Copy of Cloud Data Quality - CDQ

CDQ helps specify with a declarative syntax (like DBT or Dataform) data quality tests on top of a Big Query Dataset.
It groups these tests into several data quality dimensions, it profiles the test results and stores them directly in Big Query.

It based on SQL tests, so everything eventually is compiled in SQL statements.

Where we are coming from

To expand our understanding of the Data Quality dimensions, and to find which kind of practical tests they could include, we also looked at other references and articles.
The table below shows some of the literature we found most relevant / interesting

Topic	Lit.	Comments/ Description
Data Quality dimension: Accessibility	Proposed Metrics for Data Accessibility in the Context of Linked Open Data	• Test if dataset/table/column is accessible by query (is access denied) • Among a list of N datasets/tables/columns , how many can we access in percentage (is access denied)
Overview of Data Quality Frameworks	Methodologies for data quality assessment and improvement	General overview of different Data Quality Framework and the dimensions they use:
Categories of Data Quality dimensions	Beyond Accuracy: What Data Quality Means to Data Consumers	One of the very first articles trying to define grouping of Data Quality Dimensions
Literature list	http://dimensionsofdataquality.com/research	Nice list of articles and literature very focused on Data Quality and its dimensions
Grouping of Data Quality Dimensions	http://dimensionsofdataquality.com/content/list-underlying-concepts	Graph showing a practical grouping of Data Quality dimensions and their underlying concepts
Overview of Data Quality Frameworks	Big Data Quality Dimensions: A Systematic Literature Review (>= 2018)	
Selected Quality Dimensions	DAMA - Dimensions of Data Quality (DDQ) Research Paper (2020)	Picked best quality dimensions definitions and created subcategories of how the definitions change depending on whether the dimension is applied on column, records, metadata or other data concepts.

Most detailed Data Quality dimensions definitions (from DAMA - Dimensions of Data Quality (DDQ) Research Paper - 2020)

Since Data Quality Dimensions have tons of definitions in older and recent literature, we have picked the article that has analyzed all the previous definitions and came up with a good summary.
The *Data Concept* definition is based on Table 3 (p. 31) of the same article.

A Diagram of how the data concepts are linked together is also available., note the grouping of **Data vs Metadata** concepts

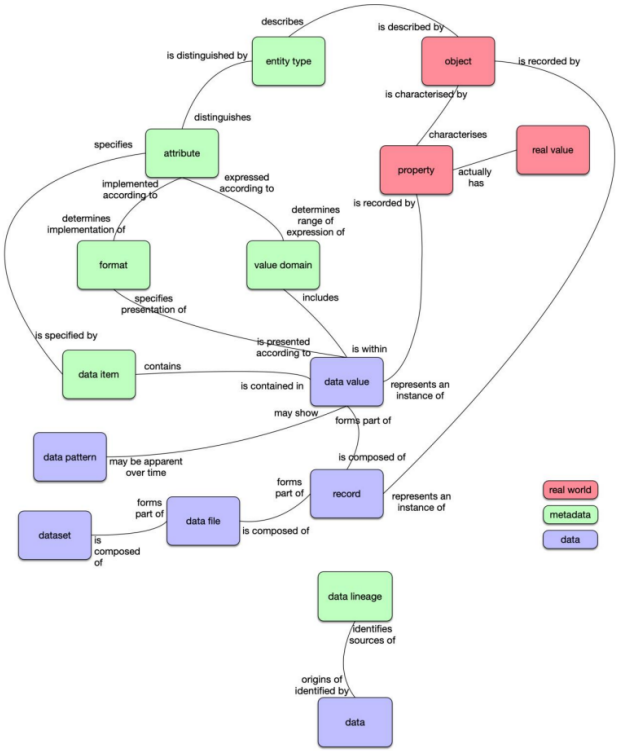


Figure 1: Data concept system DDQ

Data Concept	Definition	Relationship with other concepts
Data	A representation of facts, concepts, or instructions in a formalized manner, suitable for communication, interpretation, or processing by humans or by automatic means. (ISO 2382-4).	
Data value	The value of a data item	Is contained in data item. Forms part of record Is within value domain Represents a property of an object
Record	A logically related set of data values that represent a (real-world) object	Forms part of data file Is composed of data values
Dataset	Any organized collection of data.	Is composed of data files
Dataset availability	The degree to which a dataset can be consulted or retrieved by data consumers or processes.	Is a characteristic of a dataset.
Dataset Composition	The way in which a dataset is made up	

The definition of dimensions is based on Table 2 (p. 22) of the same article.

By "**DAMA Common?**" the article intends whether the dimension is widely used and mentioned in the literature or not

By wide? it means that the dimension is mentioned in the Company Quality documentation pages .

