Collibra Data Quality & Observability: DQ Rule Cheat Sheet

Which DQ Problem Do You Have?

Collibra Data Quality & Observability auto-discovers issues in data using a machine learning first, rules second based approach. It uses associative, unsupervised machine learning to auto generate SQL-based, explainable and adaptive data quality rules. It creates snapshots and baselines to benchmark past data, constantly learns from new data and makes predictions for typos, formatting issues, outliers, relationships and more.

DQ Problem Statement	DQ Dimension	DQ Solution Description
1. Datashapes: How many phone number formats are in this column?	Shape, Format, Conformity	Common in STRING fields or fields defined as VARCHAR you can end up with many different formats that express something like a zip code or phone number or SSN for example. It is helpful to find the majority formats and show the topN data shapes that make up the column values. This helps identify typos and strange formats. Quick demo video: PII & Data Quality
2. Row count check: Has my row count dropped on any dataset?	Behavior	It can be important to know if the volume of a dataset drops, also known as a row count drop. When a dataset suddenly has fewer rows than normal it can mean data is missing in the file or table.
3. Business DQ rules: I need to write custom DQ rules.	Custom rules	Define rules in standard SQL. There are about 105 predefined functions included but not limited to concat, length, avg, max, min, substr etc. Most common use-cases are easily covered by vanilla SQL but there are more advanced situations where window functions, grouping and time-series functions are required. These can also be expressed fairly easily with some Collibra DQ SQL extensions. Quick demo video: Quickly add rules
4. Auto-generated & adaptive DQ rules: I want DQ rules generated for me.	Adaptive rules	Out of the box unless turned off, Collibra DQ will start generating many rules for each column based on its profile over time. It compares each daily run with its baseline. It helps you avoid writing and managing thousands of conditional statements. Leverage ML to auto-generate SQL-based, non-proprietary, explainable and adaptive data quality rules. Reduce manual rule management efforts. Quick demo video: Out-of-the-box features
5. Fuzzy matching: Is Bill Gates the same as William Gates in my database?	Dupe or Uniqueness	Commonly we need to find exact duplicates or similar duplicates. This problem is not suited for conditional statements. Collibra DQ allows you to opt into any grouping of columns and find exact or similar records, ie. fuzzy matching. This can be done at the column or record level. Identify duplicate or redundant data across multiple data domains via fuzzy or exact matching. Quick demo video: Duplicates or Fuzzy Matching
6. NULL check: Do I have complete observations in a world of fill-out forms and manual data collection?	Completeness	Out of the box every column will have a NullCheck in place, the null check is generated from the columns' past behavior or descriptive statistics. You can turn it off per dataset or per column or add manually if desired.
7. Valid values: A valid FICO credit score is between 300 and 850.	Validity	One of the most common DQ rules is a valid range of values or nickname valid values. In the case of a credit_score a rule such as where credit_score between 300 and 850. Another example for string values could be where credit_providers IN ('experian', 'trans union', 'equifax'). Both Numeric and String values can be expressed as valid values.

