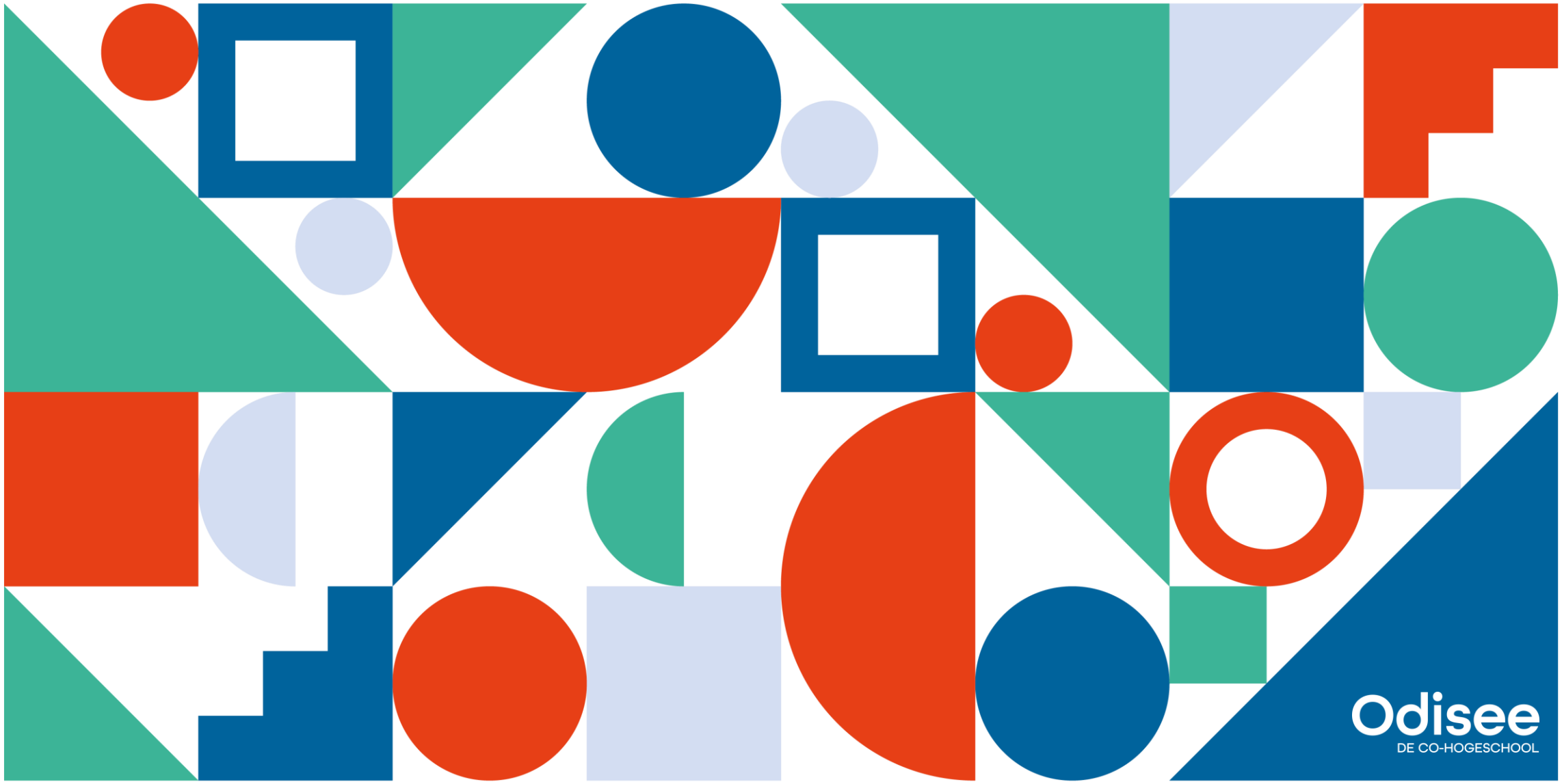

Odisee
DE CO-HOGESCHOOL

Big Data - Introduction

Jens Baetens

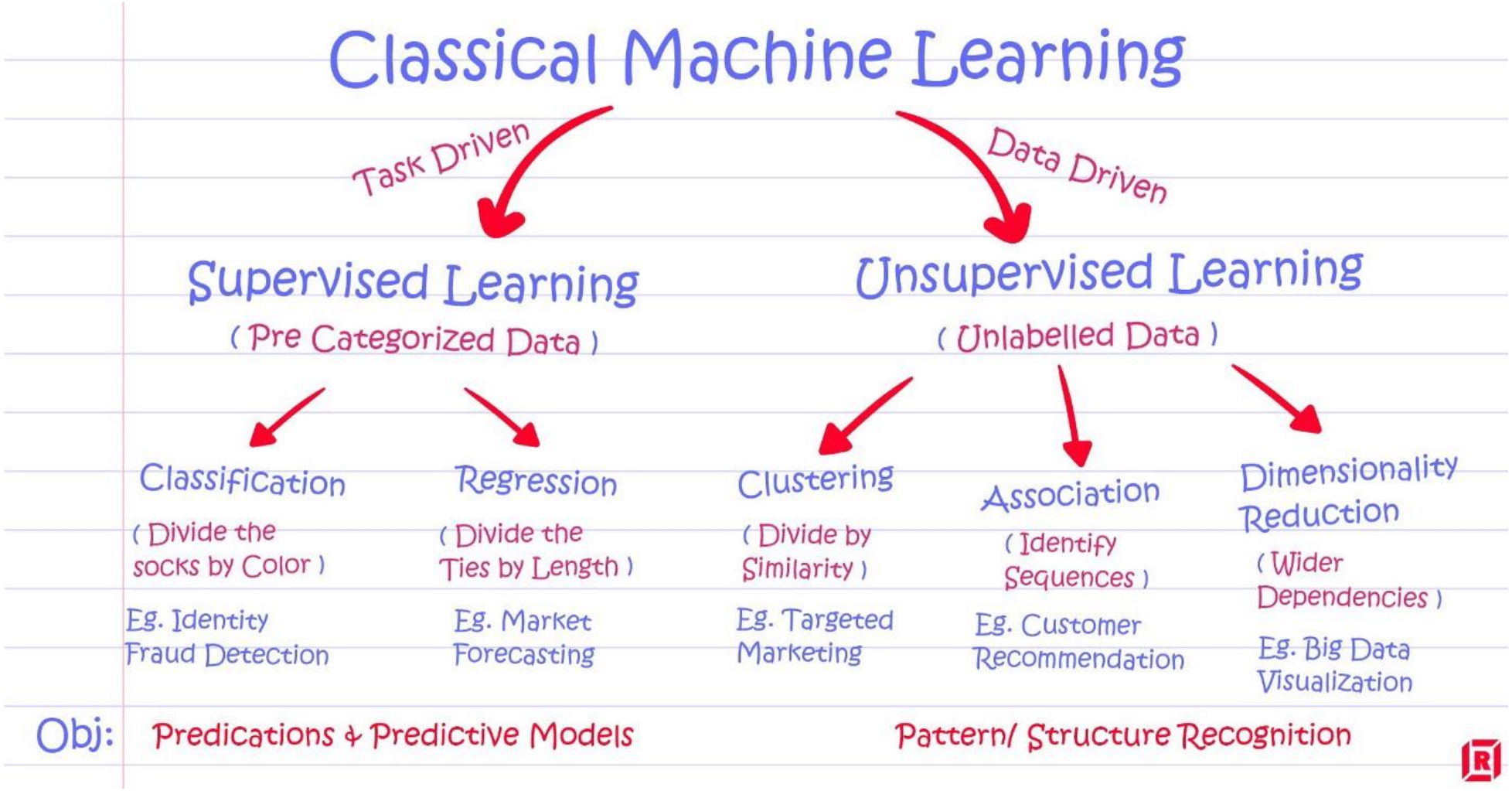
Recap – Data Science

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

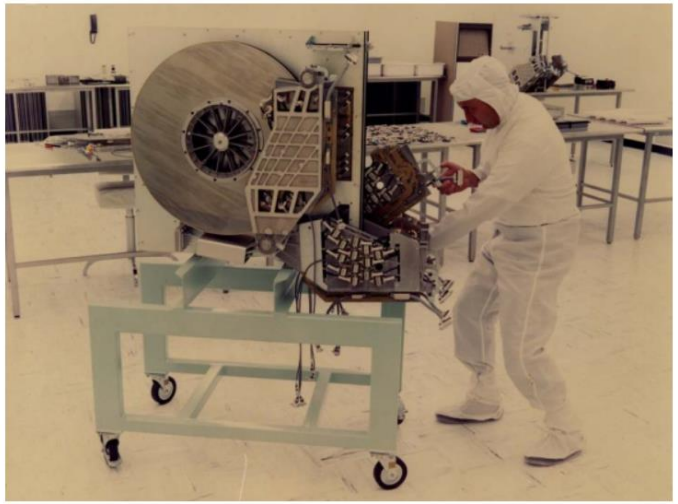
2. Is dit altijd mogelijk?

Zien jullie problemen?

Zien jullie problemen?

Grootte harde schijven?

1956: 5 MB

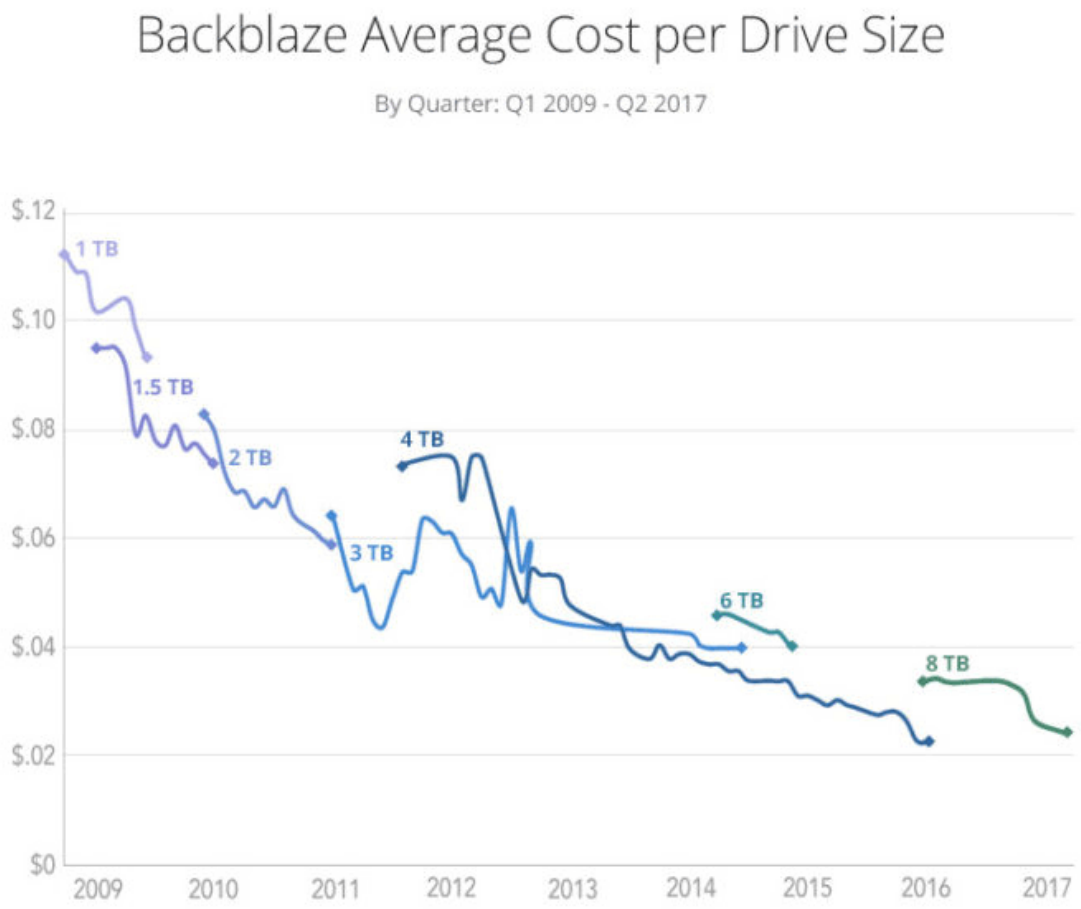


1975: 250 MB

1988: 1 GB

2019: 1 TB

Prijs?


Backblaze Average Cost per Drive Size
By Quarter: Q1 2009 - Q2 2017
$.12
$.10
$.08
$.06
$.04
$.02
$0
2009
2010
2011
2012
2013
2014
2015
2016
2017
1 TB
1.5 TB
2 TB
3 TB
4 TB
6 TB
8 TB

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 devcies – 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

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

65bn

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

Facebook

DEMYSTIFIYING 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.

