Big Data – Opdracht AP 3

Opdracht 1 Big Data, SQL

Dimitry Volker - 1661152

# Extra informatie

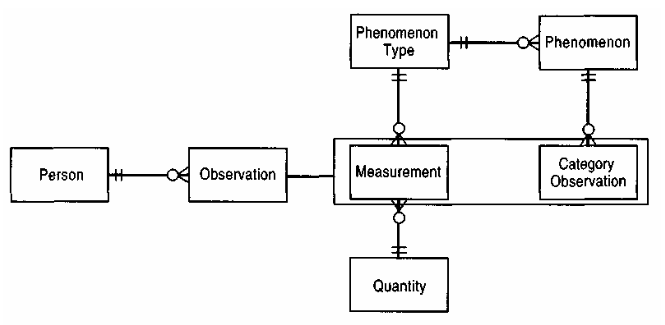
Alle diagrammen en sql scripts zijn te vinden in mijn publieke repo:

https://github.com/Yerath/BIGDATA-ASS01

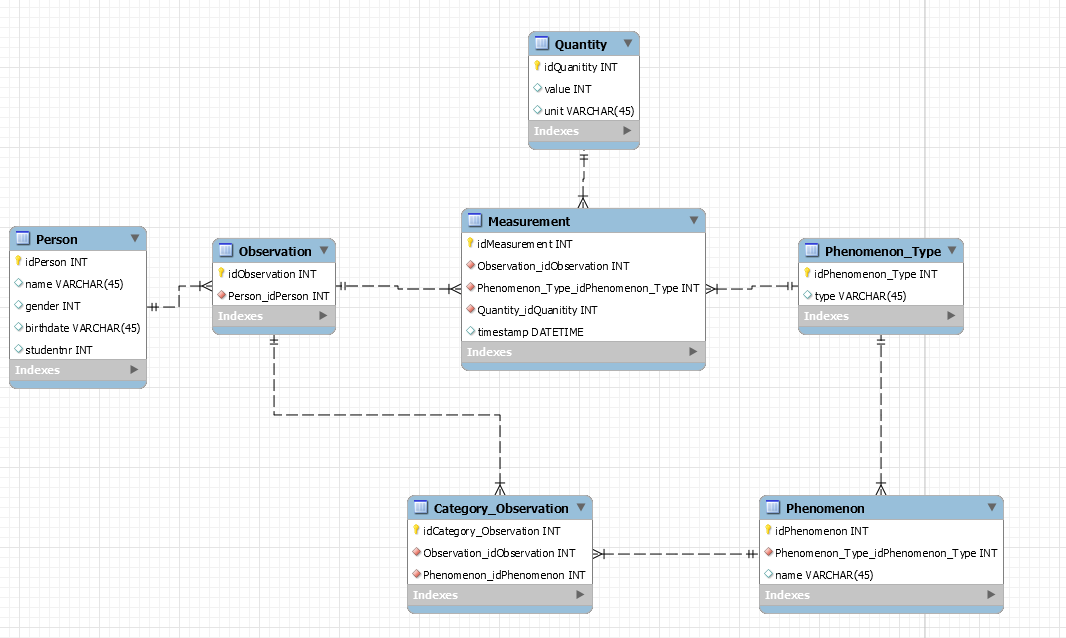
# ERD / RDBM

## Fowler

In de opdracht staat aangegeven dat de eerste oplossing gemaakt moet worden op basis van de ERD van Fowler. Deze heb ik aangehouden en hier ook de tabellen *Phenomenon* en *Category\_Observation* meegenomen. Deze zijn onnodig, want de tabellen zullen niet gevuld worden.



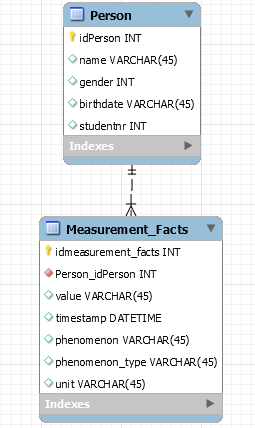
ERD Fowler



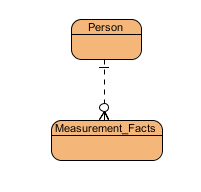
RDBM Fowler

## Sterschema

In het ontwerp van de Sterschema heb ik de naamgeving gebruikt uit het boek van Fowler. Als we kijken naar de “feiten” tabel, is het belangrijk om daar echte alle feiten op te leveren en geen onnodige dimensies te creëren. Daarom is alleen *Person* een aparte tabel geworden.

