Data Architect - Data Quality

Naveen Gurram-ext	0.1	Initial Version	11 Jan 2023
Fernando Girante	0.2	Revisions to the original text	09 May 2023

1- Introduction

1.1- Purpose of the document

This document describes how to define data quality metrics on the data loaded into GCP.

1.2- Intended Audience

This document is intended for the Data Architectures, Data Engineering and operational team.
It will be used as reference for any project or domain for the developments of the models.

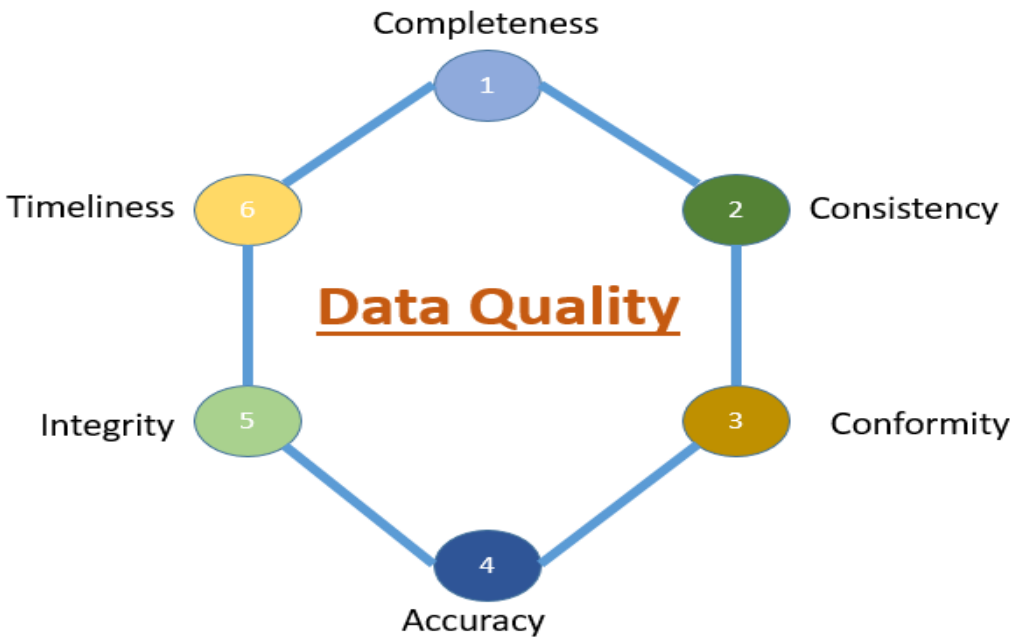
2- How it works

2.1- Description

Data quality is essential for any business to make informed decisions. Measuring the right Data Quality metrics ensures that we have most trustable data.
Different types of metrics can be used to track and report the quality of data available in GCP. Each type of KPI has it's own importance and role.

2.2- Data Quality Dimensions:

Data quality is reliant upon the following dimensions:



Completeness: Incomplete data is as dangerous as inaccurate/wrong data. Data requirements should be clearly specified based on the information needs of the organization and data collection processes matched to these requirements.

- Degree to which expected records are present, and data attributes are populated
- Degree to which duplicate entities are identified and appropriately resolved
- Degree to which values not needed for decision making are excluded

Consistency: Data consistency ensures the data is the same across the organization no matter where it appears.

- Degree to which data is synchronized across all sources
- Degree to which data is consistent between subject areas
- Degree to which data is consistent across multiple transaction entries

Conformity: Data conformity measures the data is following the set of standard data definitions like data type, size and format.

- Degree to which data values comply with the specified formats
- Degree to which duplicate entities are identified and appropriately resolved
- Degree to which values not needed for decision making are excluded

Accuracy: As the name implies, it refers to the exactness of the data. Data cannot have any erroneous elements and must convey the correct message without being misleading.