463EB

of data will be created every day by 2025

IDC

95m

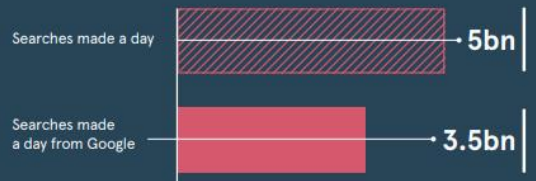
photos and videos are shared on Instagram

Instagram Business

28PB

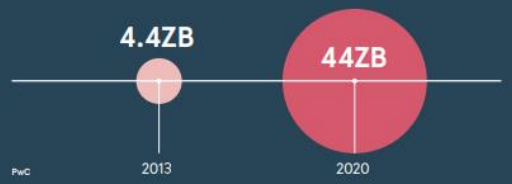
to be generated from wearable devices by 2020

Statista



Smart Insights

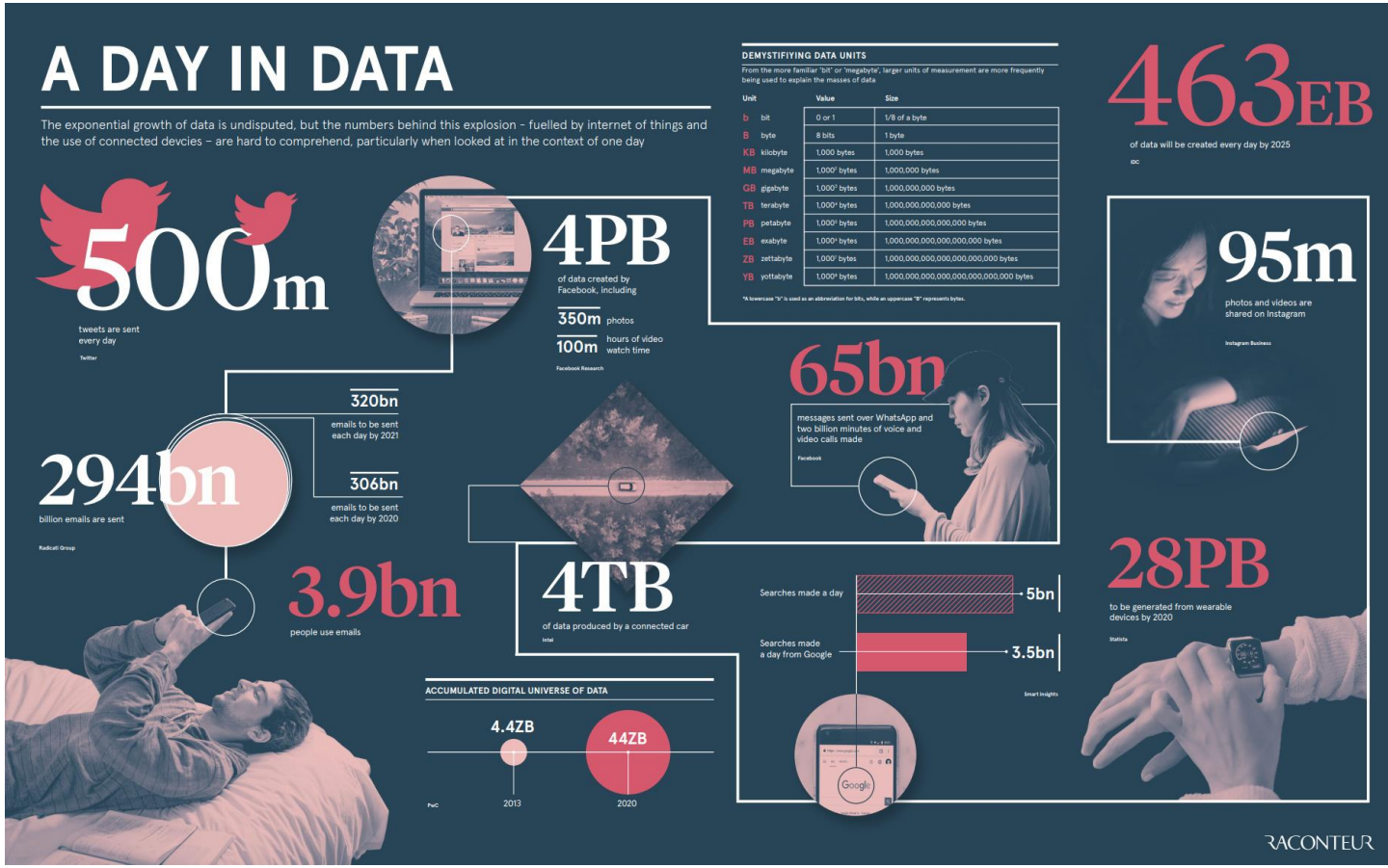
ACCUMULATED DIGITAL UNIVERSE OF DATA



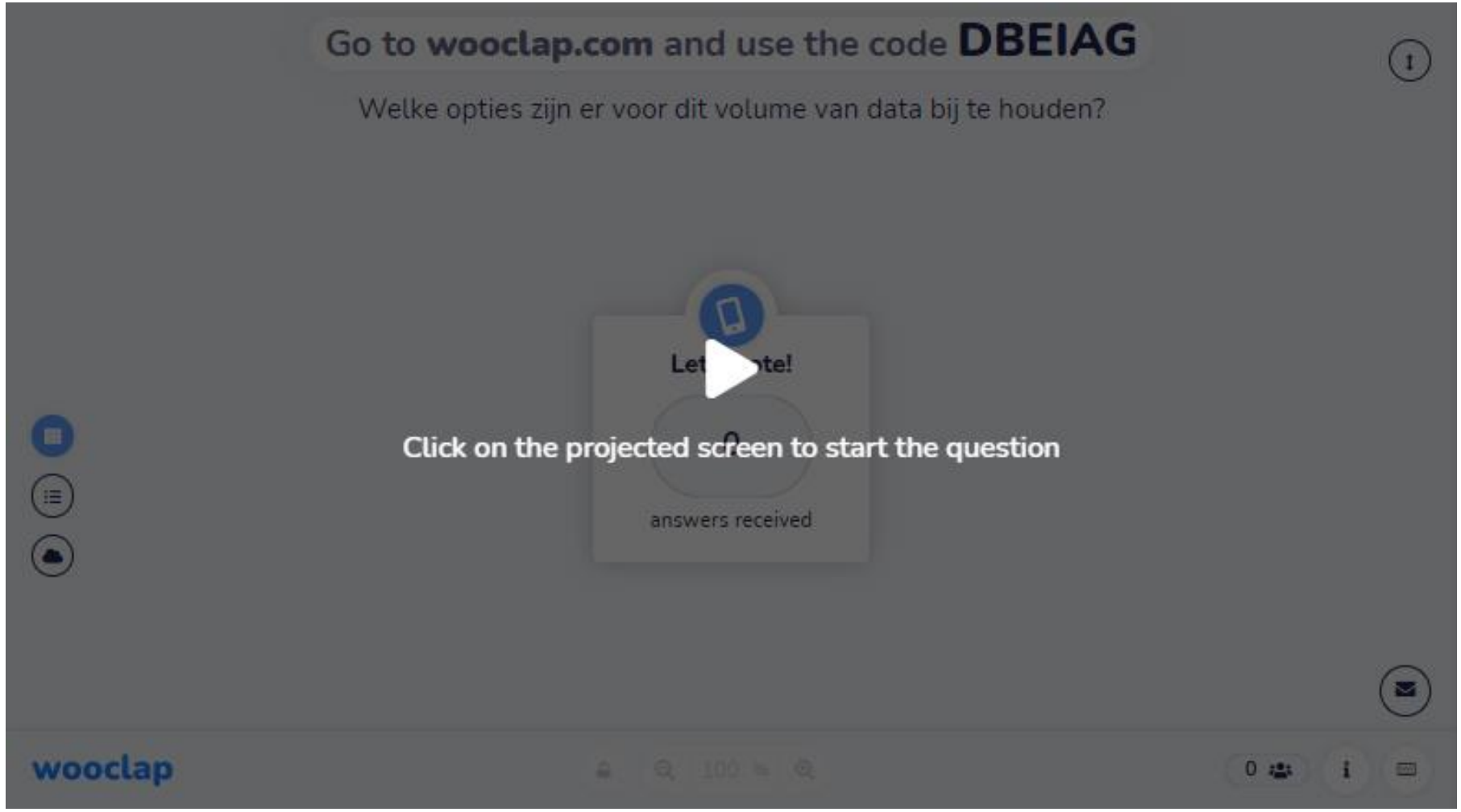
PwC

3ACONTEUR

- 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?



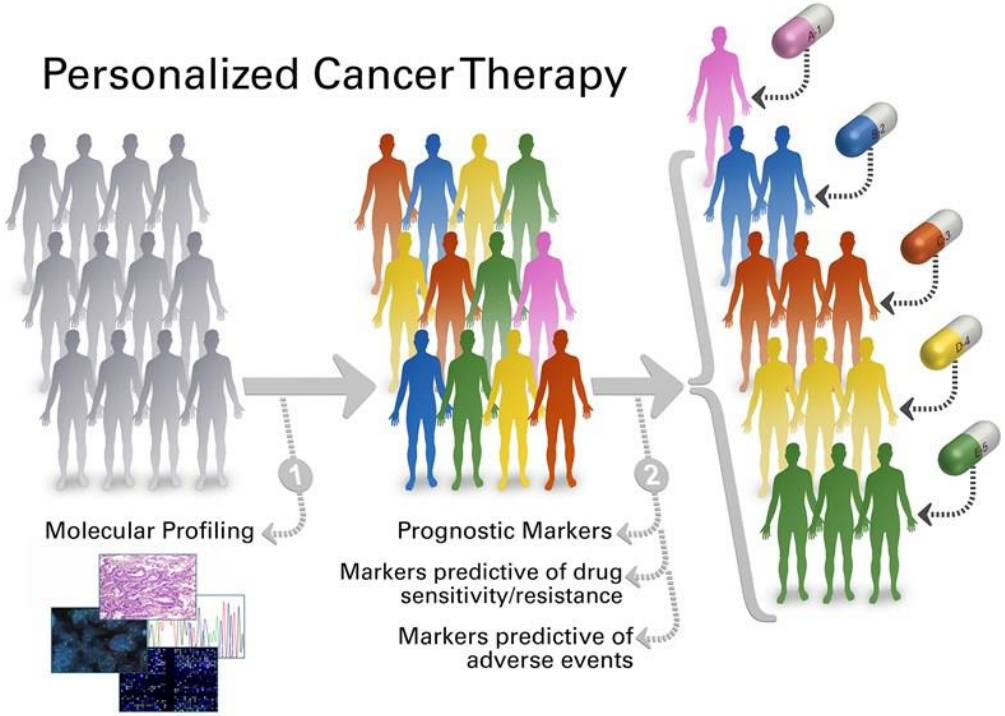
Is het mogelijk om alles lokaal bij te houden om te verwerken?

- 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

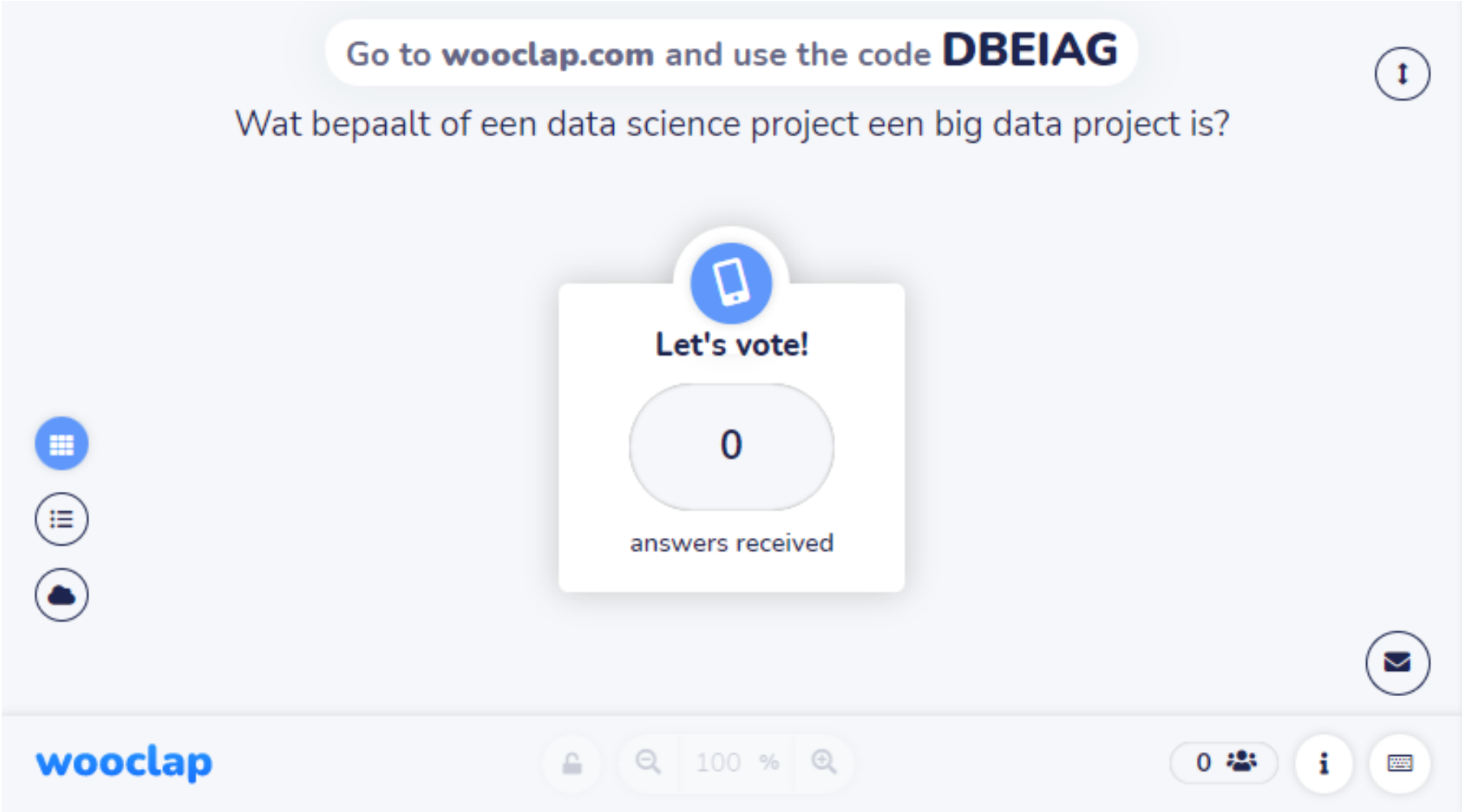
- **Big data** of **massadata**[1] zijn gegevensverzamelingen (datasets) die te groot en te weinig gestructureerd zijn om met reguliere databasemanagementsystemen te worden onderhouden. De gegevens hebben een direct of indirect verband met privégegevens van personen. [2] Big data spelen een steeds grotere 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

