

Beroepsproduct DDDQ: Casus top 2000

Auteurs

Geo Bouwmeestern, Neo Hop & Julian van Zwol

HAN University of applied sciences

Academie IT en Mediadesign

I-ADB 2024/2025 - Advanced Databases (deeltijd)

13 maart 2025

Versie 1.0

Inhoud

1	Inleiding	2
2	Onderzoek Datakwaliteit	3
2.1	Duiding Databasestructuur	3
2.1.1	Database Diagram	3
2.1.2	Toelichting	3
2.2	Controle op Schending Primary key	3
2.2.1	Check op tabel Top2000Lijst	3
2.2.2	Check op tabel SongGenre	3
2.2.3	Check op tabel Song	3
2.3	Controle op schending van Referentiële integriteit	4
2.3.1	Check on Top2000Lijst's (logically) referentie naar Song	4
2.3.2	Check on SongGenre's (logically) referentie naar Song	4
2.4	Controle op Derde Normaalvorm (3NF)	4
2.5	Controle op Integriteitregels	5
2.6	Controle op Datakwaliteitsdimensie <i>Completeness</i>	5
2.7	Controle op Datakwaliteitsdimensie <i>Accuracy</i>	5
2.7.1	Geen uniforme schrijfwijze	5
2.8	Verbeteringen m.b.t. Data Intension	5
2.9	Verbeteringen m.b.t. Data Extension	6
2.10	Overige constatering	6
	Lijst van figuren	6
	Lijst van figuren	7
3	Bijlage	8
.1	Controle op Schending Primary key	8
.2	Controle op Schending van Referentiële integriteit	8
.3	Controle op Integriteitregels	9
.4	Controle op Datakwaliteitsdimensie <i>Completeness</i>	9
.5	Controle op Datakwaliteitsdimensie <i>Accuracy</i>	10

Hoofdstuk 1

Inleiding

Hoofdstuk 2

Onderzoek Datakwaliteit

2.1 Duiding Databasestructuur

2.1.1 Database Diagram

2.1.2 Toelichting

Invulling
geven.

2.2 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 groepeerd op de (logische) primary key en vervolgens controleert of hier duplicaten tussen zitten.

2.2.1 Check op tabel Top2000Lijst

Uitvoering van de SQL select query in bijlage .1 resulteerde in geen schending van de primary key.

2.2.2 Check op tabel SongGenre

Uitvoering van de SQL select query in bijlage .1 resulteerde in geen schending van de logische primary key.

2.2.3 Check op tabel Song

Uitvoering van de SQL select query in bijlage .1 resulteerde in geen schending van de logische primary key.

Top2000Lijst			
Column Name	Condensed Type	Nullable	
editiejaar	int	No	
positie	int	No	
Artiest	nvarchar(255)	Yes	
Titel	nvarchar(255)	Yes	

SonaGenre			
Column Name	Condensed Type	Nullable	
artiest	nvarchar(255)	Yes	
titel	nvarchar(255)	Yes	
genre	nvarchar(255)	Yes	

Sona			
Column Name	Condensed Type	Nullable	
Artiest	nvarchar(255)	Yes	
titel	nvarchar(255)	Yes	
Jaar	float	Yes	
Componist	nvarchar(255)	Yes	
Speelduur	time(0)	Yes	

Figuur 2.1: Generated diagram from SSMS.

2.3 Controle op schending van Referentiële integriteit

2.3.1 Check on Top2000Lijst's (logically) referentie naar Song

Uitvoering van de SQL select query in bijlage .2 resulteerde in schending van de referentiële integriteit op **122** records.

2.3.2 Check on SongGenre's (logically) referentie naar Song

Uitvoering van de SQL select query in bijlage .2 resulteerde in schending van de referentiële integriteit op **7** records.

2.4 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.

2.5 Controle op Integriteitregels

IR1. Een song staat maximaal één keer in een editie van de Top 2000. Uitvoering van de SQL select query in bijlage .3 resulteerde in schending van de Integriteitregel op **2** records.

IR2. Elke song die in een editie van de Top 2000 staat, is vóór of in het jaar van de betreffende editie uitgebracht. Uitvoering van de SQL select query in bijlage .3 resulteerde in geen schending van de Integriteitregel.

Zijn er nog andere integriteitregels? Wordt expliciet naar verwezen in rubriek.

2.6 Controle op Datakwaliteitsdimensie *Completeness*

Uitvoering van de SQL select query in bijlage .4 resulteerde in een percentage van *96.00%* van nummers in de Top2000 lijst van 2019 waarvan voor elk gegeven attribuut een waarde beschikbaar is.

Uitvoering van de SQL select query in bijlage .4 resulteerde in een percentage van *96.00%* van nummers in de Top2000 lijst van 2019 waarvan voor elk gegeven attribuut een waarde beschikbaar is.

