JADS mkbdatalab

Data Structuur Gids

Hoe verzamel je data op een juiste manier?

Introductie

Vaak horen we dat (big) data de olie van de 21ste eeuw is. Dat kan je het gevoel geven dat je blindelings alle data moet gaan verzamelen die je mogelijk zou kunnen verzamelen. Stel jezelf echter een bedrijf voor dat de afgelopen 10 jaar veel klantgegevens heeft verzameld die ze willen analyseren. Een eerste stap voor dit bedrijf zou kunnen zijn om een gesprek te voeren met een data analist. Het bedrijf vertelt de analist dan dat ze al flink veel data verzameld hebben, dus dat het gemakkelijk zou moeten zijn om veel interessante inzichten uit de data te halen. Echter, wanneer de analist naar de beschikbare data kijkt, is de kans groot dat hij/zij zal zeggen: “Ik heb eerst een paar maanden nodig voordat we tot de gewenste inzichten kunnen komen, want jouw data is nog niet klaar voor analyse".

Deze handleiding zal je helpen begrijpen waarom het langer kan duren dan verwacht om zinvolle inzichten te halen uit uw bedrijfsgegevens en wat je als MKB’er kunt doen om je data op de juiste manier te prepareren om het analyseren ervan te vergemakkelijken. Als je het concept van schone en gestructureerde data begrijpt, kun je je eigen data gaan verbeteren. Na het opschonen en structureren van je data, is je data klaar voor exploratie en analyse!

Hoe kan deze handleiding me helpen mijn data bruikbaar te maken voor analyse?

In deze handleiding worden de volgende stappen uitgelegd, gevolgd door enkele voorbeelden en vragen die je kunnen ondersteunen bij het opschonen van je eigen data:

1. Data verzameling: Welke data moet je verzamelen? (als je dat nog niet hebt gedaan)
2. Verzamelen van de juiste data: Hoe verzamel je de juiste data?
    a. Gestructureerde data
    b. Schone data
    c. Relationele database
3. Data voorbereiden voor verdere analyse

Dataverzameling: welke data moet je verzamelen?

Dit gedeelte is vooral relevant voor het MKB dat nog geen data heeft verzameld. Allereerst, vanuit het perspectief van toekomstige analyse; hoe meer data je kunt verzamelen met weinig tot geen inspanning en middelen, hoe beter. Als je echter net begint, moet je er rekening mee houden wat de reden is om bepaalde variabelen te verzamelen en wat de extra kosten zijn. Het heeft zelden zin om de temperatuur van de dag vast te leggen tijdens het registreren van bijvoorbeeld verkooporders, maar als je geïnteresseerd bent in het analyseren van het effect van het weer op je verkopen, kan het interessant zijn. Besteed dus geen tijd en middelen aan het verzamelen van irrelevante gegevens die je sowieso niet gaat gebruiken. Door de volgende vragen te beantwoorden, kun je bepalen welke data je moet verzamelen voor je eerste analyses:

1. Welke vraag wil ik beantwoorden met de data analyse?
2. Welke data heb ik nodig om deze vraag te beantwoorden?
3. Welke relevante data kan ik verzamelen?
4. Is dat voldoende om mijn eerste vraag te beantwoorden?

Voorbeeld

Je hebt een bedrijf dat laptops verkoopt en het is je doel om je klanten beter te benaderen met gepersonaliseerde deals. Het invullen van de bovenstaande vragen zou er ongeveer zo uit kunnen zien:

1. Welke vraag wil ik beantwoorden met de data analyse?

Hoe kan ik onze klanten targeten met meer gepersonaliseerde deals op basis van hun laptopvoorkeuren?

2. Wat zijn de deelvragen om deze vraag te beantwoorden en welke gegevens heb ik nodig om mijn analyses te maken?

- Wat voor verschillende soorten klanten heb ik?
  Data die relevant is om vast te kunnen stellen of er verschillende soorten klantgroepen zijn, hiervoor is vaak enige domeinkennis nodig:
  - Land
  - Leeftijd
  - Geslacht
  - Grootte van de stad
  - Burgerlijke staat
  - Inkomen
  - Gezinsgrootte
  - Laatste aankoop (datum)
  - Aantal aankopen
  - Totaal uitgegeven geld
  - ..
- Wat kopen mijn klanten meestal?
  Order data en data die relevant is voor het onderscheiden van producttypen:
  - Productnaam
  - Merk
  - Geheugen van de laptop
  - Schermgrootte
  - Camera
  - Bluetooth
  - Aantal USB-poorten
  - …
- Kan ik een patroon zien in bepaalde klanten die bepaalde producten kopen?
  Vereist alle bovenstaande informatie

3. Welke relevante gegevens kan ik nu verzamelen?

- Klantgegevens - naam, adres, telefoonnummer, e-mail, land, leeftijd, ..
- Productgegevens - productnaam, beschrijving, specificaties, …
- Verkoopgegevens - datum, product, klant, hoeveelheid, prijs, …

4. Is dat voldoende om mijn eerste vraag te beantwoorden?

Data analyse is een iteratief proces, wat betekent dat we kunnen starten met de beschikbare data. Als blijkt dat we niet genoeg informatie hebben om tot interessante inzichten te komen, is het altijd mogelijk om meer data te verzamelen. Begin echter met zoveel mogelijk van de mogelijke en relevante databronnen.

