

# Data Foundations

## Hoofdstuk 3

### Data analytics

Hassan Haddouchi



**Waaraan denken jullie bij data analytics (data analyse)?**

# Overzicht

- Data Analytics
- Volume – Data Opslag
- Velocity – Data Verwerking
- Variety – Data Structuur en Types
- Veracity – Data Integriteit
- Value – Business Intelligence
- Labo

Extra lectuur: zie [Data Analytics Fundamentals](#) cursus van AWS

# Data analytics

Data Analytics zijn alle methodes die nodig zijn om data te analyseren.

Organisaties gebruiken data analytics om belangrijke vragen te beantwoorden en beslissingen te nemen.

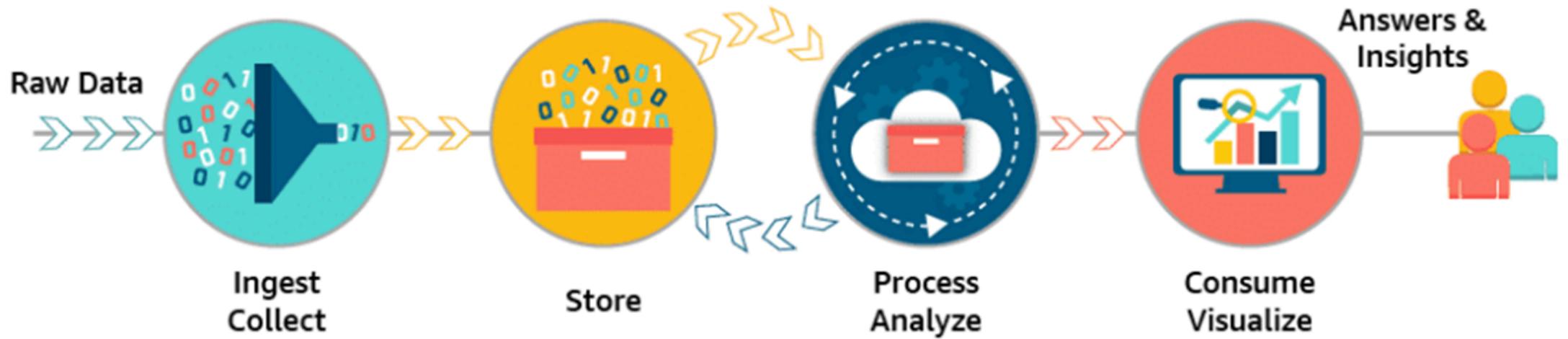
Zonder data en data analytics zouden veel beslissingen worden genomen op basis van ons buikgevoel.

Toepassingen van data analytics noemt men data producten.

# Toepassingen

- Personalisaties van klanten
- Gebruikersgedrag
- Realtime waarschuwingen
- Fraudetectie
- Financiële modellering en prognose
- Detectie van beveiligingsbedreigingen

# Componenten van data analytics



## Ingest Collect

De data van verschillende bronnen (logs, IoT, ...) en verschillende types (semi- of (on)gesubjecteerd) verzamelen en injecteren.

## Store

Een goede data infrastructuur laat veilige schaalbare en duurzame opslag toe.  
Voorbeelden: databases, een data warehouse of een data lake.

## **Process Analyze**

De data moet worden verwerkt en getransformeerd vooraleer nuttige operaties kunnen worden uitgevoerd. Voorbeelden: sorteren, aggregeren, joinen.

## **Consume Visualize**

Er zijn 2 manieren om data te consumeren: met een query of via BI tools. Query: goed voor een snelle analyse. BI tools zorgen voor visualisatie van rapporten en dashboards.

# Uitdagingen bij data

## De 5 V's



*Volume*



*Velocity*



*Variety*



*Veracity*



*Value*



**Volume:** de hoeveelheid data die wordt verwerkt – de totale grootte van de inkomende data.

**Velocity:** de snelheid waarmee de data binnenkomt. Steeds vaker wordt er naar een quasi real-time dataverwerking gestreefd.

Data kan van verschillende bronnen komen. **Variety** staat voor het aantal verschillende bronnen en de soorten bronnen.

**Veracity** staat voor hoe accuraat en betrouwbaar de data is. Dit wordt ook data integriteit genoemd.

**Value:** de mate waarin een data product erin slaagt om nuttige informatie te genereren.

# Belangrijke vragen m.b.t. data producten

1. Waar komt de data vandaan?

Databases, fileservers, streaming data, publieke datasets, ...

2. Wat zijn de opties om de data te verwerken?

Er is geen one-size-fits-all oplossing. Je kiest in functie van de noden om het gewenste resultaat te bekomen.

