed with distributions, trends, and summary statistics, and it behaves differently than discrete counts.
4. **Correlation**
   Correlation measures how strongly two variables move together, and in what direction (positive or negative). It does not prove causation, and the exam likes to test that distinction.
5. **Data cleansing**
   Data cleansing is correcting or removing inaccurate, messy, duplicate, or inconsistent data so it's usable. Typical actions include fixing formats, standardizing values, removing duplicates, and handling missing entries.
6. **Data collection**
   Data collection is how raw data is gathered from systems, people, devices, or logs. Exam questions often focus on choosing the right method and recognizing bias, gaps, or quality risks at the collection step.
7. **Data dictionary**
   A data dictionary is documentation that defines fields, meanings, data types, allowed values, and business rules for a dataset. It's a key artifact for clarity, consistency, and preventing misinterpretation across teams.
8. **Data governance**
   Data governance is the set of policies, roles, and processes that control how data is

defined, accessed, protected, and used. It typically covers ownership, standards, quality expectations, privacy, and retention.

9. **Data ingestion**
   Data ingestion is moving data into a tool, platform, or storage location for processing and analysis. It can be batch (scheduled) or streaming (near real-time), and exam questions often tie it to pipeline reliability and data freshness.

10. **Data integration**
    Data integration combines data from multiple sources into a unified view that can be analyzed consistently. This often involves matching keys, aligning definitions, resolving conflicts, and ensuring the merged output is trustworthy.

11. **Data lineage**
    Data lineage tracks where data came from, how it moved, and what transformations happened to it along the way. On the exam, lineage is tied to trust and auditability, especially when you need to explain why a number changed from source to report.

12. **Data normalization**
    Data normalization can mean organizing relational tables to reduce redundancy and improve integrity, usually by splitting data into related tables with keys. In analytics prep, it can also show up as scaling values, so watch the context clues in the question.

13. **Data pipeline**
    A data pipeline is the end-to-end flow that moves data from sources through ingestion, transformation, and storage to analysis or reporting. Exam questions often probe where failures happen, how to validate outputs, and how to keep steps repeatable.

14. **Data profiling**
    Data profiling is the process of examining a dataset to understand its structure, completeness, distributions, and anomalies. It's usually done early to spot null rates, duplicates, unexpected ranges, and inconsistent formats before deeper analysis.

15. **Data quality**
    Data quality describes how fit the data is for its intended use, commonly measured by accuracy, completeness, consistency, timeliness, and validity. The exam often frames quality as "can we trust decisions made from this data?"

16. **Data source**
    A data source is the system or origin that produces the data, such as a CRM, ERP, web analytics tool, sensor, or flat file. The exam likes to test how source limitations, latency, and definitions affect what conclusions you can safely draw.

17. **Data standardization**
    Data standardization is making data consistent in format and meaning, such as standard date formats, units of measure, naming conventions, and category labels. It prevents downstream issues like mismatched joins, double-counting, and misleading comparisons.

18. **Data stewardship**
    Data stewardship refers to the people and responsibilities that ensure data is managed correctly day-to-day, including definitions, quality checks, and issue resolution. On the exam, it often connects to accountability: who owns the fix when data is wrong?

19. **Data transformation**
    Data transformation is changing raw data into a more usable form, such as converting types, deriving new fields, mapping categories, or reshaping tables. Exam items commonly test whether a transformation could introduce errors, bias, or loss of detail.

20. **Data type**
    A data type defines what kind of values a field can hold, such as integer, decimal, date, boolean, or string. Choosing the right type matters for storage, validation, calculations, and whether sorting or comparisons behave correctly.

21. **Data validation**
    Data validation is checking whether data meets defined rules, such as allowed ranges, required fields, correct formats, or referential integrity. On the exam, validation is often the “gate” that prevents bad data from entering reports and KPIs.

22. **Descriptive statistics**
    Descriptive statistics summarize what the data looks like, using measures like mean, median, range, variance, and standard deviation. They describe patterns but don’t prove why something happened, which is a common exam distinction.

23. **Discrete data**
    Discrete data is countable and takes distinct values, like number of logins, tickets, or units sold. It’s different from continuous data because you typically analyze it as counts and rates rather than measurements on a smooth scale.

24. **Duplicate record**
    A duplicate record is an unintended repeated entry that can inflate counts, totals, and averages. Exam questions often test how duplicates happen (bad keys, merges, reprocessing) and how they distort metrics.

25. **ELT (Extract, Load, Transform)**
    ELT moves data into the destination system first, then performs transformations inside that system (often a data warehouse). It's common when the target platform has strong compute and you want to store raw data for flexibility.

26. **ETL (Extract, Transform, Load)**
    ETL transforms data before loading it into the destination, which can enforce cleaner, standardized outputs up front. Exam scenarios often compare ETL vs ELT based on performance, governance, and whether raw data must be preserved.

