



Odissee
DE CO-HOGESCHOOL

Big Data - Introductie



Jens Baetens



Structuur

- ▣ Structuur van het vak
- ▣ Evaluatie
- ▣ Recap – Data Science
- ▣ Kenmerken van Big Data
- ▣ Distributed Storage
- ▣ Distributed Computing
- ▣ Tools

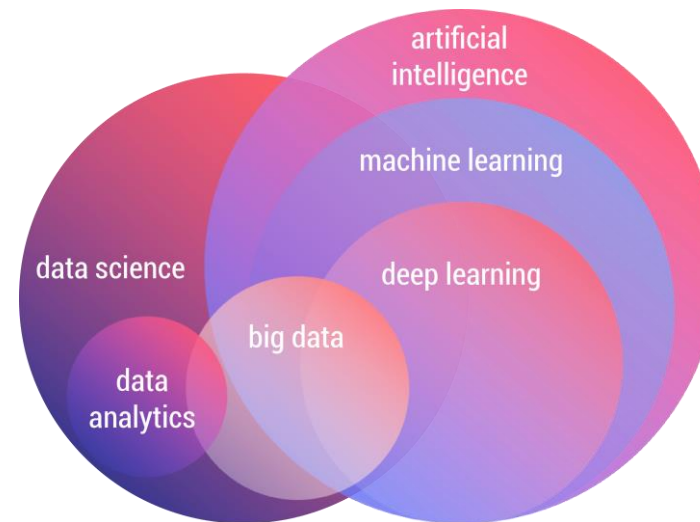


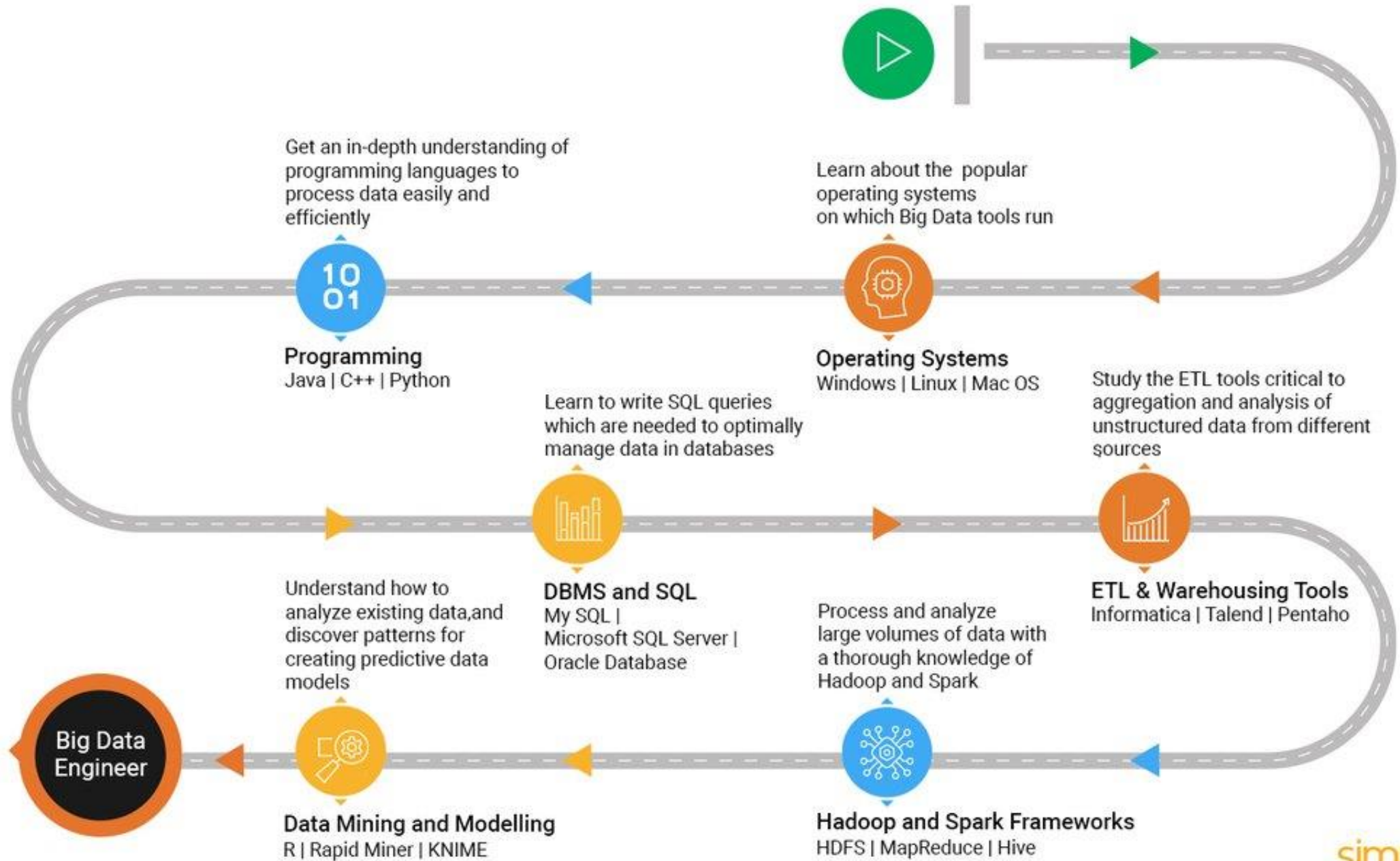


Structuur van het vak

Vakken keuzerichting

- ▣ Data Science – 5 studiepunten
- ▣ Big Data – 5 studiepunten
- ▣ Machine Learning – 6 studiepunten





Vacature Data Scientist KBC



Data Scientist (NL/ENG)

KBC Bank & Verzekering · Flemish Region, Belgium (Hybrid) 12 hours ago · 15 applicants

Together, we develop artificial intelligence and machine learning solutions that **transform our business**. We work on them using, among others, data platforms like HDFS, Spark and AWS.

What do we expect from you?

- You develop AI/ML solutions that transform our business. In order to roll them out successfully, you work closely with your colleagues in IT.
- You stay up to date with data science and AI.
- You help to strengthen the foundations within the KBC Group in the areas of knowledge and technology.
- You network with potential academic and commercial partners to set up partnerships.

- You have experience with data analytics and know how to use the more common tools for statistical analysis and machine learning.
- You have insight into learning algorithms, data-processing strategies, machine learning and natural language processing and know how to implement them in a business context.
- You're more than able when it comes to Python and data science toolsets (pandas, scikit-learn or H2O). Any experience you have with any of the following will be a benefit:
 - Learning frameworks (TensorFlow, PyTorch or Keras)
 - Big data analyses via cloud computing (AWS or MS Azure)
 - Apache Spark, Hadoop, Kubernetes
 - Cloudera
 - Git.

es

applicants. [Try Premium for free](#)

+8 more

- ▣ Eigenschappen van Big Data
- ▣ Mogelijke vormen van data
- ▣ Distributed Filesystemen
- ▣ Distributed Computing
- ▣ Cloud platforms



Verloop

- ▣ # lesblokken van 4 uur
 - ▬ online in het derde kwartaal (dinsdag voormiddag)
 - ▬ op campus in het vierde kwartaal (vrijdag namiddag)
- ▣ Evaluatie op basis van
 - ▬ 4 oefeningen
 - ▬ Project



Studiemateriaal

- ▣ Slides en voorbeeld code via github repository
- ▣ Opdrachten en project via Toledo / Github classroom
- ▣ Tip: Hou een goed overzicht bij van het Hadoop eco-systeem en waarvoor de verschillende geziene toepassingen gebruikt kunnen worden.



Afspraken

- ▣ Wees op tijd
- ▣ Vragen buiten de lessen mag steeds via Teams of mail
- ▣ Actief meewerken in de les beste leermethode



Evaluatie



Evaluatie

- ▣ 4 oefeningen: samen 50%
- ▣ Project met mondelinge verdeling: 50%
 - In groepen van max 2 studenten
 - Opgave volgt nog
 - Respecteer de tussentijdse deadlines
- ▣ Respecteer de deadlines, te laat = 0 voor die oefening
- ▣ Elkaar helpen mag maar oplossingen overnemen niet