Je kunt natuurlijk veel verschillende Key Performance Indicators (KPI's) hebben, waarvoor een andere informatievoorziening nodig is. Het is daarom geen slecht idee om veel data te verzamelen. Wel dien je in ieder geval over de data te beschikken die nodig zijn om je onderzoeksvragen te beantwoorden. Deze vragen helpen je te bepalen welke data en variabelen je nodig hebt om die vragen te beantwoorden.

Nadat we de verschillende variabelen hebben geïdentificeerd, is het belangrijk om de data op de juiste manier op te slaan. Er zijn veel opties om je data op te slaan, maar Excel is een goede plek om te beginnen als je net begint met het verzamelen van data. Als je al veel data hebt, of je hebt meer ervaring, dan kun je kijken in databases zoals MySQL, PostgreSQL of MongoDB.

Verzamelen van de juiste data: Gestructureerde data

Gestructureerde data wordt meestal gevonden in Excel datasets of relationele databases (dit wordt later behandeld). Het betekent dat alles wordt opgeslagen in kolommen of velden. Elke rij geeft een nieuw exemplaar aan (persoon, item, bedrijf of ander stuk informatie). Wanneer je data gestructureerd is, is het eenvoudiger om stukjes informatie te zoeken voor jou als persoon, maar ook voor de computer, om uiteindelijk snellere en betere analyses uit te voeren.

Ongestructureerde data is in wezen al het andere. Ongestructureerde data kan een interne structuur hebben, maar zijn niet gestructureerd via vooraf gedefinieerde datamodellen of schema's. Het kan tekstueel of niet-tekstueel zijn, en door mensen of machines gegenereerd.

Voorbeelden van ongestructureerde data:

- Tekstbestanden: tekstverwerking, spreadsheets, presentaties, e-mail, logboeken.
- E-mail: e-mail heeft een bepaalde interne structuur dankzij de metadata, en we noemen het soms semi-gestructureerd. Het berichtveld is echter ongestructureerd en traditionele analysetools kunnen het niet *parsen*.
- Sociale media: data van Facebook, Twitter, LinkedIn.
- Website: YouTube, Instagram, sites voor het delen van foto's.
- Mobiele data: tekstberichten, locaties.
- Communicatie: Chat, IM, telefoonopnames, samenwerkingssoftware.
- Media: mp3, digitale foto's, audio- en videobestanden.
- Zakelijke toepassingen: MS Office-documenten, productiviteitstoepassingen.

Voorbeeld van ongestructureerde data en gestructureerde data

Ongestructureerd
Vandaag hadden we een meeting met Martin Jones, CEO van een mogelijke leverancier, Lenovo. Ze zijn bereid om ons 10% korting te geven op elke laptop die we kopen als we deze in batches van minimaal 100 kopen.

Gestructureerd	
Datum:	21-07-2020
Type meeting:	Supplier intake
Contactpersoon:	Martin Jones
Bedrijf:	Lenovo
Korting:	10%
Minimale ordergrootte:	100

Vragen

1. Wat voor type ongestructureerde data heb jij?

   _____

2. Indien aanwezig, hoe zou je deze data om kunnen zetten naar gestructureerde data?

   _____

   _____

Verzamelen van de juiste data: Schone data

Data opschonen verwijst naar het proces waarbij het onjuiste, onvolledige, onnauwkeurige, irrelevante of ontbrekende deel van de data wordt geïdentificeerd en vervolgens naar behoefte wordt aangepast, vervangen of verwijderd. Data is essentieel voor analyse en machine learning, maar als het gaat om data uit de echte wereld, is het waarschijnlijk dat gegevens onvolledige, inconsistente of ontbrekende waarden bevatten. Daarom kan het belang van het opschonen van data niet genoeg worden benadrukt. Ongeacht het algoritme dat je gebruikt - als je gegevens slecht zijn, krijg je slechte resultaten. Professionele datawetenschappers weten dit en hebben onthuld dat het opschonen van data tot 70% (!) van de tijd in beslag neemt die aan een data science-project wordt besteed.