3. Wat wil je leren uit de data?

**Onder welke component van een data product valt het samenbrengen van verschillende databronnen?**

## **Case 1: welke V's vormen hier een uitdaging?**

Een bedrijf genereert elk uur 15 JSON files van elk 2.5 GB die op een file server worden gezet. Zodra ze klaar zijn moeten ze worden geïnjecteerd in het data analytics systeem. Deze data wordt gecombineerd met alle financiële transacties voor dezelfde periode en vergeleken met de aanbevelingen van de marketing engine. De data is reeds volledig klaargemaakt en betrouwbaar. De resultaten moeten telkens 10 minuten na elk uur worden beschikbaar gesteld aan het hogere management in de vorm van een financieel dashboard.

## **Case 2: welke V's vormen hier een uitdaging?**

Een organisatie brengt data samen die gegenereerd wordt door honderden bedrijven. Deze data wordt geleverd via verschillende kanalen: grote data files, logs van transacties en data streams. De data moet worden nagekeken en voorbereid zodat verkeerde inputs de resultaten niet vertekenen. Het kennen van de databron voor elk datapunt is cruciaal. Een groot deel van de verzamelde data is irrelevant en kan gewist worden. Het uiteindelijke doel is dat al de data gecombineerd wordt en opgeladen in het data warehouse, van waar ze verder geanalyseerd kan worden.

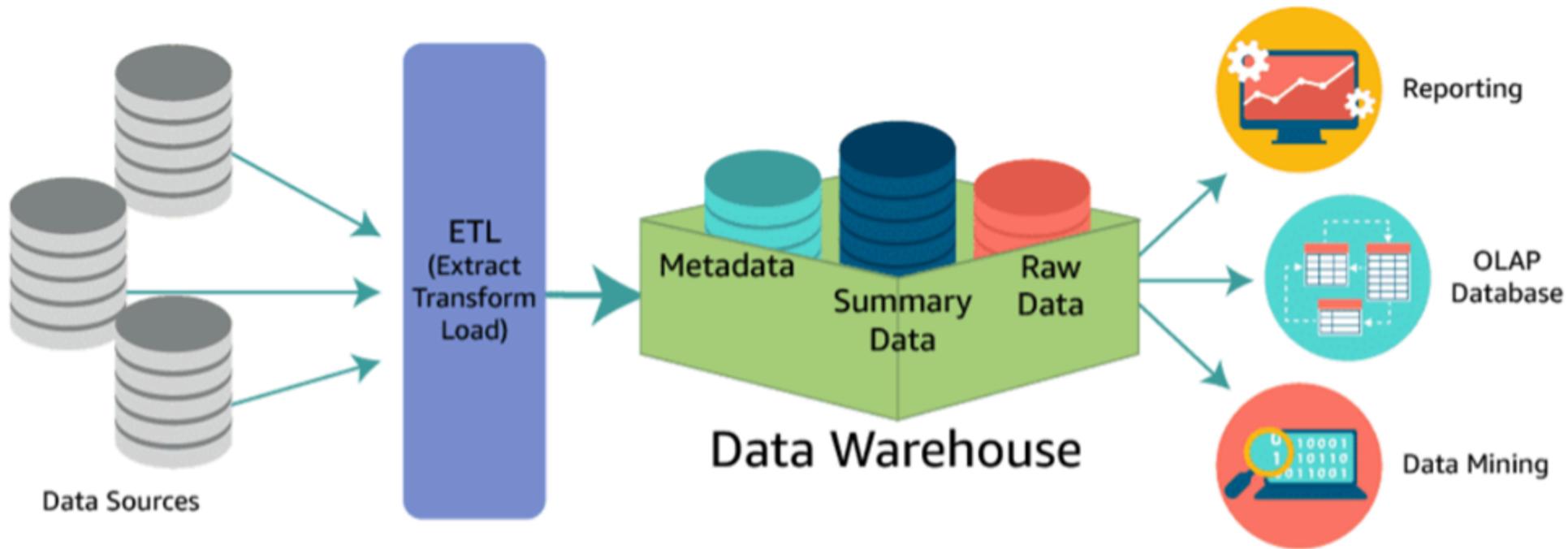
## Volume - data opslag

"Wanneer bedrijven meer data hebben dan ze kunnen verwerken en analyseren, hebben ze een volumeprobleem"

## **Elke dag worden datasets groter en diverser**

Er komen steeds nieuwe opties om data op te slaan als antwoord op deze uitdagingen. Traditionele oplossingen zoals een data warehouse zijn nog steeds populair en relevant maar de laatste tijd winnen data lakes aan populariteit.

# Data Warehouse



Een data warehouse, de onderliggende motor van data producten, is een centrale opslagplaats voor data die van één of meerdere databronnen komt (bv. transactie systemen, relationele databases, ...).