Recap – Data Science

Data Lifecycle

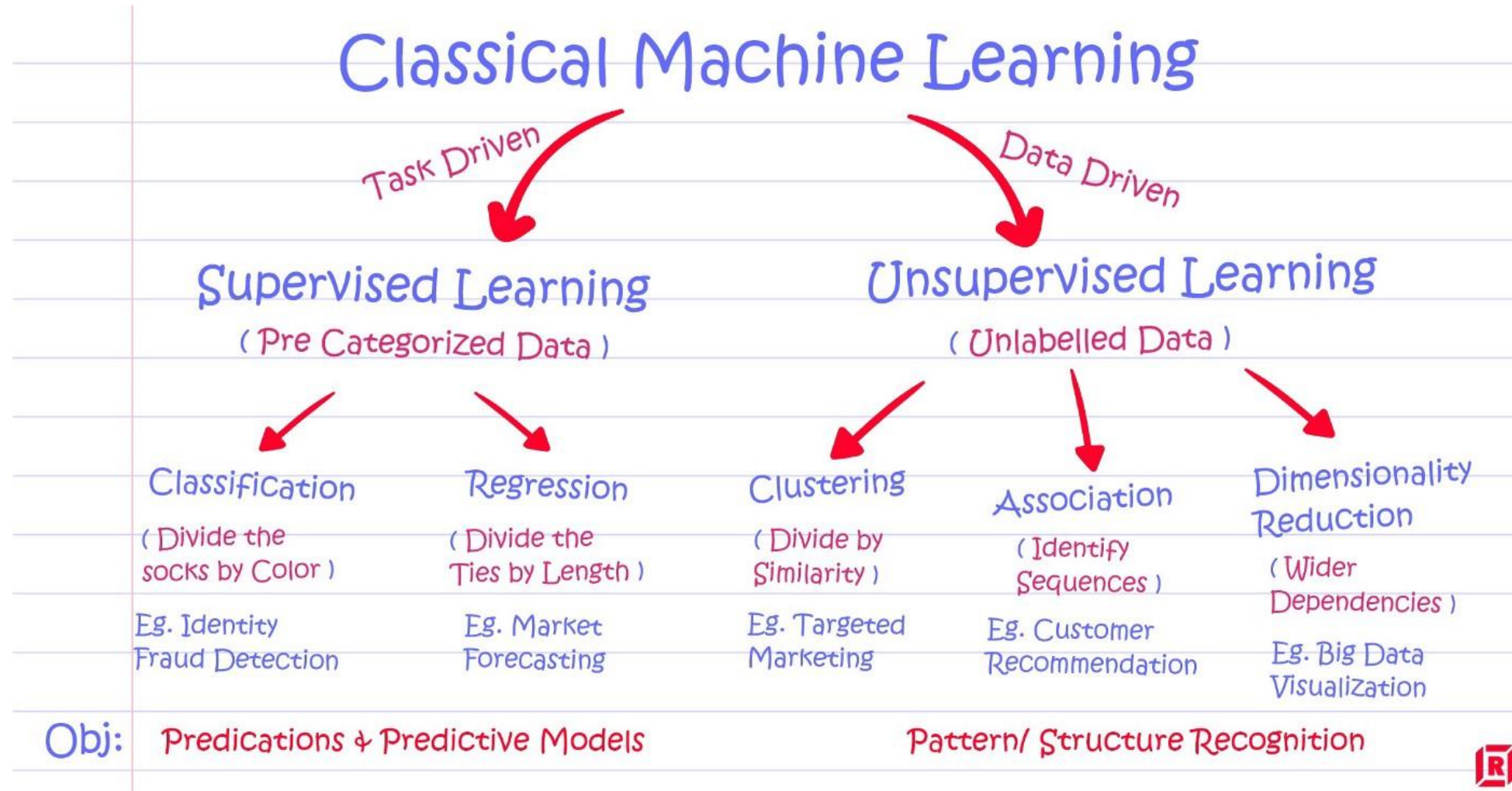




Data Cleaning & Exploration

- ▣ Bestuderen beschikbare datasets
 - ▬ Vinden van correlaties en verbanden
 - ▬ Informatie over de beschikbare data en hoe bruikbaar ze is
- ▣ Opschonen en bewerken van beschikbare data
 - ▬ Omzetten dataformaten (datums, bag of words, scaling ...)
 - ▬ Privacy van personen
 - ▬ Oplossen problemen in de data (typo's, vertalingen, ontbrekende data, ...)

Data Modelling



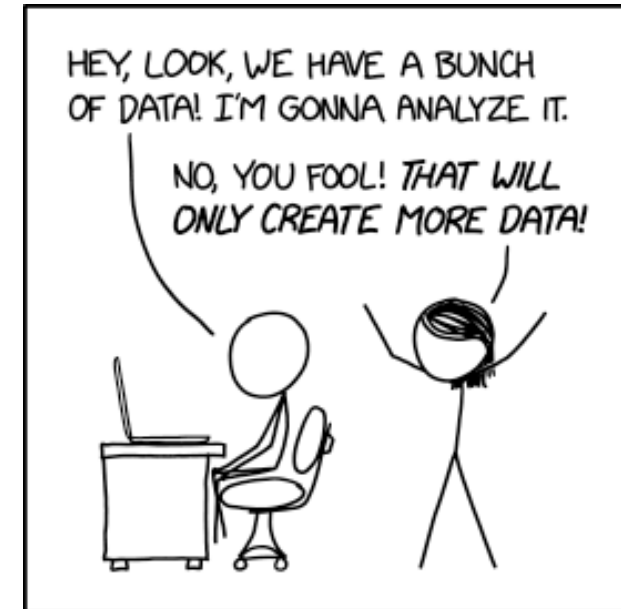


Gebruikte datasets

- ▣ Aantal honderden MB
- ▣ Csv of jpegs
- ▣ Gedownload naar harde schijf
- ▣ Volledig ingeladen in memory voor verwerking

2.

Is dit altijd mogelijk?



Zien jullie problemen?



Zien jullie problemen?

Ga naar **wooclap.com** en gebruik de code **DBEIAG**

Zien jullie problemen met de daarnet besproken methode?

Nee, te veel data

te veel of grote data

Comput er kan te zwakke specs

Nee, er is te veel

Door zwakke specs kan het

Ja, de data kan te groot zijn

Nee, size of processing

Download duurt te lang

onvoldoende

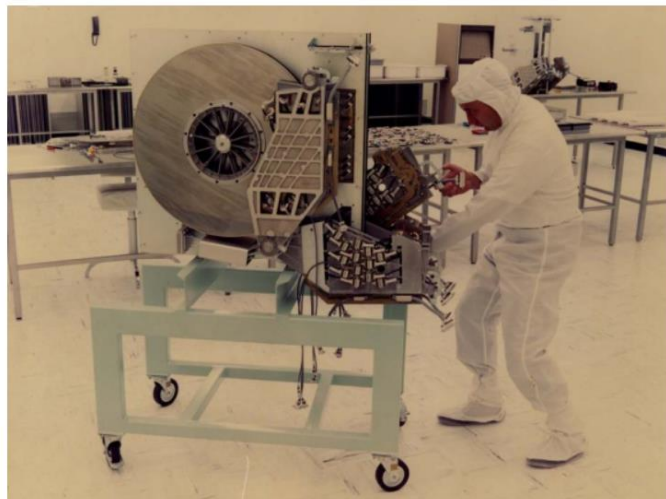
wooclap

16

Grootte harde schijven?



1956: 5 MB



1975: 250 MB



1988: 1 GB



2019: 1 TB

Prijs?

Backblaze Average Cost per Drive Size

By Quarter: Q1 2009 - Q2 2017



A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

500m

tweets are sent every day

Twitter



4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

DEMYSTIFYING DATA UNITS

From the more familiar 'bit' or 'megabyte', larger units of measurement are more frequently being used to explain the masses of data

Unit	Value	Size
b bit	0 or 1	1/8 of a byte
B byte	8 bits	1 byte
KB kilobyte	1,000 bytes	1,000 bytes
MB megabyte	1,000 ² bytes	1,000,000 bytes
GB gigabyte	1,000 ³ bytes	1,000,000,000 bytes
TB terabyte	1,000 ⁴ bytes	1,000,000,000,000 bytes
PB petabyte	1,000 ⁵ bytes	1,000,000,000,000,000 bytes
EB exabyte	1,000 ⁶ bytes	1,000,000,000,000,000,000 bytes
ZB zettabyte	1,000 ⁷ bytes	1,000,000,000,000,000,000,000 bytes
YB yottabyte	1,000 ⁸ bytes	1,000,000,000,000,000,000,000,000 bytes

*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook



294bn

billion emails are sent

Radicati Group

320bn

emails to be sent each day by 2021

306bn

emails to be sent each day by 2020

3.9bn

people use emails

4TB

of data produced by a connected car

Intel



Searches made a day

5bn

Searches made a day from Google

3.5bn

Smart Insights



ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

44ZB

PwC

2013

2020

463EB

of data will be created every day by 2025

IDC

95m

photos and videos are shared on Instagram

Instagram Business



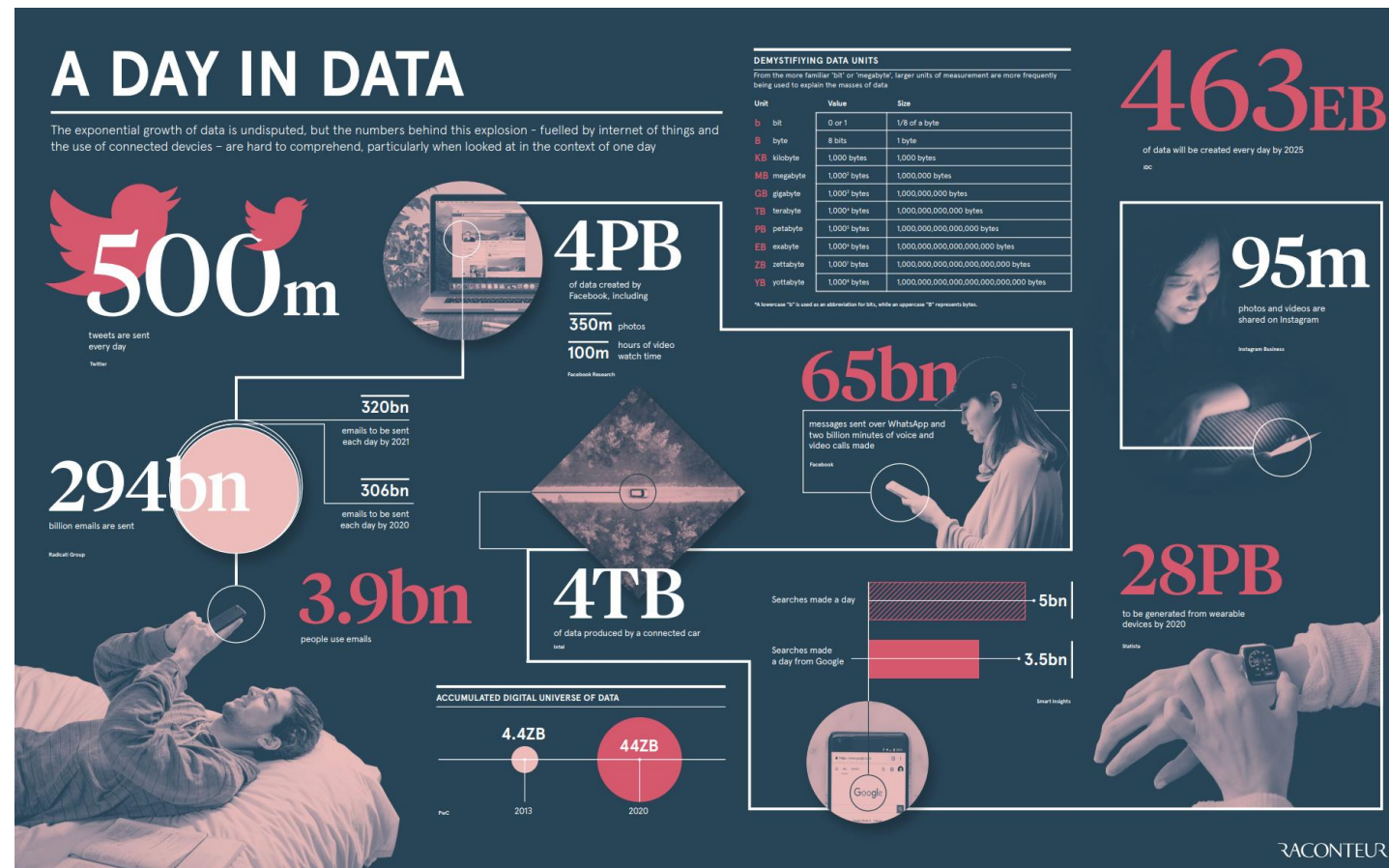
28PB

to be generated from wearable devices by 2020

Statista



- ▣ 1 PB = 125 8TB HDD's
- ▣ 1 EB = 125000 8TB HDD's
- ▣ 1 ZB = 125 000 000 8TB HDD's



Wat zijn je opties voor dit soort data bij te houden?

Ga naar **wooclap.com** en gebruik de code **DBEIAG**

Welke opties zijn er voor dit volume van data bij te houden?

Dockers ...