We tried to highlight here the ones which are "measurable", thus, we can create a test for it

DIMENSION	DAMA Common?	wide?	Data Concept	Definition	Example tests

Completeness	y	y	Records	The degree to which all required records in the dataset are present.	• Check that we have no gaps in ingested records. E.g. Check that we do not miss records: Check that in the last 30 days, each day contains at least 1000 records based on their transaction date.
			Data values	The degree to which all required data values are present.	• Check that all values for the shop_location field are filled for all the records • The media types' values (e.g. phone, sms, chat) for some type of customer contacts are not filled
			Metadata	The degree to which the metadata are fully described.	• Check all tables in Dataset have a description • Check all columns in a table have a description
Timeliness	y	y	Dataset availability	The degree to which the period between the time of creation of the real value and the time that the dataset is available is appropriate.	• Based on any field containing a timestamp (e.g. ingestion_timestamp), it checks that new data have been ingested within the last N hours
Consistency	y	y	Data Values	The degree to which data values of two sets of attributes within a record, within a data file, between data files, within a record at different points in time comply with a rule	Some checks could also be subcategorised as Plausibility • within a record: Check if the discount_percentage field is filled when the discount_flag field is TRUE • within a data file: In a sales table, a transaction containing a cancellation sale should have its cancelled_amount_eur field <= than the sale_amount_eur of the original transaction. • between data files: TBD • within a record a at different points of time: Ex 1. We have a table where records are never updated in place, instead, a new record is stored with the change. If the record contains an ID, a customerID, the transaction's location (say Malmo) and the IKEA CustomerSupport region (Skåne), one could create a rule to check whether, over time, the transaction location always matched the CustomerSupport region. This could be useful to identify special cases, where perhaps the customer has bought something in Copenhagen, but then contacted the customer support center in Sweden. Ex. 2 The running total amount spent is stored in a table per each customer. Records are always appended and never updated. This latest total should always be greater than the sum of all the previous entries. This could identify erroneous calculations
			Data values of a set of attributes of a dataset at different points in time (temporal consistency)	The degree to which the data values of a set of attributes of a dataset at different points in time comply with a rule.	•
			Data values of two sets of attributes between datasets (across datasets)	The degree to which data values of two sets of attributes between datasets comply with a rule	• see Referential Integrity • A customer interaction in the interaction_table has happened around 12:00, but the corresponding call recording in the voice_recordings table says that the call happened at 14:00 • The number of incoming customer support interactions each week, from the interactions table, matches (or is in range of) the total number of customer tickets opened in each week in the support_tickets table
			Data values of two sets of attributes between records (cross record)	The degree to which data values of two sets of attributes between records comply with a rule	• The same customer interaction appears in 2 records, A: where the interaction starts, and B: where the interaction ends. Problem is that some field's values contradict each other. In A, contact_center_location = Malmo, sales_region = Skåne in B, contact_center_location = Malmo, sales_region = Stockholm Region (perhaps the call was transferred to an employees in another region)
			Data values of two sets of attributes within a record (record level)	The degree to which data values of two sets of attributes within a record comply with a rule.	• The store_location = Paris, but the market= France South (while Paris belongs to France North)
Uniqueness	y	y	Records	The degree to which records occur only once in a data file	• No duplicates found in one field or a group of fields in a table ◦ The same store_location is present twice with 2 different IDs
Accuracy	y	y	Data values	The degree of closeness of data values to real values	• The store_location has closed but sales are still coming in from that location • Count how many times a field contains one or more values as a percentage of the total no. of records.. The test PASS/FAIL depending on exceeding a threshold.
Validity	y	y	Data Values	The degree to which data values comply with rules	• Data are valid only up to a specific time (e.g. name of a product has changed after 1/1/2022) • Data value in range or in set of values

Plausibility (could be Consistency)	n	y	Data Values	The degree to which data values match knowledge of the real world	Inspired by here.
					Could also be included under Consistency - Data values
					• Stability: The avg monthly transactions should not be 200% more than the previous month
					• Outlier Analysis:
					  ◦ The avg monthly sales should not exceed 500% of the previous month
					  ◦ Single sales should not exceed 1M euros

Other common and less common dimensions:

DIMENSION	DAMA Common?	ingka-wide?	Data Concept	Definition	Example tests

Integrity	n	n	Data values	The degree of absence of data value loss or corruption.	• A field can always be safely cast to another type. E.g. A STRING representing a number, can always be cast to a FLOAT type E.g. A STRING representing a date in a special format can always be cast to a DATETIME type
Traceability	y	n	Data	The degree to which data lineage is available	
Punctuality	y	n	Dataset availability	The degree to which the period between the actual and target point of time of availability of a dataset is appropriate	• Transactions from 1st of Jan had a target availability date of Jan 3 but arrived on the 4th transaction_date = Jan 01 (batch of transactions) target_date = Jan 03 (transactions were expected by this target date) actual_availability_date = Jan 08 (but arrived on this actual date)
Accessibility	n		Data	The ease with which data can be consulted or retrieved	Business related, needs a clear definition to have a chance to test it programmatically
Clarity	y		Metadata	The ease by which data consumers can understand the metadata of a dataset	Business related & Subjective, hard to test it programmatically
Relevancy	n		Composition of datasets	The degree to which the composition of datasets meets the needs of the data consumer	Business related, hard to test it programmatically
Currency	y	n	Data Values	The degree to which data values are up to date.	
Availability	y	n	Data	The degree to which data can be consulted or retrieved by data consumers or a process.	Check if a service user has access to the datasets or datafiles/tables
Confidentiality	n	n	Data	The degree to which disclosure of data should be restricted to authorized data consumers	
Latency	n	n	Data	The period of time between the point when the data is created and the point when it is available for use NB: Similar to timeliness, but focus on short time periods (millesecs) NB2: Timeliness VS Latency : latency focuses on short time periods (millesecs) while timeliness could be years Timeliness starts counting the time delay from when the event happens in the real world (e.g. timestamp A - customer sales happen in a shop) Latency starts counting from when the sales are actual created as data (e.g. timestamp B - imagine the sales were stored on a paper and then they are inserted in a Database) Both Timeliness (from timestamp A) and latency (from timestamp B) measure the time passed until data are available for analysis (timestamp C)	
Referential Integrity (can be Consistency)	n	n	Data Files	The degree to which data values of the primary key of one data file and data values of the foreign key of another data file are equal. NB. Can be a subset of consistency	PK-FK referential integrity

The setup

CDQ is maintained by Google and used by the Dataplex Service.

We use CDQ as an open source library, adding to it a set of extra generic tests definitions covering the various data quality dimensions.

Once triggered, CDQ will automatically save the results of its run in BigQuery, and a template ~~Cloud Data Studio~~ Grafana Dashboard can visualize the results.

Plus and minuses

Plus - Most interesting features of CDQ

1. It embeds by design the concept of Data Quality Dimensions - generic tests belong to a Data Quality Dimension and the run results can be easily grouped by these dimensions.
2. It provides a limited profiling on top of the tests results
    a. NB: The profiling is limited to the columns tested, it is not a generic and dataset-wide as defined in Data Profiling: Requirements Definition
    b. NB2 The statistical profiling is in relation to the columns being tested and depends on the type of test run
3. Stores the tests results directly in BigQuery with minimal configuration
4. Can be deployed as standalone component (no need of other tools, like DBT, Dataform etc.)
5. Developed and maintained by Google to define data quality tests in Dataplex

Minus - Some limitations

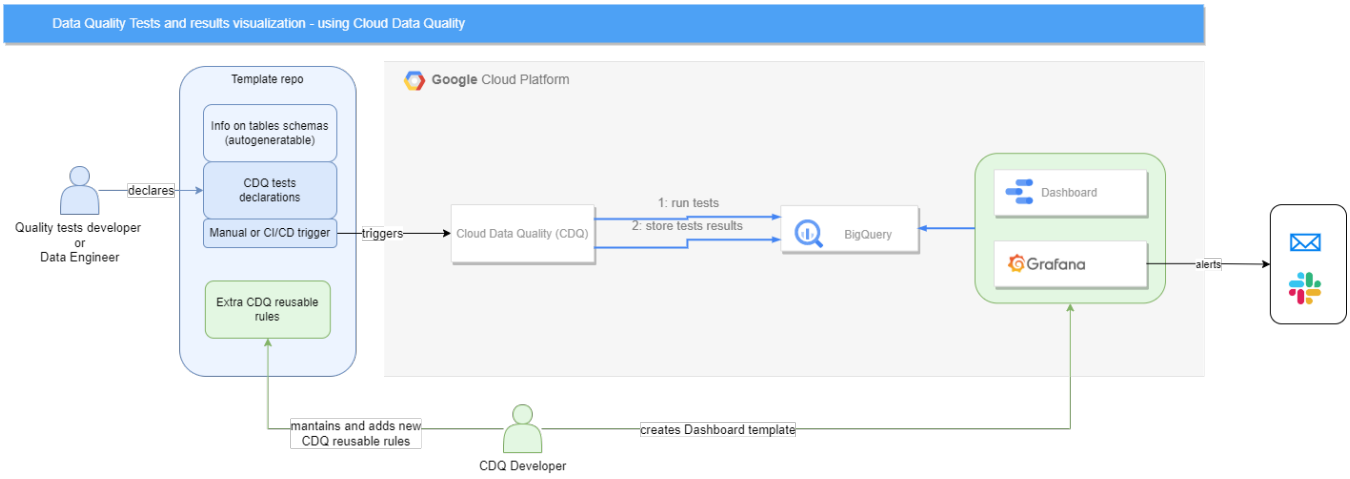
1. It has very few embedded generic tests/rules that can be used out of the box - but it's easy to add new ones.
2. Extra SQL tests can only be templated using SQL, no Javascript (Dataform) or Jinja (DBT) seems available
3. Query to extract failing records not available yet (coming soon)
4. There's no out-of-the-box way to run only a subset of checks, or to define such subsets. Though, one can select one specific rule to run at one time, so one could create several CI scripts with groups of cdq run commands grouping several checks