Zijn er nog meer referentiele integriteit regels?

2.7 Controle op Datakwaliteitsdimensie *Accuracy*

2.7.1 Geen uniforme schrijfwijze

Uitvoering van de SQL select query in bijlage .5 resulteerde in **226** unieke records. Echter hebben sommige genres verschillende schrijfwijzen en daarmee dus geen uniforme manier om genres weer te geven. Enkele voorbeelden staan hieronder:

Invulling geven.

Zijn er typefouten?

Zijn er standaard waarden zoals: 1-1-1900

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	-
trash metal	trashmetal	-

Tabel 2.1: Alternatieve schrijfwijzen

uniforme manier?

2.8 Verbeteringen m.b.t. Data Intension

Invulling geven.

2.9 Verbeteringen m.b.t. Data Extension

Invulling
geven.

2.10 Overige constatering

- The attribute 'artiest' regards a person and should therefore be it's own entity (table).
- The table 'SongGenre' is not normalized past the First Normal Form and has therefore a lot of redundant data.
- Three

Lijst van figuren

2.1	Generated diagram from SSMS.	4
-----	--------------------------------------	---

Hoofdstuk 3

Bijlage

.1 Controle op Schending Primary key

```
1  -- Check op tabel Top2000Lijst.
2  SELECT editiejaar, positie
3  FROM Top2000Lijst
4  GROUP BY editiejaar, positie
5  HAVING COUNT(*) <> 1
6
7  -- Check op tabel Song.
8  SELECT Artiest, titel
9  FROM Song
10 GROUP BY Artiest, titel
11 HAVING COUNT(*) > 1
12
13 -- Check op tabel SongGenre.
14 SELECT artiest, titel, genre
15 FROM SongGenre
16 GROUP BY artiest, titel, genre
17 HAVING COUNT(*) > 1
```

.2 Controle op Schending van Referentiële integriteit

```
1  -- Check on Top2000Lijst's (logically) referentie naar Song
2  SELECT Artiest, Titel
3  FROM Top2000Lijst as list
4  WHERE NOT EXISTS (
5      SELECT *
6      FROM Song as song
7      WHERE list.Artiest = song.Artiest
```

```

8         AND list.Titel = song.titel
9     )
10
11 -- Check on SongGenre's (logically) referentie naar Song
12 SELECT artiest, titel
13 FROM SongGenre as genre
14 WHERE NOT EXISTS (
15     SELECT *
16     FROM Song as song
17     WHERE genre.Artiest = song.Artiest
18     AND genre.Titel = song.titel
19 )

```

.3 Controle op Integriteitregels

```

1 -- IR1. Een song staat maximaal een keer in een editie van de Top
  2000.
2 SELECT editiejaar, Artiest, Titel, COUNT(1)
3 FROM Top2000Lijst
4 GROUP BY editiejaar, Artiest, Titel
5 HAVING COUNT(*) > 1
6
7 -- IR2. Elke song die in een editie van de Top 2000 staat, is voor
  of in het jaar van de betreffende editie uitgebracht.
8 SELECT song.Artiest,
9         song.titel,
10        song.Jaar,
11        list.editiejaar
12 FROM Top2000Lijst AS list
13 INNER JOIN Song AS song
14 ON list.Artiest = song.Artiest
15 AND list.Titel = song.titel
16 WHERE song.Jaar > list.editiejaar

```

.4 Controle op Datakwaliteitsdimensie *Completeness*

```

1 -- Check on value for every given property for songs in the Top2000
  list in 2019.
2 SELECT (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM Top2000Lijst WHERE
  editiejaar = '2019')) as completeness_percentage
3 FROM (
4         SELECT song.Artiest,

```

```

5         song.titel,
6         song.Jaar,
7         song.Speelduur,
8         song.Componist,
9         list.editiejaar,
10        COUNT(genre.genre) AS number_of_genres
11 FROM Top2000Lijst AS list
12 INNER JOIN Song AS song
13 ON list.Artiest = song.Artiest
14 AND list.Titel = song.titel
15 LEFT JOIN SongGenre as genre
16 ON genre.artiest = song.Artiest
17 AND genre.titel = song.titel
18 WHERE list.editiejaar = '2019'
19 GROUP BY song.Artiest,
20         song.titel,
21         song.Jaar,
22         song.Speelduur,
23         song.Componist,
24         list.editiejaar,
25         genre.artiest,
26         genre.titel
27 HAVING COUNT(genre.genre) >= 1
28 AND song.Componist IS NOT NULL
29 AND song.Speelduur IS NOT NULL
30 ) as Dataset

```

.5 Controle op Datakwaliteitsdimensie *Accuracy*

```

1 -- Geen uniforme schrijfwijze
2 SELECT DISTINCT genre
3 FROM [Top2000].[dbo].[SongGenre]

```