Data-Science 1

datamanagement

Karel de Grote Hogeschool

02/03/2023

Inhoud

- wat is datamanagement?
- voorbeelden in Python
- oefeningen

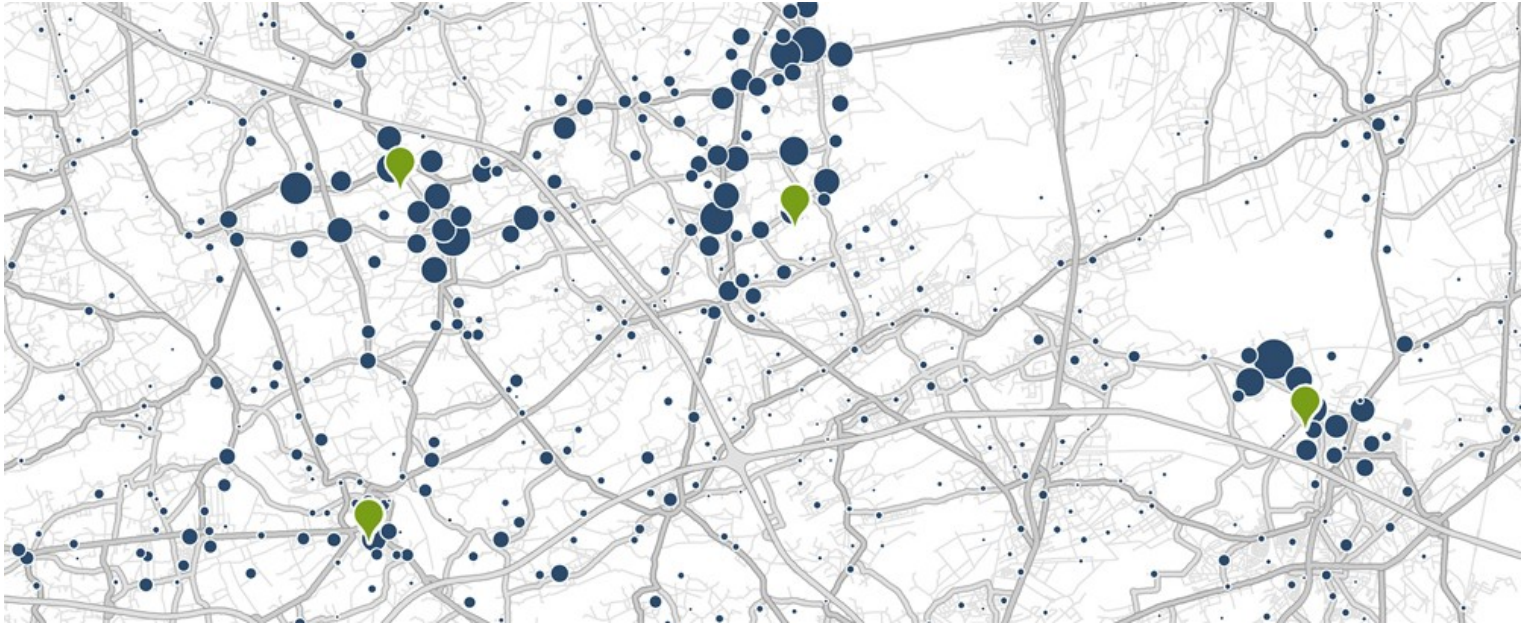
Wat is data management?

Data management

- data omzetten naar een bruikbare vorm
- data bevat meestal veel fouten
  - ontbrekende waarden
  - verkeerde waarden
  - verkeerd formaat
  - verkeerd aantal kolommen
  - ...
- data moet soms geëxtraheerd of gecombineerd worden

Voorbeeld

- bedrijf wil overzicht waar grote en kleine klanten zich bevinden: kaart met grote en kleine bollen
  - we baseren ons op het totaalbedrag van de bestellingen (exclusief BTW) van het laatste jaar



