



Grand Design Data Quality in Colibra

Datum	5/17/2023
Auteur	Alwine van Aarle
Rapport	Grand Design
Versie	1.0



Versiebeheer

Versie	Datum	Auteur	Bijzonderheden
0.1	20-1-2023	Alwine van Aarle	Initiële versie
0.2	27-3-2023	Alwine van Aarle	Versie na 1 ^e review
1.0	17-5-2023	Alwine van Aarle	Enkele wijzigingen aangebracht

Versie	Datum	Akkoord EDM	Akkoord Architect

Distributielijst

Persoon	Functie

Copyright PostNL BV. © te Den Haag

Niets uit deze uitgave mag verveelvoudigd en/of openbaar worden gemaakt (voor willekeurig welke doeleinden) door middel van druk, fotokopie, microfilm, geluidsband, elektronisch of op welke andere wijze dan ook zonder voorafgaande schriftelijke toestemming van PostNL BV. Dit rapport is enkel en alleen bedoeld voor intern gebruik voor hierboven genoemd(e) bedrijf/bedrijven.

Inhoudsopgave

1. Inleiding	6
1.1. Context	6
1.2. Doelsituatie	6
1.3. Betrokken Partijen	7
1.4. Uitgangspunten	7
1.5. Scope	7
2. Modellen en attributen	9
2.1. Contextdiagram DQ	9
2.2. Logisch datamodel	10
3. Functionaliteiten	11
3.1. Het configureren van een DQ job	11
3.1.1. Dataset configureren	11
3.1.2. Het configureren van een profile	11
3.1.3. Colibra dimensies configureren	11
3.2. Het maken en beheren van datakwaliteitsregels	12
3.3. Het exporteren van de resultaten van DQ jobs	13
3.4. Alerts	13
3.5. Scheduling	14
3.6. View and manage	14
3.6.1. Catalog	14
3.6.2. Scorecards	14
3.6.3. Pulse view	14
3.6.4. List view	14
3.6.5. Workflow	15
3.7. Interfaces	15
3.7.1. Connectie Athena	15
3.7.2. Connectie S3	15
3.7.3. Connectie flatfiles	16
3.7.4. Connectie naar PowerBI	16

4. Databeheerprocessen 17

4.1.	Rollen en rechten	17
4.2.	Eigenaarschap data en systeem	18
4.2.1.	Data eigenaarschap	18
4.2.2.	Systeem eigenaarschap	18
4.3.	Datakwaliteitsmaatregelen	19
4.3.1.	Bepalen van de randvoorwaarden	19
4.3.2.	Vastleggen rollen, aanleverspecificaties en werkinstructies	19
4.3.3.	Vaststellen rapportagebehoefte	19
4.3.4.	Opstellen van de rapportages	20
4.3.5.	Periodiek toetsen van de voorgaande stappen	20

5. Bijlagen 21

5.1.	Technische inrichting Colibra	21
5.1.1.	Systeem	22
5.1.2.	Tabel	22
5.1.3.	Veld	23
5.1.4.	Data Set	23
5.1.5.	DQ regel_Business rule	25
5.1.6.	Business rule	26
5.1.7.	Profile	27
5.1.8.	DQ job	29
5.1.9.	Colibra dimensies	30
5.1.9.1	Behavior	30
5.1.9.2	Rules/Datakwaliteitsregel	32
5.1.9.3	Outliers	33
5.1.9.4	Pattern	35
5.1.9.5	Source	36
5.1.9.6	Record	37
5.1.9.7	Schema	38

5.1.9.8	Dupes	38
5.1.9.9	Shapes	39
5.1.10.	Dashboard	41
5.1.11.	Drilldown	42
5.1.12.	Report	43

1. Inleiding

1.1. Context

Normaliter is een Grand Design een ontwerpdocument, aan de hand waarvan een solution of functionaliteit gebouwd kan worden. Hier betreft het een omschrijving van een bestaande solution, waarin bovendien geen eigen ontwerpkeuzes kunnen worden gemaakt. Het doel van dit Grand Design is daarom om inzicht te geven in het ontwerp en functionaliteit van Colibra DQ - een zogenaamde "blik onder de motorkap".

Data Quality Management bestaat uit het borgen van correcte data opvoer vooraf, monitoring achteraf en beheersing van het transport tussen producent en gebruiker. Dit Grand Design behelst de monitoring achteraf in de daarvoor aangeschafte tool, Colibra Data Quality.

Tot halverwege begin 2022 had PostNL voor DQ monitoring een module van Informatica tot haar beschikking; IDQ. Met de migratie van Informatica naar EBX als MDM-tool, ontving IDQ geen actuele data meer. Omdat IDQ bovendien niet uitblonk in gebruikersvriendelijkheid, heeft er een selectietraject plaats gevonden om een nieuwe tool te selecteren, dat eind 2021 werd afgerond.

Hiervoor zijn er gesprekken geweest met andere bedrijven over hun tooling. Hieruit, en uit de aanbevelingen van Gartner en Forrester Wave, is er een long list van mogelijke leveranciers geselecteerd. Aan de hand van demo's van deze partijen, is er met twee partijen van de short list een PoC gedaan. Eind 2021 is er vervolgens voor Colibra gekozen. Ook voor de data catalog voldeed Colibra het beste aan de requirements.

Er zijn meer tools aanwezig binnen PostNL die beschikken over Data Profiling en functionaliteit. PowerBI is daar een voorbeeld van. Aan het begin van het selectietraject is hier naar gekeken, en besloten dat PowerBI niet het antwoord op het volledige DQ-vraagstuk is.

1.2. Doelsituatie

De technische implementatie van Colibra DQ is inmiddels afgerond. Uiteindelijk is het de bedoeling dat van objecten behorende tot de masterdata, transactionele data en conditionele data van PostNL de datakwaliteit in kaart gebracht wordt, en er actief wordt gestuurd op het verhogen hiervan. Hiervoor dienen in het kader van het Data Quality Management-proces voor deze objecten datakwaliteitsregels te worden opgesteld.

Voor het meten van Data Quality zijn drie capabilities van belang; profiling, DQ monitoring en visualisatie.

Data profiling behelst een objectieve analyse van de data. Het resultaat is een overzicht met daarin het datatype, welke patronen er te vinden zijn in een kolom, en hoe vaak die terugkomen. De gemiddelde, laagste en hoogste waarden, alsmede de distributie over die waarden wordt inzichtelijk. Deze stap van het proces is belangrijk voor het leren kennen van de data. Bovendien kunnen hier al aanwijzingen voor Data Quality issues naar voren komen. Een kolom met postcodes, waarvan niet alle waarden aan het patroon 9999AA voldoen, verdient bijvoorbeeld nadere aandacht. Data profiling is geen doel op zich; het is een middel dat gebruikt kan worden om data en de bijbehorende DQ regels te ontdekken.

Bij *Data Quality monitoring* wordt aan de hand van Data Quality rules inzichtelijk gemaakt wat de kwaliteit van de data is. Een Data Quality rule is subjectief: middels rules weet de tool wat er goed dan wel fout is aan data. Een voorbeeld van een Data Quality rule die tijdens monitoring zou kunnen worden toegepast, is dat een Nederlandse postcode altijd aan het patroon 9999AA dient te voldoen.

Colibra Data Quality maakt naast de traditionele data profiling en traditionele data quality monitoring door middel van datakwaliteitsregels ook gebruik van machine learning en AI om problemen met datakwaliteit op te sporen. Deze functionaliteit is zowel van belang voor data profiling als voor data quality monitoring. Zo is het opsporen van duplicaten door de tool "slimmer" dan alleen het signaleren van exacte dubbelingen, en is

het ook mogelijk om data als goed of fout aan te merken aan de hand van de bevindingen. Meer hierover volgt in paragraaf 3.1.3.

Visualisatie in het kader van Data Quality behelst het visualiseren van de resultaten van DQ monitoring in een dashboard, middels grafieken of andere grafische weergaven. Het doel van deze capability is inzicht en sturen te verschaffen, en om de cijfers op een visueel aantrekkelijkere manier te presenteren.

De toegevoegde waarde van alle drie de capabilities is dat PostNL haar data beter kan leren kennen, inzicht krijgt in de kwaliteit van haar data, en de resultaten van de verbeterinitiatieven van de kwaliteit van haar data kan monitoren. Uit een profile of dashboard kunnen bovendien Data Quality issues naar voren komen, die vervolgens opgelost kunnen worden. Op die manier kan de datakwaliteit verbeterd worden, wat ook weer de kwaliteit van de core processen van PostNL, en haar data-producten verbetert.