Missende waarden

Dit is misschien wel de meest voorkomende eigenschap van niet schone data. Deze waarden hebben meestal de vorm van *NaN* of *None* in een tabel. Dit is bijvoorbeeld het geval wanneer je je gegevens in Excel opent en er lege cellen zijn. Ontbrekende waarden kunnen ook leiden tot modellen die *NaN*-waarden voorspellen, wat we natuurlijk niet willen. Om met ontbrekende data om te gaan, is het belangrijk om de oorzaak ervan te achterhalen. Dit zal je helpen bij het beslissen over de beste manier om ermee om te gaan.

Er zijn verschillende oorzaken van missende waarden. Sommige waarden ontbreken omdat ze niet bestaan, andere ontbreken vanwege een onjuiste verzameling of een slechte data invoer. Als iemand bijvoorbeeld vrijgezel is en een vraag is enkel van toepassing op gehuwden, bevat de vraag een ontbrekende waarde. In dergelijke gevallen zou het verkeerd zijn om een waarde voor die vraag in te vullen. Hoe ga je om met deze ontbrekende waarden? Dit hangt af van het type data dat je verzamelt:

Data Types

- **Numerieke data**: je kunt voor ontbrekende waarden een 0 of het gemiddelde invullen;
- **Tekstuele data**: tekst is meestal niet relevant voor (basis) analyse, bijv. wanneer een kolom de beschrijving van een item bevat;
- **Categorische data**: als er een rij is die niet in de beschikbare categorieën valt, kun je een nieuwe categorie "Geen categorie" aanmaken. Die kan worden gebruikt om de lege waarden op te vullen;
- **Datums**: als een datum ontbreekt, terwijl dit essentieel is voor je analyse, kun je het beste de hele rij verwijderen als je de datum niet meer kunt herinneren/achterhalen. Het verwijderen van rijen of kolommen heeft de minste voorkeur (verlies van data!), maar is een goed alternatief als je de ontbrekende waarden niet kunt invullen.

Inconsistente data

Dit gebeurt meestal bij tekstuele of categorische data. Als je bijvoorbeeld de namen van verschillende landen moet opschrijven, kan werknemer nr. 1 "nederland" schrijven, werknemer nr. 2 schrijft "Nederland" werknemer nr. 3 schrijft "Holland". Hoewel ze allemaal hetzelfde betekenen, is het anders geschreven en zal een computer ze niet als gelijk beschouwen. Als we een analyse zouden doen op basis van de demografische gegevens van je klant, zou een

resultaat zijn dat er één persoon in "nederland", "Nederland" en "Holland" woont in plaats van 3 personen in Nederland.

Dus hoe ga je om met deze inconsistente data? Zoek zelf uit hoe je de data nu verzamelt. Is het mogelijk dat je medewerkers bepaalde waarden invullen? Zo ja, worden ze beperkt door een vervolgkeuzemenu (consistente data) of is het een open tekst vak dat alle verschillende soorten waarden toestaat (inconsistente data). Als dit het laatste is, controleer dan je gegevens op inconsistenties en probeer open velden te voorkomen; kies liever voor keuzemenu's of schrijf duidelijke instructies voor het invoeren van de gegevens.

Voorbeeld

Naam	Adres	Telefoon	E-mail
Tyrone Lowe	89 St James Boulevard, EX39 0YB, Horns Cross	+44 77 6428 2654	t.lowe@gmail.com
Spears, Loren	17 Ponteland Rd, Mortimer WE, TD5 7XT	070-09974620	lorenspears@outlook.com
Maria Randall	96 Preston Rd, Mortimer West End	077 1879 1573	m.randall@icloud.com

Tabel 1. Gestructureerde, **niet** schone data

Voor naam	Achter naam	Straat	Postcode	Stad	Telefoon	E-mail
Tyrone	Lowe	89 St James Boulevard	EX39 0YB	Horns Cross	+44 77 6428 2654	t.lowe@gmail.com
Loren	Spears	17 Ponteland Rd	TD5 7XT	Mortimer West End	+44 70 0997 4620	lorenspears@outlook.com
Maria	Randall	96 Preston Rd	RG7 5DS	Mortimer West End	+44 77 1879 1573	m.randall@icloud.com

Tabel 2. Gestructureerde, schone data

Kun je de verschillen zien? Laat me je helpen:

- ‘Voornaam’ en ‘achternaam’ zijn gescheiden in kolommen om knoeien met de volgorde te voorkomen
- ‘Adres’ is opgesplitst in ‘straat’, ‘postcode’ en ‘stad’ om geknoei met de bestelling te voorkomen
- ‘Stad’ is consequent opgeschreven (Mortimer WE / Mortimer West End)
- ‘Telefoon’ is opgeschreven in overeenstemming met de landcode