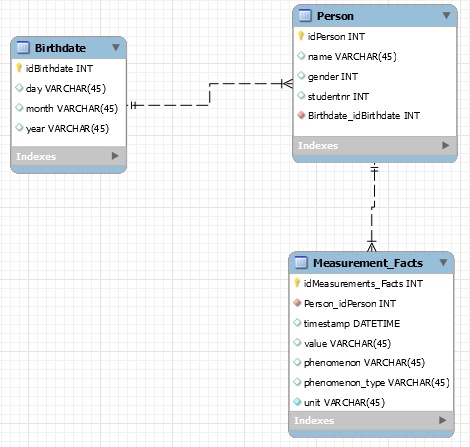


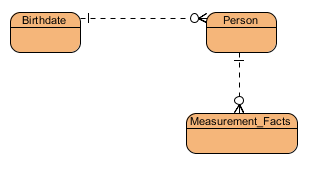
RDBM Sterschema



ERD van Sterschema

## Snowflake

Naast het sterschema kan er een dimension ook weer aftakkingen hebben. In dit probleem is het niet goed toepasbaar. Wel is het mogelijk om de geboortedatum van een persoon te normaliseren waardoor er een snowflake diagram ontstaat.



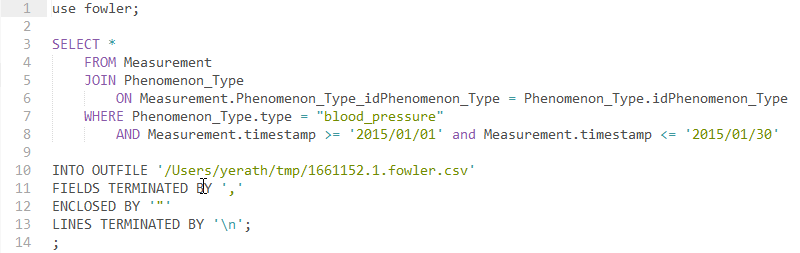
ERD van Sterschema

RDBM Snowflake

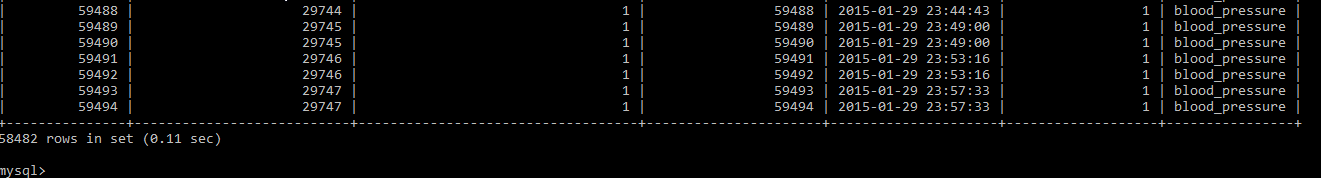
# Queries

## Fowler

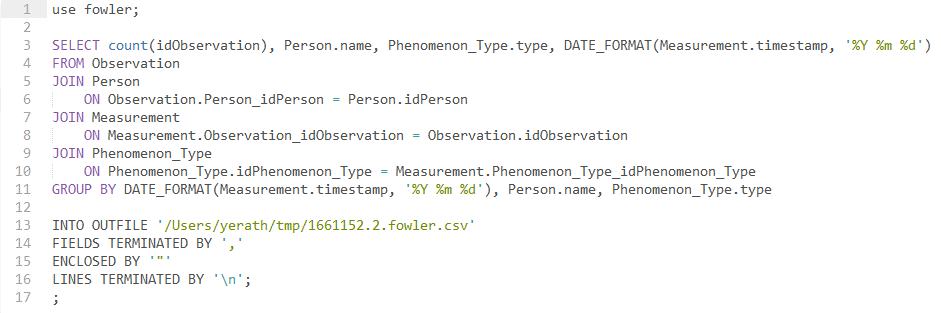
### Overzicht bloeddruk metingen



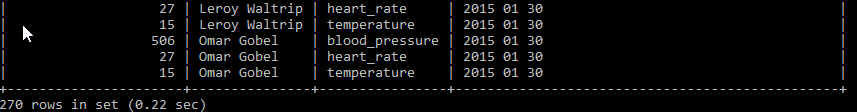
Resultaat



### Overzicht observaties



Resultaat



### Gemiddelde bloeddruk

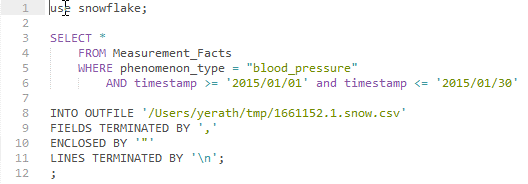


Resultaat

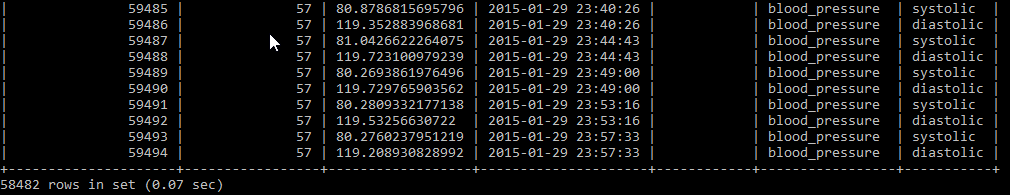
Voor deze query is alleen een CSV bestand als resultaat. Vanwege het gebruik van MySQL is deze query ontzettend traag en liep Workbench gewoon vast.

## Sterrenschema

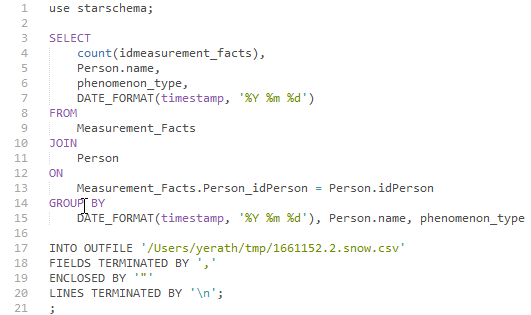