- Degree to which the data correctly reflects the real world? (Ex: Postcode in right format)
- Degree to which reported information values confirm with the true or accepted values

Integrity: Integrity ensures that all the data in a database can be traced and connected to other data

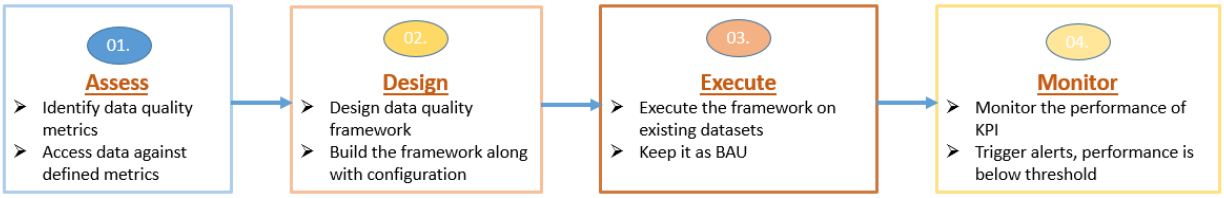
- Degree to which a defined relational constraint is implemented between two data sets

Timeliness: Data should be captured as quickly as possible after the event or activity and must be available quickly and frequently enough to support management decisions.

- Degree to which specified data values are up to date between data change and processing

2.3- Data Quality Framework:

A data quality framework is a systematic process that consistently monitors data quality, implements a variety of data quality processes and triggers the alarms as soon as the quality of a particular KPI is observed below threshold.



3- Data Model

3.1- Description

To implement a data quality monitoring framework, a data quality data mart is needed.
Described below the data mart can be used to handle the data quality rules and store the result sets.

3.2- Model

- **DQ_RULE_CONFIG:** A table to store all predefined data quality rules.
- **DQ_CONFIG:** A table to store the possible DQ_CHECK. Checks which can be executed at table and attribute level
- **DQ_RULE_AUDIT_RESULTS:** A table to store the execution results of data quality rules which can help to monitor the performance of the KPI's

DQ_CONFIG
DQ_CHECK_ID: INTEGER
DQ_CHECK_NAME: CHARACTER(0)
DQ_CHECK_DESC: CHARACTER(0)
CHECK_STAGE: CHARACTER(0)
INSERT_DATETIME: TIMESTAMP(6)
UPDATE_DATETIME: TIMESTAMP(6)

DQ_RULE_CONFIG
RULE_ID: INTEGER
RULE_NAME: CHARACTER(0)
RULE_DESCR: CHARACTER(0)
DOMIAN_NAME: CHARACTER(0)
DIMENSION_NAME: CHARACTER(0)
PROJECT_NAME: CHARACTER(0)
TABLE_NAME: CHARACTER(0)
COLUMN_NAME: CHARACTER(0)
DQ_CHECK_ID: INTEGER
FREQUENCY: CHARACTER(0)
FILTER_CONDITION: CHARACTER(0)
SQL_STATEMENT: CHARACTER(0)
SEVERITY: CHARACTER(0)
RULE_EXECUTE_STATUS_CODE: CHARACTER(0)
LOWER_THRESHOLD_VALUE: INTEGER
UPPER_THRESHOLD_VALUE: INTEGER
INSERT_DATETIME: TIMESTAMP(6)
UPDATE_DATETIME: TIMESTAMP(6)

DQ_RULE_AUDIT_RESULTS
DQ_RULE_RESULT_ID: INTEGER
RULE_ID: INTEGER
DOMIAN_NAME: CHARACTER(0)
DIMENSION_NAME: CHARACTER(0)
TABLE_NAME: CHARACTER(0)
COLUMN_NAME: CHARACTER(0)
STATUS: CHARACTER(0)
RECORD_COUNT: INTEGER
SUCCESS_PERCENTAGE: INTEGER
FAILURE_PERCENTAGE: INTEGER
INSERT_DATETIME: TIMESTAMP(6)
UPDATE_DATETIME: TIMESTAMP(6)

3.2- Tables

DQ_CONFIG:

001	DQ_CHECK_ID	Unique Identifier	Integer	E.g. 1001
002	DQ_CHECK_NAME	Name of the DQ check	String	E.g. NOT_NULL_CHECK, CUSTOM_CHECK
003	DQ_CHECK_DESC	Details about the DQ check	String	E.g. This check will be to validate the number of nulls values present in the given attribute
004	CHECK_STAGE	Details weather the check can be applied at FIELD or TABLE level	String	E.g. FIELD, TABLE
005	INSERT_DATETIME	Date & Time the record was first inserted into the table	Timestamp	E.g. 2022-12-12 17:00:21 UTC