Vragen

1. Hoe schoon is de data in jouw systemen?

   _____

2. In welke databronnen wordt data op een schone en consistente manier verzameld?

   _____

3. Voor welke databronnen geldt dit (nog) niet?

   _____

4. Welke stappen kun je zetten om de data kwaliteit van deze bronnen te verbeteren?

   _____

Verzamelen van de juiste data: Relationele database

In de bovenstaande voorbeelden noemden we verschillende databronnen zoals klantdata en productdata. Het is gebruikelijk dat organisaties meerdere tabellen hebben waarin verschillende soorten gegevens worden opgeslagen. Om een meer holistisch/rijk beeld te krijgen van de analytische vraag die je probeert te beantwoorden, kan het nodig zijn om verschillende databronnen binnen jouw bedrijf te combineren. Een relationele database is een manier om jouw data op te slaan en te ordenen in tabellen en deze te koppelen op basis van gedefinieerde relaties. Met deze relaties kunt u gemakkelijk data uit een of meer tabellen ophalen en combineren. Daarom kan een database worden gezien als een verzameling tabellen. Als u de logica achter deze relationele databases begrijpt, kan het u helpen je eigen gegevens op de juiste manier te structureren.

- Een database bevat een of meer tabellen met informatie;
- De rijen in een tabel worden *records* genoemd;
- De kolommen in een tabel worden velden of attributen genoemd;
- Een database die twee of meer gerelateerde tabellen bevat, wordt een relationele database genoemd.

In deze handleiding gaan we niet te veel in op het maken van een volledig functionerende relationele database, maar geven we eerder wat basiskennis en regels. Houd bij het verzamelen of organiseren van je data deze regels in gedachte. We zullen de regels in verschillende stappen uitleggen en eindigen met een praktijkvoorbeeld in Excel.

Stap 1. Begin met een schone en gestructureerde tabel met informatie

Stel je voor dat je verantwoordelijk bent voor het bijhouden van alle boeken die worden uitgeleend in een bibliotheek. Je vult in Tabel 3 elke keer in wanneer iemand een boek leent om alle belangrijke informatie bij te houden.

Deze tabel voldoet aan de basisbehoefte om bij te houden wie welk boek heeft geleend, maar heeft wel een aantal ernstige tekortkomingen in termen van efficiëntie, benodigde ruimte en onderhoudstijd.

Problemen met deze tabel:

- *Het steeds opnieuw in moeten voeren van bestaand informatie*
  Wanneer Bob na verloop van tijd meer boeken leent, moet je al zijn contactgegevens voor elk boek opnieuw invoeren. Dit is tijdverspilling, aangezien je zijn contactgegevens al hebt ingevoerd. Daarnaast vergroot het ook de kans op fouten.

- *Alle rijen moeten bij worden gewerkt als er een wijziging optreedt*
  Wanneer een update nodig is (bijv. Bob zijn telefoonnummer verandert), moeten alle records van Bob worden gelokaliseerd en gecorrigeerd. Als een van de records van Bob een ander telefoonnummer heeft dan de rest, is het dan een correctie, een record die over het hoofd is gezien tijdens de laatste update of een fout bij het invoeren van gegevens?

- *Veel dubbele informatie*
  Voor elk boek wordt de auteur en het jaar waarin het is gepubliceerd ingevuld. Dit is dubbele informatie en vergroot opnieuw de kans op fouten.

Deze problemen kunnen worden verminderd door onze data te **normaliseren** - met andere woorden, de informatie in meerdere tabellen onder te verdelen met als doel "een plaats te hebben voor alles en alles op zijn plaats". Elk stukje informatie hoeft maar één keer te verschijnen, waardoor het onderhouden van je data vereenvoudigd wordt en de benodigde opslagruimte afneemt.

Redenen voor normalisatie:

1. Minimaliseer dubbele gegevens
2. Minimaliseer of vermijd problemen met het wijzigen van gegevens
3. Vereenvoudig het extraheren van belangrijke gegevens

Voor-naam	Achter-naam	Adres	Telefoon	Titel	Auteur	Jaar	Inlever datum
Bob	Smith	123 Main St.	555-1212	Don Quixote	Miguel Cervantes	1605	14-07-2020
Alicia	Peterson	136 Oak St.	555-1234	Three Men in a Boat	Jerome K. Jerome	1889	16-07-2020
Bob	Smith	123 Main St.	555-1212	Things fall Apart	Chinua Achebe	1958	15-08-2020
Bob	Smith	123 Main St.	555-1212	Anna Karenina	Leo Tolstoy	1873	15-08-2020
Zayn	Murray	248 Pine Dr.	555-1248	Heidi	Johanna Spyri	1880	17-08-2020
Bob	Smith	123 Main St.	555-1212	The Old Man and the Sea	Earnest Hemingway	1952	10-09-2020

Tabel 3. Originele tabel

Stap 2. Creëer unieke rijen in je tabel, voorkom dubbele informatie

Om te voorkomen dat dubbele gegevens worden verzameld en er minder ruimte is voor fouten, zijn er enkele basisrichtlijnen die je kunt volgen:

Basisrichtlijnen

- *Elke tabel heeft een **primary key** (primaire sleutel)*
  De primary key is de waarde van een of meer kolommen waarvan de waarden uniek zijn in deze tabel, en kunnen dus worden gebruikt om verschillende rijen te identificeren, bijv. bestelnummer, personeelsnummer, factuurnummer. De primary key is de kolom waarmee je verschillende informatietabellen correct kunt verbinden.