Data comprimeren

computer kopen

datacenters cloud

Cloud

cloud

Datacenters

Verkleinen, comprimeren?

wooclap

18

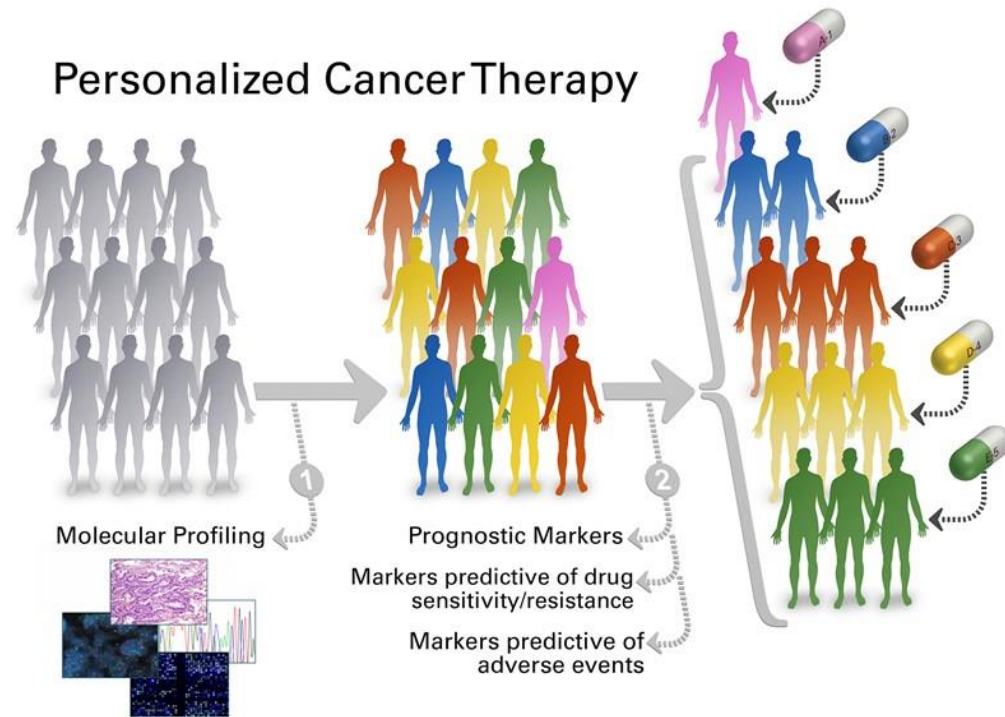


Is het mogelijk om alles lokaal bij te houden om te verwerken?

- ▣ Onmogelijk om computers te kopen die deze hoeveelheid data bijhoudt.
- ▣ RAM-geheugen nodig om data in te laden (Ook niet mogelijk)
 - > Distributed Computing
 - > Cloud Computing

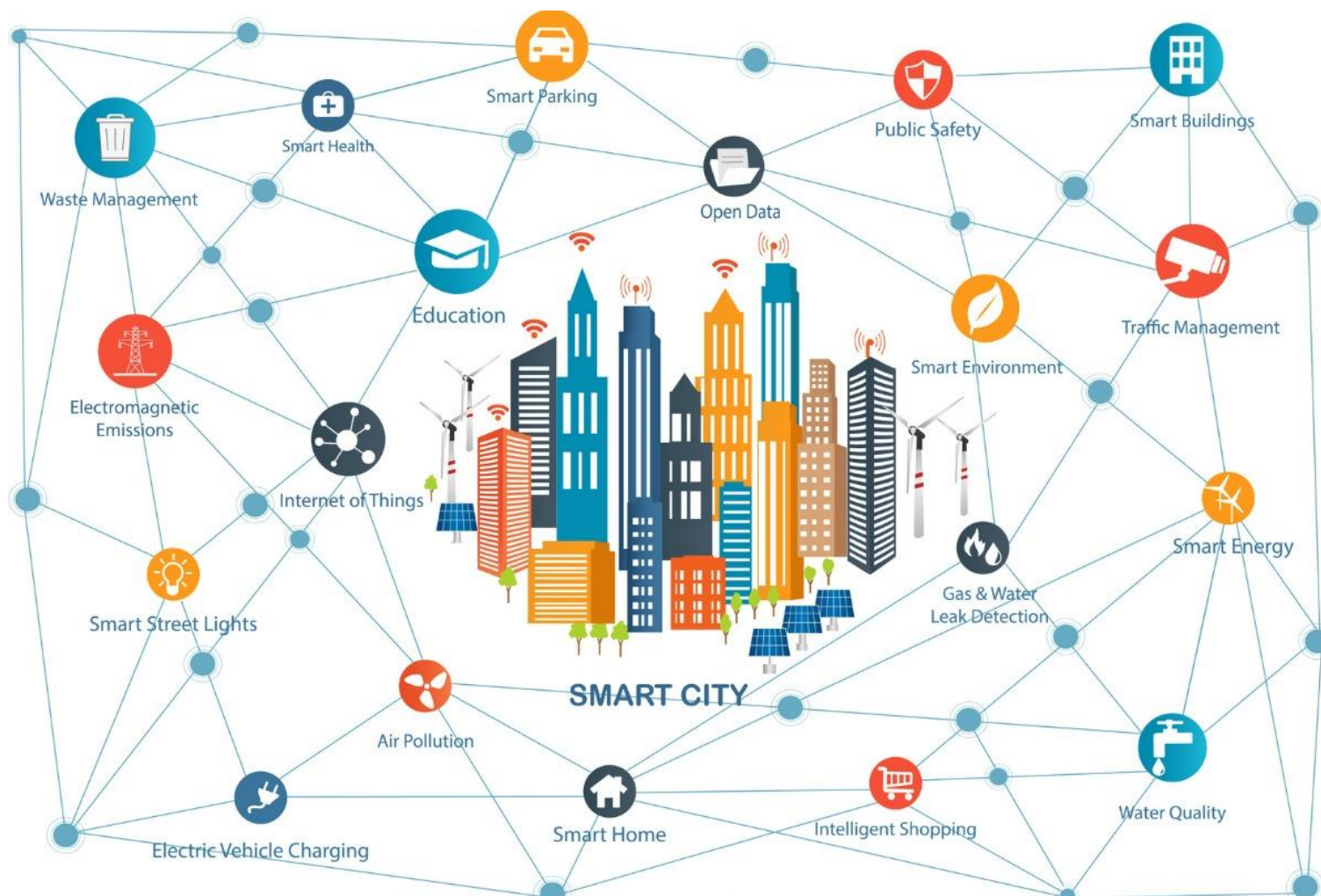
Waarom zoveel data nodig?

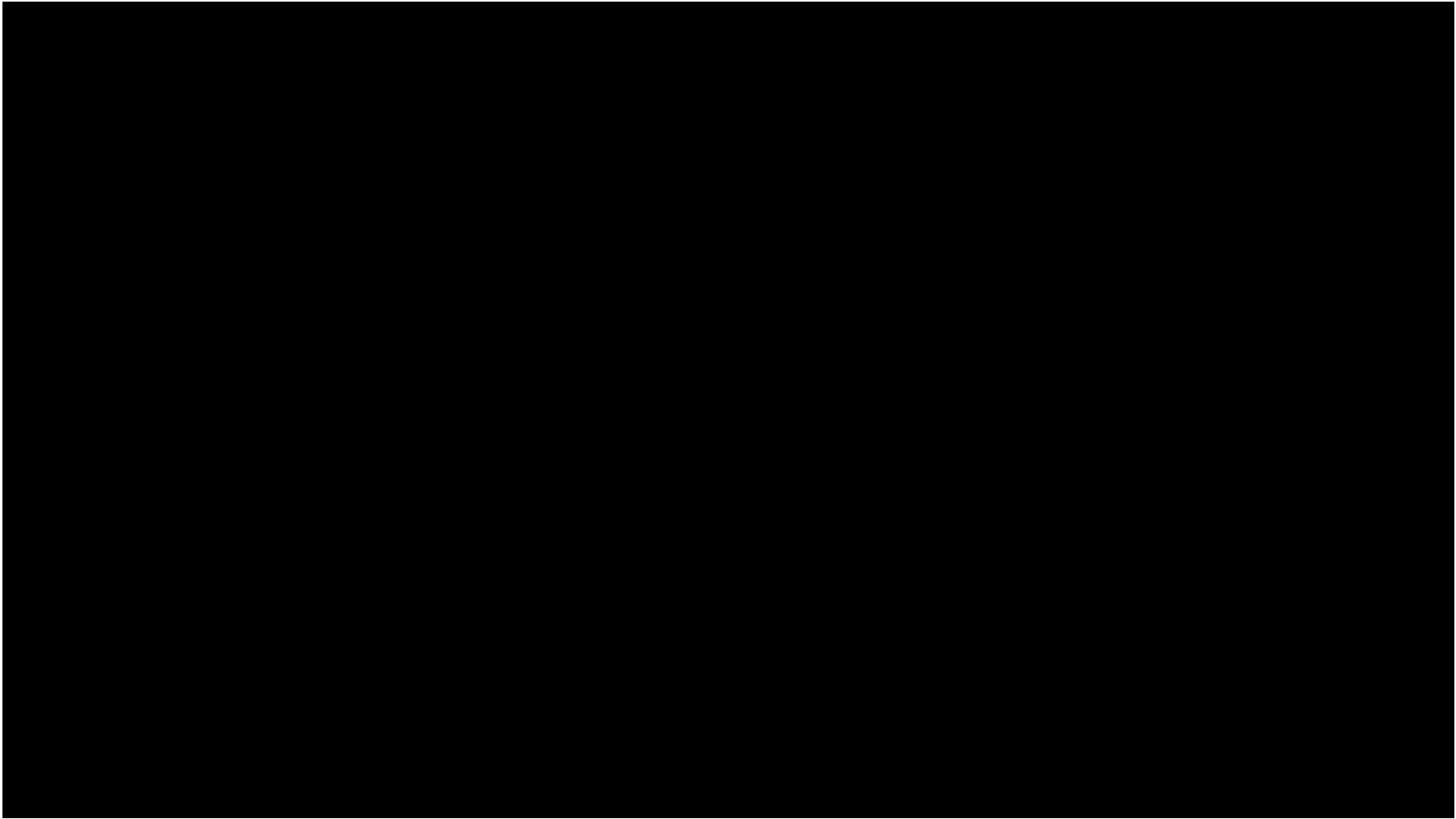
- ▣ Meer data -> betere modellen -> betere voorspellingen / verder vooruit voorspellen
- ▣ Menselijk DNA = 100 GB




Waarom zoveel data nodig?

Smart Cities





<https://www.youtube.com/watch?v=i3zx3gF9AUU>

- 
- ▣ Large Hadron Collider : 90 PBs per jaar
 - ▣ Boeing 737: 20 TB per uur per motor



3.