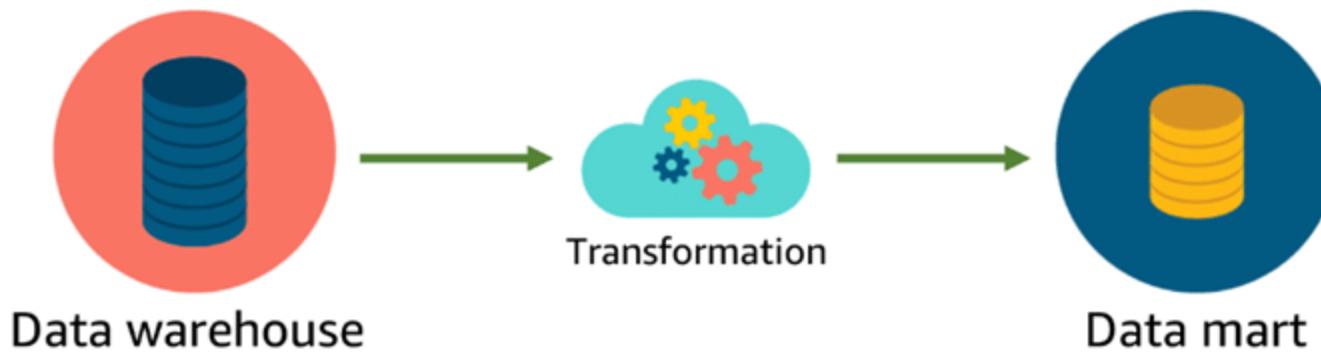
Deze databronnen worden getransformeerd naar gestructureerde data voor ze opgeslagen worden in het data warehouse.

Data wordt opgeslagen met een schema. Een schema definieert de structuur van de data met tabellen, kolommen en rijen en legt bepaalde beperkingen op om de data integriteit te waarborgen.

De data kan dan gebruikt worden via BI tools en andere data analytics toepassingen.

## Data Mart

Een data mart bevat een deel van de data van een data warehouse en concentreert typisch op 1 onderwerp of context.

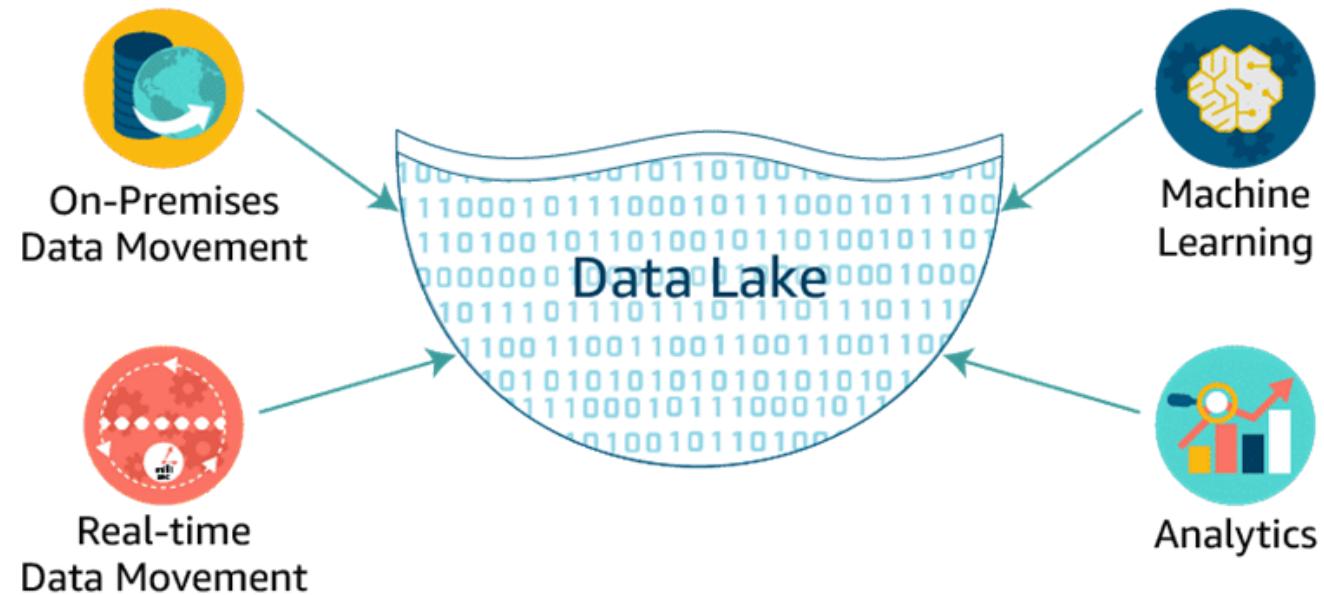


# Data warehouse

Sterktes	Zwaktes
snelle data toegang	Duur om te implementeren
Voorbereide datasets	Onderhoud kan een uitdaging zijn
Centrale opslag	Security is een uitdaging
Goede bron voor Business Intelligence	Moeilijk schaalbaar

# Data Lake

Een centrale opslagplaats die toelaat om alle soorten data van elke grootte op te slaan.



Data wordt verzameld van verschillende bronnen en in zijn originele formaat opgeslagen in het data lake.