006	UPDATE_DATETIME	Date & Time the record was last updated	Timestamp	E.g. 2022-12-12 17:00:21 UTC

DQ_RULE_CONFIG:

001	RULE_ID	Unique Identifier	Integer	E.g. 1
002	RULE_NAME	Name of the Rule	String	E.g. COMP_CD_NOT_NULL_CHECK
003	RULE_DESCR	Details about the rule	String	E.g.
004	DOMAIN_NAME	Name of the domain, the rule is related to .	String	E.g. Finance, HR, Industrial and R&I
005	DIMENSION_NAME	Dimension (Type) of DQ Check	String	E.g. Completeness, Conformity, Accuracy, Consistency, Timeliness, Integrity
006	PROJECT_NAME	Name of the GCP project	String	E.g. pcm_dev
007	TABLE_NAME	Name of the Table that holds the data on which Data Quality Rules need to execute	String	E.g. COMPANY_CODE
008	COLUMN_NAME	Column Name on which rule is executed	String	E.g. C_COMPCDE
009	DQ_CHECK_ID	Surrogate key to DQ_CONFIG.DQ_CHECK_ID attribute	Integer	E.g. 1001
010	FREEQUENCY	Indicates how frequently the check is scheduled to execute	String	E.g. Daily, Weekly
011	FILTER_CONDITION	Provide the filter condition if the checks are related to rang check. Should be populated as 'NA' for regular checks like Not Null, custom checks.	String	E.g. where comp_cd in ('A','B')
012	SQL_STATEMENT	SQL statement can be used for custom DQ checks.	String	
013	SEVERITY	Severity level of the DQ check	String	E.g. High
014	RULE_EXECUTE_STATUS_CODE	Indicates the status of the code. 'Active' means the check will be scheduled to run. Inactive means the schedule will be skipped to execute.	String	E.g. Active, Inactive
015	LOWER_THRESHOLD_VALUE	Lower threshold value can be used to check the result of the DQ check	Integer	E.g. 96
016	UPPER_THRESHOLD_VALUE	Upper threshold value can be used to check the result of the DQ check	Integer	E.g. 100
017	INSERT_DATETIME	Date & Time the record was first inserted into the table	timestamp	E.g. 2022-12-12 17:00:21 UTC
018	UPDATE_DATETIME	Date & Time the record was last updated	timestamp	E.g. 2022-12-12 17:00:21 UTC

DQ_RULE_AUDIT_RESULTS:

001	DQ_RULE_RESULT_ID	Id of the run, this information come from Talend job run	Integer	E.g. 1
002	RULE_ID	Reference to DQ_RULE_CONFOG.RULE_ID	Integer	E.g. 1
003	DOMAIN_NAME	Name of the domain, the rule is related to .	String	E.g. Finance, HR, Industrial and R&I
004	DIMENSION_NAME	Dimension (Type) of DQ Check	String	E.g. Completeness, Conformity, Accuracy, Consistency, Timeliness, Integrity
005	TABLE_NAME	Name of the Table that holds the data on which Data Quality Rules need to execute	String	E.g. COMPANY_CODE
006	COLUMN_NAME	Column Name on which rule is executed	String	E.g. C_COMPCDE

007	STATUS	Status of the execution	String	E.g. Green, Amber, Red Green: If the Success count is above upper threshold value Amber: If the Success count is between lower and upper threshold values Red: If the Success count is less than lower threshold value
008	RECORD_COUNT	Number of the records present in the table	Integer	E.g. 1543
009	SUCCESS_PERCENTAGE	Count of the records that returned by the DQ check / RECORD_COUNT	Integer	E.g. 97
010	FAILURE_PERCENTAGE	Count of the records that's not satisfied by the DQ check / RECORD_COUNT	Integer	E.g. 99
011	INSERT_DATETIME	Date & Time the record was first inserted into the table	Timestamp	E.g. 2022-12-12 17:00:21 UTC
012	UPDATE_DATETIME	Date & Time the record was last updated	Timestamp	E.g. 2022-12-12 17:00:21 UTC