- *Tabellen mogen geen subkolommen bevatten*
  Je kunt niet meerdere steden in één kolom weergeven en ze scheiden met een puntkomma. De waarde "Chicago" is bijvoorbeeld gewenst; terwijl “Chicago; Los Angeles; New York” dat niet is.

- *Een tabel mag geen herhalende kolom-groepen bevatten*
  Voorbeelden zijn Customer1Name, Customer2Name en Customer3Name. Hierin wordt de kolom groep ‘klantnaam’ herhaald.

- *Als een kolom dubbele informatie bevat, moet deze informatie in een nieuwe tabel worden opgenomen*
  Als het adres van een persoon meerdere keren in een kolom wordt herhaald, is het gevoelig voor fouten en moet het in een aparte tabel worden geregistreerd (klantgegevens).

Voorbeeld

Volgens de bovenstaande regels kunnen we de data uit Tabel 3 in twee tabellen scheiden. In plaats van alles wat we weten over een klant te herhalen wanneer hij een boek leent, geven we elke klant een uniek ID en alleen de ID herhalen wanneer we die persoon willen linken aan een record in een andere tabel. Dit helpt om te laten zien welke records in de tabel *Klant* overeenkomen met welke records in de tabel *Leen*, met andere woorden, wie heeft welk boek geleend.

Klant tabel				
Klant ID	Voornaam	Achternaam	Adres	Telefoon
1	Bob	Smith	123 Main St.	555-1212
2	Alicia	Peterson	136 Oak St.	555-1234
3	Zayn	Murray	248 Pine Dr.	555-1248

Leen tabel					
Leen ID	Klant ID	Boek titel	Auteur	Jaar	Inleverdatum
1	1	Don Quixote	Miguel Cervantes	1605	14-07-2020
2	2	Three Men in a Boat	Jerome K. Jerome	1889	16-07-2020
3	1	Things fall Apart	Chinua Achebe	1958	15-08-2020
4	1	Anna Karenina	Leo Tolstoy	1873	15-08-2020
5	3	Heidi	Johanna Spyri	1880	17-08-2020
6	1	The Old Man and the Sea	Earnest Hemingway	1952	10-09-2020

Tabel 4. Scheiding van Klant & Leen tabel

Nu kunnen de twee tabellen aan elkaar worden gelinkt op basis van de Klant ID. In de tabel *Klant* is het veld Klant ID de primary key en daarom moeten de waarden uniek blijven. De waarde "2" kan bijvoorbeeld slechts op één record voorkomen - die van Alicia - en Alicia kan slechts één klant-ID hebben - "2".

Klant ID kan niet de primary key van de *Leen* tabel zijn, omdat zoals we zien Klant ID "1" meer dan eens voorkomt en dus niet uniek is. Logisch, want we willen dat de klanten meerdere

boeken kunnen lenen. Als Klant ID de primary key was voor de *Leen* tabel, zou elke persoon slechts één boek mogen lenen en daarna mag het nooit meer boeken worden uitgeleend.

We kunnen de titel van het boek niet de primary key maken, anders zouden we een soortgelijk probleem hebben: elk boek kan maar één keer worden geleend en daarna mag niemand het ooit nog eens lenen. We kunnen de inleverdatum ook niet als primary key gebruiken, anders kan er elke dag maar één boek geleend worden. Aangezien geen van de bestaande velden als primary key gebruikt kan worden, wordt een nieuw veld toegevoegd om elke record met de naam Leen ID te identificeren.

Stap 3 – Kolommen met betrekking tot primaire sleutels

De primary key wordt gebruikt om elke rij in een tabel uniek te identificeren. Als we het hebben over kolommen die afhankelijk zijn van de primaire sleutel, bedoelen we dat om een bepaalde waarde te vinden, zoals hoeveel uur per week je werknemer Kris werkt, je eerst de primary key moet weten, zoals een Werknemer ID, om het antwoord op te kunnen zoeken. Als alle kolommen betrekking hebben op de primary key, hebben ze natuurlijk een gemeenschappelijk doel, zoals het beschrijven van een werknemer.