Big Data

Definitie - Wikipedia

- **Big data** of **massadata**^[1] zijn gegevensverzamelingen (datasets) die te groot en te weinig gestructureerd zijn om met reguliere databasemanagementsystemen te worden onderhouden. De gegevens hebben een direct of indirect verband met privégegevens van personen. ^[2] Big data spelen een steeds grotere rol. De hoeveelheid data die opgeslagen wordt, groeit exponentieel. Dit komt doordat consumenten bij sociale media in toenemende mate data opslaan in de vorm van bestanden, foto's en films (bijvoorbeeld op Facebook of YouTube, waar Facebook ook de door de gebruikers gewiste data bewaart) en organisaties, overheden en bedrijven steeds meer data over burgers produceren en opslaan, en doordat apparaten zelf data verzamelen, opslaan en uitwisselen (het zogenaamde internet der dingen). Hierdoor is er steeds meer sensordata beschikbaar. Niet alleen de opslag van deze hoeveelheden is een uitdaging, maar ook het analyseren ervan. Deze data bevatten namelijk informatie voor doeleinden zoals marketing, wetenschappelijk onderzoek, of preventief onderhoud.



Definitie - Gartner

- ▣ **Big data** is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

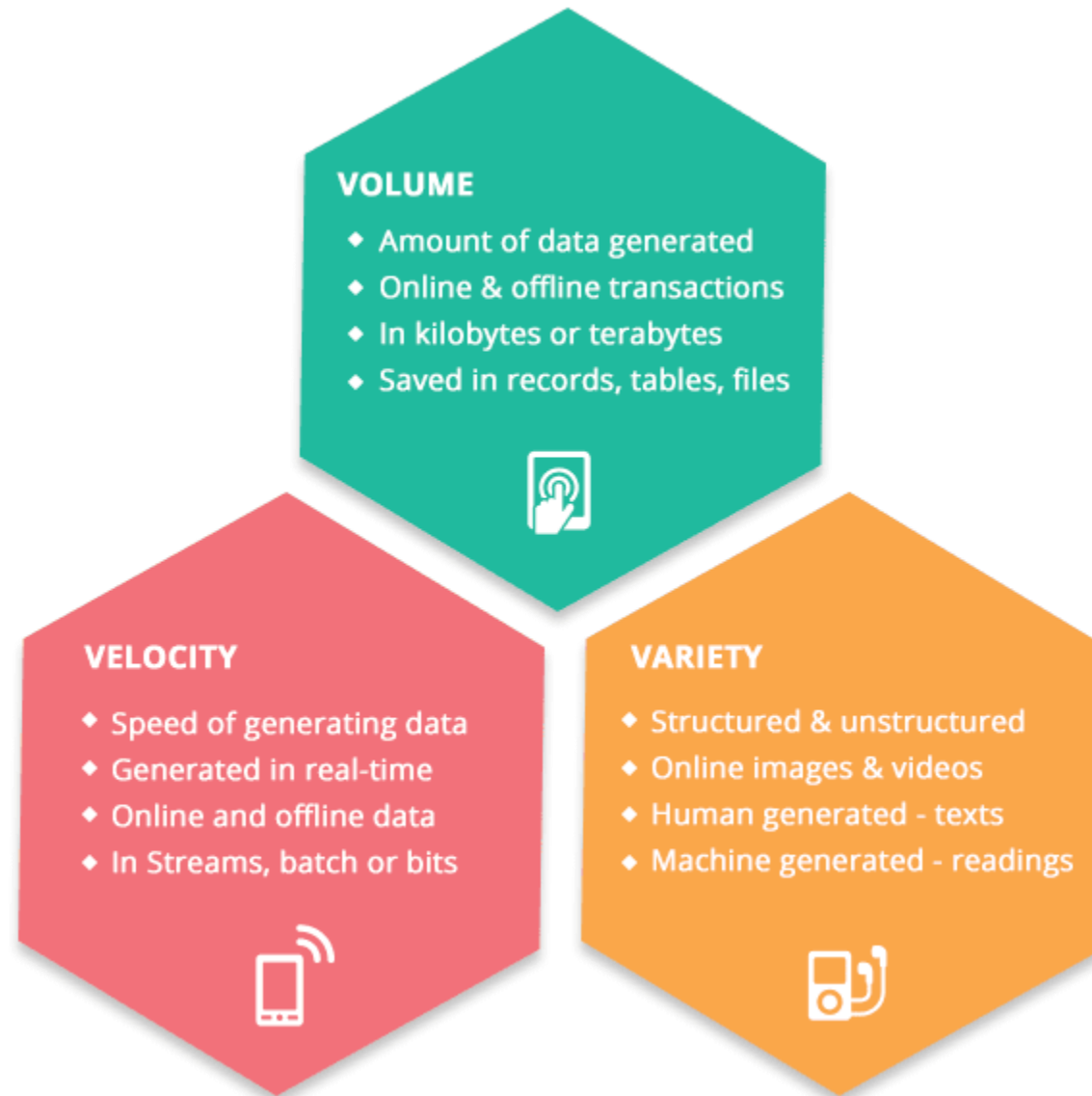


Kenmerken

Polling – Wat bepaalt wanneer een data science project een big-data project is?



De drie hoofd V's



Extra V: Veracity



THE 4 V'S OF BIG DATA

40 ZETTABYTES
of data will be created by
2020, an increase of 300
times from 2005



6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION



Volume

SCALE OF DATA

2.5 QUINTILLION BYTES
of data are created
each day



Most companies in the
U.S. have at least
100 TERABYTES
of data stored



As of 2011, the global size of
data in healthcare was
estimated to be
150 EXABYTES



**30 BILLION
PIECES OF CONTENT**
are shared on facebook
every month



Variety

DIFFERENT
FORMS OF DATA

**4 BILLION +
HOURS OF VIDEO**
are watched on
You Tube each month



4 MILLION TWEETS
are sent per day by about
200 million monthly active
users



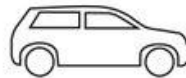
The New York Stock
Exchange captures
**1TB OF TRADE
INFORMATION**
during each trading
session



Velocity

ANALYSIS OF
STREAMING DATA

Modern cars have
close to
100 SENSORS
that monitor items such as
fuel level and tire pressure