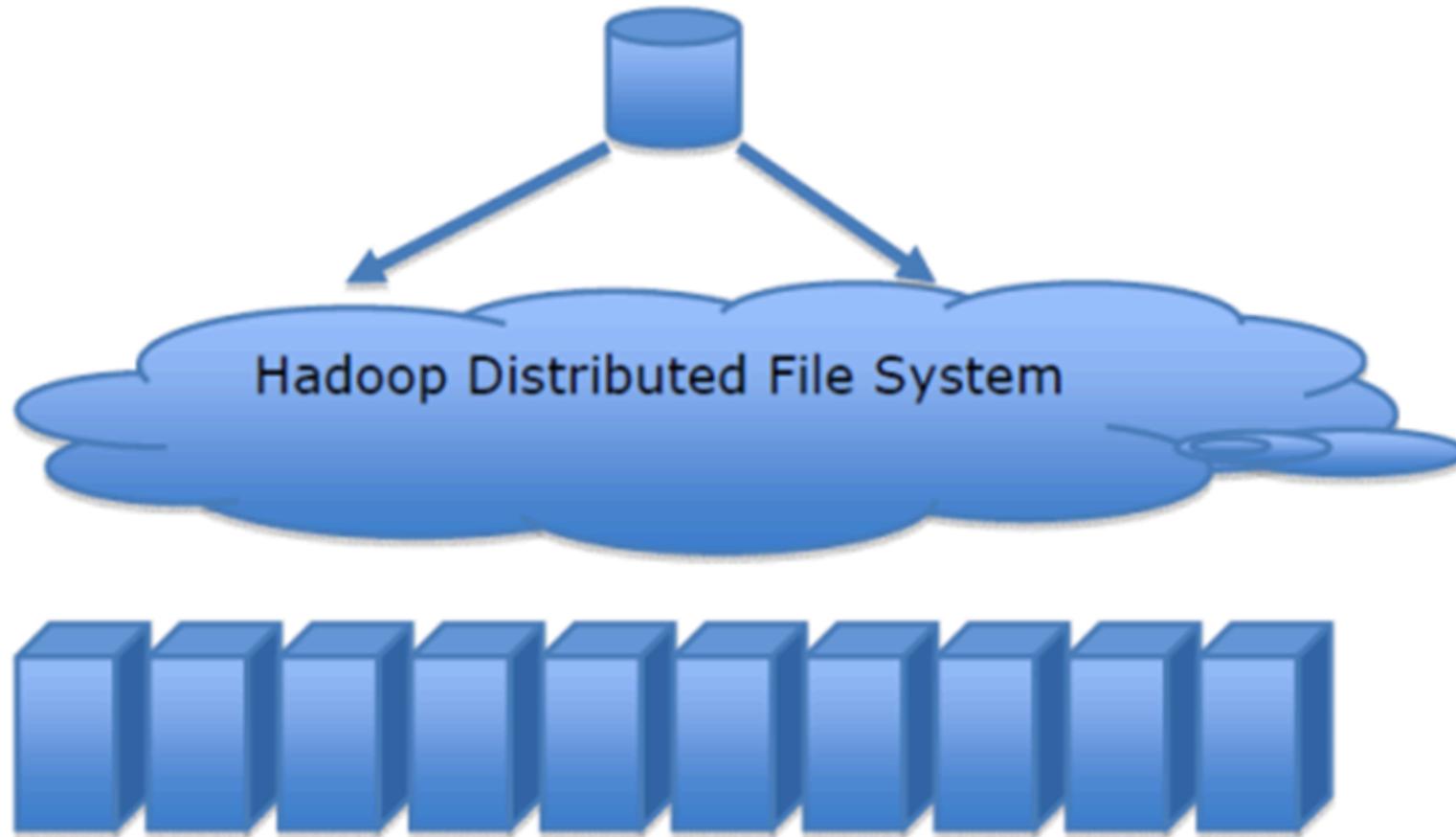
Een data lake laat toe om in real-time data te importen van verschillende databronnen.

Een data lake laat ook toe om inzichten te genereren zoals rapporten van historische data.

Je kan zowel relationele data als niet-relationele data (IoT devices, sociale media) opslaan.

# Gedistribueerde opslag

Om een quasi ongelimiteerde data opslag te voorzien worden de data in een data lake gedistribueerd over tientallen servers (bv. met Hadoop Distributed File System)



Eigenschappen	Data Warehouse	Data Lake
Data	relationele data	(niet-)relationale data en on- en semi-gestructureerde data
Schema	Tijdens implementatie	Opgesteld bij de data analyse
Prijs/performantie	Dure data / goede snelheid	Goedkope opslag
Data kwaliteit	Waarachtige data	Alle data die al dan niet is voorbereid
Gebruikers	Business analisten	Data scientists en business analysten
Analytics	Reporting, BI en visualisaties	ML en voorspellingen

**Stel dat je een gestructureerde dataset hebt die moet dienen als de referentie voor analytische queries.  
Welke soort data opslag is het beste?**

## Velocity - data verwerking en snelheid

"Als bedrijven snel inzicht nodig hebben in de data die ze verzamelen, maar de bestaande systemen eenvoudigweg niet aan de behoefte kunnen voldoen, is er sprake van een snelheidsprobleem."