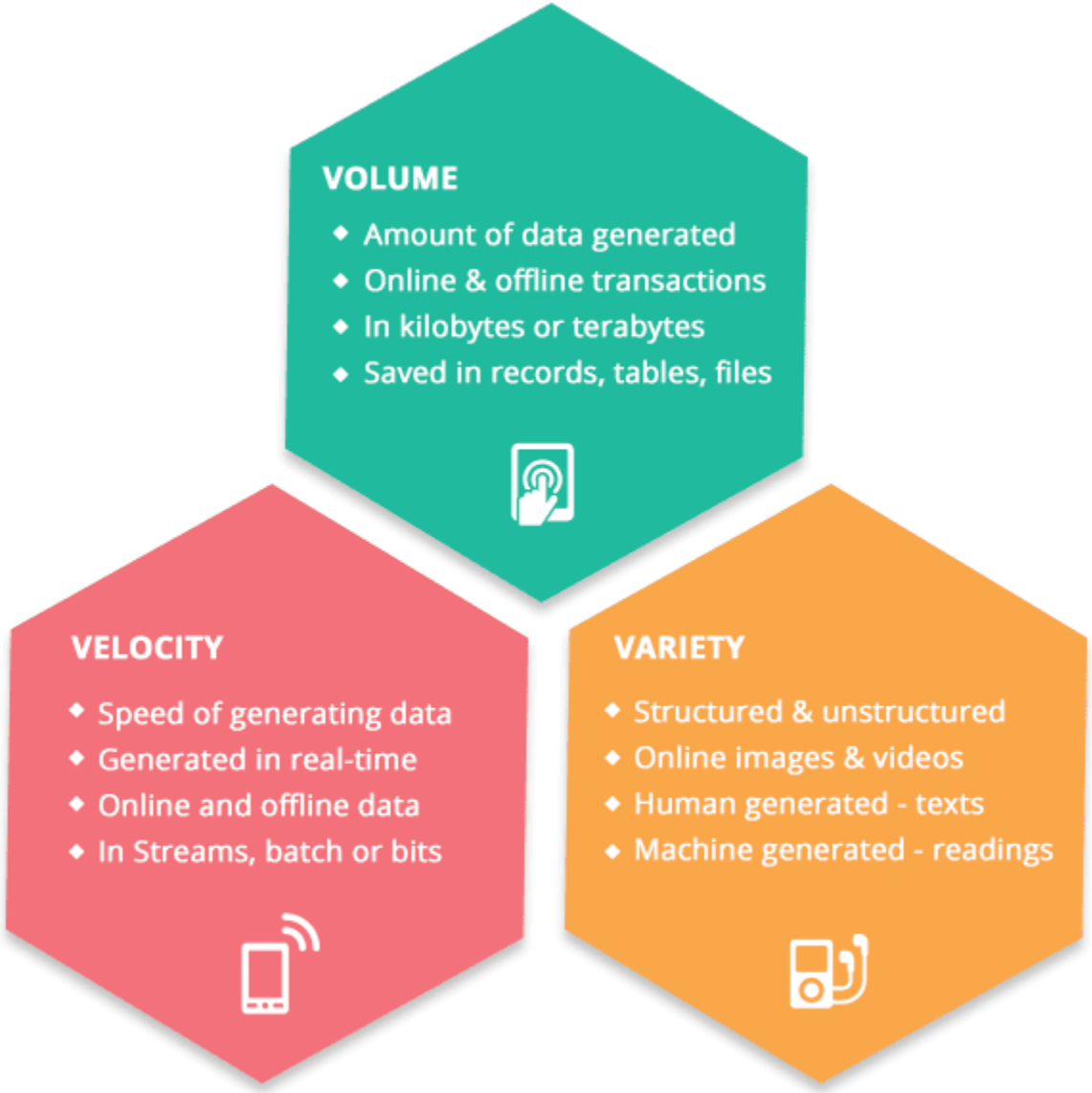
- **Big data** is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

Kenmerken

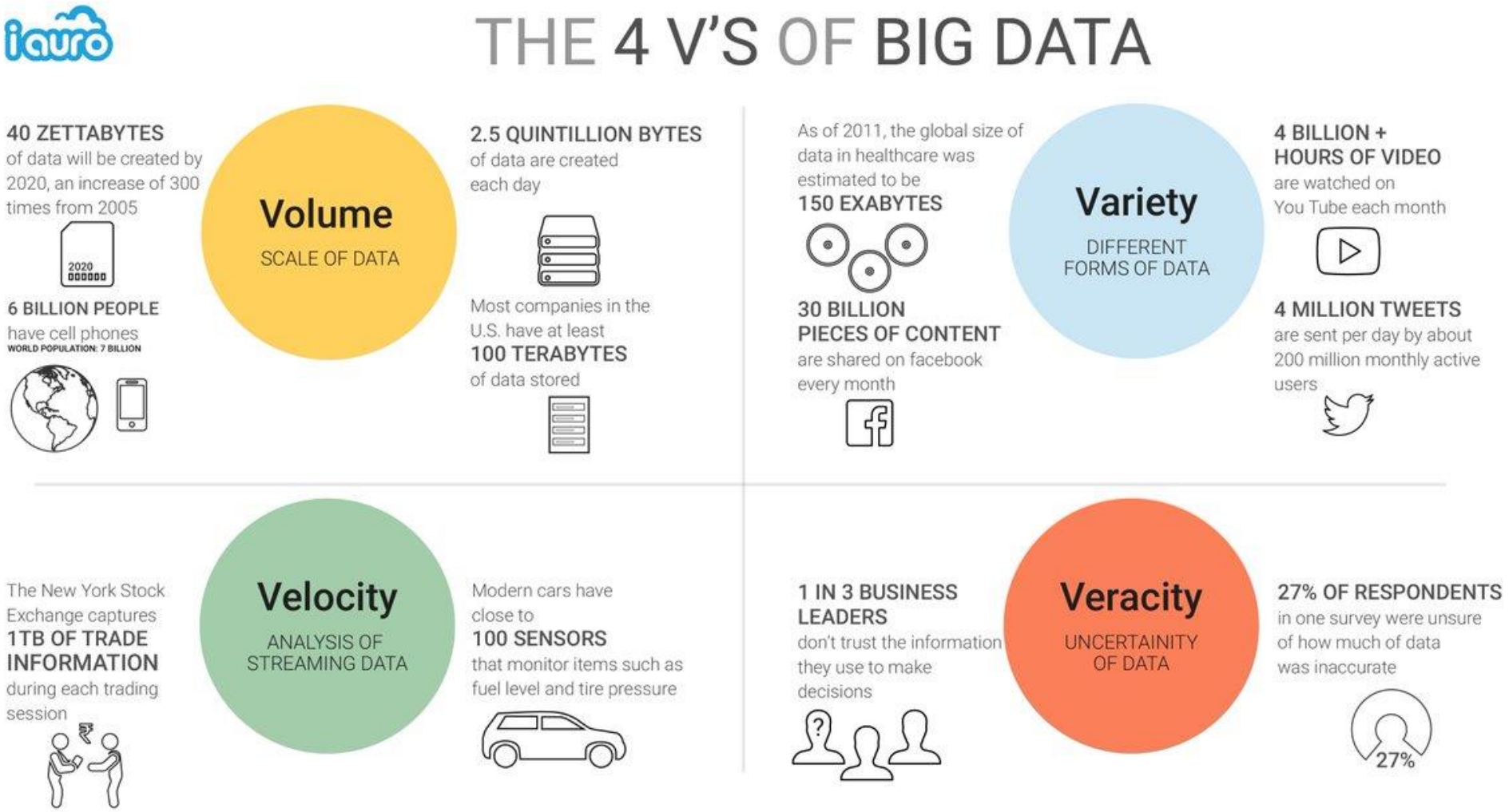
Polling – Wat bepaalt wanneer een data science project een big-data project is?



De drie hoofd V's

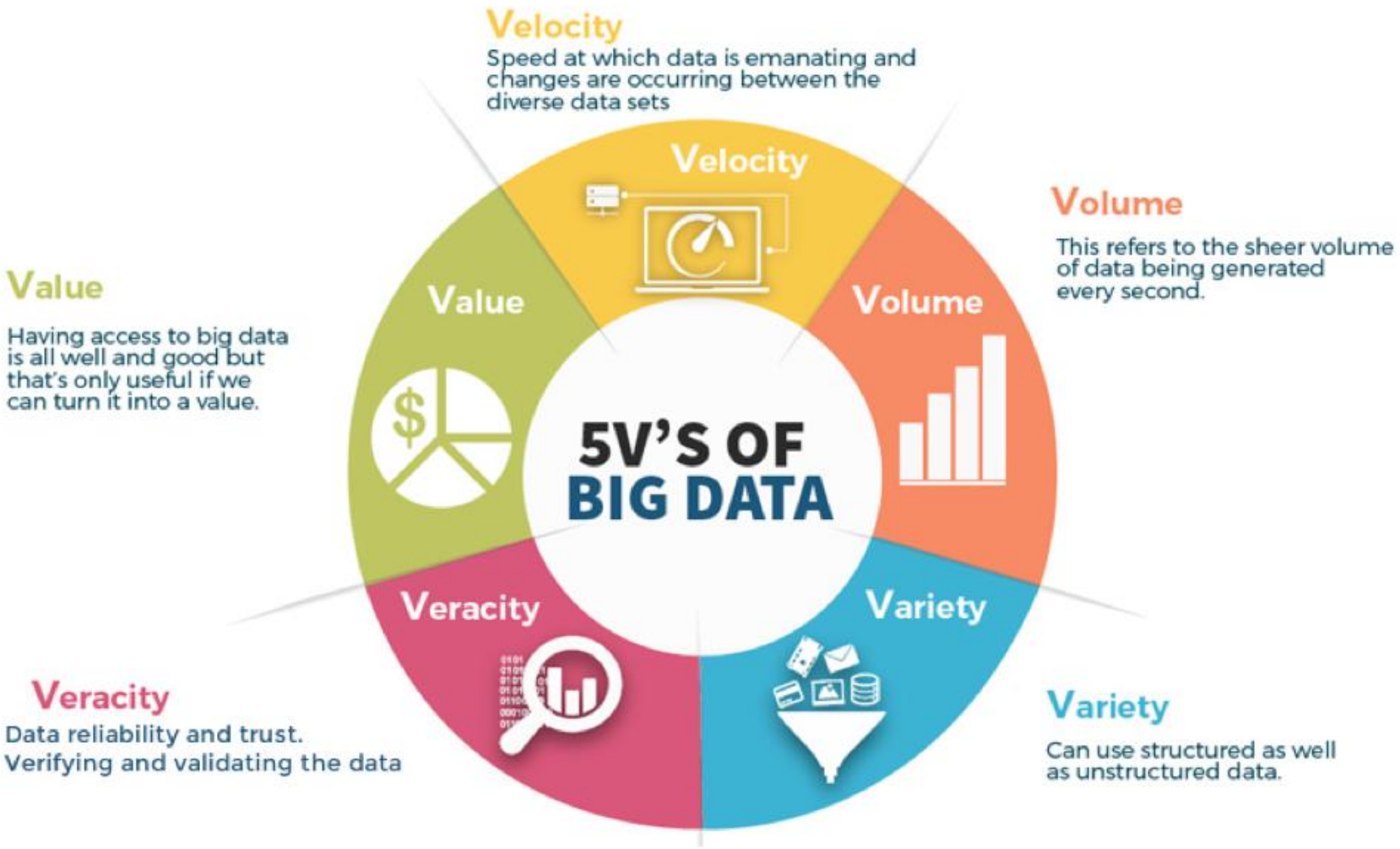


Extra V: Veracity



Reference : http://www.ibmbigdatahub.com/infographic/four-vs-big-data

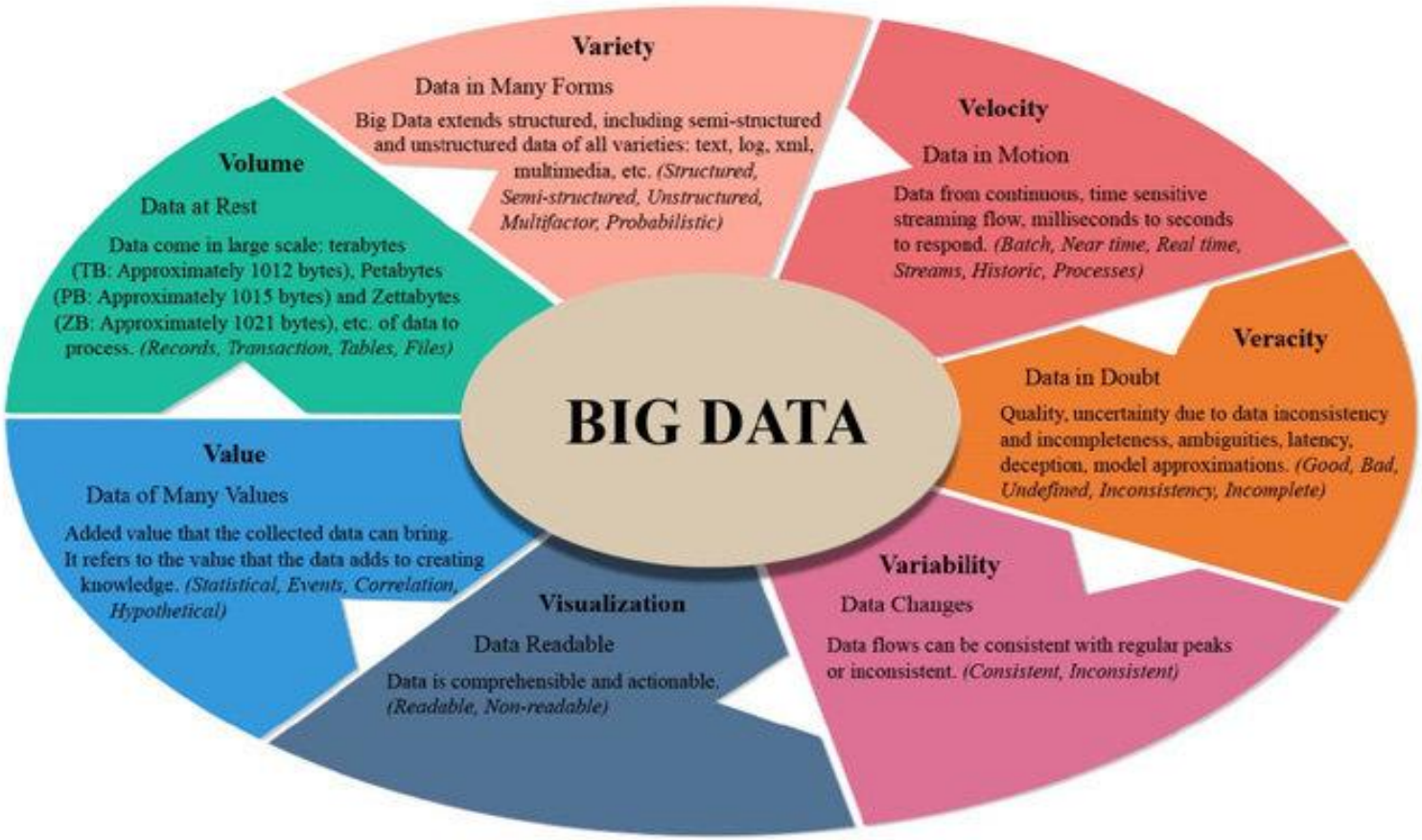
Of 5 V's: Value


Velocity
Speed at which data is emanating and changes are occurring between the diverse data sets
Volume
This refers to the sheer volume of data being generated every second.
Value
Having access to big data is all well and good but that's only useful if we can turn it into a value.
Veracity
Data reliability and trust. Verifying and validating the data
Variety
Can use structured as well as unstructured data.
Velocity
Volume
Value
Veracity
Variety
5V'S OF
BIG DATA

