|  |
| --- |
| Beroepsproduct DDDQ  Casus top 2000 |

Geo Bouwmeester, Neo Hop & Julian van Zwol

13 maart 2025

INHOUDSOPGAVE

[INLEIDING 3](#_Toc192803745)

[1 Onderzoek Datakwaliteit 4](#_Toc192803746)

[1.1 Duiding databasestructuur 4](#_Toc192803747)

[1.2 Accuracy 4](#_Toc192803748)

[1.2.1 Controle op uniforme schrijfwijze 4](#_Toc192803749)

[1.3 Completeness 5](#_Toc192803750)

[1.3.1 Controle op schending van referentiele integriteit 5](#_Toc192803751)

[1.3.2 Controle op column completeness 5](#_Toc192803752)

[1.4 Consistency 6](#_Toc192803753)

[1.4.1 Controle op schending primary key 6](#_Toc192803754)

[1.4.2 Controle op derde normaalvorm (3NF) 7](#_Toc192803755)

[1.4.3 Controle op integriteitsregels 7](#_Toc192803756)

[1.5 Schema kwaliteit 7](#_Toc192803757)

[1.6 Data opschonen 7](#_Toc192803758)

[2 verbeterde versie 7](#_Toc192803759)

[2.1 Verwoorden van feiten 7](#_Toc192803760)

[2.2 Sorteren verwoordingen 7](#_Toc192803761)

[2.3 Analyse van verwoordingen 7](#_Toc192803762)

[2.4 CDM 7](#_Toc192803763)

[3 Lijst van figuren 7](#_Toc192803764)

[Bijlage A. 7](#_Toc192803765)

INLEIDING

[Klik hier als je tekst wilt invoeren, maak gebruik van de stijlen op de tab MIJN DOCUMENT voor het typen van (genummerde) koppen.]

# Onderzoek Datakwaliteit

## Duiding databasestructuur

[Klik hier als je tekst wilt invoeren.]

Afbeelding met tekst, schermopname, Lettertype, nummer

Door AI gegenereerde inhoud is mogelijk onjuist.

Figuur : Gegenereerde diagram vanuit SSMS o.b.v. originele database.

## Accuracy

### Controle op uniforme schrijfwijze

Uitvoering van de SQL select query hieronder resulteerde in **226** "unieke" records. Echter hebben sommige genres verschillende schrijfwijzen en daarmee dus geen uniforme manier om genres weer te geven.

-- Geen uniforme schrijfwijze

**SELECT** **DISTINCT** genre

**FROM** **[**Top2000**].[**dbo**].[**SongGenre**]**

Enkele voorbeelden van alternatieve schrijfwijzen terwijl hetzelfde bedoeld wordt staan hieronder:

|  |  |  |
| --- | --- | --- |
| Optie 1 | Optie 2 | Optie 3 |
| alternatieve rock | alternatieve rock | - |
| elektronisch | elektronische muziek | elektronische muziek |
| hard rock | hardrock | - |
| Keltisch | Keltische muziek | - |
| kerstlied | kerstmuziek | - |
| klassiek | klassieke muziek | - |
| poppunk | popunk | - |
| psychedelic rock | psychedelische rock | - |
| rock-'n-roll | rock-'n-roll | - |

## Completeness

### Controle op schending van referentiele integriteit

Uitvoering van de volgende SQL select query resulteerde in schending van de referentiële integriteit op **122** records:

-- Check on Top2000Lijst's (logische) referentie naar Song

**SELECT** Artiest**,** Titel

**FROM** Top2000Lijst **as** list

**WHERE** **NOT** **EXISTS** **(**

**SELECT** **\***

**FROM** Song **as** song

**WHERE** list**.**Artiest **=** song**.**Artiest

**AND** list**.**Titel **=** song**.**titel

**)**

Uitvoering van de volgende SQL select query resulteerde in schending van de referentiële integriteit op **7** records:

-- Check on SongGenre's (logische) referentie naar Song

**SELECT** artiest**,** titel

**FROM** SongGenre **as** genre

**WHERE** **NOT** **EXISTS** **(**

**SELECT** **\***

**FROM** Song **as** song

**WHERE** genre**.**Artiest **=** song**.**Artiest

**AND** genre**.**Titel **=** song**.**titel

**)**

### Controle op column completeness

Uitvoering van de volgende SQL select query resulteerde in een percentage van **96.00%** van nummers in de Top2000 lijst van 2019 waarvan voor elk gegeven attribuut een waarde beschikbaar is:

-- Check on value for every given property for songs in the Top2000 list in 2019.