Zodra je het doel van een tabel hebt vastgesteld, kijkt je naar elk van de kolommen van de tabel en vraag je je af: **"Beschrijft deze kolom wat de primary key identificeert?"**

- Als je 'ja' antwoordt, is de kolom afhankelijk van de primary key en hoort deze in de tabel;
- Als je "nee" antwoordt, moet de kolom naar een andere tabel worden verplaatst.

Voorbeeld

Als we naar onze *Leen* tabel in Tabel 4 hierboven kijken, zien we dat er nog wat aanvullende informatie (auteur, jaar) wordt verzameld over een boek, terwijl dit niet het doel van de tabel is. Het doel van de *Leen* tabel is puur en alleen het loggen van wie welk boek heeft geleend. Daarom maken we in Tabel 5 hieronder een nieuwe tabel met de naam *Boek*. Nu hebben we de optimale databasestructuur voor de gegeven dataset die aan alle richtlijnen voldoet. Deze structuur maakt snelle analyses mogelijk en minimaliseert de kans op fouten.

Klant tabel				
Klant ID	Voornaam	Achternaam	Adres	Telefoon
1	Bob	Smith	123 Main St.	555-1212
2	Alicia	Petersohn	136 Oak St.	555-1234
3	Zayn	Murray	248 Pine Dr.	555-1248

Boek tabel			
Boek ID	Boek titel	Auteur	Jaar
1	Don Quixote	Miguel Cervantes	1605
2	Three Men in a Boat	Johanna Spyri	1880
3	Things fall Apart	Chinua Achebe	1958
4	Anna Karenina	Leo Tolstoy	1873
5	Heidi	Johanna Spyri	1880
6	The Old Man and the Sea	Earnest Hemingway	1952

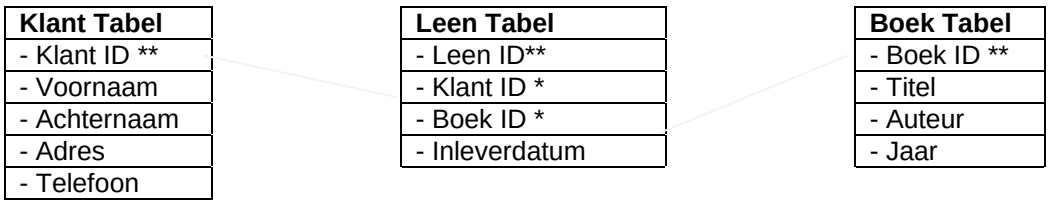
Leen tabel			
Leen ID	Klant ID	Boek ID	Inleverdatum
1	1	1	14-07-2020
2	2	2	16-07-2020
3	1	3	15-08-2020
4	1	4	15-08-2020
5	3	5	17-08-2020
6	1	6	10-09-2020

Tabel 5. Optimale databasestructuur

Stap 4. Hoe krijg je toch een duidelijk overzicht van al je data?

Diagram

Stel je voor dat er veel verschillende tabellen in je database staan, dan wil je waarschijnlijk een duidelijk overzicht krijgen van al deze tabellen en hoe ze gerelateerd zijn. We kunnen een diagram maken dat deze relaties illustreert en de **primary** (**) en **foreign** (*) keys toont. Foreign keys zijn primary keys van een andere tabel, zoals Klant ID in de *Leen* tabel. Dit is vooral handig als je veel kolommen in veel verschillende tabellen verzamelt.



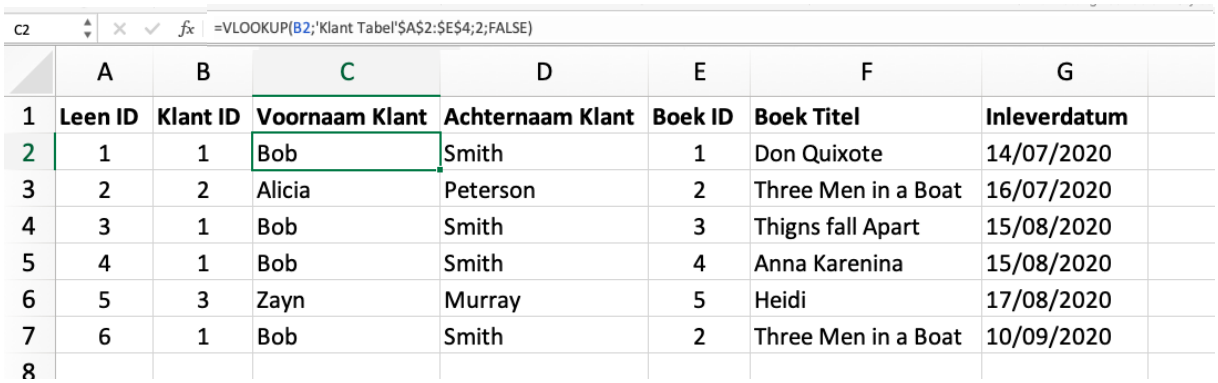
Figuur 1. Diagram van Relationale Database

Overzichtstabel

Hoewel Tabel 5 de beste databasestructuur visualiseerd, is het niet mogelijk om in één oogopslag te zien welke klant welk boek heeft geleend in de *Leen* tabel. Mocht je dit belangrijk vinden, is het niet per se een slechte zaak om één tabel te hebben die alle relevante informatie bevat. In plaats van deze informatie handmatig in te vullen, is het echter mogelijk om deze zodanig te koppelen aan de tabel *Boek* of *Klant* dat deze automatisch wordt bijgewerkt in alle huidige en eerdere velden wanneer er iets verandert in een van de onderliggende tabellen.

