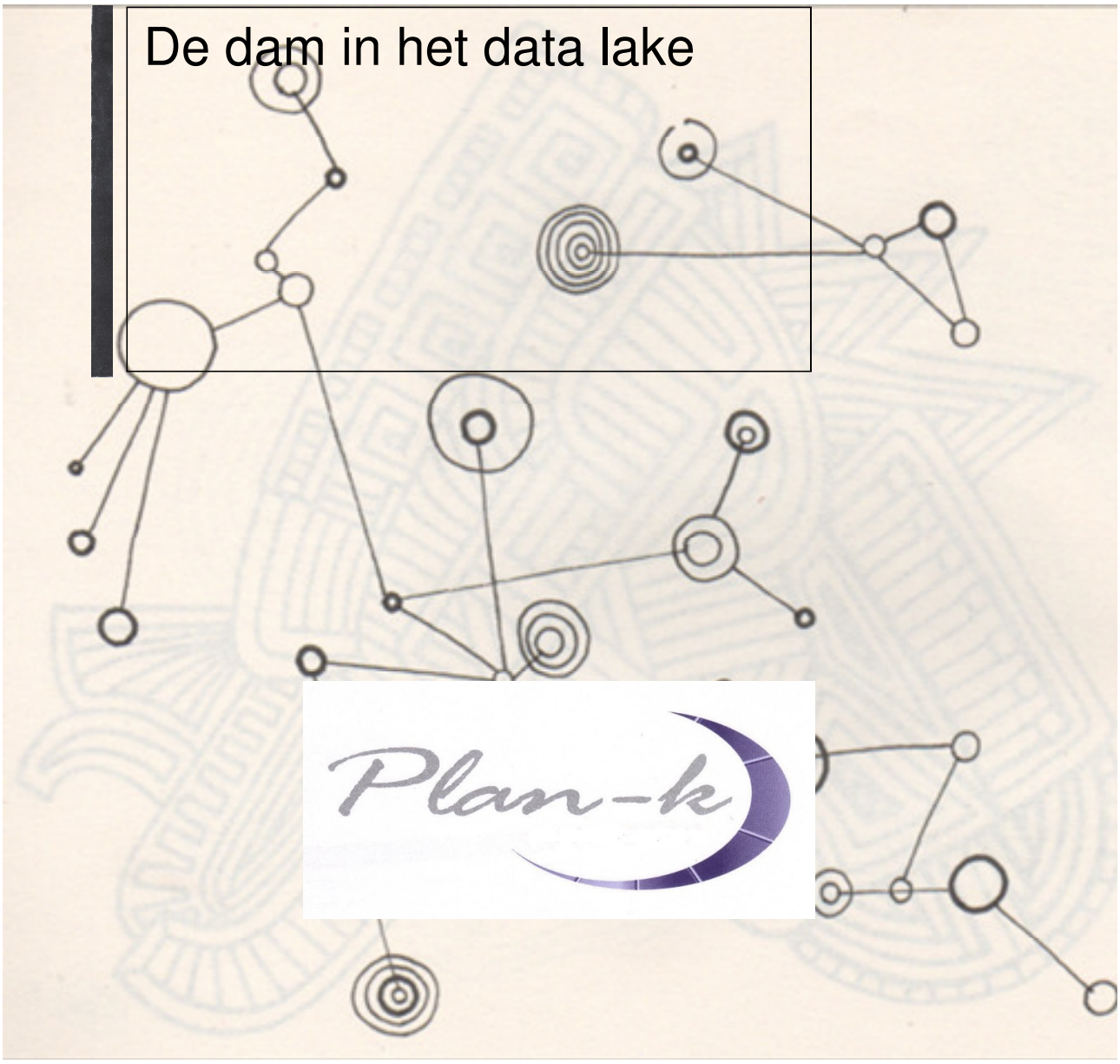
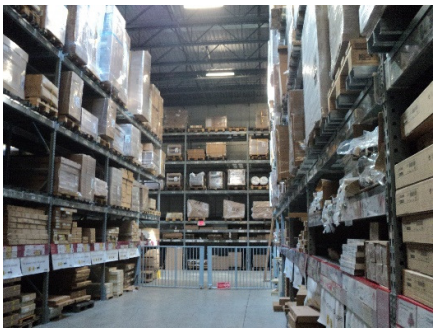
# De dam in het data lake



## Datawarehousing

Het is inmiddels een paar decennia geleden dat datawarehousing ontstond. Zoals je verwacht van een pakhuis worden daarin zaken -in dit geval data- opgeslagen op een gestructureerde manier en op zo'n manier dat we die terug kunnen vinden, wetende wat het is. Juist die structuur in een datawarehouse helpt ons interpreteren wat erin staat en is een sterke pijler onder het bereiken van het doel: inzicht in de materie (welke dat ook moge zijn).



De structuur van een datawarehouse beperkt ons echter ook: het vraagt dat brondata al een duidelijke structuur heeft en vaak moet die brondata ook nog ingedikt worden om goed en wel in een datawarehouse opgenomen te kunnen worden (deze stap is bekend als translatie). Alleen als we vooraf een helder beeld hebben van het doel waarmee we die gegevens verzamelen, kunnen we die op de juiste manier in het datawarehouse opslaan. Hierdoor zijn we feitelijk gedwongen om de analyse op de data uit te voeren voordat we die gegevens in het datawarehouse laden. Dit proces van gegevens uit de bron halen (extractie), bewerken voor opslag (translatie) en opnemen in het datawarehouse (laden) staat bekend als ETL.

## Datawarehousing na de digitale transformatie

De digitale transformatie heeft gezorgd voor een enorme toename in het gebruik van digitale middelen, gevoed door de exponentiële toename van de mogelijke toepassingen. Dat gebruik vraagt en levert veel gegevens op. Bovendien is het veel makkelijker geworden om te meten en observeren. En dat dan we dan ook veel en vaak.
Niet alleen de hoeveelheid gegevens neemt toe, ook het aantal vormen waarin data ontstaat neemt sterk toe. Niet langer hebben we vooral te maken met databases, maar ook met meer of minder gestructureerde gegevensvormen als mail, foto's, elektronische documenten en video. De architectuur van een datawarehouse is niet voldoende in staat dit soort gegevens te bevatten.

## Data lakes

Om invulling te geven aan de moderne informatiebehoefte is het concept data lake ontstaan. Hierbij werd de gegevensstructuur zoals die in datawarehouses zit losgelaten, zodat ook ongestructureerde gegevens opgeslagen kunnen worden. De translatie-stap die we kennen uit de datawarehousing wordt hierbij nagenoeg volledig overgeslagen; de gegevens stromen

van de bron direct het data lake in, net zoals water een meer in stroomt. De data houden de originele vorm. Het is dus zeker niet zo dat de inhoud van een data lake homogeen is, zoals dat voor een meer met water wel geldt. In die zin is de term data lake misleidend.
Dat gebrek aan structuur in gegevens is zowel de kracht als de zwakte van data lakes. Niets gaat verloren in de vertaling, maar interpretatie van de gegevens in het lake vraagt wel iemand die kennis heeft van data en van tooling waarmee een analyse uitgevoerd kan worden.
Ter vergelijk: de dagelijkse gebruiker van een datawarehouse is een business professional,

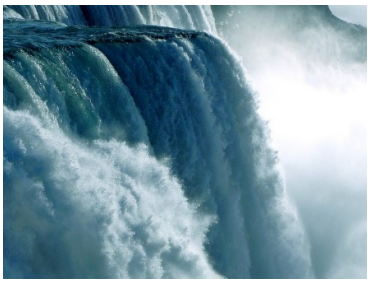
die van een data lake is een data scientist. De reden hiervoor is dat de translatie pas plaatsvindt op het moment dat gegevens uit het data lake gebruikt worden. Een data lake stelt zowel eisen aan de omvang van de opslag, als aan de wijze van opslag, waarbij metadatering en realtime translatie een belangrijke rol spelen.

## Privacy kopje onder


Een extra reden om niet zomaar gegevens in het data lake te laten vloeien is privacy-bescherming; persoonsgegevens mogen onder de AVG niet zomaar opgeslagen worden. En juist het feit dat een data lake opgezet wordt om ongestructureerde data op te slaan maakt het lastig om vast te stellen of gegevens onder die wetgeving opgeslagen mogen worden. Om compliance te kunnen garanderen is daardoor eigenlijk een analyseslag op gegevens noodzakelijk, voordat ze het meer in kunnen stromen. Het maken van afspraken met aanleverende partijen of het gebruik van machine learning zouden hierin kunnen helpen. Het ongecontroleerd binnen laten stromen van data levert echter een risico op dat beheerst moet worden.

## Zondvloed

Een van de nadelen van een data lake is de enorme hoeveelheid data die hierin verzameld kan worden; het is mooi om “alles” te hebben in je data lake, maar hoe voorkom je overspoeld te worden door de zondvloed? Bovendien brengt het bewaren van een grote hoeveelheid data nu eenmaal meer kosten met zich mee dan het bewaren van een dataset die gericht is op een specifiek doel. Ook is het zo dat de analyse op gegevens iemand vraagt die zowel kennis heeft van data als van het gebruik van de tooling waarmee die data geanalyseerd worden.



## Bezint eer ge begint

Uit bovenstaande is een aantal overwegingen te halen die kunnen bepalen of een informatiebehoefte een data lake vraagt of dat een datawarehouse die kan invullen:

- **Behoefte aan flexibiliteit**. Datawarehousing stuurt in sterke mate aan op het maken van een schifting in de gegevens voordat ze opgeslagen worden. Als vooraf helder is wat de informatiebehoefte is, hoeft dit geen probleem te zijn, maar dit leidt hoe dan ook tot een sterke beperking van de flexibiliteit in het beantwoorden van informatievragen. Een data lake houdt wat dat aangaat alles open.
- **Beschikbaar budget**. De omvang van een data lake zal veel groter zijn dan dat van een datawarehouse. Weliswaar worden data lakes ingericht op het beheersen van de kosten die hieruit voortkomen, maar desondanks zullen de totale lopende kosten meestal aanzienlijk groter zijn dan bij een datawarehouse, ook omdat het dagelijks gebruik ervan meer kennis en vaardigheden vraagt dan het gebruik van een datawarehouse. De potentie van een data lake is groter (aangezien er meer mogelijkheden zijn), tegelijkertijd is er minder inzicht in

welke resultaten behaald kunnen worden en de mate waarin die resultaten de investering rechtvaardigen.

- **Analysecapaciteit**. Bij de inrichting van een datawarehouse kan een aantal standaardrapportages of -kubussen gedefinieerd worden die relatief weinig onderhoud vragen en die dagelijks gebruikt kunnen worden zonder dat veel technische kennis nodig is; alleen tijdens de inrichting is die kennis nodig. Bij een data lake zorgt het gebrek aan structuur in de gegevens ervoor dat voor de beantwoording van bijna iedere vraag de kennis van een data-analist nodig is.
- **Structuur van brondata**. Datawarehouses bestaan bij de structuur van de gegevens die erin opgeslagen zitten. Als een aanzienlijk deel van de relevante brondata ongestructureerd is (zoals beeld- of geluidsmateriaal), dan zal het vaak niet haalbaar zijn die in een datawarehouse op te slaan.

## Dam that lake

Er zijn zowel argumenten voor een oplossing op basis van een data lake als op basis van een datawarehouse aan te dragen. Het onderscheid hoeft echter niet zo hard te zijn; hoewel een data lake geschikt is om ongestructureerde gegevens te bevatten, is er geen enkel bezwaar tegen structuur aan te brengen in het meer. Op deze wijze kan gezocht worden naar een oplossing waarin flexibiliteit gecombineerd wordt met toegankelijkheid. Bijvoorbeeld door een deel van de gegevens te structureren en te gebruiken als basis voor standaardrapportages, terwijl tegelijkertijd de gegevens beschikbaar zijn voor diepere analyse. Waarbij opgemerkt kan worden dat altijd de afweging gemaakt moet worden of investeringen in data wel de potentie hebben zichzelf terug te verdienen. Data zijn geduldig, net als data-experts. Maar ze kosten wel geld.

## Tot slot

Hoe meer gegevens, hoe groter de kans dat je daaruit zakelijke kansen kunt extraheren. Organisaties als facebook, Amazon en Google hebben enorm veel data tot hun beschikking en zullen bij de analyse daarvan ongetwijfeld interessante informatie boven tafel krijgen. Hun financiële armslag zorgt ervoor dat ze de keuze kunnen maken om serieus te investeren in dit soort ontwikkelingen, zelfs als vooraf niet duidelijk is wat het op zal leveren. Voor kleinere organisaties ligt het doen van zo'n investering veel minder voor de hand. Data biedt kansen, de grootste voor de grootverzamelaars met diepe zakken.

Roberto Lambooy