**1 IN 3 BUSINESS
LEADERS**
don't trust the information
they use to make
decisions



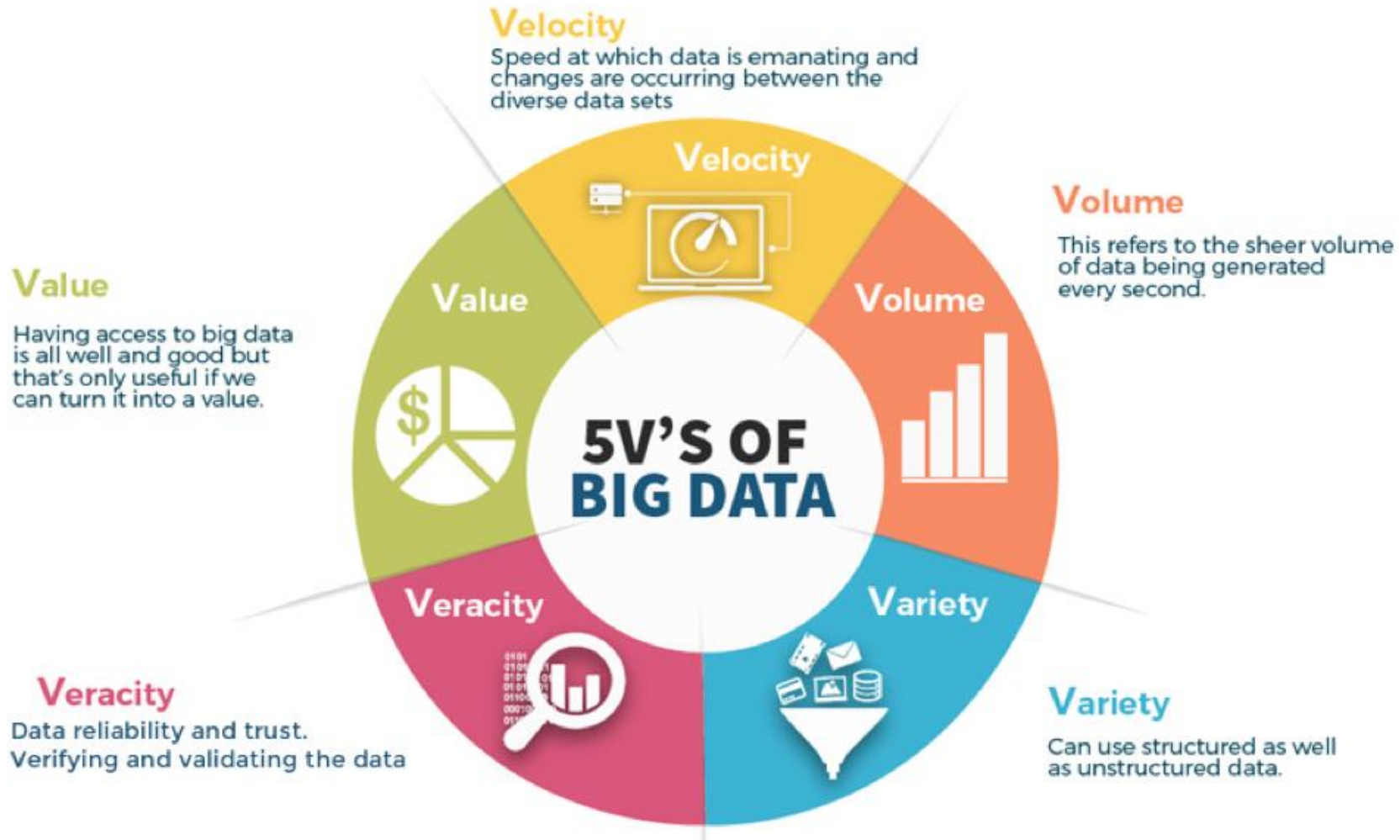
Veracity

UNCERTAINTY
OF DATA







27% OF RESPONDENTS
in one survey were unsure
of how much of data
was inaccurate



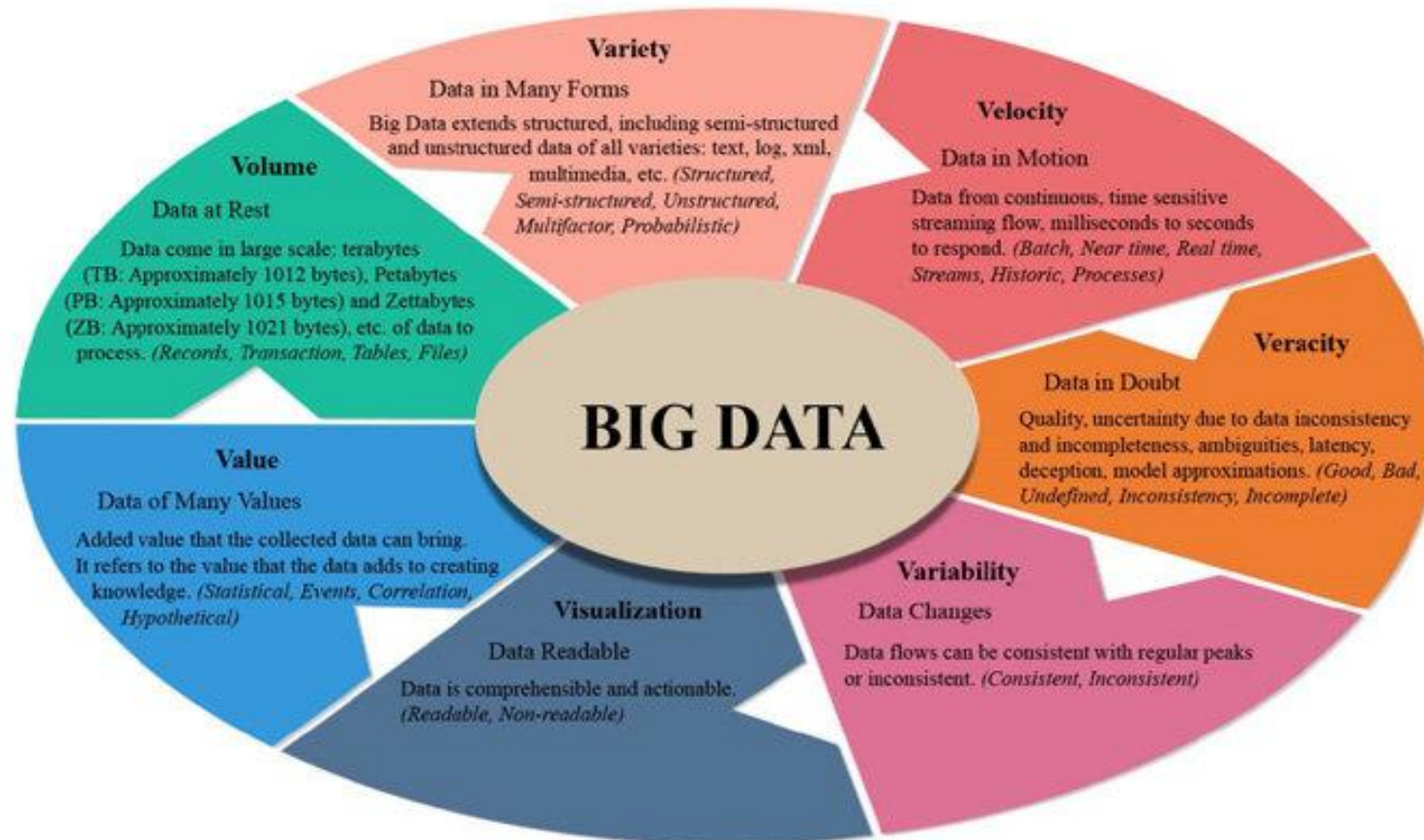
Of 5 V's: Value



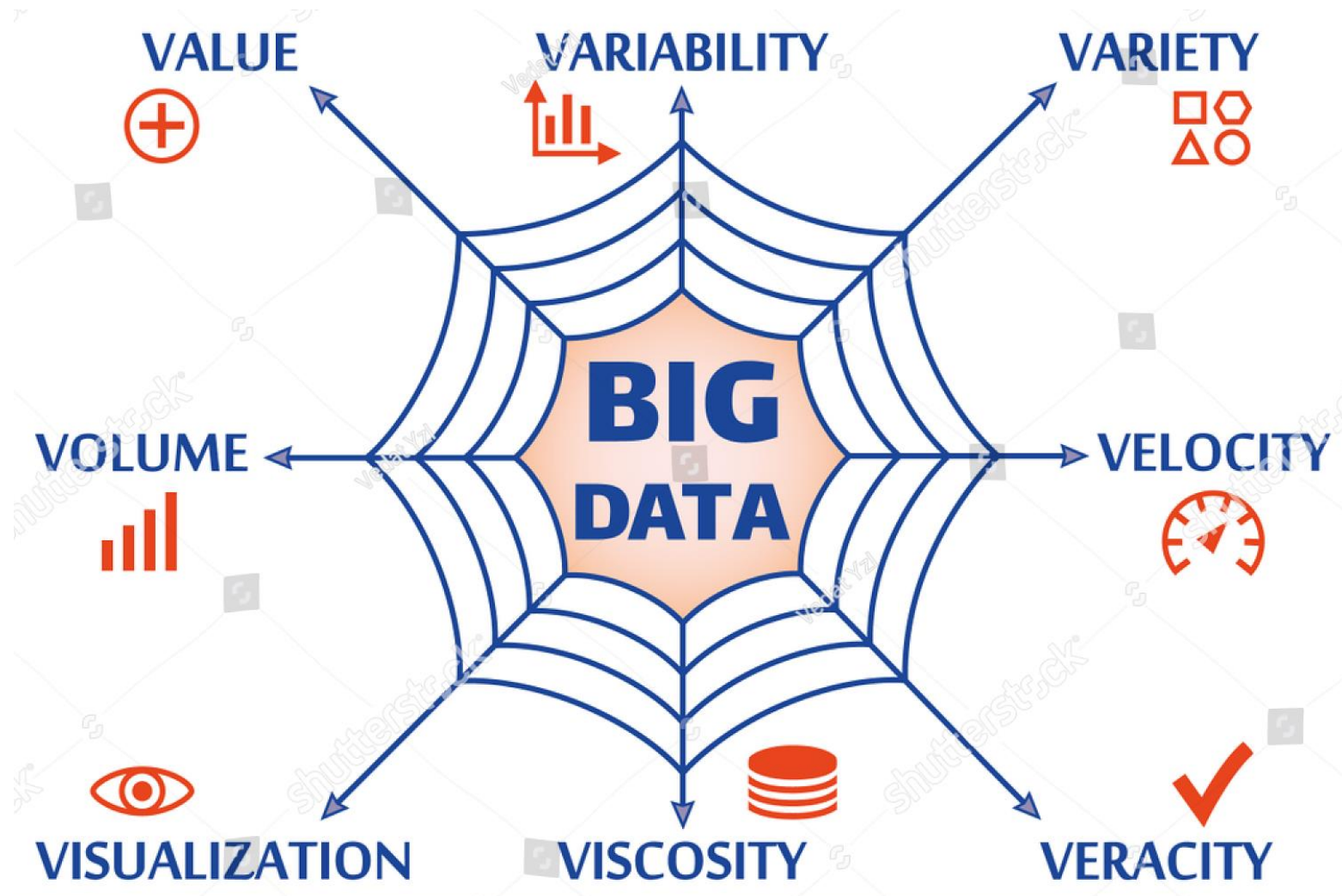
Of 6? Variability

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.
					

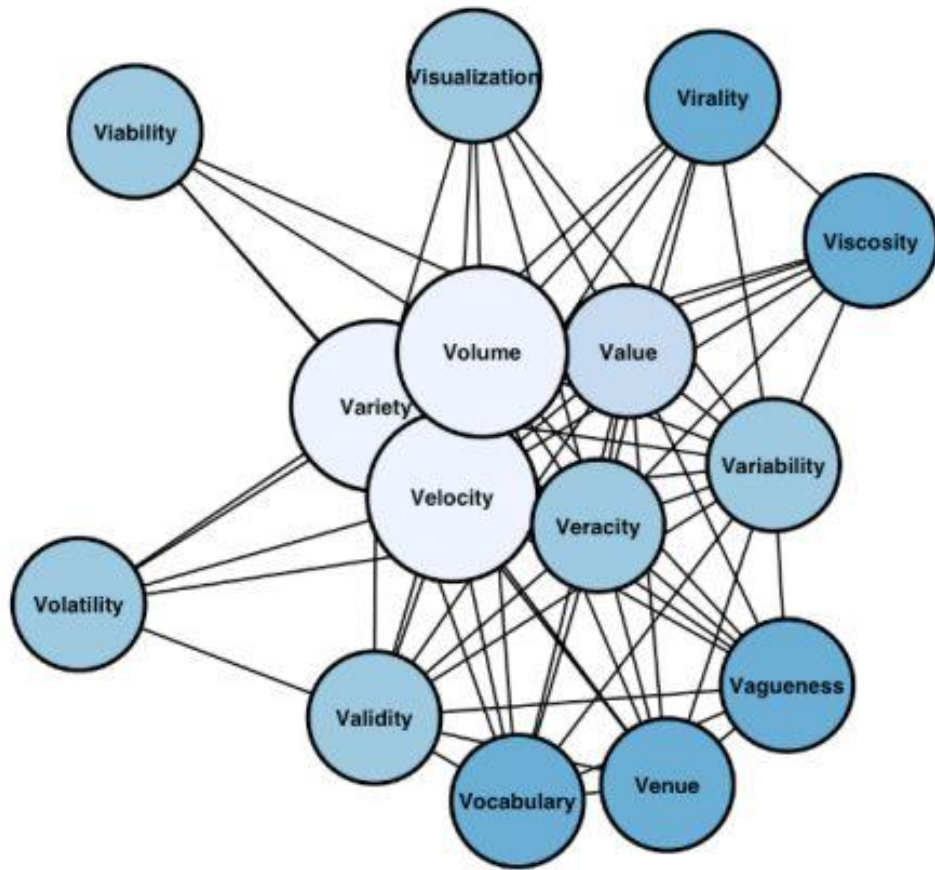
Of 7? Visibility



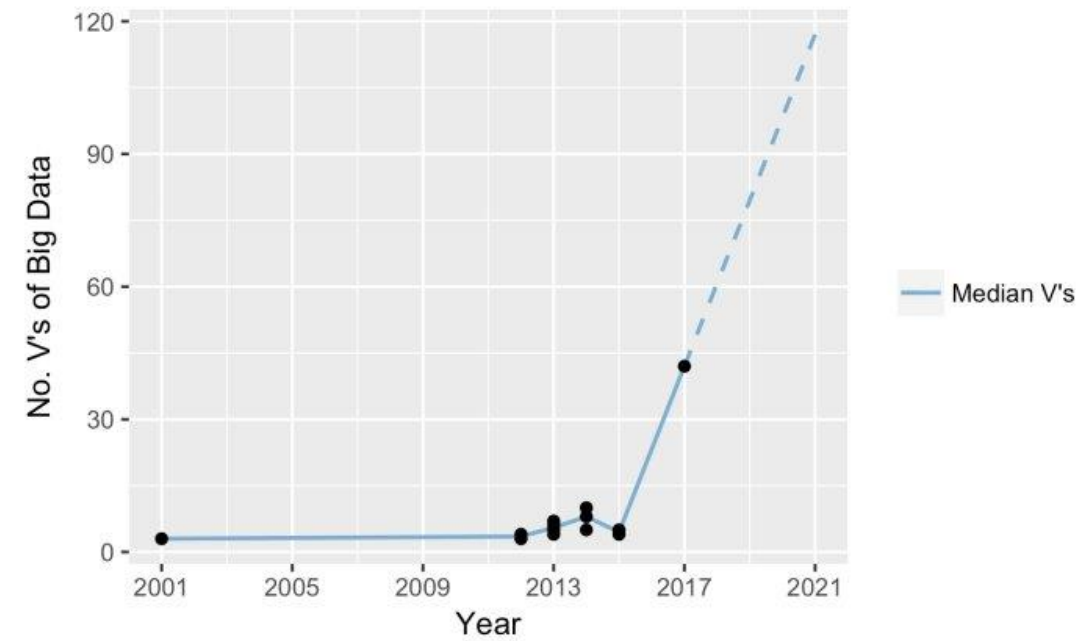
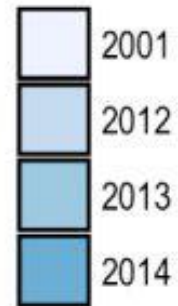
Of 8? Viscosity



10 V's of meer?



First Occurrence



3. Soorten data

Structured data

- ▣ Vast data formaat in tabel vorm met rijen en kolommen
- ▣ Alle formaten vooraf vastgelegd
- ▣ Excel files, Sql-database, csv, ...