Excel Voorbeeld

Stel je voor dat de *Klant* tabel, de *Boek* tabel en de *Leen* tabel allemaal verschillende tabbladen zijn in Excel en dat je graag een leesbaar overzicht van de *Leen* tabel wilt hebben.



	A	B	C	D	E	F	G
1	**Leen ID**	**Klant ID**	**Voornaam Klant**	**Achternaam Klant**	**Boek ID**	**Boek Titel**	**Inleverdatum**
2	1	1	Bob	Smith	1	Don Quixote	14/07/2020
3	2	2	Alicia	Peterson	2	Three Men in a Boat	16/07/2020
4	3	1	Bob	Smith	3	Thigns fall Apart	15/08/2020
5	4	1	Bob	Smith	4	Anna Karenina	15/08/2020
6	5	3	Zayn	Murray	5	Heidi	17/08/2020
7	6	1	Bob	Smith	2	Three Men in a Boat	10/09/2020
8							

Figuur 2. Overzichtstabel (in Excel)

Zoals je kunt zien, wordt er een formule gebruikt om de naam van de klant en de boektitel te weergeven. Deze handige Excel formule wordt hieronder uitgelegd. Door de formule te gebruiken, kunnen we wijzigingen aanbrengen in de tabel *Boek* en de tabel *Klant* (in een apart tabblad in dezelfde file opgeslagen) die automatisch worden bijgewerkt in deze overzichtslijst.

Als je echter veel gegevens hebt, kan dit een zeer tijdrovende werkwijze zijn. Je kunt dit dus het beste gebruiken als het geen invloed heeft op je kerntaken in Excel. Als het wel invloed heeft op je taken en je beschikt over een enorme hoeveelheid gegevens, is het misschien een idee om verder te kijken naar database beheersystemen zoals MySQl, PostgreSQL of MongoDB.

- Formule die wordt gebruikt:
  = VLOOKUP(B2; 'Klant Tabel'!$A$2:$E$4; 2; FALSE)
- Syntax:
  =VLOOKUP(lookup waarde, tabel, col_ kolomindex, [range_lookup])
  - lookup waarde - de waarde waarnaar moet worden gezocht in de eerste kolom (primary key) van een tabel
  - tabel - de tabel waaruit een waarde moet worden opgehaald
  - kolom_index - de kolom in de tabel waaruit een waarde moet worden opgehaald. In het bovenstaande voorbeeld geldt kolom A is 1, kolom B is 2, etc.
  - range_lookup - [optioneel] TRUE = geschatte match (standaard). FALSE = exacte match.

Data voorbereiden voor analyse

Hieronder volgt een korte samenvatting van de bovenstaande stappen, aangevuld met een nieuwe stap (stap 3). Deze nieuwe stap is gerelateerd aan het omzetten van tekstuele data naar numerieke data (voorkeur van computers).

Zorg ervoor dat je een kopie van de originele dataset gebruikt, zodat je geen wijzigingen aanbrengt in je originele data.

Stap 1 – Check voor missende waarden

Open je data in Excel en controleer of elke cel informatie bevat. Is er een kolom met veel ontbrekende waarden?

- Probeer de lege cellen met de juiste informatie te vullen
- Is de informatie onbekend, maar de kolom wel relevant? Vul een numerieke waarde dan met een 0 of het gemiddelde van de kolom.
    - Bevat een van de kolommen veel missende waarden, maar wil je weten of de informatie aanwezig is of niet? Wil je bijvoorbeeld weten of een klant een website heeft of niet, dan kunt u een kolom "Website beschikbaar?" Toevoegen. Als er een website is vul je de waarde 1 in, zo niet vul je 0 in. Hierdoor ga je perfect om met je ontbrekende data en haal je toch meer informatie uit de beschikbare data.
- Is de informatie onbekend en de kolom niet relevant? Bijvoorbeeld; de beschrijving van een product. Dan kun je de kolom verwijderen.

Stap 2 – Check voor inconsistenties

Is de verzamelde data consistent?