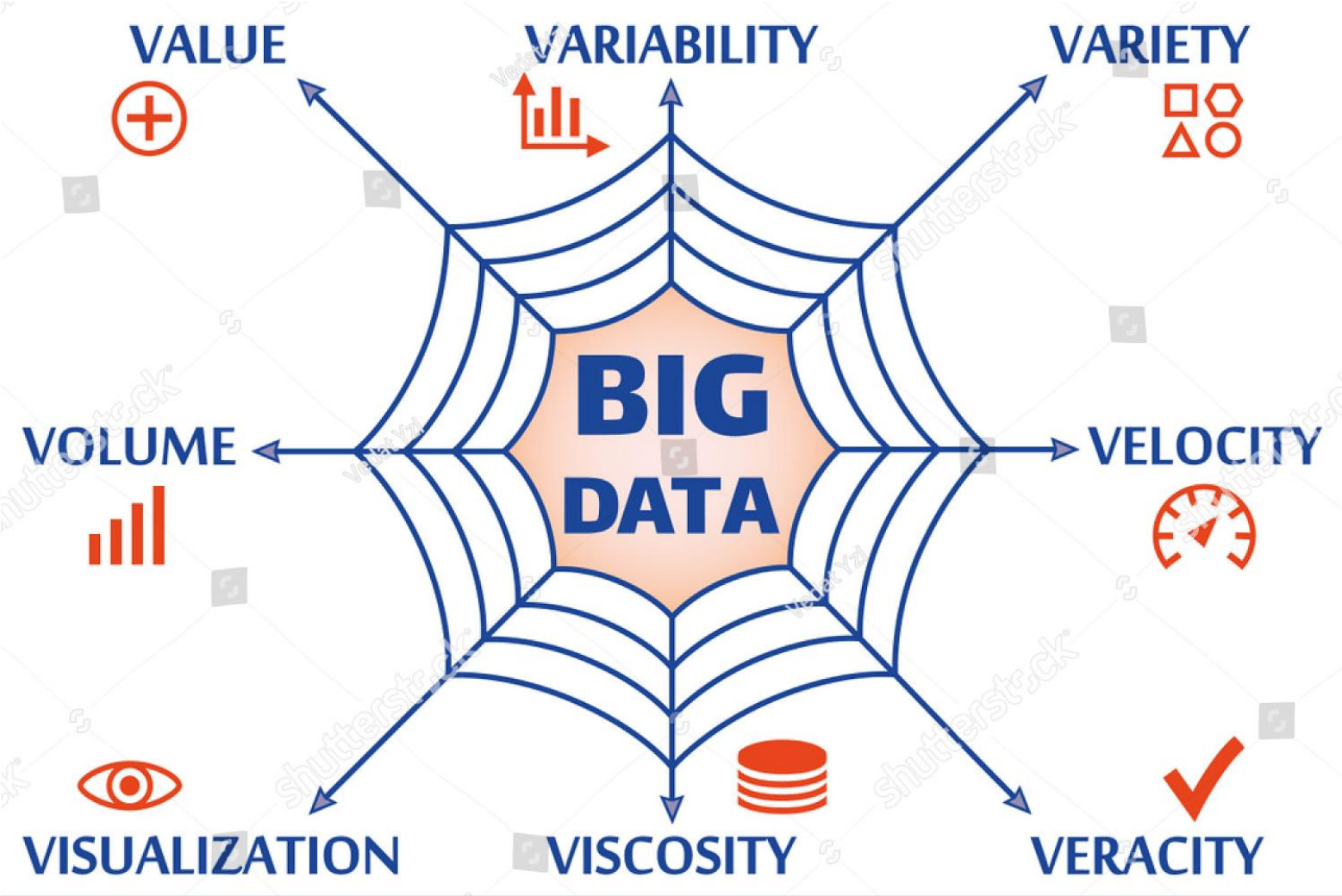
Of 6? Variability

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.

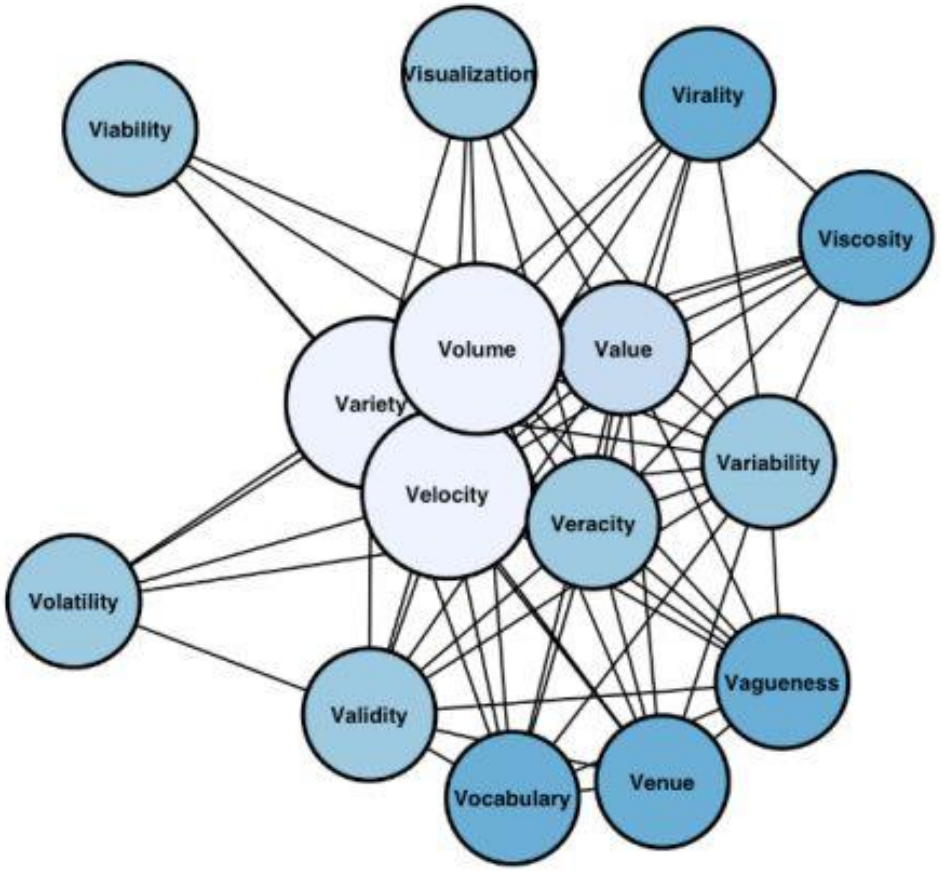
Of 7? Visibility



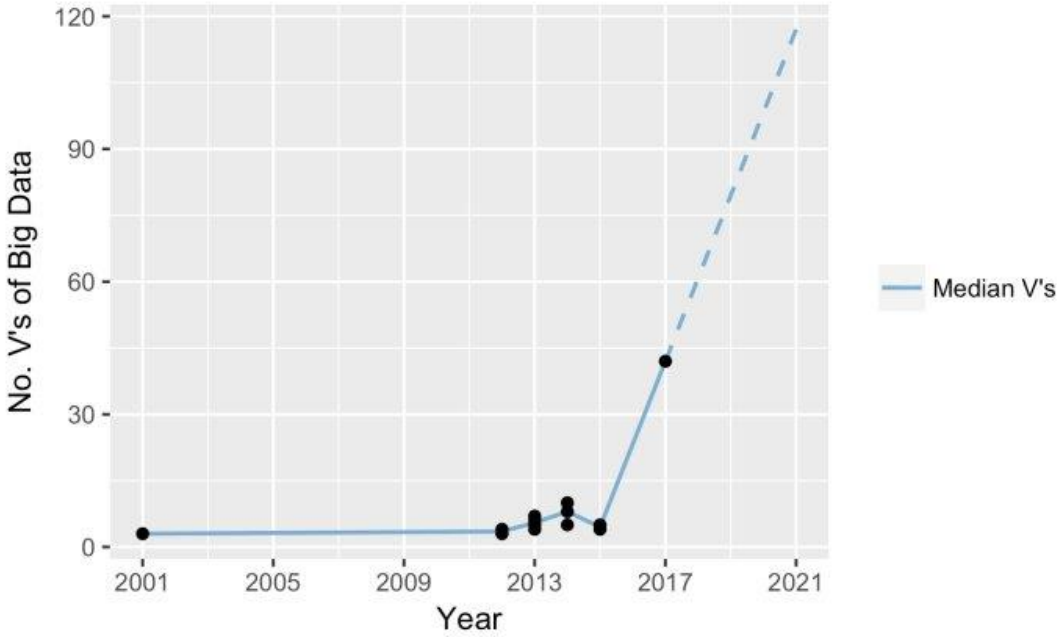
Of 8? Viscosity



10 V's of meer?


Visualization
Virality
Viability
Viscosity
Volume
Value
Variety
Variability
Velocity
Veracity
Volatility
Vagueness
Validity
Vocabulary
Venue


First Occurrence
2001
2012
2013
2014


120
90
60
30
0
No. V's of Big Data
2001
2005
2009
2013
2017
2021
Year
Median V's

3. Soorten data

Structured data

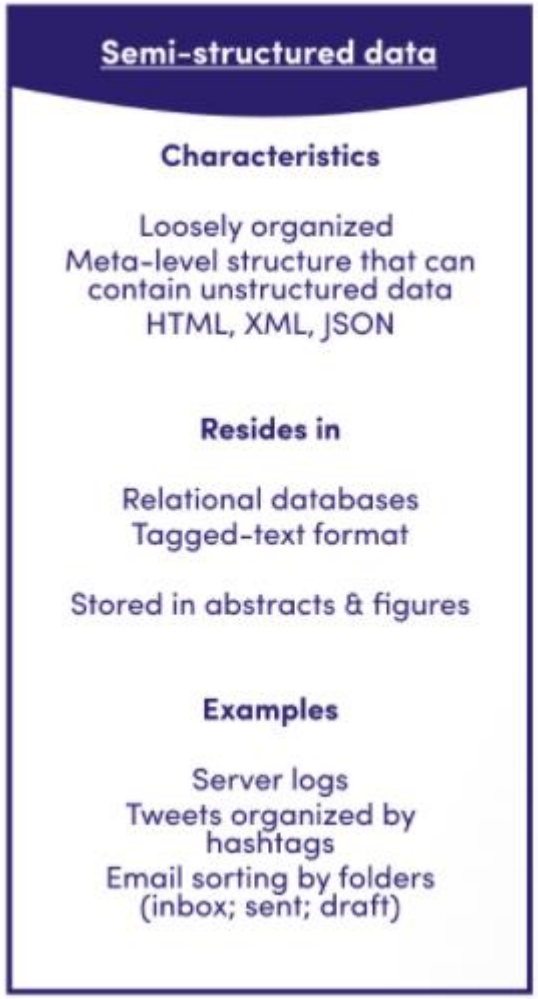
- Vast data formaat in tabel vorm met rijen en kolommen
- Alle formaten vooraf vastgelegd
- Excel files, Sql-database, csv, …

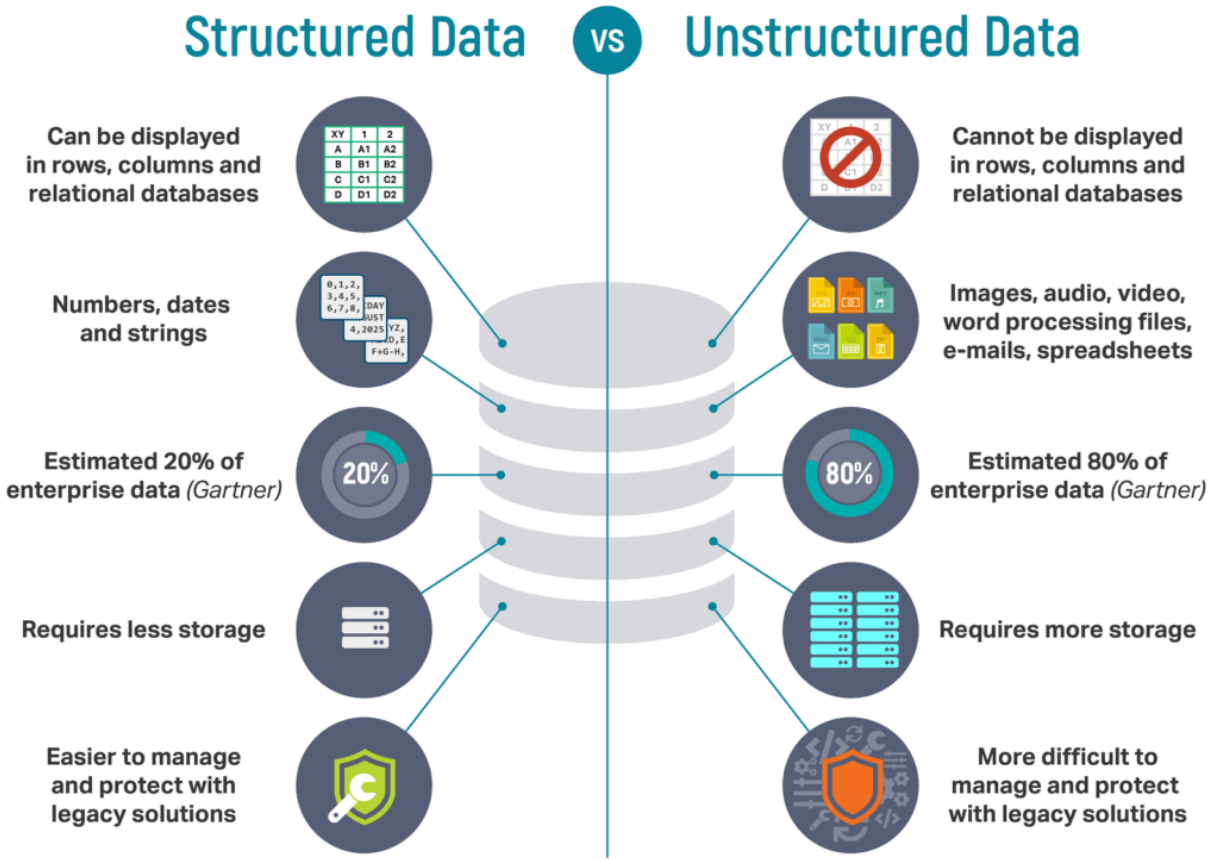
Unstructured data

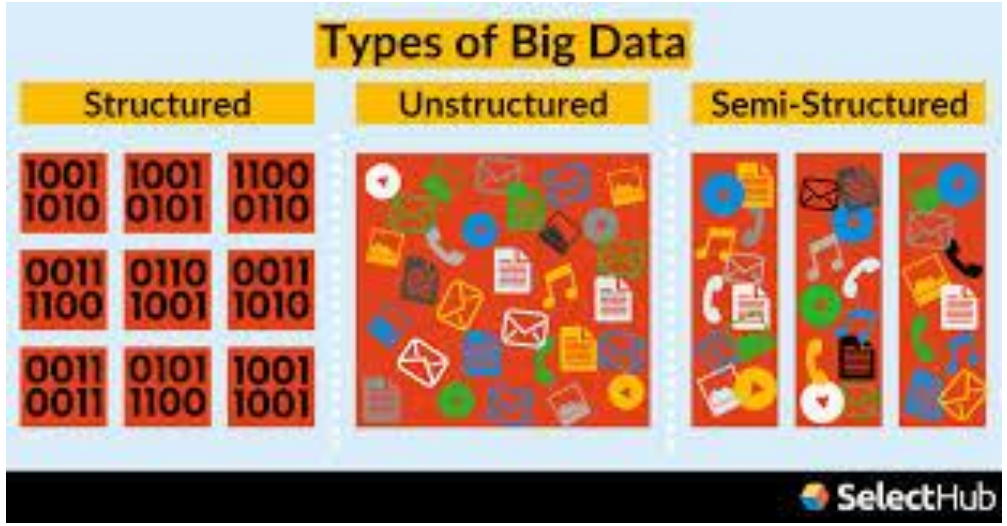
- Geen vaste structuur in de data
- Moeilijk om in te zoeken
- Foto’s, video’s, audio, tekst …