Het is de bedoeling dat iedere DMO de objecten waar het verantwoordelijk voor is, op kwaliteit gaat meten en sturen. EDM zal hierin een ondersteunende rol spelen. Op termijn zullen ook de platform-teams aan DQ monitoring kunnen doen. Dit is op moment van schrijven nog niet uitgewerkt.

1.3. Betrokken Partijen

Voor het beantwoorden van onderhavig vraagstuk is onder andere gesproken met de volgende partijen:

- EDM (Manager EDM, Data Architect, Data Management Consultant, Data Quality Analyst).
- DMP (Solution Consultant, Data Engineer).
- DMO's (Data Manager, Data Steward).
- DAP (Data Engineer).
- Collibra.

Indien Collibra ook toegepast gaat worden voor decentrale kwaliteitsmetingen, dan zal er met aanvullende partijen gesproken moeten worden. Voor nu is de scope de DMO's als eigenaars van de data, die daarmee ook verantwoordelijk zijn voor de kwaliteit ervan.

1.4. Uitgangspunten

- Collibra DQ maakt om te beginnen verbinding met de bronnen op het data lake. Er zal vooralsnog geen individuele connectie naar de verschillende bronsystemen komen.
- Iedere DMO is zelf verantwoordelijk voor het monitoren van de kwaliteit van de data-objecten waar het eigenaar van is. EDM biedt ondersteuning en beheert de tool.
- De resultaten van de DQ scores mogen binnen het bedrijf gedeeld worden.
- Procedures en richtlijnen zijn beschikbaar.

1.5. Scope

In Collibra DQ zal structureel de kwaliteit van de data in bronnen in kaart worden gebracht, waarvoor de benodigde stappen van het DQM-proces zijn doorlopen en die beschikbaar zijn op het data lake.

Daarnaast zijn ad-hoc analyses van flat-files ook mogelijk. Dit kan helpen bij het onderzoeken van DQ issues.

Hiervoor zullen de volgende functionaliteiten moeten worden ingericht:

- Het configureren van een DQ job. Dit behelst:
 - Dataset configureren.
 - Het configureren van een profile.

-
- Colibra dimensies configureren.
 - Het maken en beheren van datakwaliteitsregels in Colibra.
 - Het exporteren van de resultaten van DQ jobs in Colibra, ten behoeve van het bouwen van dashboards in PowerBI.
 - Alerts.
 - Scheduling
 - View and manage.
 - Interfaces.

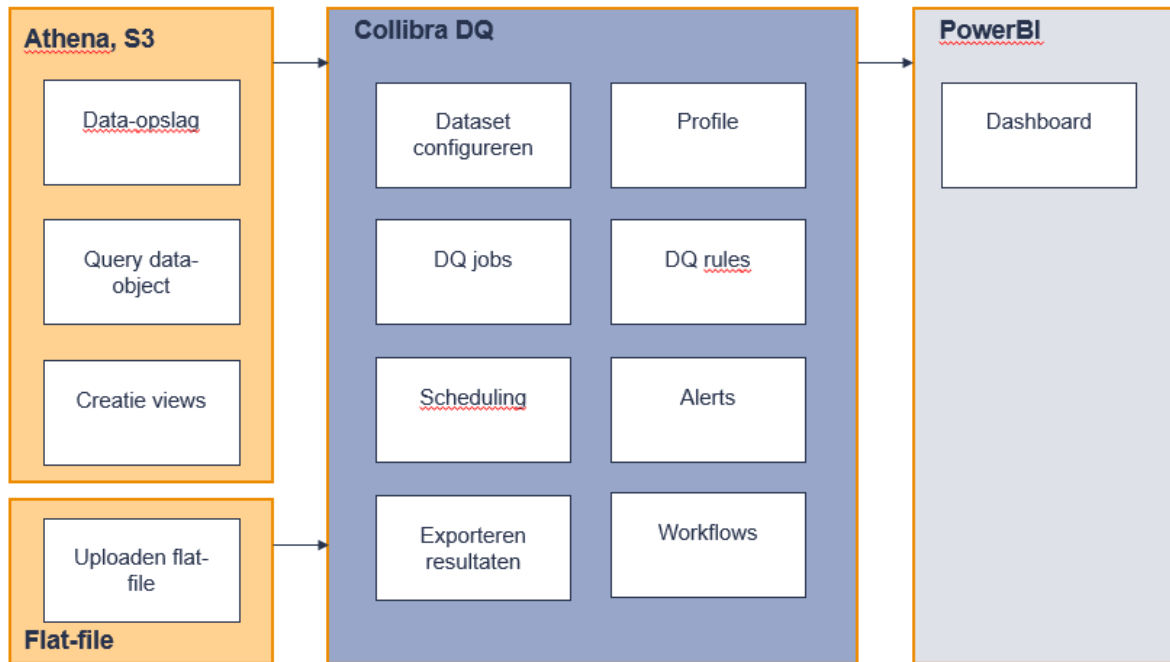
Op dit moment zijn voor dit document de volgende onderwerpen niet in scope:

- De connectie tussen Colibra DQ en Colibra DC (dit volgt wanneer beide systemen gevuld zijn en verbonden kunnen worden).
- De connectie tussen Colibra DQ en visualisatie-tool PowerBI.
- De toekomstige veranderingen voor Colibra DQ in het kader van data lake 2.0 en DES (DES Exploration Services).
- Welke oplossing er zal worden gekozen voor HR-data. Deze staat niet op het data lake, en het is op dit moment nog de vraag of Colibra de oplossing is voor het meten van de kwaliteit van de data van HR.

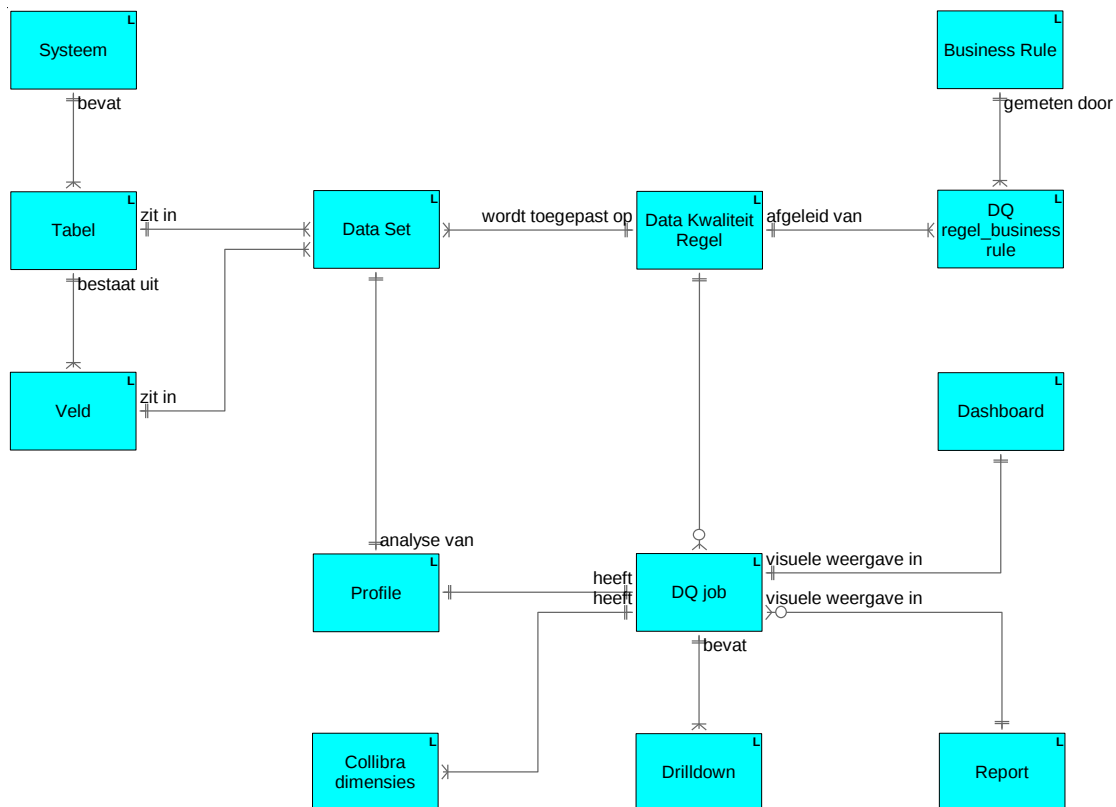
Deze onderwerpen zullen in een later stadium worden uitgewerkt, als er meer bekend is over de verschillende onderwerpen.