Unstructured data

- ▣ Geen vaste structuur in de data
- ▣ Moeilijk om in te zoeken
- ▣ Foto's, video's, audio, tekst ...

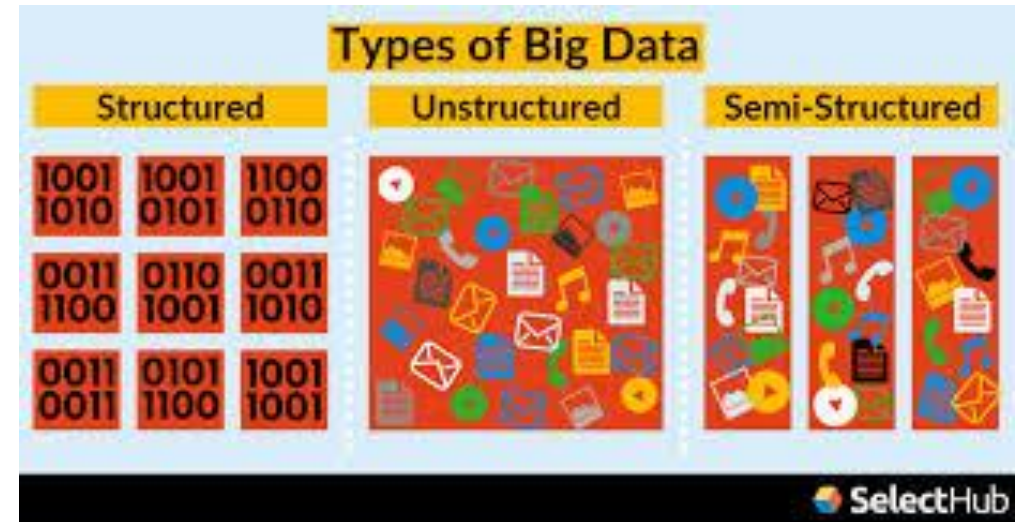
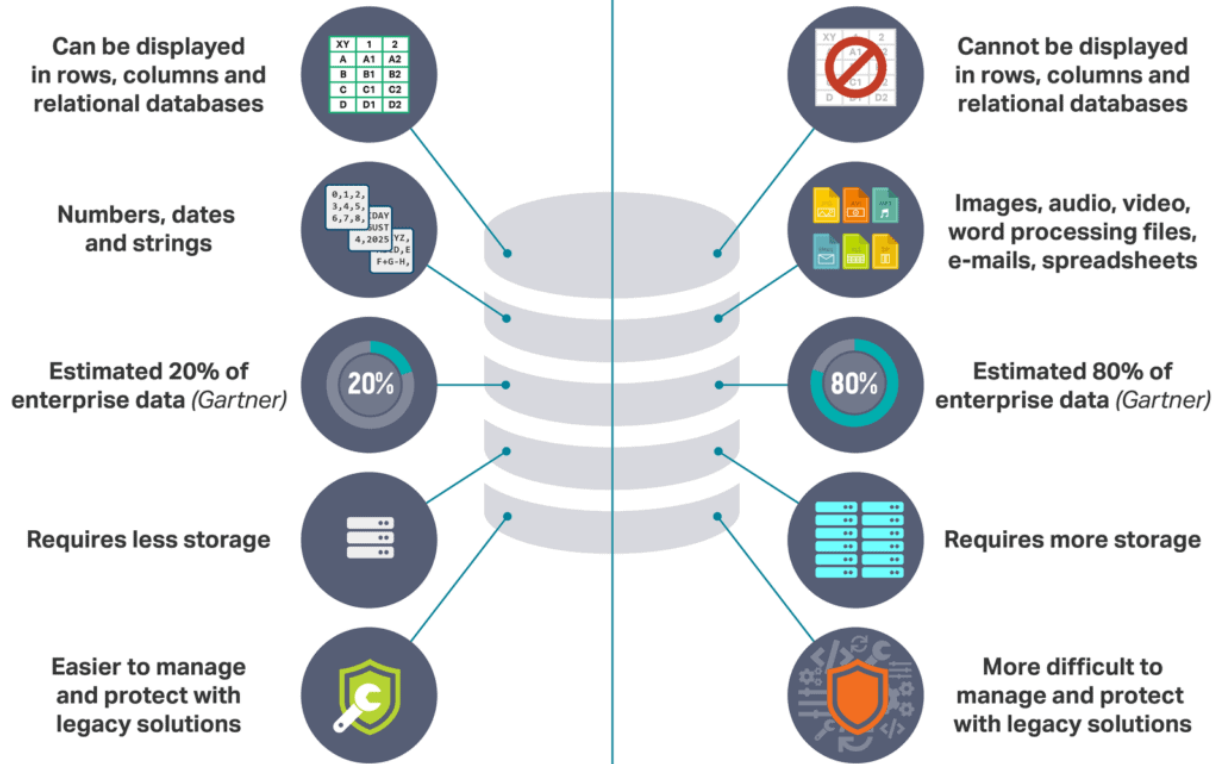


Semi-Structured data

- ▣ Licht-georganiseerde data
- ▣ Tags/metadata verzorgt de structuur
- ▣ Html, xml, json, ...



Structured Data vs Unstructured Data



Door wie is de data geproduceerd?

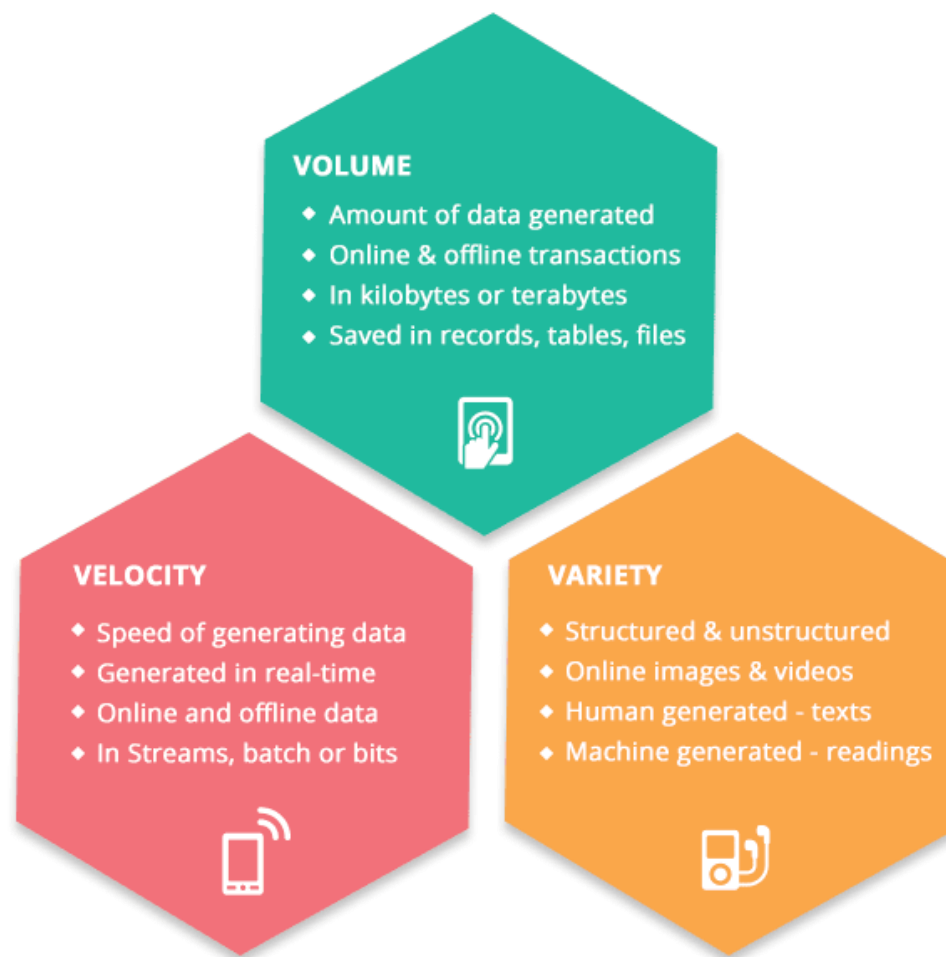




Kritiek op Big Data

- ▣ Op de onderliggende theorie:
 - ▢ Toekomst gelijkaardig aan het verleden
 - ▢ Context afhankelijk
- ▣ V-model focust op schaalbaarheid en rekenkracht, niet op verklaarbaarheid
- ▣ Grote datasets en analyses bestaan reeds decennia, niet zo nieuw als veel denken
- ▣ Buzzword om aandacht te trekken naar je product
- ▣ Privacyschendingen, datalekken, controles, ...

Hoe kan je omgaan met deze problemen?