Semi-Structured data

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




Structured Data
vs
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

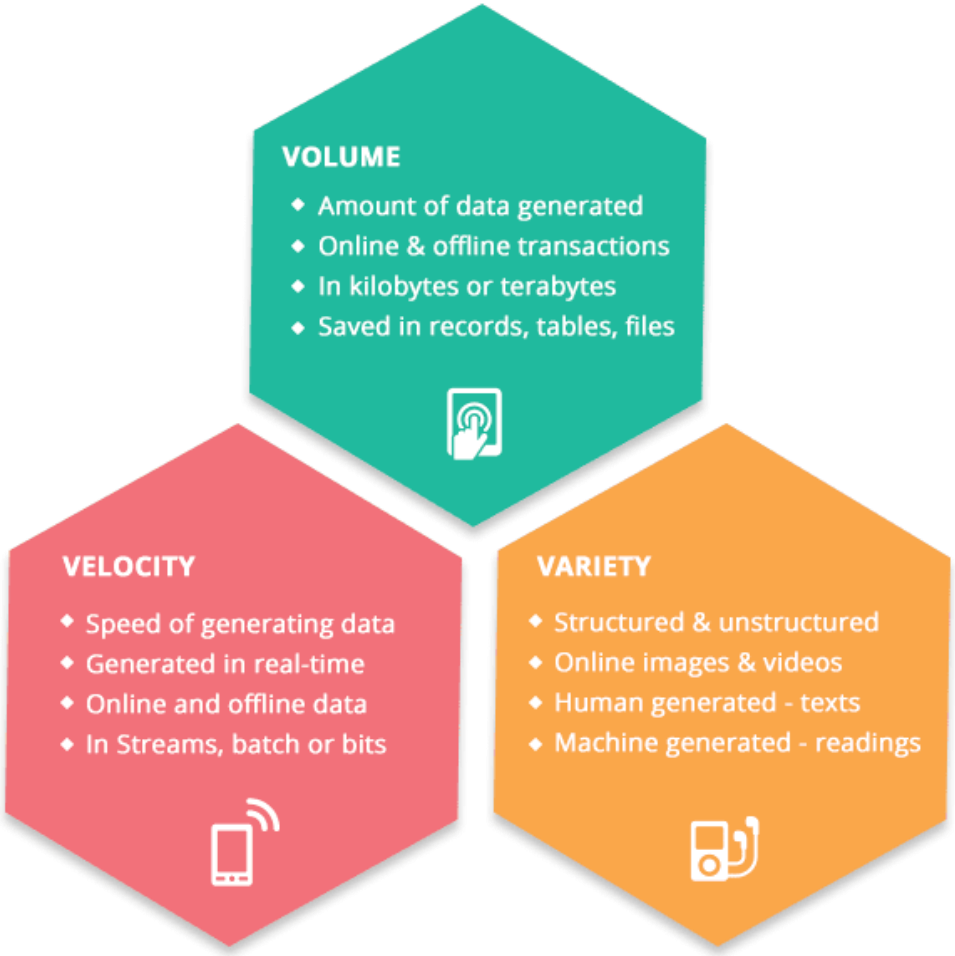

Types of Big Data
Structured
Unstructured
Semi-Structured
SelectHub

Door wie is de data geproduceerd?

Kritiek op Big Data

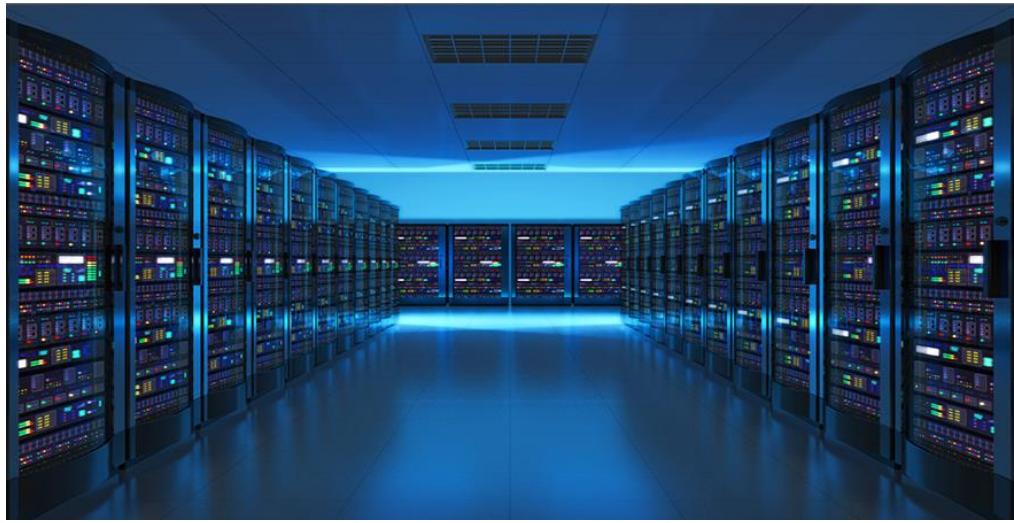
- Op de onderliggende theorie:
  - Toekomst gelijkaardig aan het verleden
  - Context afhankelijk
- V-model focust op schaalbaarheid en rekenkracht, niet op verklaarbaarheid
- Grote datasets en analyses bestaan reeds decennia, niet zo nieuw als veel denken
- Buzzword om aandacht te trekken naar je product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

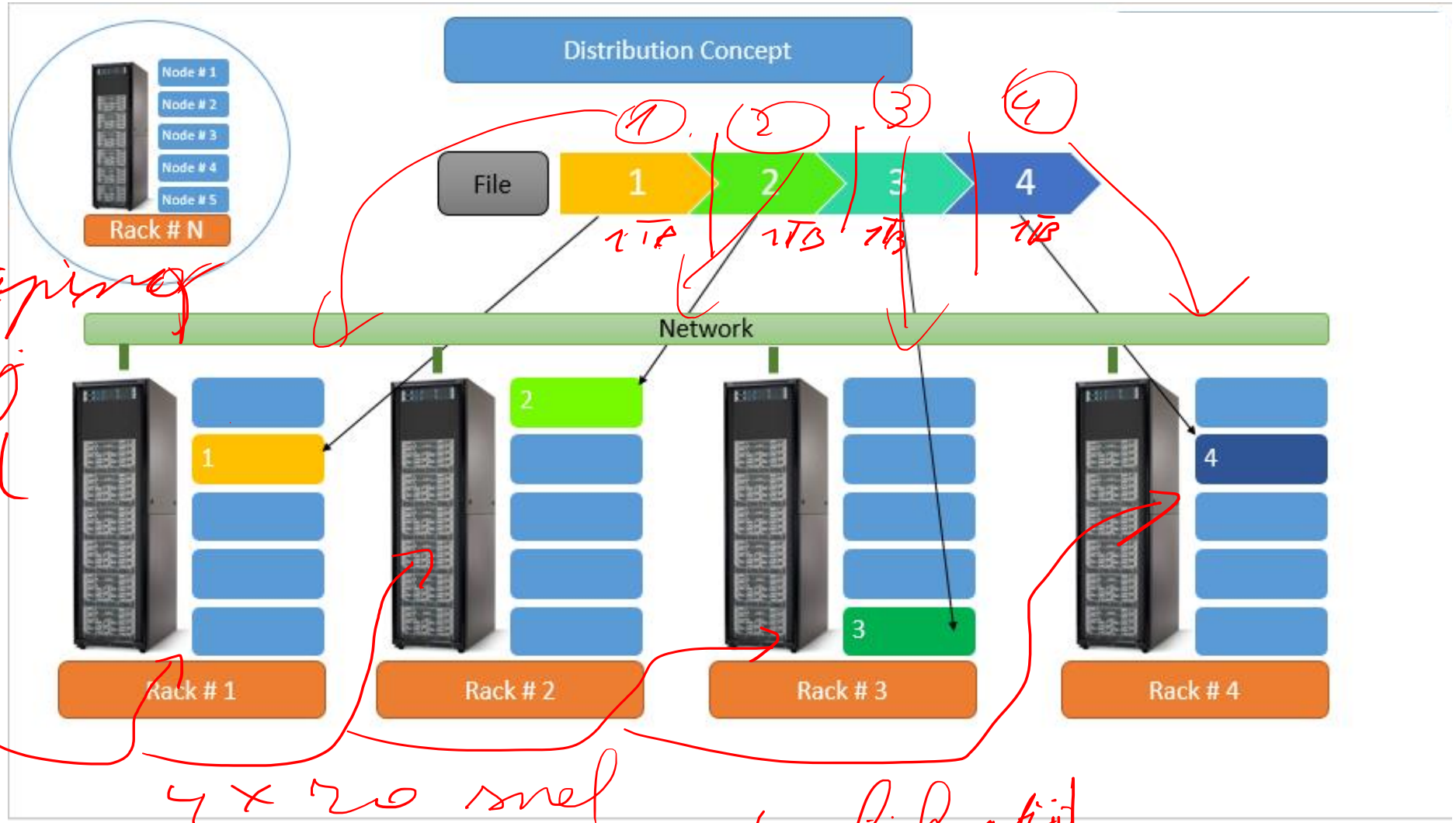
Hoe kan je omgaan met deze problemen?