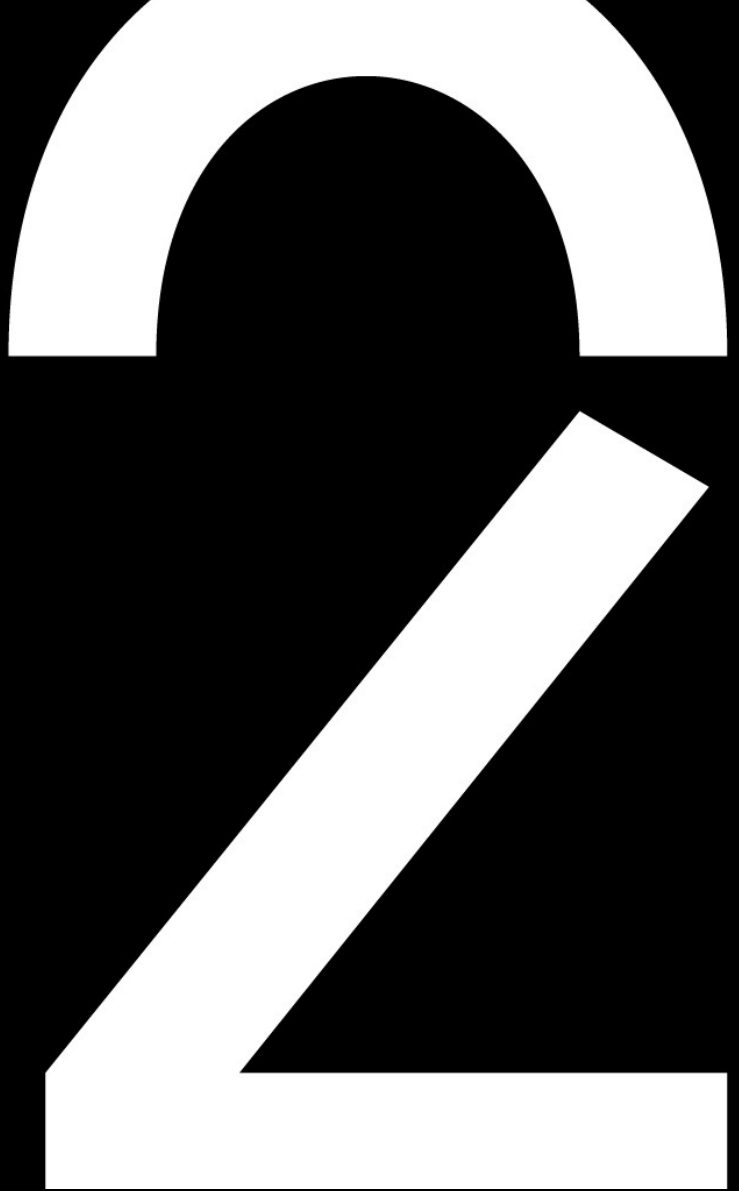
Problemen

- bedrijf is een fusie van 3 bedrijven
  - klein bedrijf gebruikt excel files en een paar binaire bestanden
  - ander bedrijf gebruikt Oracle
  - laatste bedrijf gebruikt SAP en Cobol
- als je de data bij elkaar zet, zijn er veel ontbrekende waarden
- encodering van de data is verschillend: Cobol gebruikt EBCDIC, Oracle gebruikt ASCII, de excel bestanden en de data in SAP zijn unicode (UTF-8 gecodeerd).
- er zijn ook een paar binaire bestanden: die kunnen in little of big endian gecodeerd zijn.

Problemen

- de klanten hebben soms meerdere adressen
- sommige bestanden hebben enkel factuurdatum, anderen ook besteldatum
- sommige bestanden bevatten dezelfde klanten maar soms lichtjes verschillend (schrijffouten)
- sommige bedragen zijn inclusief btw en andere zonder. We willen zonder. Er worden verschillende btw tarieven gebruikt
- datums zijn soms Amerikaans gestockeerd (maand/dag/jaar), soms aantal dagen sinds 1/1/1970, soms Europees (dag/maand/jaar)
- ...

Voorbeelden in Python



Voorbeelden in Python

- zie notebook

Oefeningen

Oefeningen

- enquête:
  - lees het bestand van de enquête in
  - maak de kolomnamen korter (zoek op)
  - zet alle datatypes juist
- zie Canvas