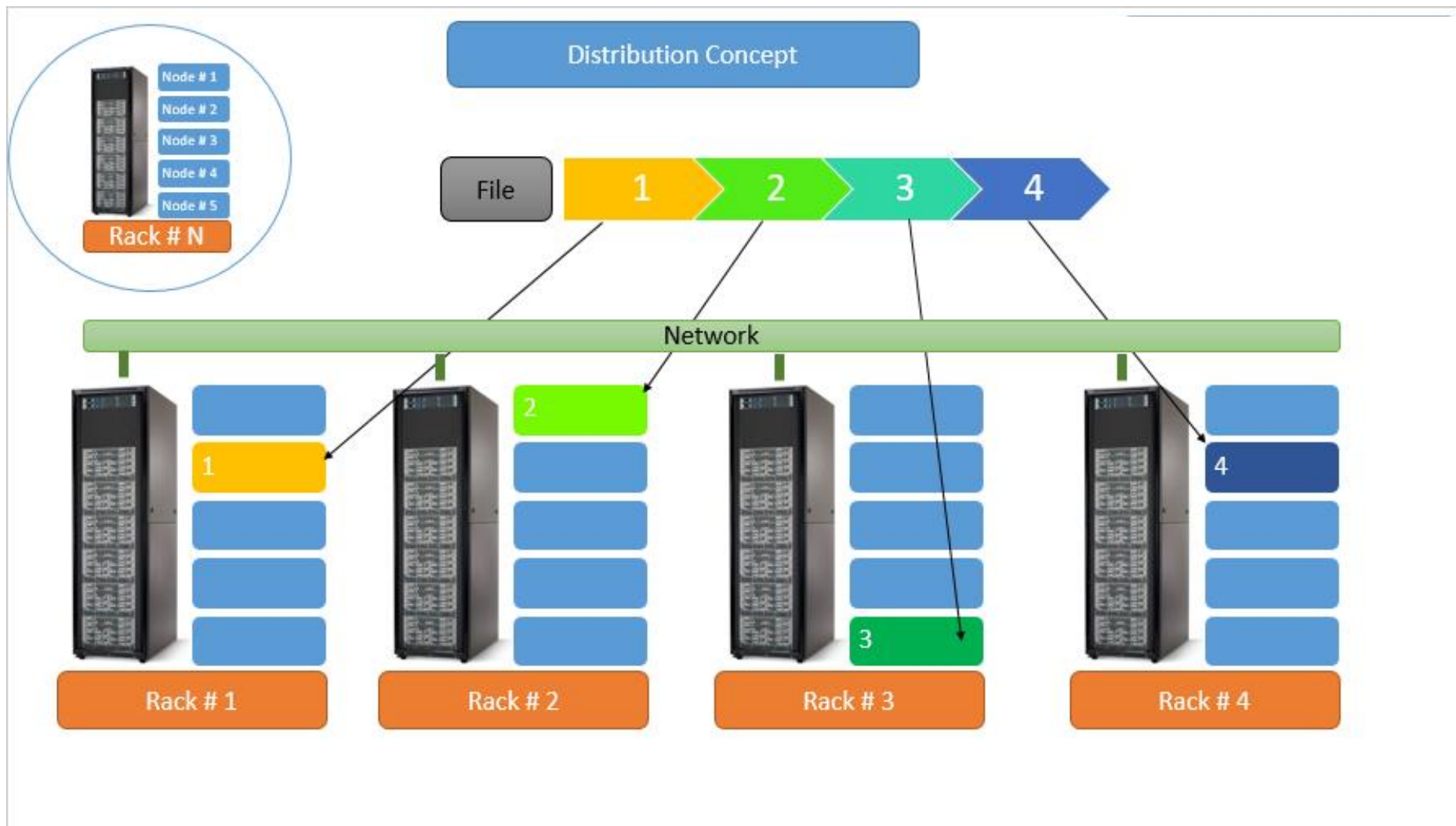




Distributed storage

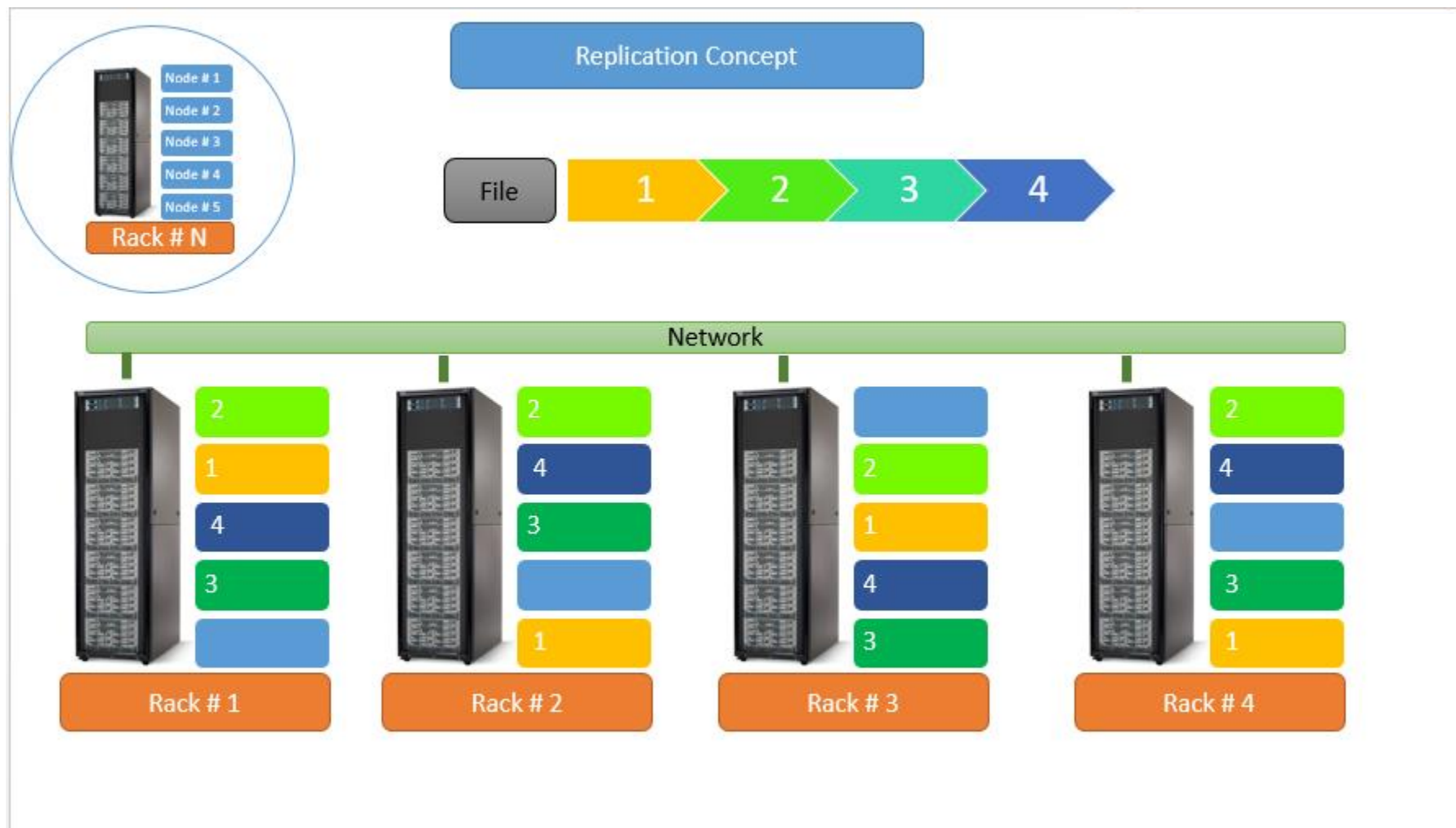
Van Pc -> Rack -> Datacenter -> Cloud





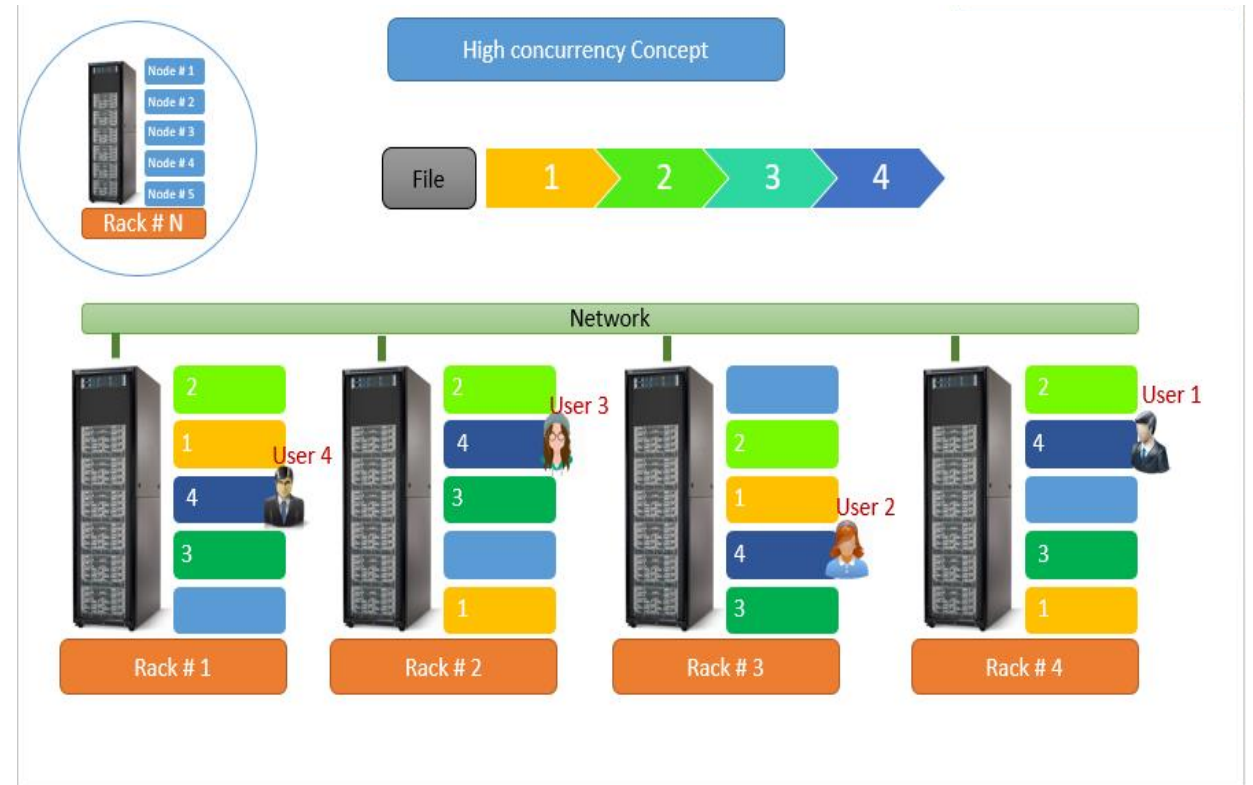
Replication

▣ Fault tolerance



Voordelen

- ▣ Schaalbaar
- ▣ Fout tolerant
- ▣ Nodige rekenkracht ook verdeeld
 - Concurrency
- ▣ Goedkoper
 - Minder gespecialiseerde computers
 - Commodity clusters





Nadelen

- ▣ Meer management van welke data op welke server zit nodig
- ▣ Replication of data maakt het nodig om synchronisatie te doen
 - ▬ Wat bij geografisch verspreide data?
 - ▬ Wat bij uitvallen van server/ datacenter / ...?

Doet dit jullie aan een andere techniek denken?