**De snelheid waarmee data wordt gegenereerd blijft stijgen.**

Deze data moet worden verzameld, verwerkt, geanalyseerd en opgeslagen.

We concentreren op 2 uitdagingen:

- Het verzamelen van de data
- Het verwerken van de data

We maken een onderscheid tussen batch en streaming.

# Batch

Een batch data verwerking is de automatische uitvoering van een aantal programma's.

De data wordt eerst verzameld in batches en naar het verwerkingsysteem gestuurd als er bepaalde condities vervuld zijn (bv. een specifieke tijd).

De resultaten gaan naar een opslaglocatie waarvan ze verder verwerkt kunnen worden.

# Streaming

Steeds meer data wordt in real-time verzameld en hoe sneller de data wordt verwerkt hoe sneller de inzichten beschikbaar zijn.

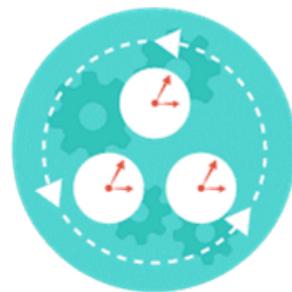
Het systeem dat de data verzameld wordt de **producer** genoemd en het systeem dat de verwerking doet is de **consumer**.

## Batch Processing

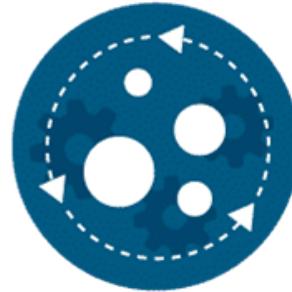


Large Bursts  
of Data

*Scheduled*



*Periodic*



## Stream Processing



Tiny Bursts  
of Data

*Near Real-Time*



*Real-Time*



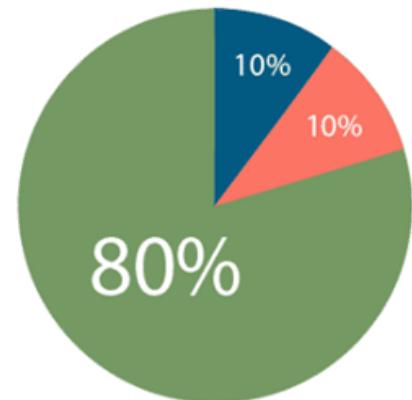
<b>Eigenschap</b>	<b>Batch data verwerking</b>	<b>Streaming data verwerking</b>
Data scope	Verwerking van alle data of een groot deel van de dataset	Data uit een bepaalde tijdsvenster
Data grootte	Grote data batches	Individuele data elementen
Latency	Minuten tot uren	Seconden tot microseconden
Analyse	Complexe analyses	Simple functies en aggregaties

## Variety - datastructuur en types

"Wanneer een bedrijf overweldigd wordt door het grote aantal databronnen dat geanalyseerd moet worden en je geen systemen kunt vinden om de analyses uit te voeren, dan heb je een diversiteitsprobleem."

# Databronnen

- Gestruktureerde data: tabellen en kolommen.
- Semi-gestruktureerde data: key-value paren.
- Ongestruktureerde data: foto's, video's, tekst, ...



Structured Data - 10% - Tabular Data

Semistructured Data - 10% - CSV, XML, JSON Files

Unstructured Data - 80% - Everything Else

## Gestructureerde data

Opgeslagen in tabellen, vaak met een database management system (DBMS). Kunnen ook als bestanden (flat files) zonder structuur worden opgeslagen.

Georganiseerd volgens een relationeel data model met relaties.

Nadeel: weinig flexibiliteit.

Voorbeelden: MySQL, PostgreSQL, MS SQL Server, Oracle, ...



Structured Data



Amazon RDS



Amazon Aurora



MySQL



MariaDB



PostgreSQL

SQL Server

ORACLE

## Semi-gestructureerde data

Opgeslagen in elementen in een bestand.

Georganiseerd volgens elementen en de bijhorende attributen.

Voordeel: flexibel en schaalbaar.

Voorbeelden: CSV, JSON, XML,  
Amazon Neptune, ...



**Semistructured Data**



## Ongestructureerde data

Opgeslagen als bestanden zonder op voorhand gedefinieerde structuur.

Voorbeelden: e-mails, foto's, video's, clickstream data, ...



**Unstructured Data**



## Flat files

- Ontbrekende waarden.
- Kennen Ahmed en John elkaar?
- Duplicaten in waarden.
- Onduidelijke waarden (Paris in Texas of in Frankrijk?)