8. Change detection: Tell me when something suddenly changes in my data.	Behavior (data drift)	Any column, schema or cell value that suddenly breaks its past trend. Would require thousands of conditional statements and their ongoing management. By default Collibra DQ behavioral analytics is turned on for automatic change control. Quick demo video: Data Drift & Shift Detection
9. PII or sensitive data discovery: Do I have sensitive data in my datasets?		Ability to filter and select datasets by PII, MNPI, PHI etc. During the daily profiling Collibra DQ discovers and tags sensitive data (i.e. SSN). The software keeps an index of all columns that contain sensitive data for interactive filtering from the UI. Automatically understand the semantic schema of your data to classify and mark sensitive data. Quick demo video: Discovery with DQ Rule Enforcement
10. Run DQ on files: I need to be able to run DQ rules on files.		Many DQ frameworks do not cover files. Collibra DQ operates on files with equal parity as database tables. Any rule that works on a table will also work the same on a file.
11. Real-time DQ analysis: I need to run DQ rules on streaming and sensor data (Kafka).		My data does not exist in a table or file but in a Kafka Topic. I want to detect real-time and unsupervised anomalies in streaming data (constantly flowing messages, jsons, avro or batch data) and sensor data (a standard time-series with signal, time and value). Quick demo video: Collibra Data Quality Kafka
12. Outliers: I need to detect outliers per grouping.		Sometimes, basic column level outliers do not solve the issue. One needs to create a subgroup like stock “symbols” and find outliers over time across all symbols. "I don’t care that I have penny stocks and stocks that trade over $200,000 a share, I care if any one particular stock breaks its past trend or baseline. This is applied when a user wants to find egregious numeric values relative to the population. This identifies invalid entries that are numerically out of place. This also allows for grouping on sub-dimensions to identify anomalies on subsets of data. Quick demo video: Outlier Detection
13. Conditional DQ rules: Do I have any rows with dates before 1970?	Conditional rule	Sometimes we just need to express a simple rule. In this example we would click add rule and type d_date < 1970 Any record that contains a d_date before 1970 will produce a break record.
14. Complex DQ rules: I need to join two datasets before I can write a rule.	Complex rule	It is a common need to reference another dataset as a lookup or validation point. Loan credit rates might be stored in a table updated each minute with new loan rates while loan applications might come in one by one or in batches. Quick demo video: DQ Rule Builder Join Example
15. Pipeline testing: I need DQ in my data pipeline.	DQ pipeline	I already have a data pipeline in python or scala or spark and want to control the DQ operations. Some call this an ETL pipeline, making this ETLQ. Collibra DQ supports inline spark commands or REST API commands depending on the need.
16. Distribution rule: Am I missing data for a subset of my dataset?		Track valid values as well as valid distribution of values Equifax -> 400 | TransUnion -> 300 | Experian -> 7. Quick demo video: Shape Detection Distribution Rule
17. Validate source: I need to compare two tables.	Source - Validating source to target accuracy	It is common to need validation when loading data from a file into a database table or from a source database into a target database to identify missing records, values and broken relationships across tables or systems. Quick demo video: Cross Table Validations
18. Pulse View: I’d like to see a heatmap of where all my data errors exist.	Pulse view	Visualize a blindspot heatmap by time, business units and scheduled jobs.
19. Cross-column anomalies: The state and zip code don't belong to each other in my dataset.	Patterns	Define relationships not rules for sophisticated use cases. This is used for identifying cross-column anomalies. Commonly used for hierarchical and parent/child mismappings. Quick demo video: Misclassified Data Cross Column Anomalies
20. Schema drift detection: Is there any change in my database structure?	Schema columns added or dropped	Fields, columns, and types can be added, removed, or changed.

DQ Dimension	Collibra DQ Sub Type	Collibra DQ Type/Class	Collibra DQ Feature	Examples
Completeness	NULL EMPTY	PROFILE BEHAVIOR RULES OUTLIER PATTERN SOURCE RECORD SCHEMA SHAPES	100% per column is 100% complete. All missing sub percentages will report % complete against the dataset.	Fname Kirk [NULL] Brian
Uniqueness	Exact match Fuzzy match	PROFILE BEHAVIOR RULES SOURCE RECORD DUPES	Define exact match or fuzzy match or identify a PK. Collibra DQ Dupe Feature	Kirk haslbeck Kirk hasslback
Validity	Format Data shift Cardinality (Distribution shift) Type shift	RULES OUTLIER PATTERN SOURCE SHAPES	Collibra DQ unsupervised learning detects infrequent shapes and data type shifts.	OCT-20 2020-10-01
Accuracy	Numeric Categorical	PROFILE BEHAVIOR RULES OUTLIER PATTERN SOURCE RECORD SCHEMA DUPES SHAPES	Collibra DQ anomaly engine detects values outside of normal or expected range.	Sym, price Goog, 1,535.0 Goog, 15.35
Timeliness	Load time	BEHAVIOR RULES SOURCE RECORD SCHEMA	Collibra DQ Behavior Feature	8:01pm, 8:02pm, 9:30pm
Consistency	Rows, schema, cell_values	RULES OUTLIER PATTERN SOURCE SHAPES	Collibra DQ Validate Source Feature	‘USA’ -> ‘USA ’