- Zijn telefoonnummers op dezelfde manier geschreven? (Denk aan +3161042455 versus 0611042455)
- Worden categorische gegevens consistent verzameld? Bijv. je kunt niet zowel Mortimer WE als Mortimer West End hebben, dit zou hetzelfde geformuleerd moeten worden
    - Je kunt de unieke waarden van een kolom in Excel controleren met behulp van de formule =SORT(UNIQUE(column)). Dit kan helpen om je inconsistenties te vinden en daarmee je data op te schonen.

Stap 3 – Creëer zoveel mogelijk numerieke waarden als mogelijk

Computers hebben moeite met het vinden van patronen in tekstuele data. Wil je bijvoorbeeld de prijs van een product voorspellen, dan is de totale omschrijving van het product niet erg handig. Wat we echter wel kunnen vragen, is of er een beschrijving is of niet. Deze vraag vertaalt zich onmiddellijk in een waar of niet waar antwoord, dat in cijfers kan worden geschreven als 0 (niet waar) en 1 (waar). Op deze manier verlies je niet alle informatie, maar is de data toch numeriek.

Voor categorische waarden, zoals landen, merken of geslacht, bevat de kolom vaak tekstwaarde zoals Nederland, Apple of Vrouw. Nogmaals, dit is moeilijk te begrijpen voor een computer. Het is ook mogelijk om deze data om te zetten in numerieke waarden zoals hierboven. Dit keer stellen we onszelf de vraag “Wonen ze in Nederland?”, “Is het merk Apple?”, “Is het geslacht vrouwelijk?”. Het antwoord op deze vragen is wederom niet waar (0) of waar (1). Een voorbeeld hoe u deze landen in Excel naar 0 of 1 kunt vertalen, wordt hieronder uitgewerkt. Dit process heet *one-hot-encoding*, ook wel *dummy encoding* genoemd.

- Let op: als je een geslacht enkel in man en vrouw uitdrukt, is het voldoende om enkel de vraag “Is het geslacht vrouwelijk?” of “Is het geslacht mannelijk?” op te nemen in je data. Namelijk, is de persoon geen vrouw (antwoord is 0), dan is deze automatisch man.

Excel voorbeeld

B2 =IF(A2=K2;1;0)

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Original column	Country_BE	Country_DK	Country_FR	Country_GE	Country_NL	Country_SW	Country_UK					
2	Germany	0	0	0	1	0	0	0			Belgium		
3	United Kingdom	0	0	0	0	0	0	1			Denmark		
4	France	0	0	1	0	0	0	0			France		
5	France	0	0	1	0	0	0	0			Germany		
6	Denmark	0	1	0	0	0	0	0			Netherlands		
7	Sweden	0	0	0	0	0	1	0			Sweden		
8	Netherlands	0	0	0	0	1	0	0			United Kingdom		
9	United Kingdom	0	0	0	0	0	0	1					
10	Sweden	0	0	0	0	0	1	0					
11	Germany	0	0	0	1	0	0	0					
12	France	0	0	1	0	0	0	0					
13	Belgium	1	0	0	0	0	0	0					

- In kolom A zien we de originele kolom waarin iemand zijn woonplaats wordt opgeslagen. Elke rij geeft een klant aan en de waarden in kolom A vertellen ons waar de klant vandaan komt.
- Rechts in kolom K zien we een lijst met unieke waarden van de landkolom (A), de originele kolom.
    - Deze unieke lijst wordt opgehaald met de formule = SORT(UNIQUE (A2: A20))
- De kolommen B t/m H geven aan of de klant uit deze rij uit een van die landen komt.
    - Kolom B stelt dus de vraag "Komt deze klant uit België of niet?"
    - De gebruikte formule is = IF(A2 = K2; 1; 0), wat zich vertaalt naar: Als de waarde in A2 (de oorspronkelijke landkolom, in dit geval Duitsland) gelijk is aan de waarde in K2 (België), noteer een 1, als ze niet gelijk zijn, noteer dan een 0.
- Kun je de formule achterhalen die in cel F10 zou staan?

Stap 4. Sla je data op als CSV file

Zodra de meeste data numeriek is en alle cellen zijn gevuld met consistente gegevens, kan het bestand worden opgeslagen. Sla het bestand op als een .csv bestand (comma separated value bestand, een veelgebruikt bestandsformaat voor data analyse) en je bestand is klaar voor verkenning!

- In Excel: File > Save As >

Aan de slag samen met het JADS MKB Datalab

Wil jij graag samen met het JADS MKB Datalab toewerken naar een eerste stap op weg naar een meer datagedreven organisatie? Neem dan contact met ons op via ons e-mailadress of telefoonnummer. Wil je graag meer informatie over het JADS MKB Datalab? Neem dan een kijkje op onze website.

www.jadsmkbdatalab.nl info@jadsmkbdatalab.nl 0628525716