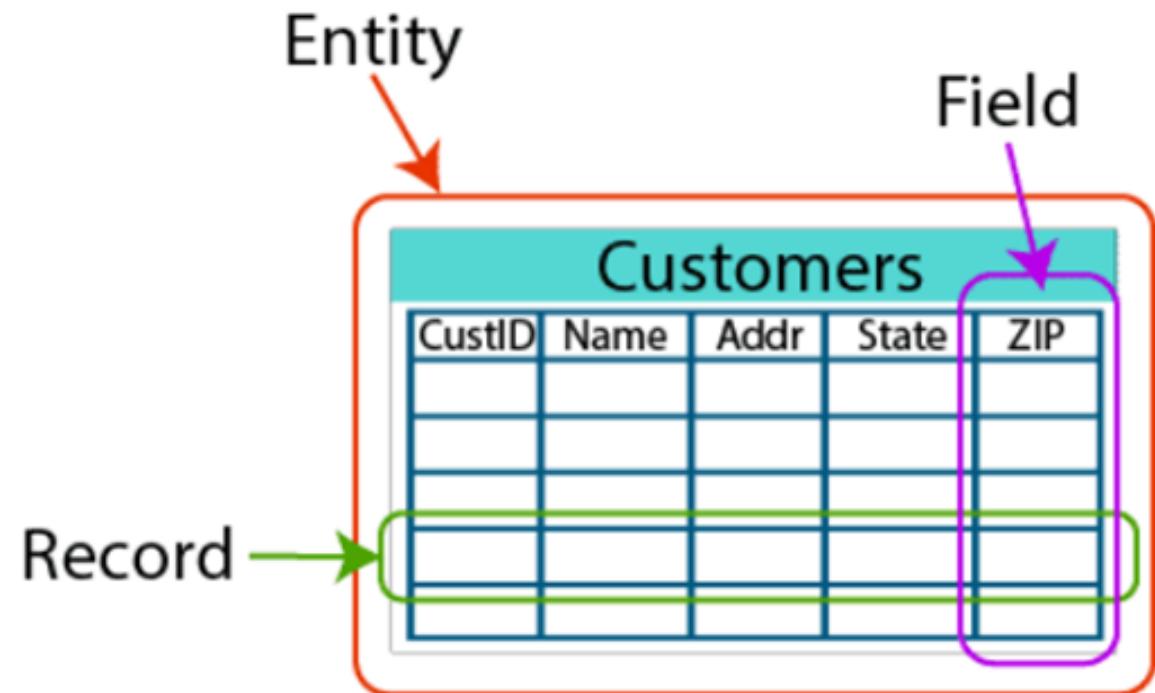
Full name	City	Age	Product	Team	Qty	Favorite
Maya Collier	Paris	34	Cap	Seahawks	2	-
Ahmed Aziz	Paris	29	Cap	Broncos	1	Yes
John Schafer	Lubuk	-	Socks	Broncos	1	Yes
John Schafer	Baltimore	-	Cap	Ravens	1	No
Isabell Hawthorn	Denver	-	Tshirt	Broncos	3	Yes
Willis Millar	Baltimore	29	Cap	Ravens	1	No

## Relationale database

Lost de uitdagingen op van flat files.

Relationale databases groeperen data in tabellen gebaseerd op objecten zoals personen of plaatsen.