### Overzicht bloeddruk metingen



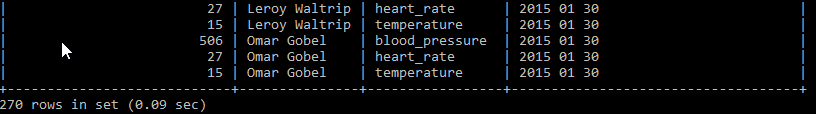
Resultaat



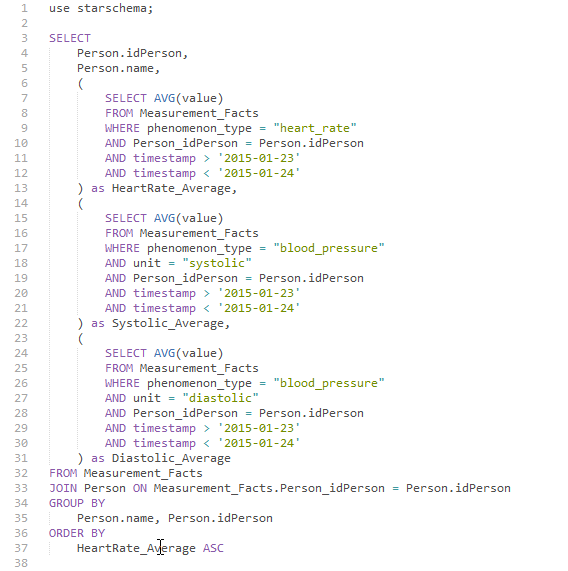
### Overzicht observaties



Resultaat



### Gemiddelde bloeddruk



Resultaat

Voor deze query is alleen een CSV bestand als resultaat. Vanwege het gebruik van MySQL is deze query ontzettend traag en liep Workbench gewoon vast.

## Snowflake

Omdat alleen de birthdate van een persoon is genormaliseerd, zijn de queries van de Snowflake database niet anders dan die van het sterschema.

# Advies

Met het huidige probleem is het aan te raden om het sterschema te gebruiken. Dit heeft een aantal voordelen ten opzichte van Fowler zijn ontwerp.

* Weinig tabellen en veel op één plek
* Queries zijn sneller omdat er minder (bijna geen) JOINS worden gebruikt.
* CSV Data is niet heel gecompliceerd en heeft ook bijna geen relaties.

Zoals eerder aangegeven, is het snowflake idee niet goed toe te passen. Het normaliseren van de geboortedatum is eigenlijk alleen gedaan voor de oefening.