2. Modellen en attributen

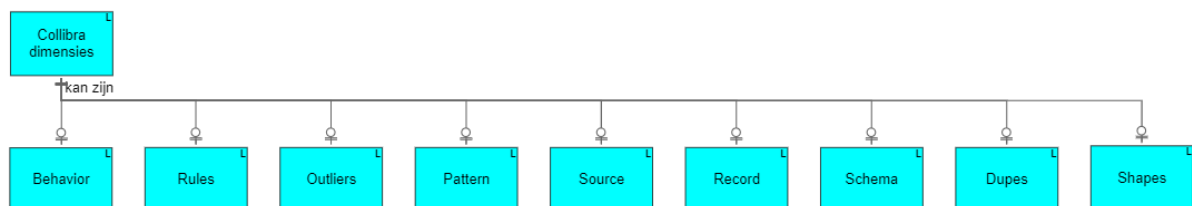
2.1. Contextdiagram DQ



2.2. Logisch datamodel



Collibra dimensies kunnen de volgende verschijningsvormen hebben:



3. Functionaliteiten

3.1. Het configureren van een DQ job

3.1.1. Dataset configureren

Een dataset (custom query) kan gebaseerd zijn op de inhoud van een hele tabel – waar Collibra verbinding mee heeft – of een in Athena gemaakte view of een los bestand. Hiervan kan in Collibra dan weer een selectie van de kolommen worden gemaakt. Dit gebeurt via het aanvinken van de kolommen, of door zelf een SQL-query te schrijven. Het is hierbij ook mogelijk om het aantal rijen te beperken en key-velden te bepalen.

3.1.2. Het configureren van een profile

In Collibra DQ is het mogelijk om een profile te maken, gebaseerd op een tabel, een view of een bestand. Hiervoor kan zowel de hele bron gebruikt worden, als een selectie van gefilterde rijen en kolommen. Het doel van een profile is het verkennen van de data in de bron.

Een profile kan naar de volgende opties kijken: null-waardes, lege velden, unieke waarden, de maximale waarde, de minimale waarde, het rekenkundige gemiddelde, het aantal rijen en laadtijd. Daarnaast kan er gekeken worden naar de relaties, semantische detectie, een analyse middels histogram, een profile pushdown, detectie van datavormen, en data shape granular,

3.1.3. Collibra dimensies configureren

Collibra hanteert 9 dimensies voor het meten van datakwaliteit. Om verwarring met de DAMA-dimensies (volledigheid, validiteit, uniciteit, nauwkeurigheid, tijdigheid en consistentie) te voorkomen, spreken we hier van de Collibra dimensies.



Van deze dimensies zijn alleen Pattern, Outlier, Dupes, Source en Rules zelf te configureren. De overige dimensies draaien automatisch.

Voor Pattern, Outlier, Dupes en Source bestaat de configuratie uit het aan- en uitzetten van deze dimensies, het selecteren van kolommen waar deze dimensies naar moeten kijken en het bepalen van de gevoeligheid. Voor Pattern en Outlier kan ook bepaald worden hoe veel dagen er terug moet worden gekeken. Voor de configuratie van rules, zie hieronder.

Door de Machine Learning die Colibra toepast, presenteert de tool onder “Shapes” zelf afwijkende waarden aan de gebruiker, die vervolgens gevalideerd dienen te worden. Eenmaal gevalideerd, zal dezelfde waarneming bij een volgende run niet meer als waarneming worden gepresenteerd. Zolang een waarneming niet gevalideerd is, heeft die een negatieve invloed op de score van de DQ job. Als mogelijke acties bij een waarneming heeft de gebruiker Validate, Invalidate en Resolve.

Actions

Validate: Assign Issue or Acknowledge (no assignee) this is a valid observation

Invalidate: Pass Value, adjust rule to be more lenient or enter manual tolerance

Resolved: Valid observation, fixed at the source, should not appear in future runs

3.2. Het maken en beheren van datakwaliteitsregels

Na het profileren van een dataset, en het verkennen van de resultaten van de profile, kunnen er rules worden toegevoegd aan de DQ job. Colibra DQ bevat van zichzelf al enkele standaard DQ rules. Een gebruiker kan die zelf ook toevoegen.

Het bouwen van een rule gebeurt in de query builder of in Spark SQL. In de query builder selecteert de gebruiker de elementen van de rule en schrijft Colibra DQ zelf de bijbehorende SQL-query. Colibra onderscheidt simpele, complexe en dataset-specifieke rules. Eventueel kunnen in een rule ook nog objecten aan elkaar gekoppeld worden middels een join.

Er wordt een onderscheid gemaakt tussen simpele, freeform SQL en “preset” regels:

Rule type	Description	Example
Simple rules	Simple rules are used when you want to filter a condition on a single column in a single table.	City = 'Baltimore'
Freeform SQL rules	Freeform SQL rules are used when you want to apply a condition across multiple tables/columns and generally when more flexibility or customization is desired.	select * from dataset where name = 'Colibra'
Preset rules	Preset rules are used for quickly adding strict condition checks. Commonly used conditions are available to add to any data set columns.	

Simpele regels bevatten alleen de conditie van een regel, die in freeform SQL na de “Where”-statement volgt.

Het is optioneel om een gewicht en een DQ-dimensie aan de regel toe te kennen. Met betrekking tot het gewicht kan Low, Medium of High of custom gekozen worden. De DQ-dimensies van DAMA zijn standaard beschikbaar, eventuele nieuwe DQ-dimensies kunnen nog toegevoegd worden.

Iedere regel heeft een eigen score. Geeft één van de regels een score van 0, dan is de score van de DQ job ook gelijk 0.

Datakwaliteitsregels kunnen generiek worden toegepast, maar ook specifiek zijn voor één dataset. Daarnaast kunnen stat rules worden toegepast op de metadata.

Alle standaard regels staan in de “rule library”. Hier kunnen ook custom regels aan worden toegevoegd.

3.3. Het exporteren van de resultaten van DQ jobs

Het is mogelijk om op meerdere plekken in Collibra DQ een export te draaien:

- Van het overzicht met alle DQ jobs.
- Van de volgende resultaten van een DQ job:
 - De rules.
 - De outliers.
 - De patterns.
 - De source.
 - De record.
 - De schema.
 - De dupes.
 - De shapes.
- Van de drilldown van incorrecte records.

Dit kan in CSV, Excel, PDF or door middel van een print-opdracht. De gebruiker kan zelf selecteren welke elementen er mee moeten worden genomen in de export.

3.4. Alerts

Het is mogelijk om gebruikers een e-mailbericht te sturen als aan bepaalde voorwaarden voldaan is voor een bepaalde dataset. Een voorbeeld hiervan is als de score van een regel onder de 75 zakt, of als een DQ job niet gedraaid kon worden. Standaard zijn alerts op scores en het aantal regels aanwezig:

Example Conditions	
Type	Condition
SCORE	score < 75
ROWCOUNT	rc = 0
RUN TIME	totalTimeInSeconds > 20
RUN TIME	totalTimeInHours > 20
RUN TIME	totalTimeInMinutes > 20
STALE	daysWithoutData >= 1
STALE	runsWithoutData > 2
STALE	daysSinceLastRun > 1
To use a rule condition as an alert	
Rule Name	Condition
Low_HighRule	Low_HighRule > 0 (must > 0 to ruleName)

Aanvullende alerts kunnen door de gebruiker geconfigureerd worden. Het bericht dat een gebruiker dan ontvangt, kan door de schrijver van het alert zelf geschreven worden.

Bij Batch Alerts gaat één bericht naar meerdere gebruikers.

3.5. Scheduling

Als een DQ job eenmaal gedraaid heeft, kan er scheduling aan toe worden gevoegd, waarmee de DQ job op bepaalde tijden met een bepaalde frequentie automatisch draait. Naast het kiezen van een tijdstip, kan ook bepaald worden of een DQ job dagelijks, maandelijks of per kwartaal draait. Tenslotte kan ook een selectie gemaakt worden van de dagen/maanden/kwartalen/het moment waarop de DQ job gedraaid moet worden.

Ook hier heeft de gebruiker de mogelijkheid om velden de (de)selecteren, of zelf een query te schrijven.

Scheduling van rapporten dient volgens de te maken afspraken gebruikt te worden. Het moet niet de bedoeling zijn dat resources verloren gaan aan het uitvoeren van de taken waardoor het systeem niet meer beschikbaar is voor “normaal” gebruik.