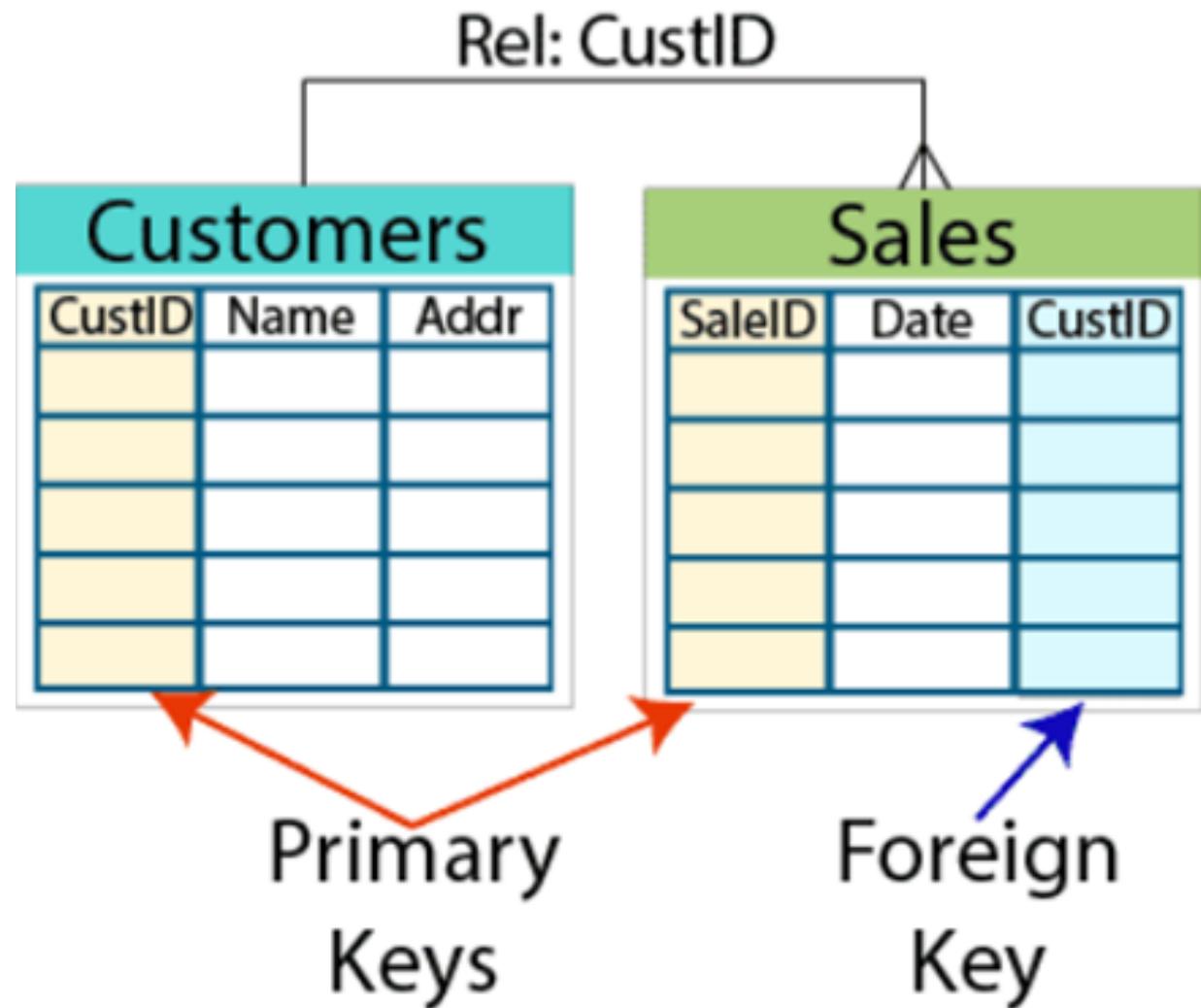
Kolommen geven eigenschappen weer, een rij (of een instantie, record) geeft een element van de groep weer.



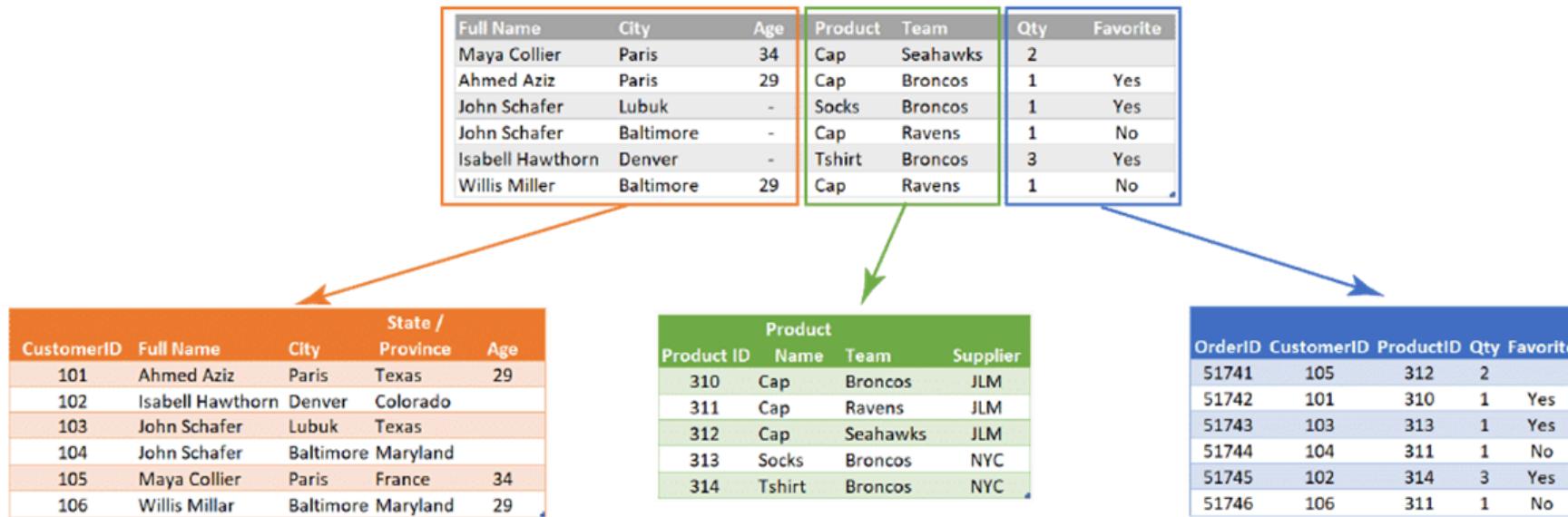
## Relationele database

Relaties worden weergegeven met keys.