**SELECT** **(COUNT(\*)** **\*** 100.0 **/** **(SELECT** **COUNT(\*)** **FROM** Top2000Lijst **WHERE** editiejaar **=** '2019'**))** **as** completeness\_percentage

**FROM** **(**

**SELECT** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

song**.**Speelduur**,**

song**.**Componist**,**

list**.**editiejaar**,**

**COUNT(**genre**.**genre**)** **AS** number\_of\_genres

**FROM** Top2000Lijst **AS** list

**INNER** **JOIN** Song **AS** song

**ON** list**.**Artiest **=** song**.**Artiest

**AND** list**.**Titel **=** song**.**titel

**LEFT** **JOIN** SongGenre **as** genre

**ON** genre**.**artiest **=** song**.**Artiest

**AND** genre**.**titel **=** song**.**titel

**WHERE** list**.**editiejaar **=** '2019'

**GROUP** **BY** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

song**.**Speelduur**,**

song**.**Componist**,**

list**.**editiejaar**,**

genre**.**artiest**,**

genre**.**titel

**HAVING** **COUNT(**genre**.**genre**)** **>=** 1

**AND** song**.**Componist **IS** **NOT** **NULL**

**AND** song**.**Speelduur **IS** **NOT** **NULL**

**)** **as** Dataset

De select queries in Bijlage A zijn gebruikt om de controle uit te voeren per attribuut. De resultaten staan boven de queries in de comments.

## Consistency

### Controle op schending primary key

Voor de controle op schending van de (logische) primary keys is gebruikt gemaakt van een query die records binnen de betreffende tabel groepeert op de (logische) primary key en vervolgens controleert of hier duplicaten tussen zitten.

Uitvoering van de volgende SQL select query resulteerde in geen schending van de primary key:

-- Check op tabel Top2000Lijst.

**SELECT** editiejaar**,** positie

**FROM** Top2000Lijst

**GROUP** **BY** editiejaar**,** positie

**HAVING** **COUNT(\*)** **<>** 1

Uitvoering van de volgende SQL select query resulteerde in geen schending van de logische

primary key:

-- Check op tabel Song.

**SELECT** Artiest**,** titel

**FROM** Song

**GROUP** **BY** Artiest**,** titel

**HAVING** **COUNT(\*)** **>** 1

Uitvoering van de volgende SQL select query resulteerde in geen schending van de logische

primary key:

-- Check op tabel SongGenre.

**SELECT** artiest**,** titel**,** genre

**FROM** SongGenre

**GROUP** **BY** artiest**,** titel**,** genre

**HAVING** **COUNT(\*)** **>** 1

### Controle op derde normaalvorm (3NF)

Volgens de theorie is een tabel in Eerste Normaalvorm (1NF) als elke cel een atomaire waarde bevat. Elke cel voor alle kolommen voor elke tabel bevat een redelijke atomaire (= ondeelbare) waarde, behalve de kolom 'Componist' in tabel 'Song'. Deze bevat een door komma's gescheiden lijst met waarden die de Eerste Normaalvorm schendt. Dit betekent dat de tabel niet genormaliseerd is en zeker niet in Derde Normaalvorm.

### Controle op integriteitsregels

*IR1. Een song staat maximaal één keer in een editie van de Top 2000.*

Uitvoering van de volgende SQL select query resulteerde in schending van de integriteitregel op **2** records:

-- IR1. Een song staat maximaal een keer in een editie van de Top 2000.

**SELECT** editiejaar**,** Artiest**,** Titel**,** **COUNT(**1**)**

**FROM** Top2000Lijst

**GROUP** **BY** editiejaar**,** Artiest**,** Titel

**HAVING** **COUNT(\*)** **>** 1

*IR2. Elke song die in een editie van de Top 2000 staat, is vóór of ìn het jaar van de betreffende editie uitgebracht.*

Uitvoering van de volgende SQL select query resulteerde in geen schending van de integriteitregel:

-- IR2. Elke song die in een editie van de Top 2000 staat, is voor of in het jaar van de betreffende editie uitgebracht.