Distributed storage

Van Pc -> Rack -> Datacenter -> Cloud





Striping van big Raid

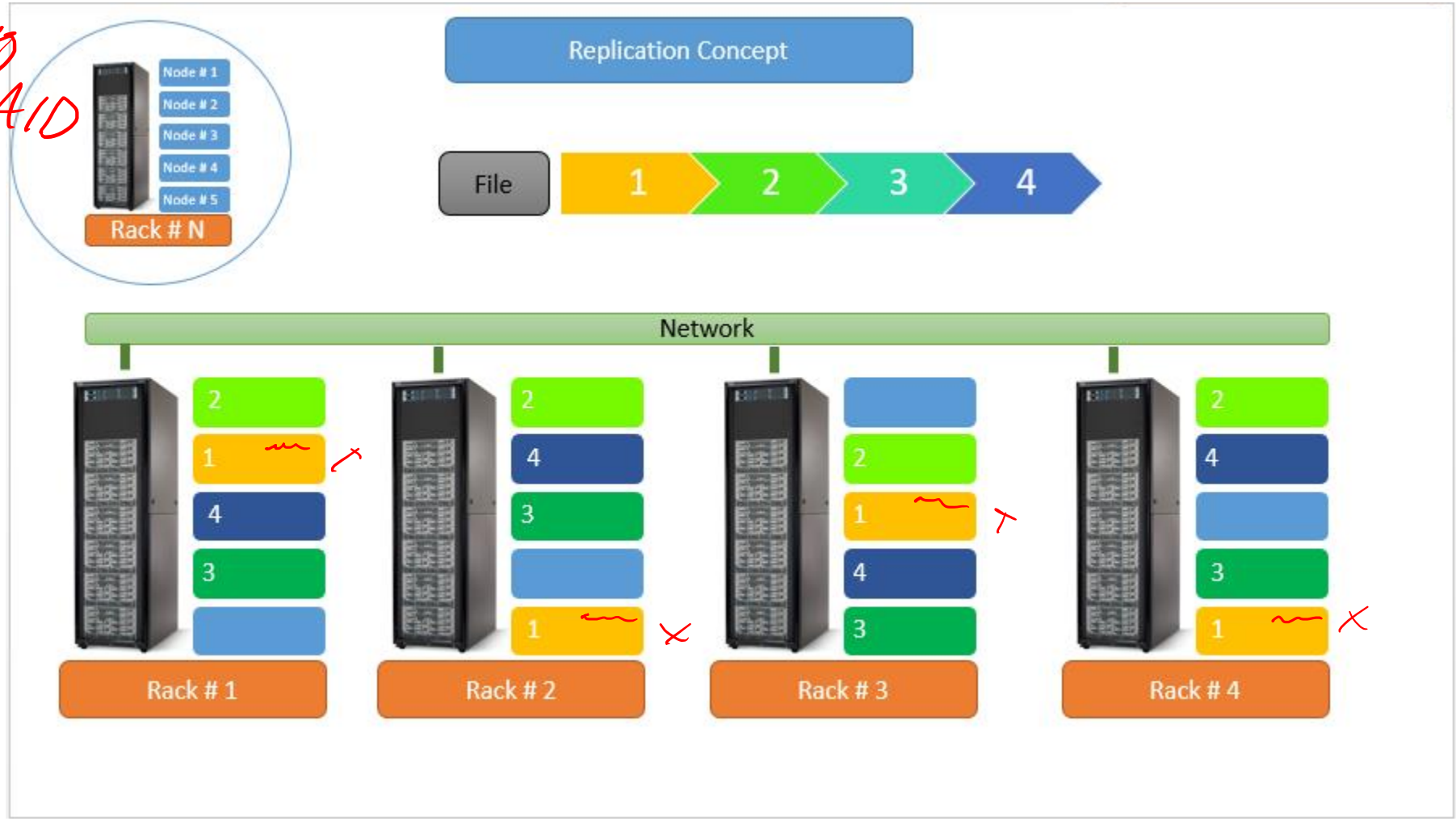
4 x 20 snel

→ 4 servers tegelijkertijd

Replication

↳ vergelijking van bij RAID

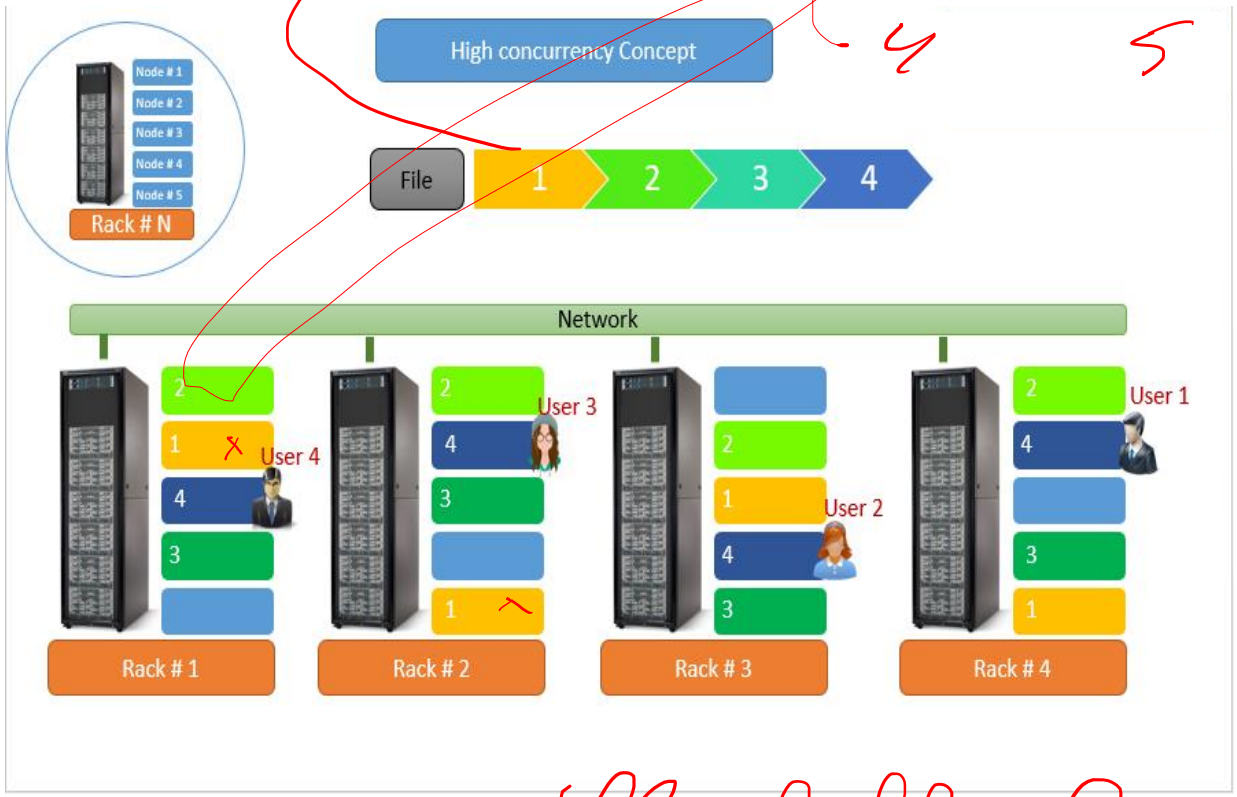
- Fault tolerance



Voordelen

- Schaalbaar → horizontaal schaalbaar
- Fout tolerant
- Nodige rekenkracht ook verdeeld
  - Concurrency → parallel

- Goedkoper
  - Minder gespecialiseerde computers
  - Commodity clusters



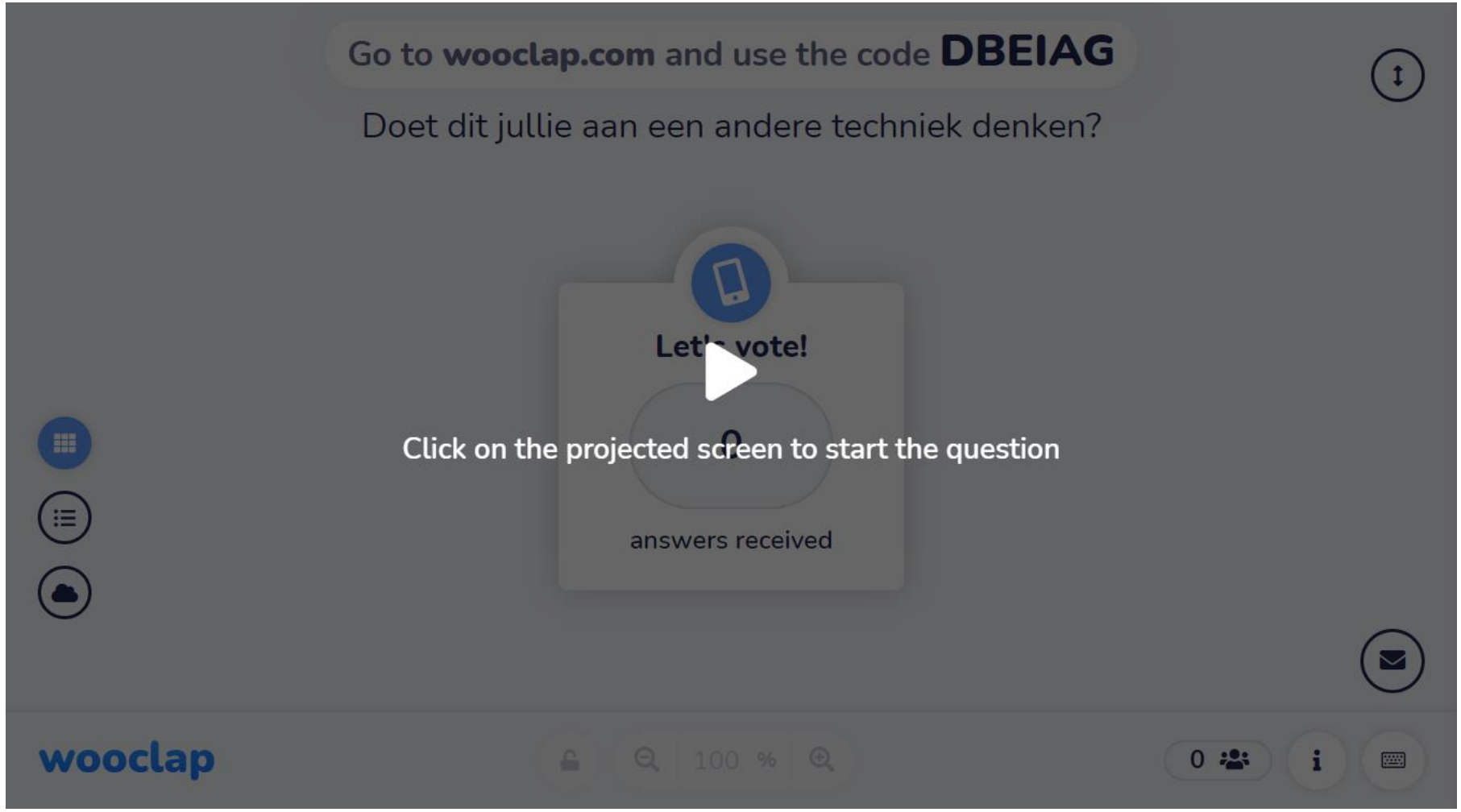
4 users willen lezen
↳ zonder replica's → wachten op elkaar
↳ met " → parallel

Nadelen

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

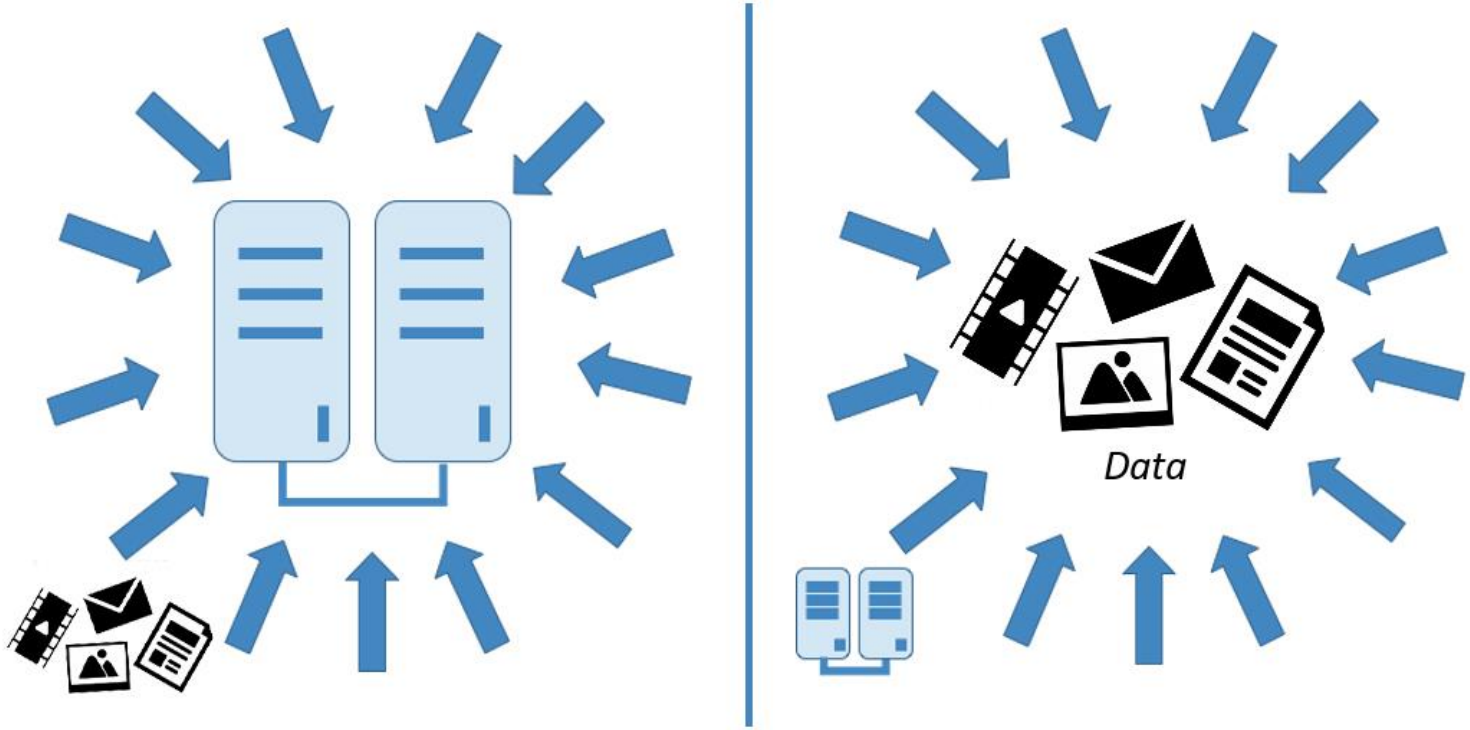
↳ lazy / eventual

Doet dit jullie aan een andere techniek denken?



Bring computing to data

Computing to data



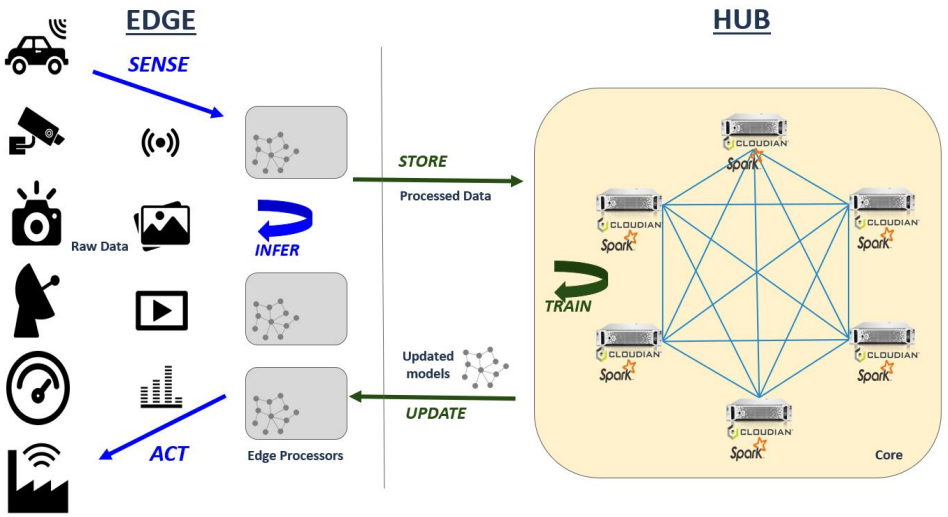
Code / Programma niet gemakkelijk te migreren
Rekeneenheid moet krachtig zijn

Local

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

Distributed

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

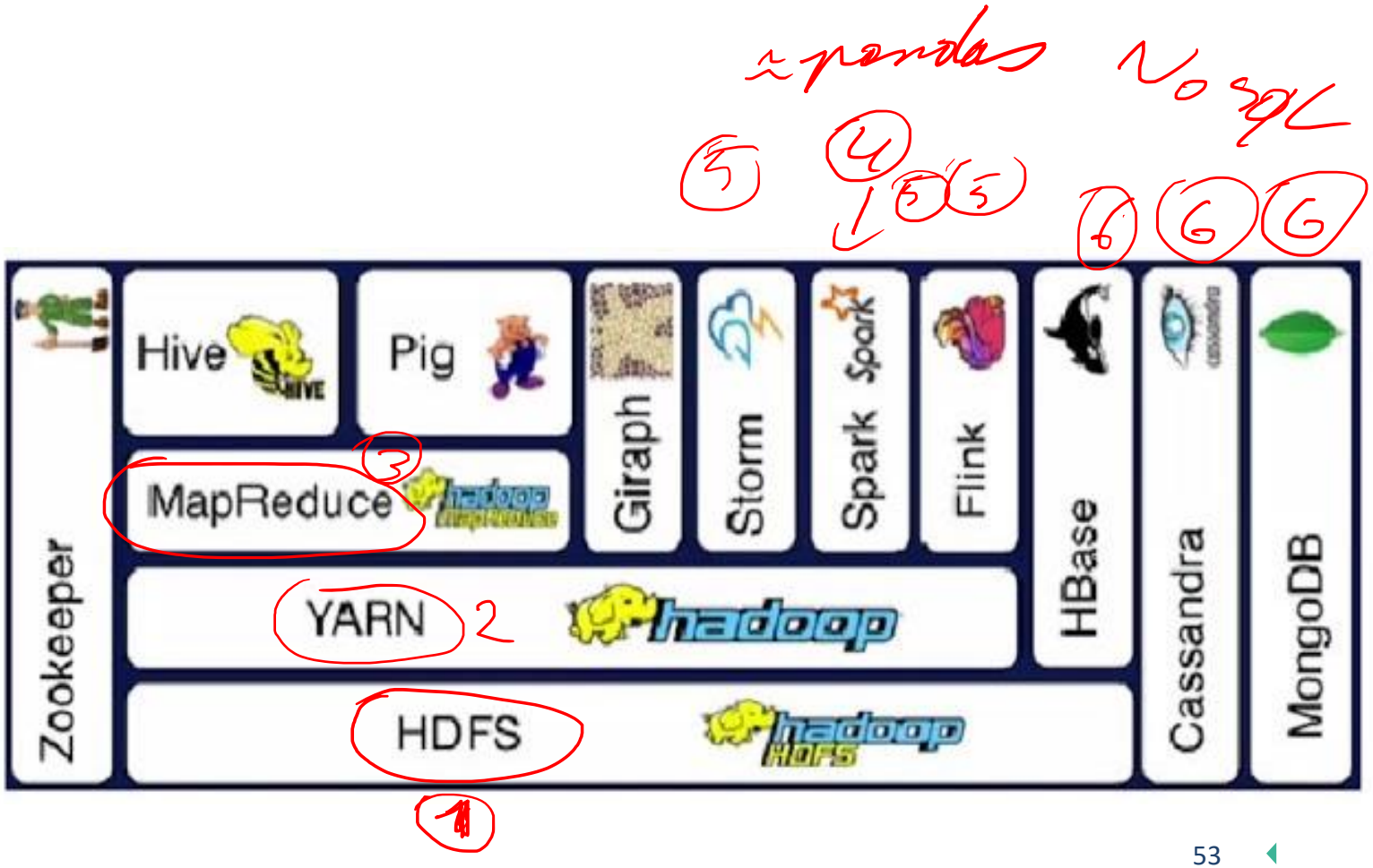
Hadoop Ecosystem

Hadoop

- Gebaseerd op Google File System (2003)
- Ontwikkeld door Apache
- Open source
- Uitgegroeid tot omgeving met veel verschillende applicaties