3.6. View and manage

3.6.1. Catalog

Op de catalog-pagina in Collibra DQ wordt automatisch een overzicht bijgehouden van de datasets die zich in Collibra DQ bevinden, met het type bron, de “parent folder”, de tabelnaam, eventuele meta tags, de gebruikte server en de file paths. Op dit scherm kan men filteren tussen de verschillende data categorieën (inc tags voor privacy-gevoelige data); de business units; de run mode; het type bron; het type data class; telling van de kolommen; telling van de rijen en het aantal DQ scans.

3.6.2. Scorecards

Onder scorecards is een algeheel overzicht van alle gedraaide scorecards te vinden. Hiermee is in één oogopslag te zien wat de historische kwaliteit van een dataset is alsmede de consistentie in het runnen van een DQ job.

3.6.3. Pulse view

In de Pulse view kan de gebruiker zien welke DQ jobs wanneer handmatig dan wel automatisch gedraaid zijn. Hiermee krijgt de gebruiker inzicht in de frequentie waarmee de DQ jobs draaien, maar ook of ze goed draaien door middel van de labels. Op dit scherm kan men filteren op connecties, gebruikers, en de frequentie van de scheduling. Hier wordt ook een algeheel cijfer gegevens aan de “gezondheid” van de DQ jobs die gedraaid worden.

3.6.4. List view

In de List view wordt een overzicht gegeven van alle DQ issues. Hierbij wordt per issue de dataset, het type issue, een omschrijving, aantal keer dat het issue is voorgekomen, de datum van ontdekking en de datum van een update. De gebruiker kan deze issues rangschikken, sorteren en opzoeken middels de filtering.

3.6.5. Workflow

Onder workflows is de assignment queue te vinden. Hier kunnen bevindingen uit gedraaide DQ jobs worden toegewezen aan gebruikers, om het issue op te lossen. Bij een internal assignment kan de bevinding alleen gelinkt worden aan een andere gebruiker van Collibra DQ. Bij een external assignment kan ook gelinkt worden aan iedere gebruiker met een Service Now account.

Voor het linken van een bevinding aan een interne gebruiker is het nodig dat de toewijzende gebruiker toegang heeft tot een data-set en de "train"-rol heeft. Voor het linken aan een externe gebruiker, moet men de admin-rol of de connection manager-rol hebben. Voor de gebruiker aan wie de bevinding wordt toegewezen, is het niet vereist dat ze die rol hebben.

De gebruiker aan wie de bevinding is toegewezen, kan de bevinding vervolgens valideren (de bevinding is inderdaad een bevinding), invalideren (de bevinding is geen bevinding en kan genegeerd worden), en oplossen (de bevinding is opgelost). Niet al deze acties zijn beschikbaar voor iedere dimensie in Collibra:

Feature	Available actions
Behaviors	Validate, Resolve
Rules	Validate, Resolve
Outliers	Validate, Invalidate, Resolve
Pattern	Validate, Invalidate, Resolve
Source	Validate, Invalidate, Resolve
Record	Validate, Resolve
Dupes	Validate, Invalidate, Resolve

Op dit moment is deze functionaliteit niet geconfigureerd.

3.7. Interfaces

3.7.1. Connectie Athena

Niet alle data die op het lake beschikbaar is, is ook automatisch in Collibra te vinden. Collibra heeft toestemming nodig, in de vorm van een technische policy, voordat het bij een bepaalde bron in het lake kan. Van de bronnen waar Collibra wel bij kan, wordt er alleen als er een query gedraaid wordt, op basis van die query een kopie van de data in de bron gemaakt en opgeslagen in Collibra DQ.

Als er iets verandert in de brondata, dan verandert er dus pas iets in Collibra als er door scheduling of handmatig een job gedraaid wordt, die een relatie heeft met de veranderde data.

Er wordt geen initial full load uit Athena gehaald, en dat is ook niet wenselijk. Hiermee zou het hele data lake naar Athena worden gehaald, waardoor Collibra DQ veel langzamer zou werken. Bovendien kan, zoals hierboven gezegd, Collibra op dit moment niet bij alle bronnen op het lake.

3.7.2. Connectie S3

Via de connectie met S3 kunnen flatfiles ingelezen worden. Hiervoor moet bijvoorbeeld een csv-bestand worden ingeladen in een S3-bucket. De ondersteunde file types hiervoor zijn CSV, ORC, Parquet, AVRO,

JSON, XML, DELTA en Other. Verschillende Character Sets worden ondersteund, waaronder UTF-8. Het is mogelijk om een bepaald aantal regels over te slaan, de header al dan niet mee te nemen, de delimiter te bepalen en het schema te by-passen. Daarnaast zijn er verschillende datum-formats beschikbaar. Bij het configureren van een bestand via de S3-connectie, is meteen al een data preview beschikbaar.

Na het inladen van het bestand, is het mogelijk om op dezelfde manier als bij de connectie met Athena de dataset en de DQ-job te configureren.

3.7.3. Connectie flatfiles

Het is ook mogelijk om een flatfile direct in Collibra in te lezen, die dan tijdelijk (tot middernacht) opgeslagen wordt. Deze optie is nu niet geconfigureerd. Flatfiles kunnen via de S3-connectie ingelezen worden. Het gaat hier om de functionaliteit om met een willekeurig S3 bucket te kunnen communiceren. Het platform heeft zelf geen eigen S3 Bucket.

3.7.4. Connectie naar PowerBI

Op dit moment is nog te weinig bekend over de connectie tussen Collibra DQ en PowerBI. Hier zal in een later stadium meer informatie over worden toegevoegd.

4. Databeheerprocessen

4.1. Rollen en rechten

De volgende rollen voor de gebruikers van Collibra DQ zijn geconfigureerd:

ID	PostNL	Role	Access Description
1	1) Admin	ROLE ADMIN	Modify any access, config settings, connections, role delegation
2	1) Admin 2) DQ Analyst	ROLE DATA GOVERNANCE MANAGER	Ability to manage (create / update / delete) Business Units and Data Concepts
3	1) Admin 2) DQ Analyst	ROLE USER MANAGER	Create / modify users, add users to roles
4	1) Admin 2) DQ Analyst	ROLE COLLIBRA ROLE MANAGER	Create roles, edit role mappings to users / AD groups / datasets
5	1) Admin 2) DQ Analyst 3) DMO	ROLE DATASET MANAGER	Create / modify datasets to roles, masking of dataset columns
6	1) Admin 2) DQ Analyst 3) DMO	ROLE COLLIBRA CHECK	Only role that can run DQ scans if Owlcheck security is enabled
7	1) Admin 2) DQ Analyst 3) DMO	ROLE DATASET TRAIN	Only role that can train datasets if dataset train security is enabled
8	1) Admin 2) DQ Analyst 3) DMO	ROLE DATASET RULES	Only role that can add / edit / delete rules if dataset rules security is enabled
9	1) Admin 2) DQ Analyst	ROLE VIEW DATA	Controls which users can access the DQ SQL editor to run the SQL against the database
10	1) Admin 2) DQ Analyst 3) DMO	ROLE DATA PREVIEW	Only role that can view source data if data preview security is enabled
11	1) Admin 2) DQ Analyst 3) DMO 4) Viewer	ROLE PUBLIC	Public: Access to scorecards, no dataset access when dataset security is enabled

Bron: [link](#).

Deze rollen zullen nog gewijzigd worden in Q2, omdat er voor iedere DMO een eigen rol zal komen. Op die manier kunnen we toegang tot bepaalde data afschermen voor alleen de leden van die DMO, die die data in mag zien.

In onderstaande autorisatiematrix staat in detail beschreven welke rol welke rechten heeft per functionaliteit. De rechten hebben betrekking op wat een gebruiker kan doen met data objecten die onder een gegeven functionaliteit vallen.

De rechten, en hoe ze geduid zijn in de matrix:

- Create-rechten (C)
- Read-rechten (R)
- Update-rechten (U)
- Delete-rechten (D)

Functionaliteit↓/Rol→	Admin	EDM	DMO*	Viewer
Access management	CRUD			
Connecties	CRUD			
Instellingen configureren	CRUD			
Gebruikers	CRUD	R		
Rollen	CRUD	R		
Business Units	CRUD	CRUD		
Dataset	CRUD	CRUD	CRUD**	
DQ jobs	CRUD	CRUD	CRUD**	
DQ rules	CRUD	CRUD	CRUD**	R
Dashboard	CRUD	CRUD	CRUD**	R
Export	CRUD	CRUD	CRUD**	Alleen van resultaten, niet van data.
Migratie A naar P	CRUD	?	?	

*Onder de rol DMO vallen zowel de Data Manager als de Data Steward.

**Alleen voor eigen domein/datasets.

4.2. Eigenaarschap data en systeem

4.2.1. Data eigenaarschap

De verantwoordelijkheid voor het (laten) bouwen en beheren van datakwaliteitsrapportages ligt bij de betreffende data-eigenaar (bijvoorbeeld data manager of platform owner). Waar het gaat om overkoepelende dashboards, zoals bijvoorbeeld ten behoeve van de DGB, EDM, of het Digital Transformation Dashboard, ligt de verantwoordelijkheid bij EDM.

4.2.2. Systeem eigenaarschap

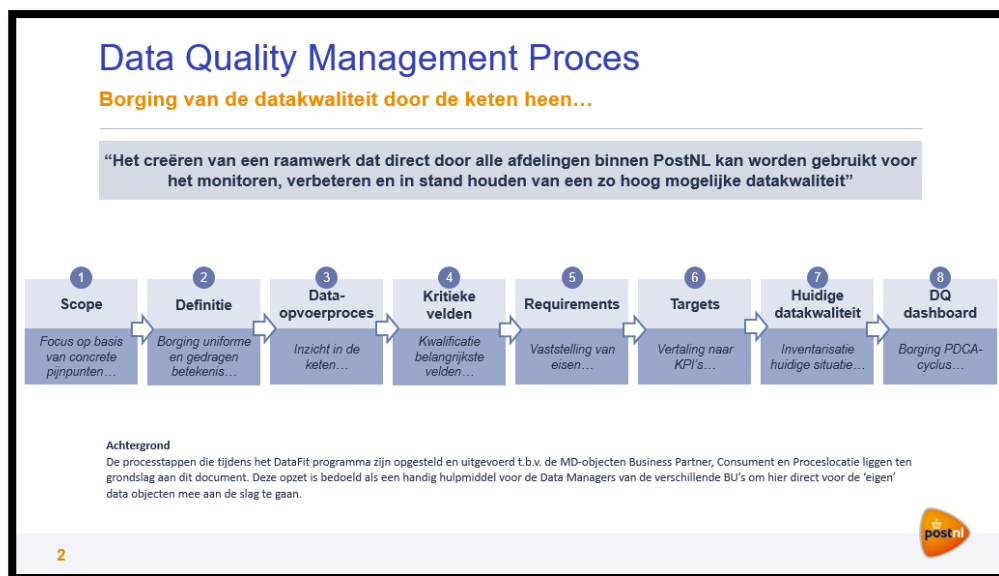
Het (technische) beheer van het systeem wordt gedaan door de Solution Consultant van het DMP Platform.

4.3. Datakwaliteitsmaatregelen

Het volgende bevat maatregelen om ervoor te zorgen dat de kwaliteit van de vulling in Collibra hoog te houden.

4.3.1. Bepalen van de randvoorwaarden

Datakwaliteitsregels kunnen het resultaat zijn van het doorlopen van het Data Quality Management Proces:



In het kader van het DQM-proces worden een aantal stappen doorlopen om op een gedegen wijze tot een specifieke set DQ regels te komen. Het is immers niet de bedoeling dat Collibra DQ gevuld gaat worden met alle mogelijke DQ-regels op alle velden binnen de organisatie, maar om waarde toe te voegen met wat gemeten wordt.

Het is ook mogelijk om ad-hoc analyses te doen, waarbij niet het hele DQM-proces wordt doorlopen.

Wel is het een voorwaarde van het maken van een nieuwe rapportage, dat de eigenaar bekend is, dat er regels bekend zijn, dat bekend is wie de kennis-houder van de data is, en dat er afspraken zijn gemaakt over het bouwen en beheren van de rapportage.

4.3.2. Vastleggen rollen, aanleverspecificaties en werkinstructies

Vastgelegd dient te worden wie (rol/functie + afdeling) de data aanlevert, in welke frequentie, in welk formaat, op welke locatie, en op welke manier.

Middels werkinstructies voor de bouwers wordt erop gestuurd dat handmatige aanpassingen van de rapportages op uniforme wijze gebeurt.

4.3.3. Vaststellen rapportagebehoefte

Om er op structurele basis voor te zorgen dat de rapportages goed worden vormgegeven, dient te worden bepaald welke kwaliteitscontroles plaatsvinden. Hierbij kan naar één of meer van de volgende momenten worden gekeken:

Initieel bouwen van de rapportages

Voordat de rapportages initieel in gebruik wordt genomen, dient te worden gecontroleerd of de regels succesvol worden toegepast. Hierbij dient zowel naar de datakwaliteitsindicatoren te worden gekeken, als naar het verwachte aantal (foute) records.

Periodiek inlezen van de data

Zodra we een update vanuit de bron op de data ontvangen, dient te worden gecontroleerd of de rapportages aan de gestelde randvoorwaarden voldoen.

Periodiek muteren van de data

Zodra we zelf een update op de rapportages hebben doorgevoerd, dient te worden gecontroleerd of deze mutaties aan de gestelde randvoorwaarden voldoen, en of ook hier geen fouten zijn gemaakt.

Exporteren van de data

Voordat de rapportage daadwerkelijk beschikbaar wordt gesteld aan de gebruikers, dient te worden gecontroleerd of de export succesvol is uitgevoerd. Ook hiervoor geldt dat zowel naar de genoemde datakwaliteitsindicatoren dient te worden gekeken, als naar het verwachte aantal (foute) records.

4.3.4. Opstellen van de rapportages

Op basis van de onder 4.4.3 vastgestelde behoeften dienen de rapportages i.s.m. de data manager, ondersteund door EDM opgesteld te worden. Hierbij moeten de datakwaliteitsregels zorgvuldig opgesteld worden: de focus ligt bij de kritische data-velden en logische regels. Er wordt bijvoorbeeld niet gecheckt op de volledigheid van velden waarvan de vulling niet verplicht is.

4.3.5. Periodiek toetsen van de voorgaande stappen

De hiervoor genoemde stappen dienen periodiek te worden getoetst:

- Zijn de initieel vastgestelde randvoorwaarden nog steeds van toepassing?
- Kloppen de rollen, aanleverspecificaties en werkinstructies nog?
- Is de rapportagebehoefte nog steeds hetzelfde?
- Zijn de nu beschikbare rapportages nog actueel en correct?

Beschreven dient te worden in welke frequentie en door wie deze vragen periodiek getoetst dienen te worden..

5. Bijlagen

5.1. Technische inrichting Collibra

In dit hoofdstuk worden de Collibra tabellen en hun omschrijving genoemd, gevolgd door een tabel waarin de bijbehorende velden worden toegelicht.

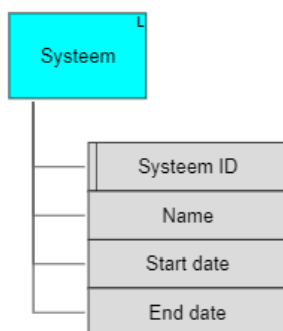
Omdat dit een omschrijving betreft van wat er al in Collibra DQ zit, en het geen instructie voor een bouwteam betreft, zijn enkele velden leeg gelaten.

Attribuutregel kolom Type: Staat voor gegevenstype en kan de volgende waarden hebben:

Waarde / Value	Omschrijving	Description
A	Veld dat letters en cijfers kan bevatten	Field that can contain letters and numbers
A (e)	Veld dat letters en cijfers kan bevatten met een keuzelijst aan opties (enumeration)	Field that can contain letters and numbers with a dropdown choice list (enumeration)
AM	Als A maar in meer talen (komt niet voor in database)	Same as A but in more languages (does not exist in database)
N	Veld dat alleen cijfers kan bevatten	Field that can only contain numbers
Rel	Relatie naar andere entiteit (veld bevat verwijzing naar die andere entiteit)	Relationship to other entities (field contains reference to that other entity)
D	Veld bevat een aanduiding voor een dag (Datum)	Field contains a day indication (Date)
DT	Veld bevat een aanduiding voor een dag (Datum) en een moment op die dag (Tijdstip)	Field contains an indication for a day (Date) and a moment on that day (Time)
T	Veld bevat een aanduiding van een moment binnen een niet nader bepaalde dag (Tijdstip)	Field contains an indication of a moment within an unspecified day (Time)
B	Veld bevat een boolean waarde 0 1 of J N of T F	Field contains a boolean value 0 1 or Y N or T F
C	Veld bevat een optie lijst	Field contains option list

5.1.1. Systeem¹

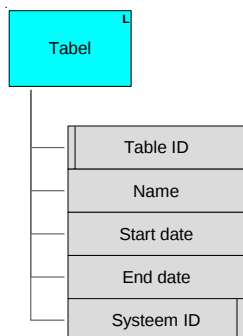
Definitie: Het systeem is de applicatie die de bron is voor de data.



Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Systeem ID	Systeem ID		PK		N	N
Name	Name				A	N
Start Date	Start Date				D	N
End Date	End Date				D	J

5.1.2. Tabel

Definitie: Een tabel is een eenheid van opslag van gegevens.



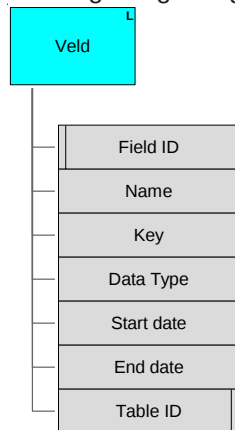
Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Object ID	Object ID		PK		N	N
Name	Name				A	N
Start Date	Start Date				D	N

¹ De tabellen en velden van Collibra DQ zijn gemodelleerd in het sjabloon van een logisch datamodel. Ook de kolomnamen zijn daarop geïnspireerd. Het blauwe blokje met L is dus geen logische entiteit maar een tabel; Attribuutnaam staat voor veldnaam.

End Date	End Date				D	J
Systeem ID	Systeem ID		FK		N	N

5.1.3. Veld

Een veld geeft betekenis aan een tabel. De eigenschappen van een veld beschrijven de kenmerken en het gedrag van gegevens die aan dat veld zijn toegevoegd.



Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Attribute ID	Attribute ID		PK		N	N
Name	Name				A	N
Key	Key				B	N
Data Type	Data Type				C	N
Start Date	Start Date				N	N
End Date	End Date				D	J
Object ID	Object ID		FK		N	N

5.1.4. Data Set

Definitie: Een dataset (custom query) is een selectie van data uit één of meerdere bronnen, welke als view opgeslagen kan worden in Athena, of direct in Collibra geconfigureerd kan worden.

Data Set

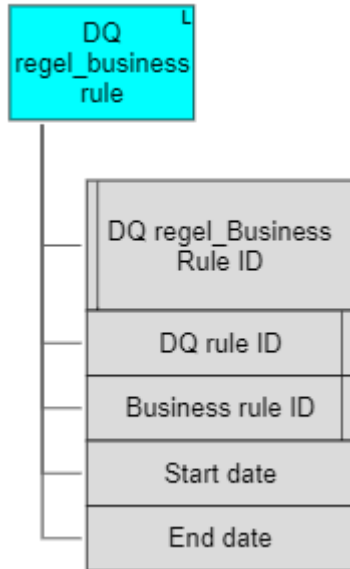
Data Set ID
Datum aangemaakt
Column
Datum gewijzigd
Link ID
Data Type
Pick Column
Comparison Operator
Range date
IsDate
Group By
Transform
Transform Query
Limit (Rows)
Escape Character
Query
Object ID
Attribuut ID
Start date
End date

Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
---------------	--------------	--------------	-----	-----	------	-----------

Data Set ID	Data Set ID		PK		N	N
Add link back to source	Add link back to source				B	N
Column	Column				C	N
Key	Key				B	N
Link ID	Link ID				B	J
Data Type	Data Type				C	N
Pick Column	Pick Column				C	J
Comparison Operator	Comparison Operator				C	J
Range Date	Range Date				D	J
IsDate	IsDate				B	J
Group By	Group By				C	J
Transform	Transform				B	J
Transform query	Transform query				A	J
Limit (Rows)	Limit (Rows)				N	J
Escape Character	Escape Character				C	J
Query	Query				A	N
Start Date	Start Date				D	N
End Date	End Date				D	J

5.1.5. DQ regel_Business rule

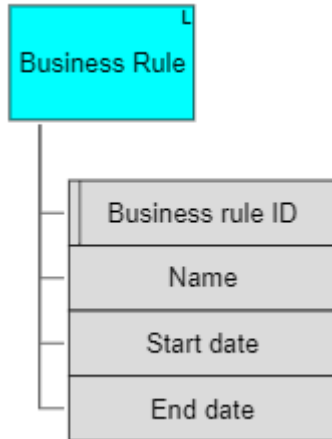
Definitie: Deze tabel is de koppeltabel tussen de tabellen met business rules en datakwaliteitsregels.



Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
DQ regel_Business Rule ID	DQ regel_Business Rule ID		PK		N	N
DQ regel ID	DQ regel ID		FK		N	N
Business rule ID	Business rule ID		FK		N	N
Start Date	Start Date				D	N
End Date	End Date				D	J

5.1.6. Business rule

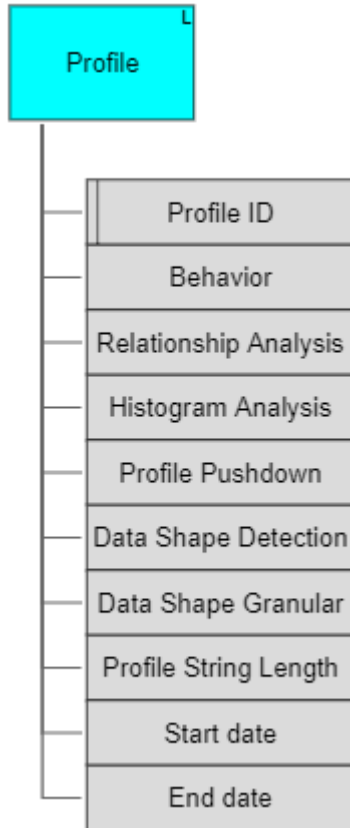
Definitie: Een business rule is een regel waaraan de data dient te voldoen. Hieruit kunnen één of meerdere datakwaliteitsregels worden afgeleid. Een business rule kan ook als technische restrictie op een veld geconfigureerd zijn.



Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Business Rule ID	Business Rule ID		PK		N	N
Name	Name				A	N
Start Date	Start Date				D	N
End Date	End Date				D	J

5.1.7. Profile

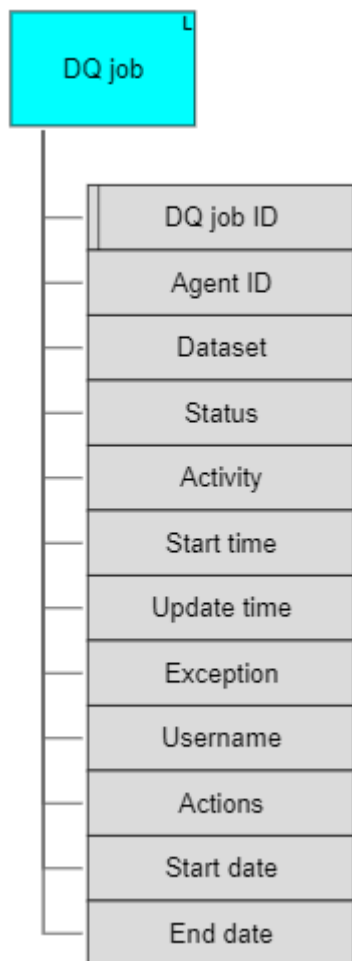
Definitie: Data profiling behelst een objectieve analyse van de data. Het resultaat is een overzicht met daarin het datatype, welke patronen er te vinden zijn in een kolom, en hoe vaak die terugkomen.



Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Profile ID	Profile ID		PK		N	N
Behavior	Behavior				B	N
Relationship Analysis	Relationship Analysis				C	N
Histogram Analysis	Histogram Analysis				C	N
Profile Pushdown	Profile Pushdown				C	N
Data Shape Detection	Data Shape Detection				C	N
Data Shape Granular	Data Shape Granular				C	N
Profile String Length	Profile String Length				B	N
Start Date	Start Date				D	N
End Date	End Date				D	J

5.1.8. DQ job

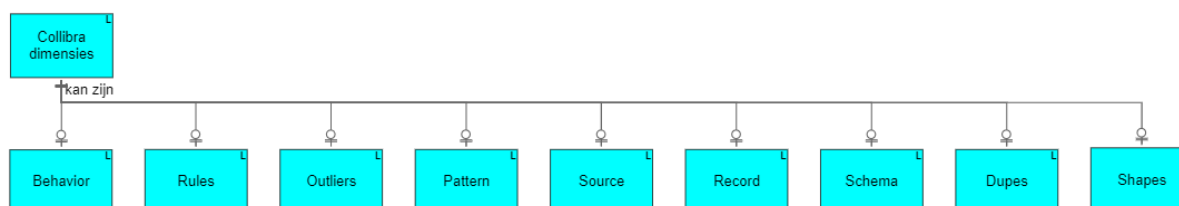
Definitie: Een DQ job in Collibra is een collectie van de profile, de Collibra dimensies van DQ, en de DQ rules die tezamen een score geven over de kwaliteit van de dataset.



Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
DQ Job ID	DQ Job ID		PK		N	N
Agent ID	Agent ID				N	N
Dataset	Dataset				A	N
Status	Status				C	N
Activity	Activity				?	J
Start Time	Start Time				D	N
Update Time	Update Time				D	N
Exception	Exception				A	J
Username	Username				C	N

Actions	Actions				C	J
Start date	Start date				D	N
End date	End date				D	J

5.1.9. Collibra dimensies



5.1.9.1 Behavior

Definitie: behavior is de waarneming van veranderingen in de data.

Behavior

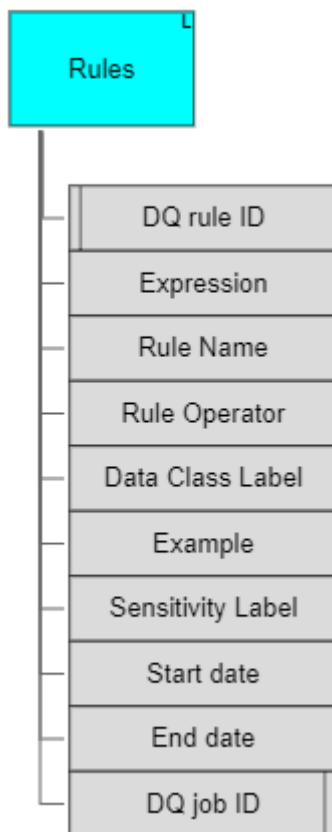
Behavior ID
Run time
Blind spot
Type
Baseline
Today
Percent change
Delta percent change
Zscore
Description
Score
Action
Status
Start date
End date
DQ job ID

Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Behavior ID	Behavior ID		PK			
Runtime	Runtime					
Blind spot	Blind spot					
Type	Type					
Baseline	Baseline					
Today	Today					
% Change	% Change					

Delta % Change	Delta % Change					
Zscore	Zscore					
Description	Description					
Score	Score					
Action	Action					
Status	Status					
DQ Job ID	DQ Job ID		FK			

5.1.9.2 Rules/Datakwaliteitsregel

Definitie: Een datakwaliteitsregel is een stelling over de data, aan de hand waarvan de data als goed of slecht gekwalificeerd kan worden.



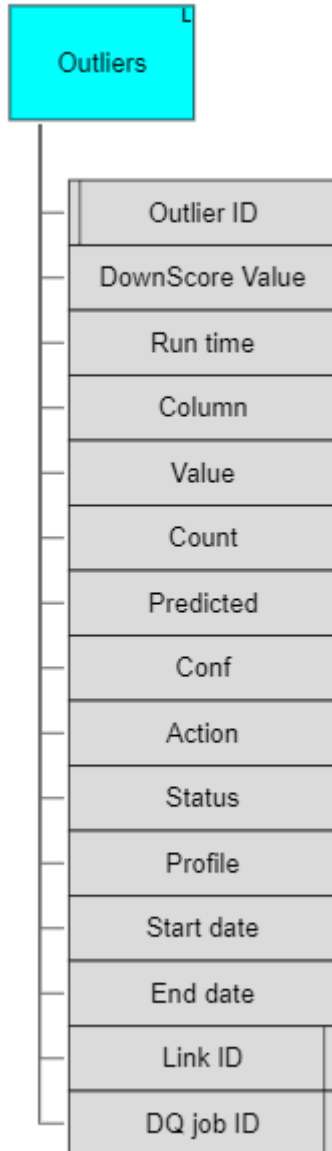
Attributen nog modelleren na check

Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
DQ rule ID	DQ rule ID		PK		N	N
Data Set Name	Data Set Name				A(e)	N
DQ rule type	DQ rule type				C	N

Column	Column				C	N
Expression	Expression				A	N
Rule Name	Rule Name				A	N
DQ dimension	DQ dimension				C	J
Category	Category				A	J
Description	Description				A	J
Preview Limit	Preview Limit				N	J
Run Time Limit	Run Time Limit				N	J
Score	Score				C	N
Points	Points				N	N
Percent	Percent				N	N
Generic Rule Type	Generic Rule Type				C	N
Is Regex	Is Regex				B	N
Rule Operator	Rule Operator				C	N
Where	Where				A	N
Data Class Label	Data Class Label				A	N
Example 1	Example 1				A	J
Example 2	Example 2				A	J
Example 3	Example 3				A	J
Sensitivity label	Sensitivity label				C	J
Column Type	Column Type				C	J
Column % Match	Column % Match				N	N
Allow Null	Allow Null				B	N
Allow Empty	Allow Empty				B	N
Start Date	Start Date				D	N
End Date	End Date				D	J
DQ Job ID	DQ Job ID		FK			

5.1.9.3 Outliers

Definitie: Outliers zijn records die a-typisch zijn in een kolom, die buiten de “upper” en “lower bound” limieten van een groep data vallen.

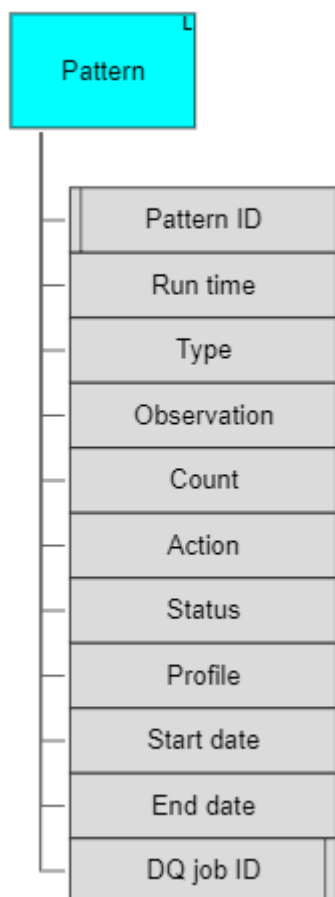


Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
DownScore Value	DownScore Value					
Runtime	Runtime					
Key	Key		PK			
Column	Column					
Value	Value					
Count	Count					
Predicted	Predicted					
Conf	Conf					

Action	Action					
Status	Status					
Profile	Profile					
Start Date	Start Date			D	N	
End Date	End Date			D	J	
Link ID	Link ID		FK			
DQ Job ID	DQ Job ID		FK			

5.1.9.4 Pattern

Definitie: een pattern is een relatie tussen gerelateerde kolommen, en geeft waarden weer die of historisch of relationeel onverwachts zijn.

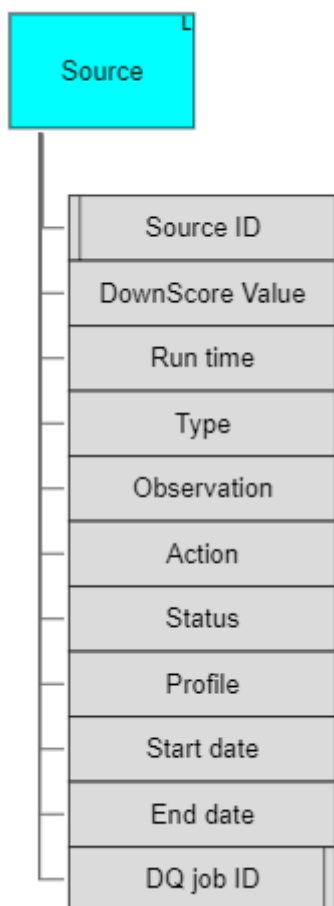


Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Pattern ID	Pattern ID		PK			
Runtime	Runtime					
Type	Type					

Observation	Observation					
Count	Count					
Action	Action					
Status	Status					
Profile	Profile					
Start Date	Start Date			D	N	
End Date	End Date			D	J	
DQ Job ID	DQ Job ID		FK			

5.1.9.5 Source

Definitie: de source geeft weer of een record consistent voorkomt in twee verschillende bronnen.