**SELECT** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

list**.**editiejaar

**FROM** Top2000Lijst **AS** list

**INNER** **JOIN** Song **AS** song

**ON** list**.**Artiest **=** song**.**Artiest

**AND** list**.**Titel **=** song**.**titel

**WHERE** song**.**Jaar **>** list**.**editiejaar

## Schema kwaliteit

Naamgeving

Engels nederlands

## Data opschonen

# verbeterde versie

## Verwoorden van feiten

## Sorteren verwoordingen

## Analyse van verwoordingen

## CDM

# Lijst van figuren

[Figuur 1: Gegenereerde diagram vanuit SSMS o.b.v. originele database. 4](#_Toc192798876)



-- 100% of the songs in the Top2000 list in 2019 has a value for duration

**SELECT** **(COUNT(\*)** **\*** 100.0 **/** **(SELECT** **COUNT(\*)** **FROM** Top2000Lijst **WHERE** editiejaar **=** '2019'**))** **as** completeness\_percentage

**FROM** **(**

**SELECT** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

song**.**Speelduur**,**

song**.**Componist**,**

list**.**editiejaar**,**

**COUNT(**genre**.**genre**)** **AS** number\_of\_genres

**FROM** Top2000Lijst **AS** list

**INNER** **JOIN** Song **AS** song

**ON** list**.**Artiest **=** song**.**Artiest

**AND** list**.**Titel **=** song**.**titel

**LEFT** **JOIN** SongGenre **as** genre

**ON** genre**.**artiest **=** song**.**Artiest

**AND** genre**.**titel **=** song**.**titel

**WHERE** list**.**editiejaar **=** '2019'

**GROUP** **BY** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

song**.**Speelduur**,**

song**.**Componist**,**

list**.**editiejaar**,**

genre**.**artiest**,**

genre**.**titel

**HAVING** song**.**Speelduur **IS** **NOT** **NULL**

**)** **as** Dataset

-- 98.60% of the songs in the Top2000 list in 2019 has a value for composer

**SELECT** **(COUNT(\*)** **\*** 100.0 **/** **(SELECT** **COUNT(\*)** **FROM** Top2000Lijst **WHERE** editiejaar **=** '2019'**))** **as** completeness\_percentage

**FROM** **(**

**SELECT** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

song**.**Speelduur**,**

song**.**Componist**,**

list**.**editiejaar**,**

**COUNT(**genre**.**genre**)** **AS** number\_of\_genres

**FROM** Top2000Lijst **AS** list

**INNER** **JOIN** Song **AS** song

**ON** list**.**Artiest **=** song**.**Artiest

**AND** list**.**Titel **=** song**.**titel

**LEFT** **JOIN** SongGenre **as** genre

**ON** genre**.**artiest **=** song**.**Artiest

**AND** genre**.**titel **=** song**.**titel

**WHERE** list**.**editiejaar **=** '2019'

**GROUP** **BY** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

song**.**Speelduur**,**

song**.**Componist**,**

list**.**editiejaar**,**

genre**.**artiest**,**

genre**.**titel

**HAVING** song**.**Componist **IS** **NOT** **NULL**

**)** **as** Dataset

-- 97.25% of the songs in the Top2000 list in 2019 has a value for genre

**SELECT** **(COUNT(\*)** **\*** 100.0 **/** **(SELECT** **COUNT(\*)** **FROM** Top2000Lijst **WHERE** editiejaar **=** '2019'**))** **as** completeness\_percentage

**FROM** **(**

**SELECT** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

song**.**Speelduur**,**

song**.**Componist**,**

list**.**editiejaar**,**

**COUNT(**genre**.**genre**)** **AS** number\_of\_genres

**FROM** Top2000Lijst **AS** list

**INNER** **JOIN** Song **AS** song

**ON** list**.**Artiest **=** song**.**Artiest

**AND** list**.**Titel **=** song**.**titel

**LEFT** **JOIN** SongGenre **as** genre

**ON** genre**.**artiest **=** song**.**Artiest

**AND** genre**.**titel **=** song**.**titel

**WHERE** list**.**editiejaar **=** '2019'

**GROUP** **BY** song**.**Artiest**,**

song**.**titel**,**

song**.**Jaar**,**

song**.**Speelduur**,**

song**.**Componist**,**

list**.**editiejaar**,**

genre**.**artiest**,**

genre**.**titel

**HAVING** **COUNT(**genre**.**genre**)** **>=** 1

**)** **as** Dataset

Bohemian Rhapsody van Queen is uitgebracht in 1975.

Somebody To Love van Queen is uitgebracht in 1976.

Somebody To Love van George Michael is uitgebracht in 1993.

Imagine van John Lennon is uitgebracht in 1971.

All For Love van Bryan Adams is uitgebracht in 1994.

ET: Song att release\_year

ID: att song\_name + ET Artist

ID: artist\_name

Bohemian Rhapsody van Queen duurt 00:05:57.

Changes van 2Pac duurt 00:04:30.