Een foreign key is een veld dat de waarden van de primaire key in een andere tabel gebruikt en creëert een relatie.



# Van een flat file naar een relationele database



# OLTP vs OLAP

Twee algemene manieren om data te organiseren in een relationele database.

- Online Transaction Processing (OLTP): hier is snelheid van het inladen van data het voornaamste doel.
- Online Analytical Processing (OLAP) (data warehouses): met als doel snel data uitlezen en gebruiken.

OLAP databases worden vaak gevoed door OLTP databases om de data te organiseren voor analytische doeleinden.

Eigenschappen	OLTP	OLAP
Gebruik	Constant opladen van transacties	Periodische grote updates, complexe queries
Voorbeelden	Boekhouding database, online verkoop transacties	Rapportering, ondersteuning management
Type	Operationele data	Geconsolideerde data
Data behoud	Korte termijn (2-6 maanden)	Lange termijn (2-5 jaar)
Opslag	Gigabytes (GB)	Terabyte (TB)/petabytes (PB)
Gebruikers	Veel	Enkele
Bescherming	Robuuste data bescherming	Periodische bescherming

## Niet-relationele databases

Met documentopslag in niet-relationele databases worden semi-gestructureerde en ongestructureerde data opgeslagen in de vorm van files, zoals JSON en XML.

Sterktes	Zwaktes
Flexibiliteit	Geen ACID compliance
Data type moet niet op voorhand gekend zijn	Geen queries met verschillende files
Makkelijk schaalbaar	

## Key-value databases

Key-value databases slaan ongestructureerde data op in de vorm van key en value paren. De data wordt opgeslagen in een tabel en de waarden in de tabel komen overeen met een specifieke key. De waardes kunnen gelijk welk type zijn.

Sterktes	Zwaktes
Erg flexibel	Onmogelijk om waardes te queryen
Grote variatie aan data types mogelijk	Updates of aanpassingen zijn erg moeilijk
Keys kunnen makkelijk doorgegeven worden aan ander systeem	Niet alle objecten passen bij de key-value structuur

# Quiz

**Wat is één V van de 5 V's van data-analyse en wat wil deze zeggen?**

**Wat is het verschil tussen gestructureerde, semi-gestructureerde en ongestructureerde data?**

**Kan je van elke een voorbeeld geven?**

# **Wat is het doel van een data warehouse?**

# Labo

# Opzet

Je zal werken op Google Cloud om data-analyse processen te doorlopen en relevante tools te gebruiken.

We zullen opnieuw in 2 delen werken. Wanneer je elk deel tot een goed einde hebt gebracht zal je een skill badge op het platform van Google ontvangen.

Via Digitap kan je voor elk deel een screenshot indienen dat aantoont dat je de badge hebt verdiend. Een screenshot met het overzicht van alle voltooide stappen is ook toegestaan (zie voorbeeld op volgende slide).

Er zal voor elk deel een aparte uploadzone worden voorzien.

Dien voor elk deel van dit labo een screenshot zoals hieronder. *Let hieronder niet op de namen van de subtaken, die kunnen niet overeenkomen met de subtaken van dit labo.*

Google Cloud Skills Boost

**Exploring Your Ecommerce Dataset with SQL in Google BigQuery**  
In this lab, you learn to use BigQuery to find data, query the data-to-insights public dataset, and write and execute queries.  
★★★★★ 30 minutes Introductory Free

**Troubleshooting Common SQL Errors with BigQuery**  
In this lab, you use BigQuery to troubleshoot common SQL errors, query the data-to-insights public dataset, use the Query Validator, and troubleshoot syntax and logical SQL errors.  
★★★★★ 50 minutes Introductory Free

**Explore and Create Reports with Data Studio**  
In this lab, you learn how to connect Google Data Studio to Google BigQuery data tables, create charts, and explore the relationships between dimensions and measures.  
★★★★★ 40 minutes Introductory Free

Wim Casteels  
wim.casteels@ap.be  
200 Credits

Settings

Sign Out

Privacy · Terms

# Deel 1

## Introduction to Data Analytics on Google Cloud

Via [deze link](#) kan je het eerste deel van ons labo terugvinden.

In deze oefening zal je de basisprincipes leren van data-analyse op Google Cloud.

Je zal data verzamelen, opslaan, verkennen en visualiseren met Google Cloud-tools. Je zal kennismaken met data-analyse-tools zoals BigQuery en Looker.

## Deel 2

### Weather Data in BigQuery

In dit lab analyseer je historische weerswaarnemingen van NOAA met behulp van de tool BigQuery.

Einde hoofdstuk 3 - Data analytics