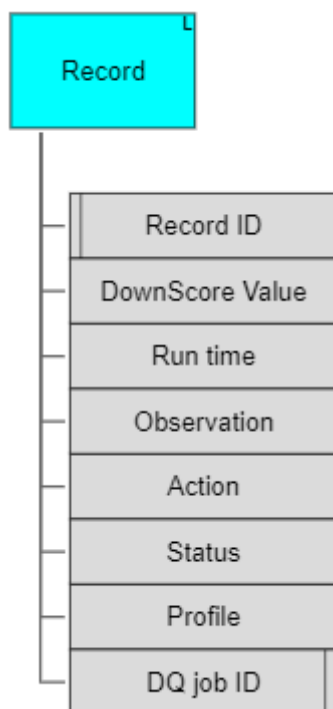


Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Source ID	Source ID		PK			
DownScore Value	DownScore Value					
Runtime	Runtime					

Type	Type					
Observation	Observation					
Action	Action					
Status	Status					
Profile	Profile					
Start Date	Start Date			D	N	
End Date	End Date			D	J	
DQ Job ID	DQ Job ID		FK			

5.1.9.6 Record

Definitie: de record checkt op dezelfde key records dagelijks terugkomen in de data.

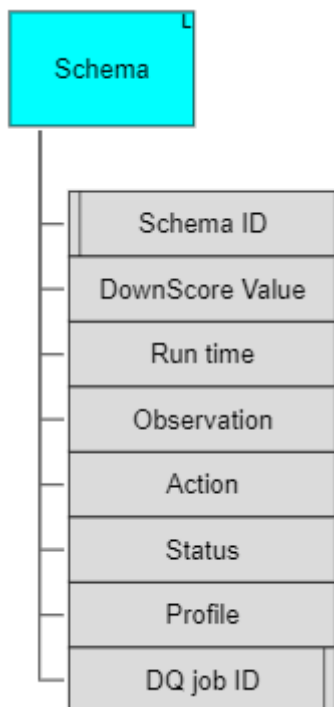


Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Record ID	Record ID		PK			
DownScore Value	DownScore Value					
Runtime	Runtime					
Observation	Observation					
Action	Action					
Status	Status					
Profile	Profile					

DQ Job ID	DQ Job ID		FK			
-----------	-----------	--	----	--	--	--

5.1.9.7 Schema

Definitie: de schema checkt of de tabelstructuur en kolomnamen onveranderd zijn.

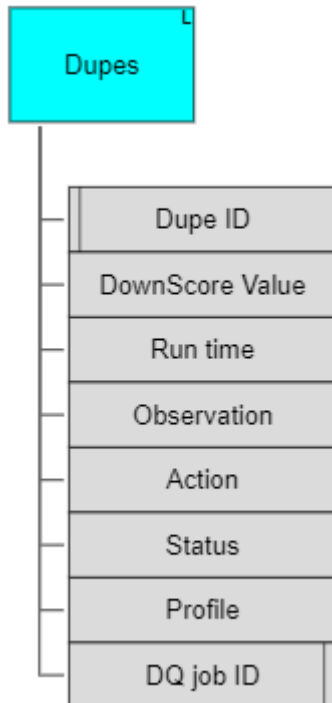


Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Schema ID	Schema ID		PK			
DownScore Value	DownScore Value					
Runtime	Runtime					
Observation	Observation					
Action	Action					
Status	Status					
Profile	Profile					
DQ Job ID	DQ Job ID		FK			

5.1.9.8 Dupes

Definitie: de dupe checkt voor dubbelingen in de data.

Nader in te vullen i.v.m. geen data.



Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Dupe ID	Dupe ID		PK			
DownScore Value	DownScore Value					
Runtime	Runtime					
Observation	Observation					
Action	Action					
Status	Status					
Profile	Profile					
DQ Job ID	DQ Job ID		FK			

5.1.9.9 Shapes

Definitie: de shape checkt of het format van de data in string-kolommen juist is.

Shapes

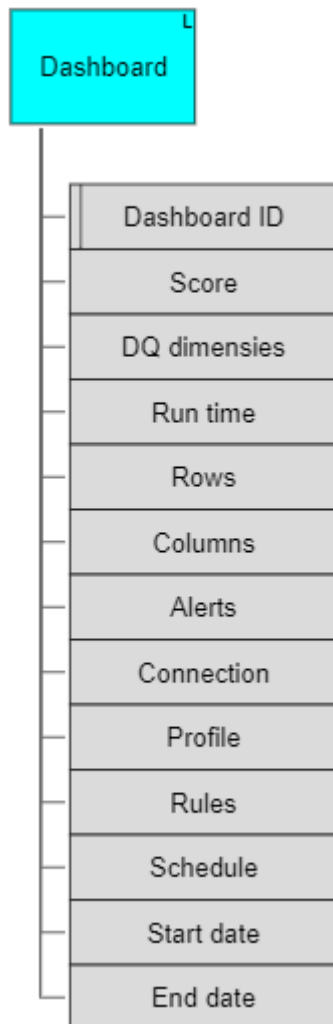
Shape ID
DownScore Value
Run time
Column
Schema
Shape
Count
Rowcount
Percent
Shapes/Col
Action
Status
Profile
DQ job ID

Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Shape ID	Shape ID		PK			
DownScore Value	DownScore Value					
Runtime	Runtime					
Column	Column					
Schema	Schema					
Shape	Shape					
Count	Count					
Rowcount	Rowcount					
Percent	Percent					

Shapes/Col	Shapes/Col					
Action	Action		C			
Status	Status					
Profile	Profile					
DQ Job ID	DQ Job ID		FK			

5.1.10. Dashboard

Definitie: een dashboard is een grafische weergave van de scores van een DQ job.

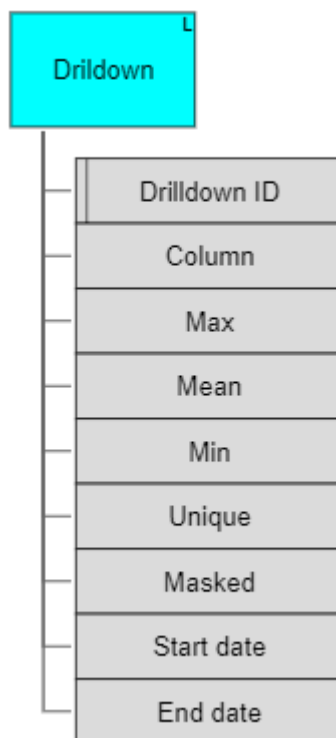


Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Dashboard ID	Dashboard ID		PK		N	N
Score	Score				N	N
DQ dimensies	DQ dimensies				A	N

Run Time	Run Time				N	N
Rows	Rows				N	N
Columns	Columns				N	N
Alerts	Alerts				N	N
Connection	Connection				A	N
Profile	Profile				A	N
Rules	Rules				A	N
Schedule	Schedule				A	N
Start date	Start date				D	N
End date	End date				D	J

5.1.11. Drilldown

Definitie: De drilldown geeft een lijst van incorrecte records die niet aan een datakwaliteitsregel voldoen.



Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Drilldown ID	Drilldown ID		PK		N	N
Column	Column				A	N
Max	Max				A	N
Mean	Mean				A	N

Min	Min				A	N
Unique	Unique				N	N
Masked	Masked				B	N
Start date	Start date				D	N
End date	End date				D	J

5.1.12. Report

Definitie: Een document met informatie die is georganiseerd in een verhalende, grafische of tabelvorm. Rapporten kunnen betrekking hebben op specifieke perioden, gebeurtenissen, voorvallen of onderwerpen.

Attributen nog modelleren na check

Attribuutnaam	Engelse naam	Omschrijving	Key	Len	Type	Optioneel
Report ID	Report ID		PK		N	N
Report Type	Report Type				C	N
Data Set	Data Set				A	N
Score	Score				N	N
Score Trend	Score Trend				?	N
Field Name	Field Name				A	N
Row Count	Row Count				N	N
Rows Trend	Rows Trend				?	N
Pass/Fail	Pass/Fail				N	N
Passing Trend	Passing Trend				?	N
Null	Null				N	N
Empty	Empty				N	N
Data Type	Data Type				N	N
Distribution	Distribution				N	N
Schema Evolution	Schema Evolution				N	N
Custom Rule	Custom Rule				N	N
Row Count	Row Count				N	N
Total	Total				N	N
Check Type	Check Type				C	N
Check Value	Check Value				C	N
Break Value	Break Value				N	N