Who does what ?

CDQ developer: develops and maintains a set of generic tests and defines a dashboard template to visualize the results.

Quality tests developer: defines which tests apply to his/her dataset/tables/columns and declares the bindings (i.e on which table and column the test needs to be run on)

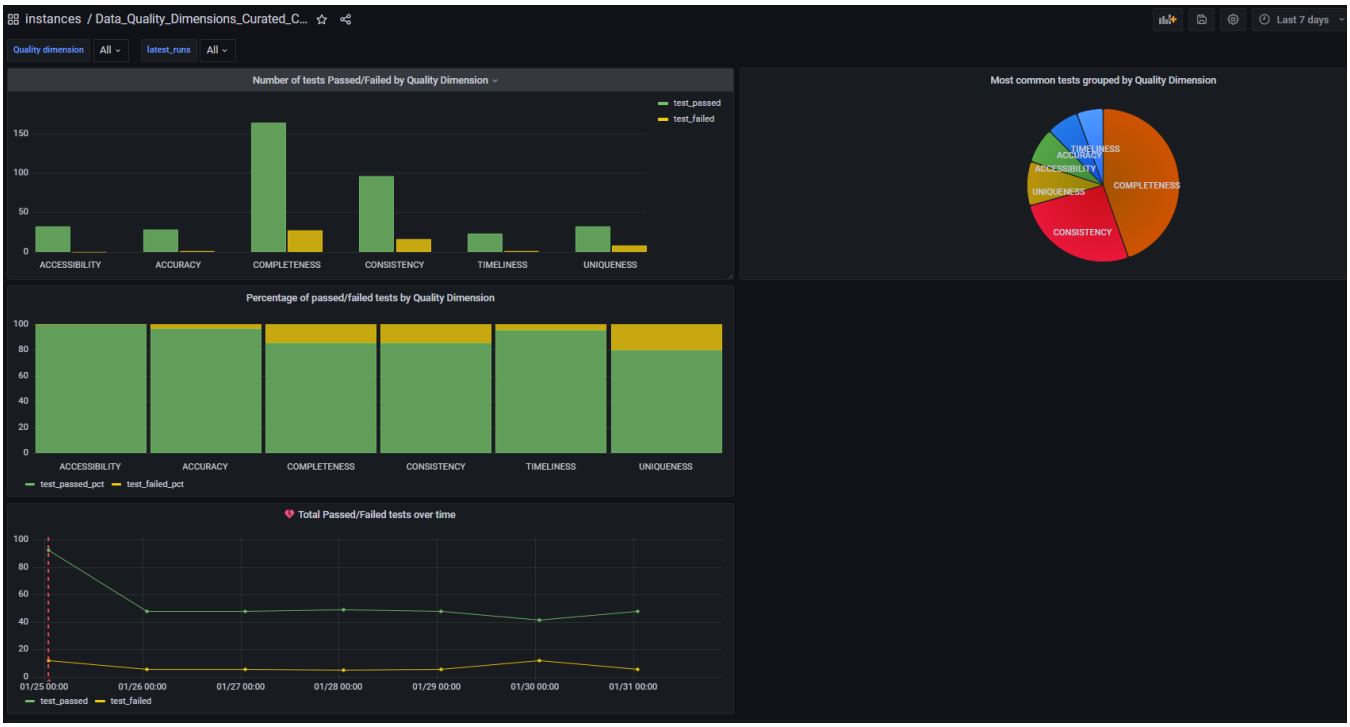
CDQ: Upon being triggered, it runs the tests, calculates some tests results statistics profiling and saves the results in BigQuery.



Visualizing the tests results by quality dimensions

We created a simple Data Studio dashboard on top of the CDQ results table to visualize the tests results aggregated by quality dimensions. Dimensions could later be ranked and a final data quality score could be obtained.

Dashboard : TODO - Add new Grafana link when available externally ~~Link (you need access to dashboard itself and the underlying data)~~



archived questions