ET: Song att duration

MATCH

Predicate: <song\_name> van <artist\_name> duurt <duration>

In the end van Linkin Park heeft als genre nu metal

In the end van Linkin Park heeft als genre rapcore

Let it be van The Beatles heeft als genre rock

With or without you van U2 heeft als genre rock

ET: Song ET: Genre

MATCH ID: att genre\_name

QUESTION: Artiest opslaan of aparte entiteit?

Aan Somebody To Love van George Michael wordt meegewerkt door Queen.

Aan All For Love van Bryan Adams wordt meegewerkt door Rod Stewart.

Aan All For Love van Bryan Adams wordt meegewerkt door Sting.

ET: Song ET: Artist

MATCH MATCH

Somebody To Love van George Michael is gecomponeerd door Freddie Mercury.

Summer of '69 van Bryan Adams is componeerd door Bryan Adams

Summer of '69 van Bryan Adams is componeerd door Jim Vallance

ET: Song ET: Composer

MATCH ID: att composer\_name

In de lijst van 2017 staat Bohemian Rhapsody van Queen op nummer 1.

In de lijst van 2017 staat Imagine van John Lennon op nummer 16.

In de lijst van 2017 staat The Sound Of Silence van Simon & Garfunkel op nummer 28.

In de lijst van 2017 staat Somebody To Love van Queen op nummer 66.

In de lijst van 2017 staat Somebody To Love van George Michael op nummer 135.

In de lijst van 2015 staat Imagine van John Lennon op nummer 1.

In de lijst van 2015 staat Bohemian Rhapsody van Queen op nummer 2.

In de lijst van 2015 staat Somebody To Love van Queen op nummer 65.

In de lijst van 2015 staat Somebody To Love van George Michael op nummer 517.

In de lijst van 2013 staat All For Love van Bryan Adams op nummer 1987.

ET: Top2000ListEntry att position

ID: ET Top2000List + ET Song

ID: att edition\_year MATCH

Top 40 is a chart in NL

Singles Chart is a chart in UK

ET: Chart att country\_code

ID: chart\_name

The song Seven Nation Army by The White Stripes reached on the NL Top 40 as its highest position 22.

The song Seven Nation Army by The White Stripes reached on the UK Singles Chart as its highest position 7.

The song Seven Nation Army by The White Stripes reached on the USA Billboard Hot 100 as its highest position 75.

The song Clint Eastwood by Gorillaz reached on the NL Top 40 as its highest position 27.

The song Clint Eastwood by Gorillaz reached on the UK Singles chart as its highest position 4.

The song Clint Eastwood by Gorillaz reached on the USA Billboard Hot 100 as its highest position 57.

ET: SongOnChart att highest\_ranking\_position

ID: ET Song + ET: Chart

MATCH MATCH

The song Seven Nation Army by The White Stripes stayed on the NL Top 40 for 5 weeks.

The song Seven Nation Army by The White Stripes stayed on the UK Singles Chart for 9 weeks.

The song Seven Nation Army by The White Stripes stayed on the USA Billboard Hot 100 for 1 week.

The song Clint Eastwood by Gorillaz stayed on the NL Top 40 for 6 weeks.

The song Clint Eastwood by Gorillaz stayed on the UK Singles Chart for 17 weeks.

The song Clint Eastwood by Gorillaz stayed on the USA Billboard Hot 100 for 11 weeks.

ET: SongOnChart att duration

MATCH

There is a Top 10 ranking for the municipality of Arnhem.

There is a Top 10 ranking for the municipality of Lingewaard.

There is a Top 10 ranking for the municipality of Renkum.

ET Top 10 Ranking ET: Municipality

ID: municipality\_code

QUESTION: Wordt een liedje van één editie één keer per jaar afgespeeld?

QUESTION: Komt de positie van een liedje van een editie overeen de positie van het liedje in het afspeelschema?

Dust in The Wind van Kansas van editie 2022 wordt op 31-12-2022 ergens vanaf 00:00 u afgespeeld

Ne Me Quitte Pas van Jacques Brel van editie 2022 wordt op 31-12-2022 ergens vanaf 00:00 u afgespeeld

--------------------------------------------------------------------- -------------

ET: PlayScheduleEntry att start\_time\_window

ID: ET EditionEntry + att play\_date

MATCH

Dust in The Wind van Kansas van editie 2022 wordt op 31-12-2022 ergens tot aan 01:00 u afgespeeld

Ne Me Quitte Pas van Jacques Brel van editie 2022 wordt op 31-12-2022 ergens tot aan 01:00 u afgespeeld

--------------------------------------------------------------------- -------------