Ga naar **wooclap.com** en gebruik de code **DBEIAG**

Doet dit jullie aan een andere techniek denken?

bittorrent
Clusters van goedkope
RAID
RAID
blockchain

wooclap

5



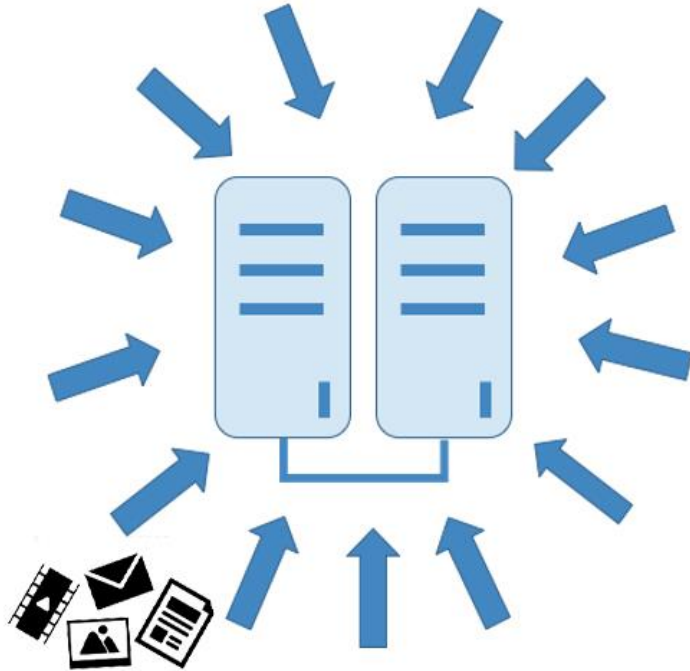
Belangrijke termen voor distributed storage

- ▣ Horizontale schaalbaarheid
- ▣ Duplicatie/replica's
- ▣ Fault-Tolerance

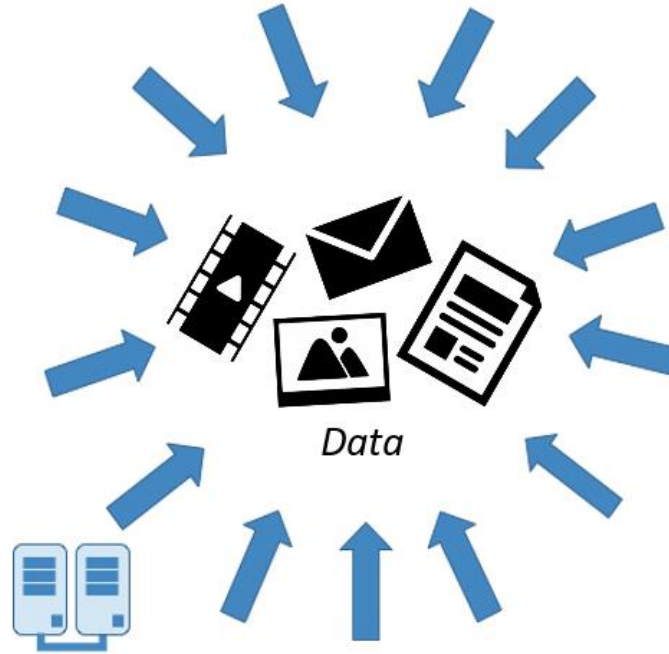


Bring computing to data

Computing to data



Code / Programma niet gemakkelijk te migreren
Rekeneenheid moet krachtig zijn



Code / Programma gemakkelijk te migreren
Volume van data is groot en schaalbaar
Rekenkracht is verspreid
Berekeningen gebeuren asynchroon en verspreid

Computing to data

▣ Sneller

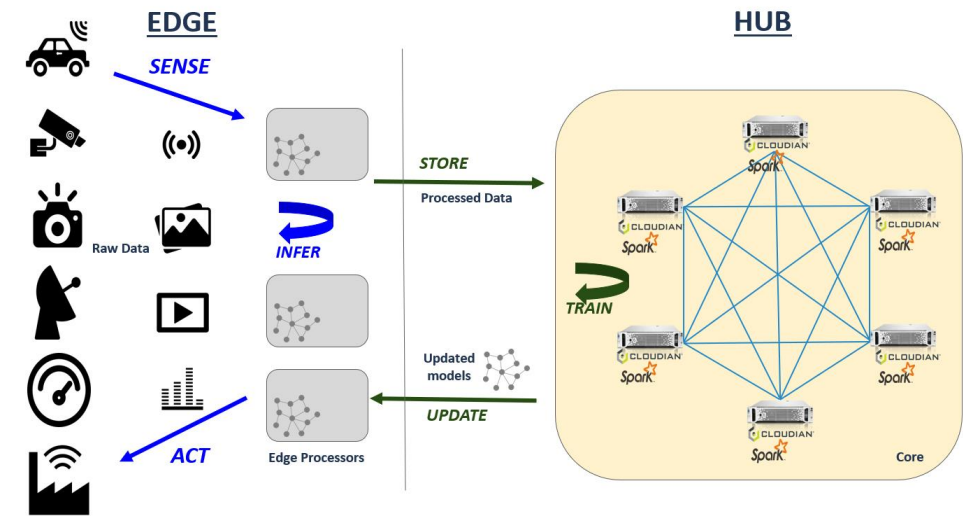
- ▬ Geen transmissietijd voor real-time beslissingen

▣ Goedkoper

- ▬ Verplaatsen data is kostelijk, schaalbaarheid ook belangrijk voor kosten te beperken

▣ Veiliger

- ▬ Data is gemakkelijker te onderscheppen bij verplaatsen



Veel gebruikte concepten bij distributed computing

▣ Divide-And-Conquer

- ▬ Elke node in de cluster voert maar berekeningen uit op een klein stukje van de data

▣ Lazy-Evaluation

- ▬ Je code bouwt een stappenplan op. Enkel als er een eind-resultaat verwacht wordt, worden er berekeningen uitgevoerd
 - ▣ Onnodige operaties worden vermeden, optimalisatie van de beschikbare rekencapaciteit

▣ Eventually-Consistent

- ▬ Updates van data zijn non-blocking, replica's kunnen tijdelijk nog de oude data bevatten



Tools

Containers

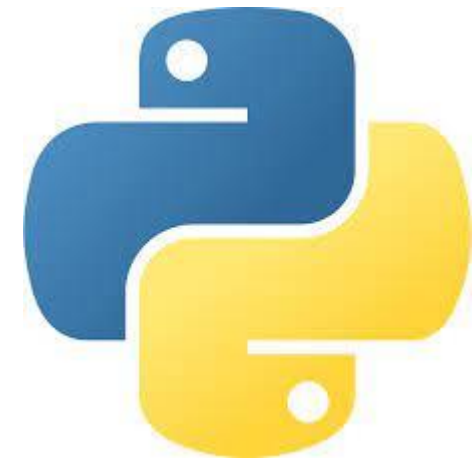
■ Docker desktop – linux containers

- ▬ Zie installatie.pdf op Toledo om de containers voor dit vak te gebruiken
- ▬ 2 Versies afhankelijk van je computer
 - Default – cluster met 4 datanodes
 - Light – cluster met 1 datanode



Python

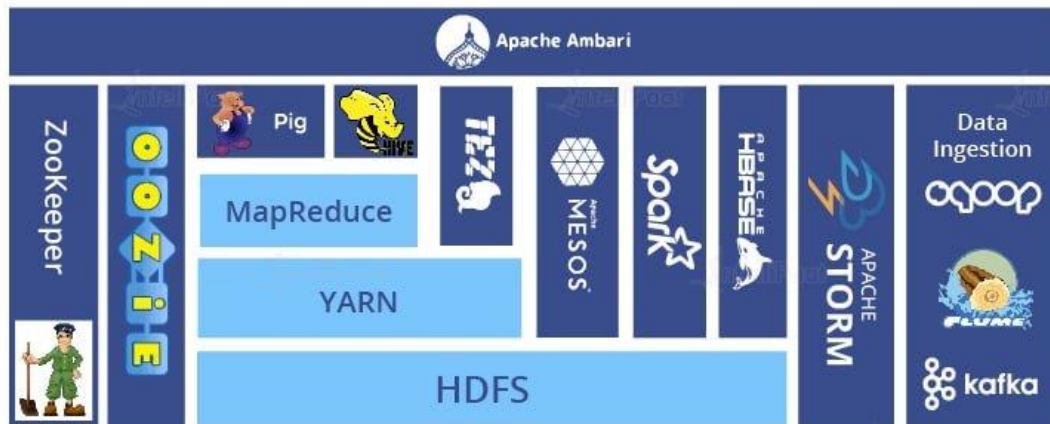
- ▣ Scripting programmeertaal
- ▣ Te installeren via academic software of anaconda
- ▣ Bevat een reeks handige packages



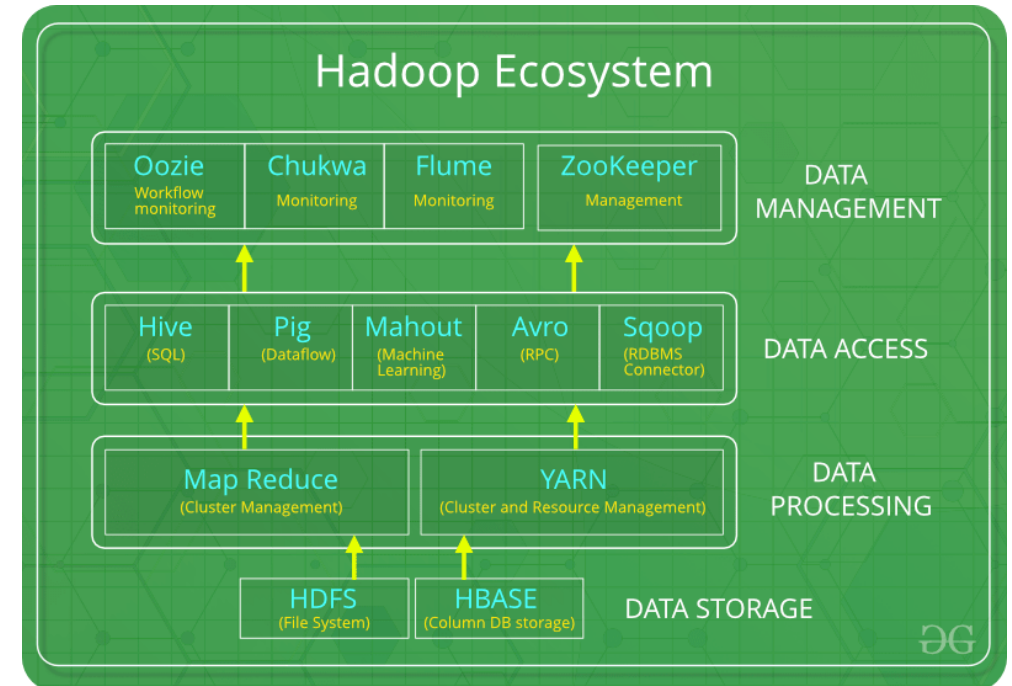
Hadoop

- ▣ Distributed file system met daarboven op functionaliteiten voor distributed applicaties / computing /

Core Hadoop Ecosystem



Hadoop Ecosystem



NoSQL databases

- ▣ Allerlei implementaties maar vooral MongoDB