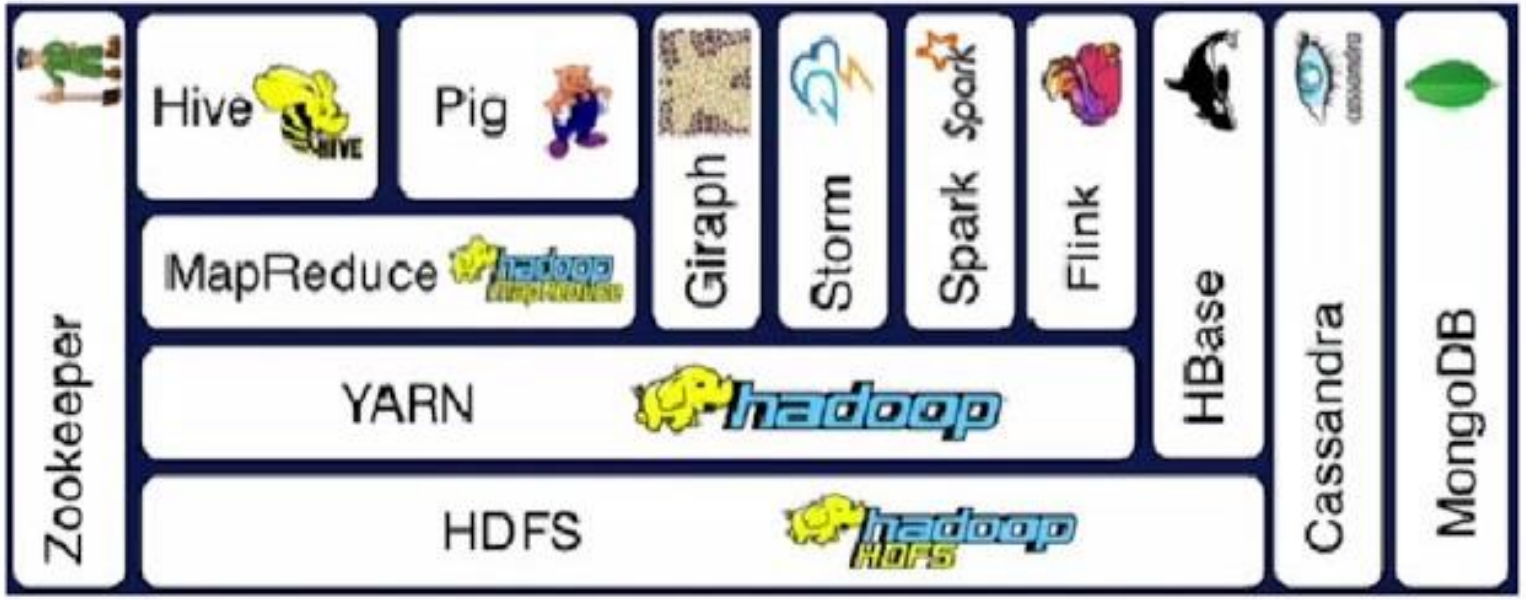
Hadoop Ecosystem

- HDFS – core functionality
- Distributed File System
- Op HDD



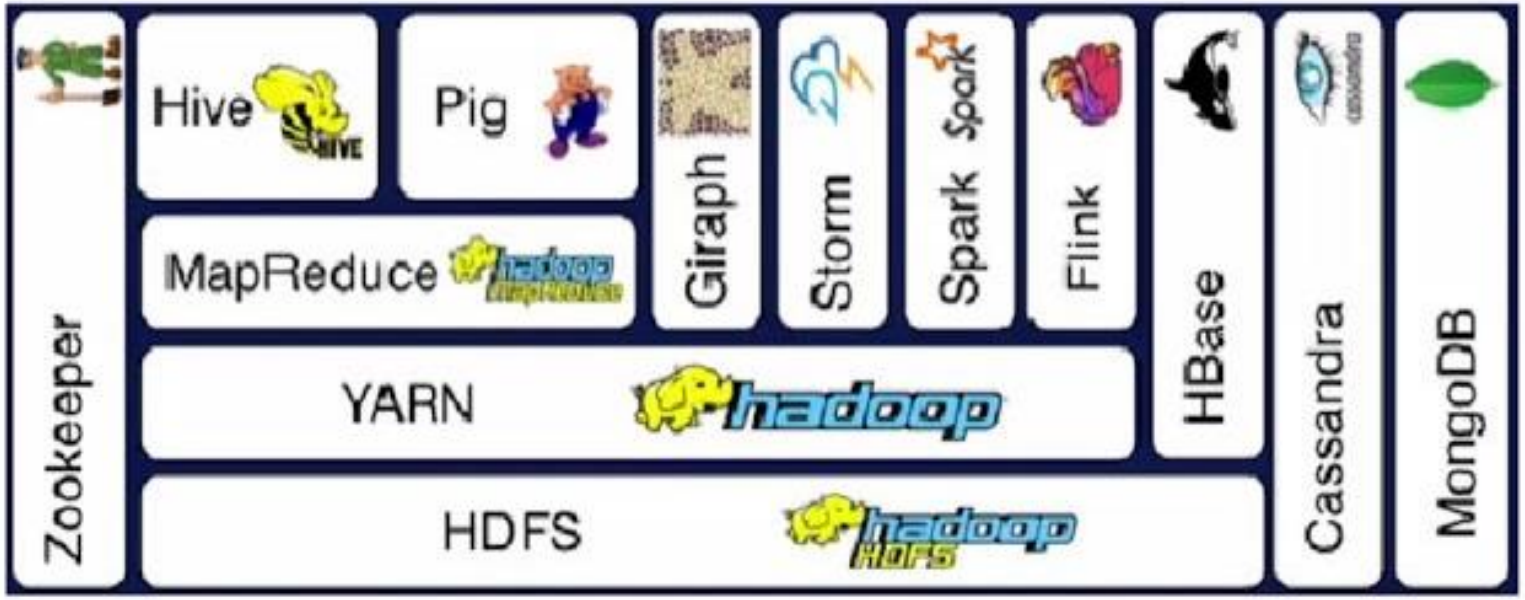
Hadoop Ecosystem

- YARN – Yet Another Resource Manager
- Beheer van computing power
- Welke code op welke node



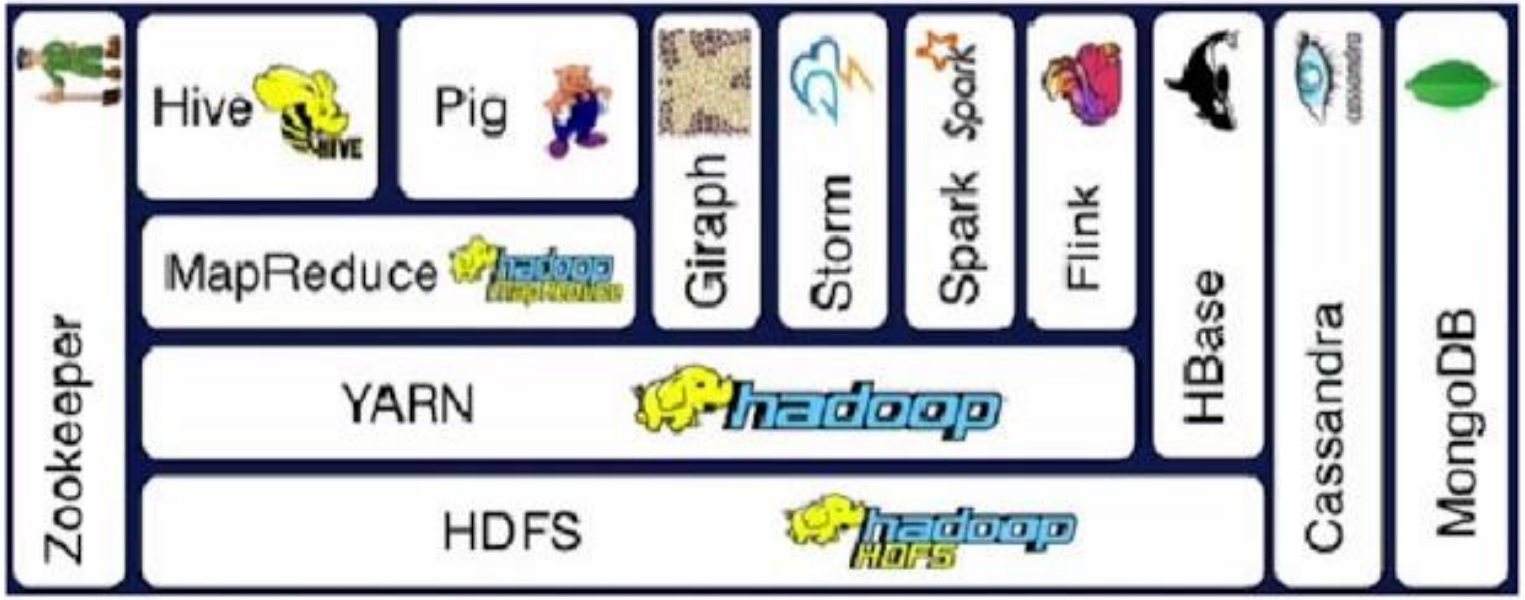
Hadoop Ecosystem

- MapReduce
- Distributed Computing
- Ontwikkeld door Google
- 2 fases
  - Mapping (Divergeren)
  - Reduce (Convergeren)