ET: PlayScheduleEntry att end\_time\_window

MATCH

DJ met code JVI heeft als voornaam Jeroen

DJ met code JKV heeft als voornaam Jeroen Kijk

--------------- ---------------------

ET: DJ att

ID: att code

DJ met code JVI heeft als tussenvoegsel van

DJ met code JKV heeft als tussenvoegsel in de

--------------- -------------------

ET: DJ att prefix

MATCH

DJ met code JVI heeft als achternaam Inkel

DJ met code JKV heeft als achternaam Vegt

--------------- ---------------

ET: DJ att last\_name

MATCH

DJ met code JVI heeft in 2021 vanaf 00:00 u nummers gedraaid

DJ met code JKV heeft in 2021 vanaf 08:00 u nummers gedraaid

DJ met code JWR heeft in 2021 vanaf 08:00 u nummers gedraaid

ET: DJInYear att start\_time

ID: ET DJ + (?)edition\_year

MATCH

DJ met code JVI heeft in 2021 tot aan 02:00 u nummers gedraaid

DJ met code JKV heeft in 2021 tot aan 10:00 u nummers gedraaid

DJ met code JWR heeft in 2021 tot aan 10:00 u nummers gedraaid

ET: DJInYear att start\_time

ID: ET DJ + (?)edition\_year

MATCH

- waarom code: geeft meer context en zijn niet veel verschillende dj's, makkelijkere weergave in ui en minder joins. = char!

- check op of er 2000 liedjes in staan per editie

- moet in 2021 was er een editie als feittype?

- kan een lijst hergebruikt worden in een ander jaar?

- zo ja, kan er dus een schema zijn van een jaar zonder dat er een top2000lijst van datzelfde jaar is. --> in dit geval kan ik niet edition\_year van DJInYear gebruiken van de EditionYear

- zo niet, heeft ieder jaar een schema van de top2000lijst van hetzelfde jaar --> t.b.v. de datakwaliteit kan ik dan beter edition\_year van de ET Edition gebruiken omwille van consistentie.

- is het afpseelschema hetzelfde als de top2000list?

De luisteraar met email marco.engelbart@hotmail.com heeft voornaam Marco.

De luisteraar met email jan.van.vliet@gmail.com heeft voornaam Jan.

------------------------------------------------- ----------------------------------

ET Listener att first\_name

ID: att email

De luisteraar met email jan.van.vliet@gmail.com heeft tussenvoegsel van.

De luisteraar met email piet.de.bruin@hotmail.com heeft tussenvoegsel de.

--------------------------------------------------- --------------------------------

ET Listener att prefix

MATCH

De luisteraar met email marco.engelbart@hotmail.com heeft achternaam engelbart.

De luisteraar met email jan.van.vliet@gmail.com heeft achternaam Vliet.

-------------------------------------------------- ---------------------------------

ET Listener att last\_name

MATCH

De luisteraar met email marco.engelbart@hotmail.com heeft geslacht man.

De luisteraar met email jan.van.vliet@gmail.com heeft geslacht vrouw.

--------------------------------------------------- ----------------------------------

ET Listener att gender

MATCH

De luisteraar met email marco.engelbart@hotmail.com valt in de leeftijdscategorie 51-55 jaar.

De luisteraar met email jan.van.vliet@gmail.com valt in de leeftijdscategorie 56-60 jaar.

-------------------------------------------------------- ---------------------------

ET Listener att age\_category

MATCH

De luisteraar met email marco.engelbart@hotmail.com heeft postcodecijfers 5343.

De luisteraar met email jan.van.vliet@gmail.com heeft postcodecijfers 1234.

--------------------------------------------------- ---------------------------------

ET Listener ET postal\_code

MATCH ID: att postal\_code\_numbers

De luisteraar met email marco.engelbart@hotmail.com heeft een stemlijst voor de top 2000 van 2020 ingestuurd.

De luisteraar met email jan.van.vliet@gmail.com heeft een stemlijst voor de top 2000 van 2020 ingestuurd.

---------------------------------------------------------------------- -----------------

ET VoteList ET Top2000List

ID: ET Listener + ET Top2000List MATCH

Het nummer Mystery Train van Elvis Presley staat op de stemlijst van de luisteraar met email marco.engelbart@hotmail.com voor de top 2000 van 2020.

Het nummer Dust My Broom van Elmore James staat op de stemlijst van de luisteraar met email marco.engelbart@hotmail.com voor de top 2000 van 2020.

------------------------------------------------------------------------------------------

ET VoteListEntry

ID: ET VoteList + ET Song