27. **Foreign key**
    A foreign key is a field in one table that references the primary key in another table, creating a relationship between the two. It supports consistent joins and helps enforce referential integrity so records don't point to "missing" entities.

28. **Interval scale**
    An interval scale has equal spacing between values, but no true zero, such as Celsius temperature or calendar years. You can add and subtract meaningfully, but ratios (like "twice as much") don't make sense.

29. **Join (inner/left/right/full)**
    A join combines rows from two tables based on a matching key, and the join type determines which unmatched rows are kept. The exam often tests whether your join will drop records (inner) or preserve them from one side (left/right) or both (full).

30. **KPI (Key Performance Indicator)**
    A KPI is a defined metric used to track performance toward a goal, like churn rate, conversion rate, or average resolution time. On the exam, the trick is usually about picking a KPI that matches the decision being made and has clean, reliable inputs.

31. **Mean**
    The mean is the arithmetic average, found by adding values and dividing by the number of values. It's sensitive to outliers, so a few extreme numbers can pull the mean away from what's "typical."

32. **Median**
    The median is the middle value when data is sorted (or the average of the two

middle values for an even count). It's more resistant to outliers than the mean, which is why it's often better for skewed data like income or response times.

33. **Metadata**
    Metadata is "data about data," such as field names, definitions, data types, refresh timestamps, or source system details. On the exam, metadata is key for traceability and for preventing confusion when two datasets use the same word but mean different things.

34. **Missing data (nulls)**
    Missing data occurs when a value is absent, unknown, or not recorded, often represented as NULL. Exam questions commonly test what to do with nulls (remove, impute, flag) based on why the values are missing.

35. **Mode**
    The mode is the most frequently occurring value in a dataset. It's especially useful for categorical data, like the most common customer segment or the most frequent error code.

36. **Nominal data**
    Nominal data is categorical data with no natural order, like colors, product types, or countries. You can count and group it, but you can't do meaningful "greater than" comparisons between categories.

37. **Normalization**
    Normalization (in relational design) organizes data to reduce duplication and update anomalies, usually by splitting repeated data into related tables. On the exam, the benefit is cleaner joins and more reliable updates, at the cost of needing more joins to analyze.

38. **Ordinal data**
    Ordinal data has an order, but the spacing between values isn't guaranteed to be equal, like satisfaction ratings (poor, fair, good) or severity levels (low, medium, high). You can rank it, but treating the gaps as numeric can mislead analysis.

39. **Outlier**
    An outlier is a value that's unusually far from most other values in the dataset. It can be a real event (like a rare spike) or a data error, and the exam often asks you to decide which is more likely based on context.

40. **Primary key**
    A primary key uniquely identifies each row in a table, such as customer_id or

order_id. If the key isn't truly unique or is missing, joins and counts can break in subtle ways, which is a common exam trap.

41. **Qualitative data**
    Qualitative data describes qualities or characteristics, often in words, such as customer feedback comments or issue descriptions. On the exam, it's usually tied to categorization, thematic analysis, and careful handling so you don't over-quantify narrative text.

42. **Quantitative data**
    Quantitative data is numeric and represents quantities you can measure or count, like revenue, time, or units sold. It supports arithmetic operations and is commonly used for statistical summaries and trends.

43. **Ratio scale**
    A ratio scale has equal intervals and a true zero, such as dollars, age, distance, or duration. Because zero means "none," ratios make sense, like "twice as long" or "half the cost," which is a frequent exam concept.

44. **Referential integrity**
    Referential integrity means relationships between tables stay consistent, so foreign keys correctly point to existing primary keys. When it's violated, joins can produce missing matches, orphan records, and incorrect totals.

45. **Sampling**
    Sampling is selecting a subset of a population to analyze so you can estimate patterns without processing everything. Exam questions often test representativeness and bias, like whether a sample accidentally excludes an important segment.

46. **Schema**
    A schema is the structure and definition of data, including tables, fields, data types, and relationships. On the exam, schema changes can break pipelines, joins, and reports if they aren't tracked and communicated.

47. **Semi-structured data**
    Semi-structured data has some organization but doesn't fit neatly into rows and columns, like JSON, XML, or log events. It often requires parsing and flattening before analysis, and the exam may test when that's appropriate.

48. **Standard deviation**
    Standard deviation measures how spread out values are around the mean. A higher

standard deviation means more variability, which matters when comparing consistency across processes or groups.

49. **Structured data**
    Structured data is organized into a fixed format, typically rows and columns with defined fields, like a relational database table. It's easier to query and validate because rules and types are usually explicit.

50. **Union**
    A union stacks rows from two datasets with compatible columns into one combined dataset. On the exam, the key is that union adds rows (not columns), and mismatched schemas or duplicated rows can cause problems.