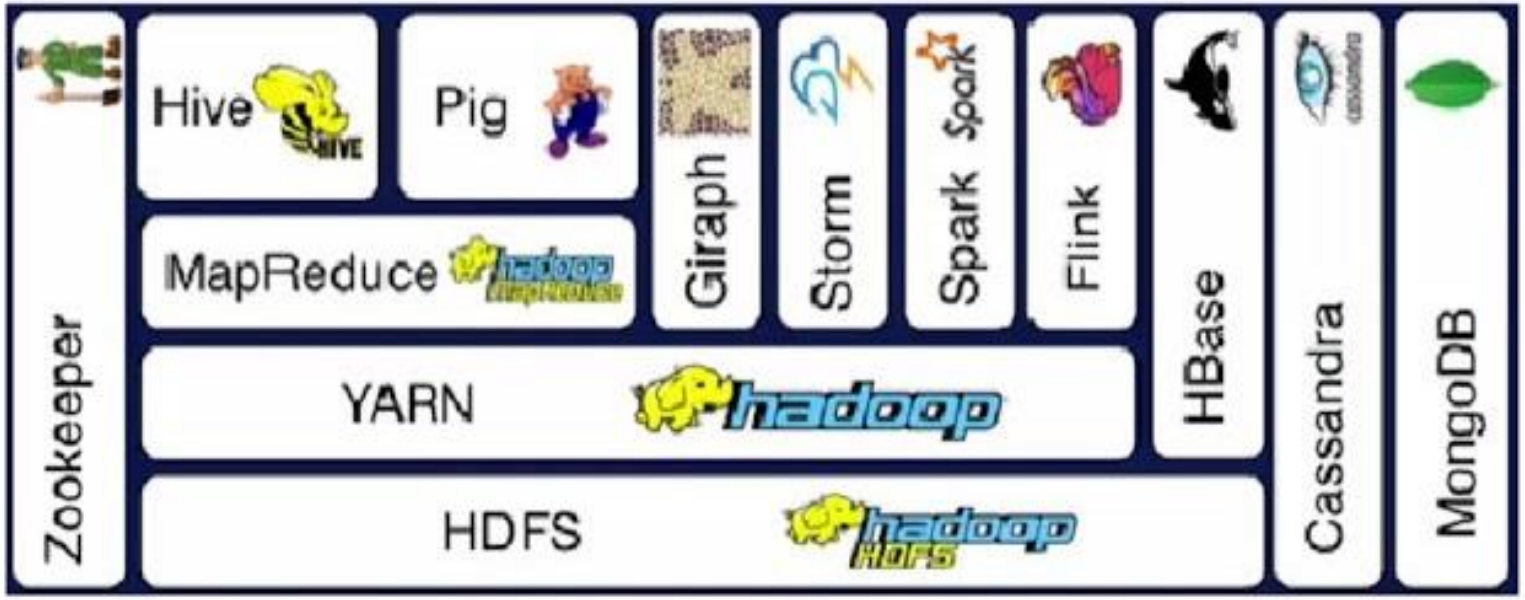
Hadoop Ecosystem

- Zookeeper
- Beheren van alle applicaties die lopen op de verschillende nodes



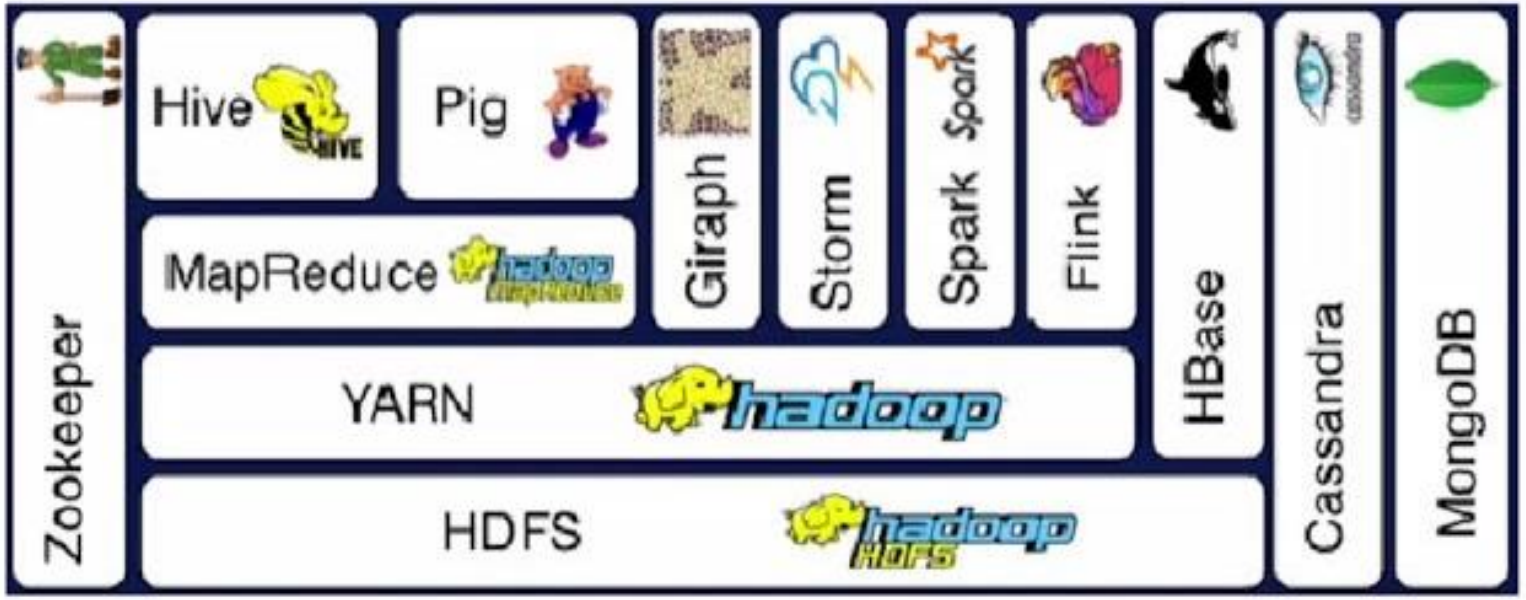
Hadoop Ecosystem

- Hive
- Distributed Datawarehouse
- Sql-like
- Queries via MapReduce



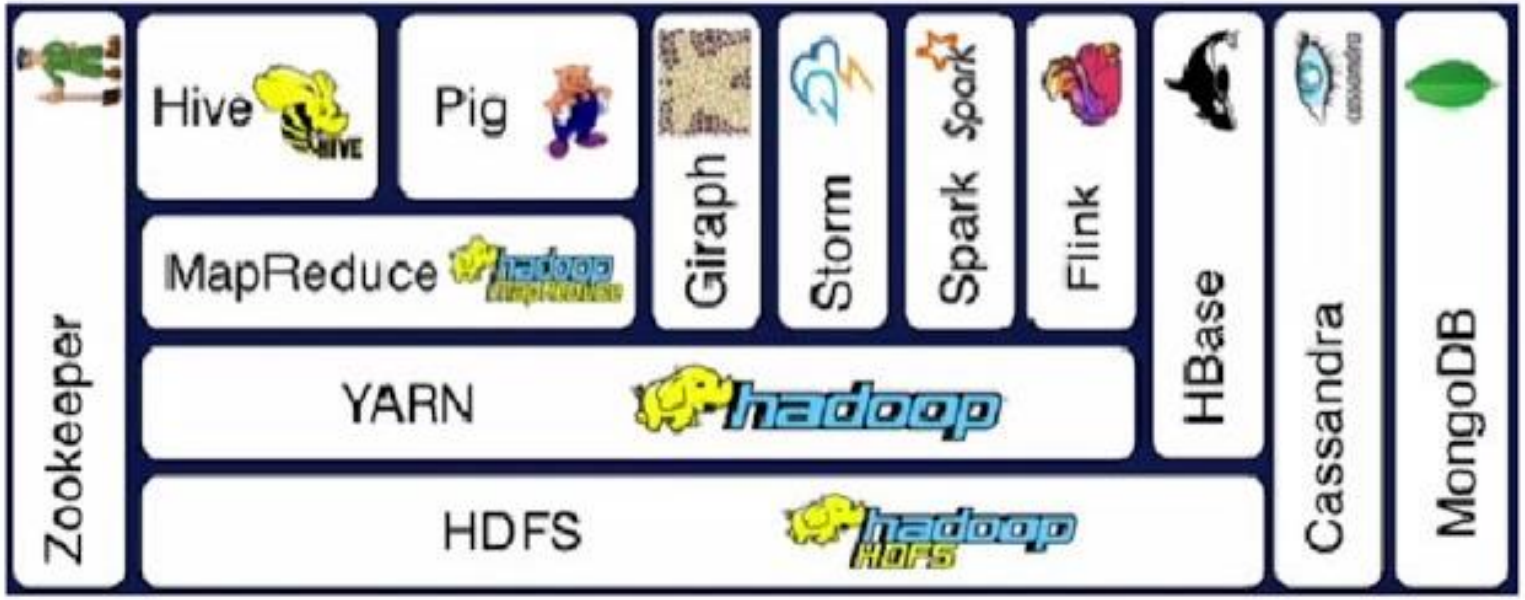
Hadoop Ecosystem

- Pig
- Data analysis
- Using MapReduce/Spark/…
- Taal: Pig Latin



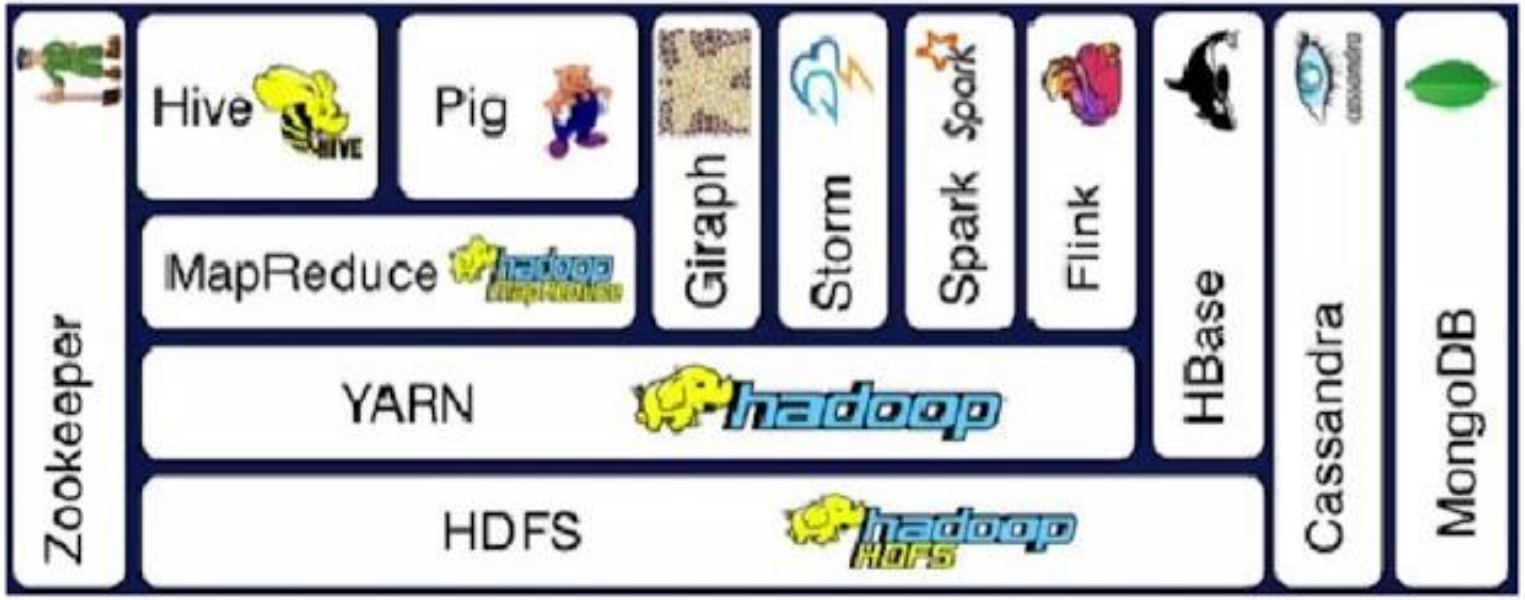
Hadoop Ecosystem

- Giraph
- Bestuderen van een graaf
- Social graph
  - Facebook
  - Twitter
  - …
- Gebruikt geen mapreduce



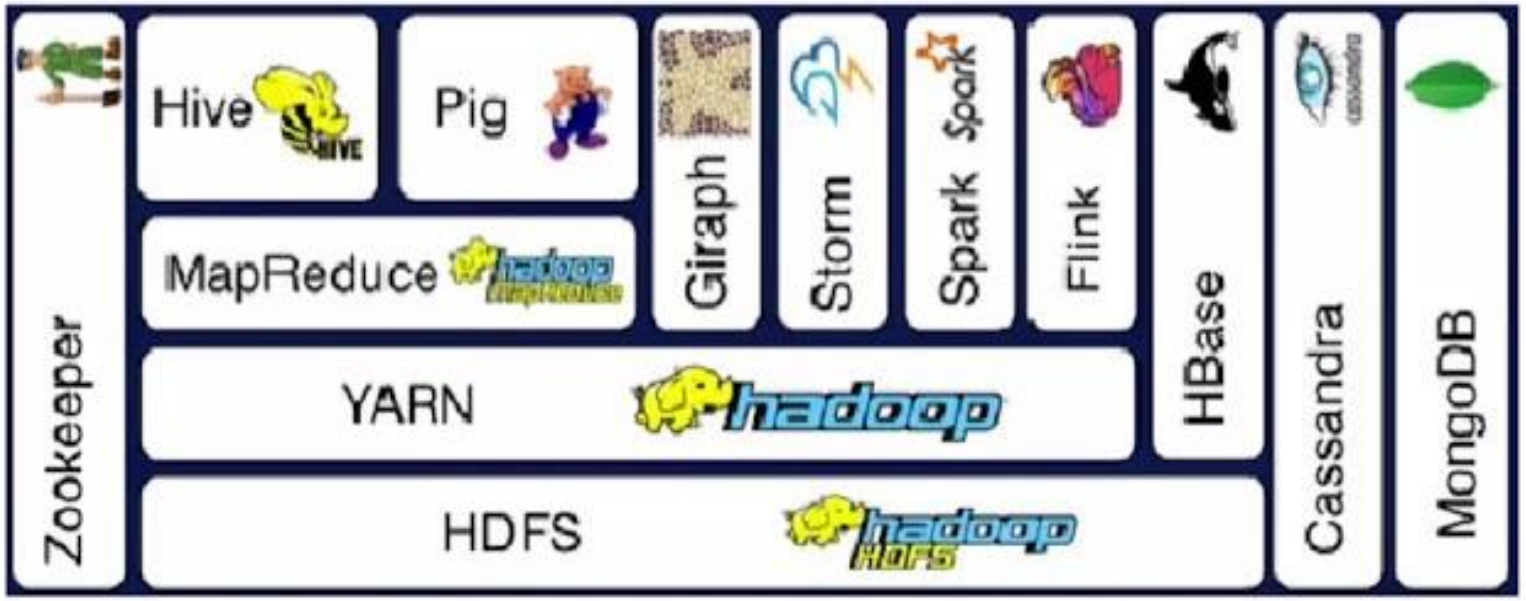
Hadoop Ecosystem

- Storm / Flink
- Verwerken van data streams – continue inkomende datastromen
  - Classificeren
  - Opslaan
  - …



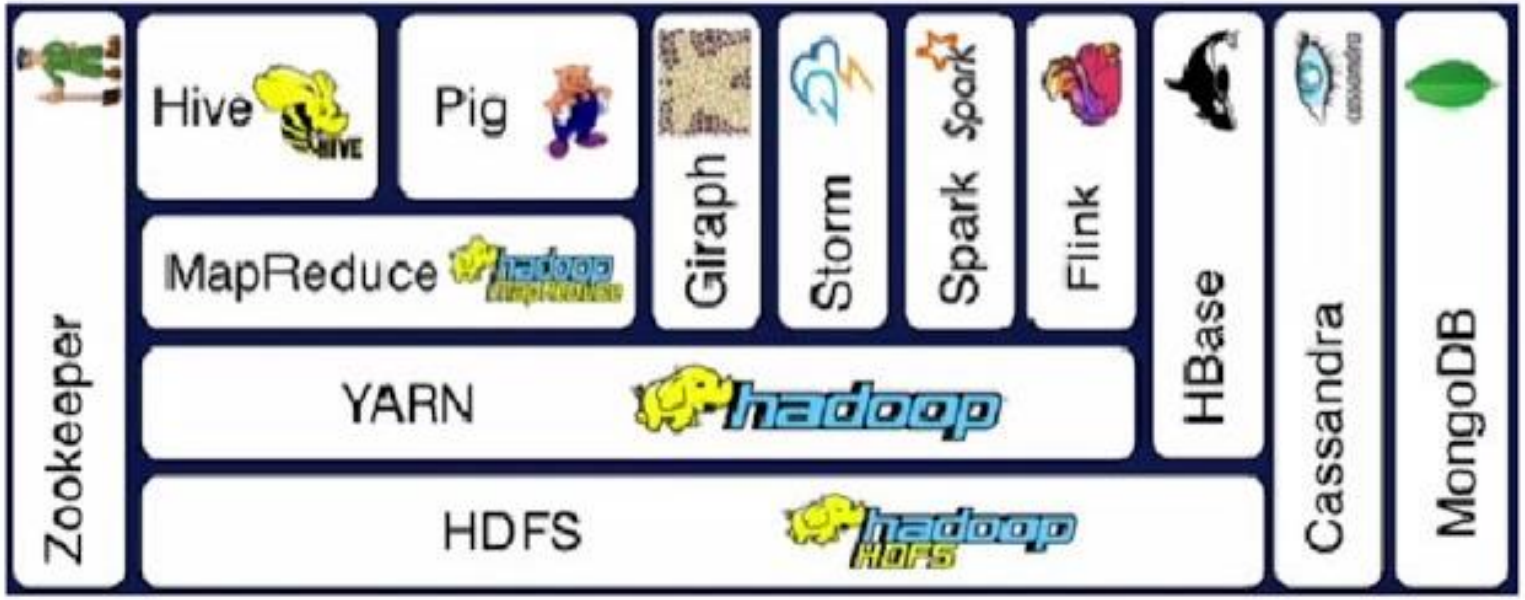
Hadoop Ecosystem

- Spark
- Alternatief voor MapReduce
- Computing in Ram
- Op Hadoop/Cloud/…
- Gebruikt voor
  - SQL (Spark SQL)
  - Streaming (Spark Streaming)
  - Machine Learning (MLlib)
  - Graph analysis (GraphX)



Hadoop Ecosystem

- HBase
- Distributed NoSQL Database
- Geen SQL maar in JAVA



Hadoop Ecosystem

- Cassandra / Mongo DB
- Maken geen gebruik van HDFS
- NoSql databases
- Stand-alone solutions