Use Case1:

Not Null validation rule:
Verify the company code C_COMPCDE attribute is populated without null values.
As the rules is at a attribute level it will have below entry in DQ_CONFIG table

DQ_CHECK_ID	DQ_CHECK_NAME	DQ_CHECK_DESC	CHECK_STAGE
1001	NOT_NULL_CHECK	This check will validate the number of NULL values in any given attribute	FIELD

Changes to DQ_RULE_CONFIG table:
The rule is "Active" and scheduled to run on a daily basis as the RULE_EXECUTE_STATUS_CODE field is set as "Active". The Threshold pass percentage is set to 100 (Both Lower and Upper threshold pass percentage is set to 100), meaning that even one record does not meet the criteria, Even one record is populated as NULL, the rule will fail and the STATUS in DQ_RULE_AUDIT_RESULTS table will be set as RED.
As the CHECK is related to NOT NULL, Both FILTER_CONDITION and SQL_STATEMENT field values can be set as NA

DIMENSION_NAME	PROJECT_NAME	TABLE_NAME	COLUMN_NAME	DQ_CHECK_ID	FREQUENCY	SEVERITY	RULE_EXECUTE_STATUS_CODE	LOWER_THRESHOLD_VALUE	UPPER_THRESHOLD_VALUE
	predict-credit-mgt-v2-dev	COMPANY_CODE	C_COMPCDE	1001	Daily	High	Active	100	100

Use Case2:

Range check validation rule:
Verify the company code C_AUTHMA attribute has got values ECo or SCo only.
As the rules is at a attribute level it will have below entry in DQ_CONFIG table

DQ_CHECK_ID	DQ_CHECK_NAME	DQ_CHECK_DESC	CHECK_STAGE
1002	RANGE_CHECK	This check will be used to validate the field is with in the range or list of values	FIELD

Changes to DQ_RULE_CONFIG table:
The rule is "Active" and scheduled to run on a daily basis as the RULE_EXECUTE_STATUS_CODE field is set as "Active". The Threshold pass percentage is set to 100% (GREEN) if the success percentage is 100% and the status will be set to AMBER if the success percentage is >=98% and <100% and RED if it is <98%.
As the CHECK is related to RANGE CHECK, SQL_STATEMENT field values can be set as NA but FILTER_CONDITION field will be populated as "C_AUTHMA in ('ECo','SC0')"

DIMENSION_NAME	PROJECT_NAME	TABLE_NAME	COLUMN_NAME	DQ_CHECK_ID	FREQUENCY	SEVERITY	RULE_EXECUTE_STATUS_CODE	LOWER_THRESHOLD_VALUE	UPPER_THRESHOLD_VALUE
	predict-credit-mgt-v2-dev	COMPANY_CODE	C_AUTHMA	1002	Daily	High	Active	98	100

Use Case3:

Custom check validation rule:
Verify the referential integrity between credit_mgmt code.C_COMPCDE attribute and company_code.C_COMPCDE.
As the rules is at a attribute level it will have below entry in DQ_CONFIG table

DQ_CHECK_ID	DQ_CHECK_NAME	DQ_CHECK_DESC	CHECK_STAGE
1003	CUSTOM_CHECK	This check will be used to carry custom checks at field level	FIELD

Changes to DQ_RULE_CONFIG table:
The rule is "Active" and scheduled to run on a weekly basis as the RULE_EXECUTE_STATUS_CODE field is set as "Active". The Threshold pass percentage is set to 100% (GREEN) if the success percentage is equal or above 95% and the status will be set to AMBER if the success percentage is >=92% and <95% and RED if it is <92%.
As the CHECK is related to CUSTOM CHECK, SQL_STATEMENT field will be populated with the query to be executed.

DIMENSION_NAME	PROJECT_NAME	TABLE_NAME	COLUMN_NAME	DQ_CHECK_ID	FREQUENCY	SEVERITY	RULE_EXECUTE_STATUS_CODE	LOWER_THRESHOLD_VALUE	UPPER_THRESHOLD_VALUE
	predict-credit-mgt-v2-dev	COMPANY_CODE	C_AUTHMA	1003	Weekly	High	Active	92	95

4- Talend Jobs

Reference: