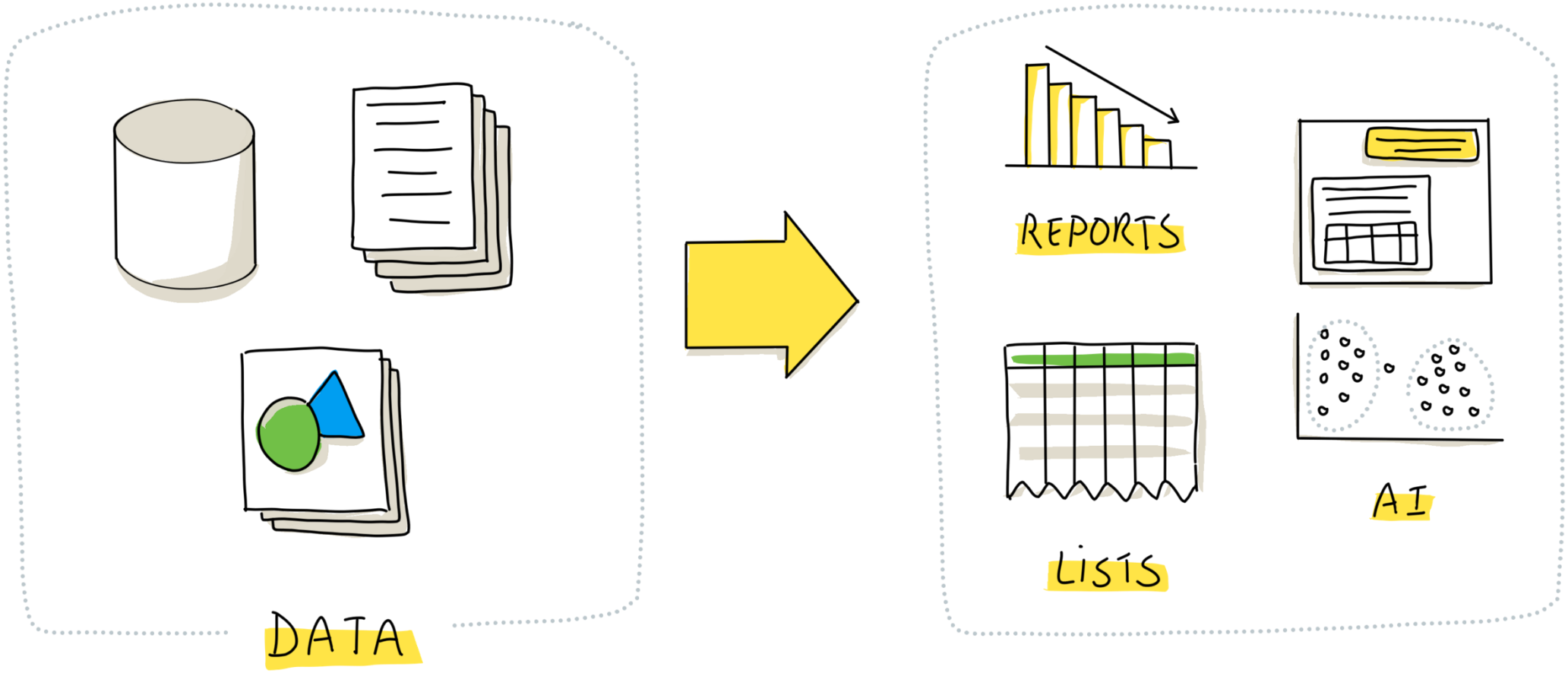
DEEL 9

DATA MANAGEMENT

Why **Data Management** in the context of **AI**?



Data Management

- Defining Data
- Data Producers
- Big Data Vs
- The Bigger Picture
- Data Management
- Data Technology

Data Management

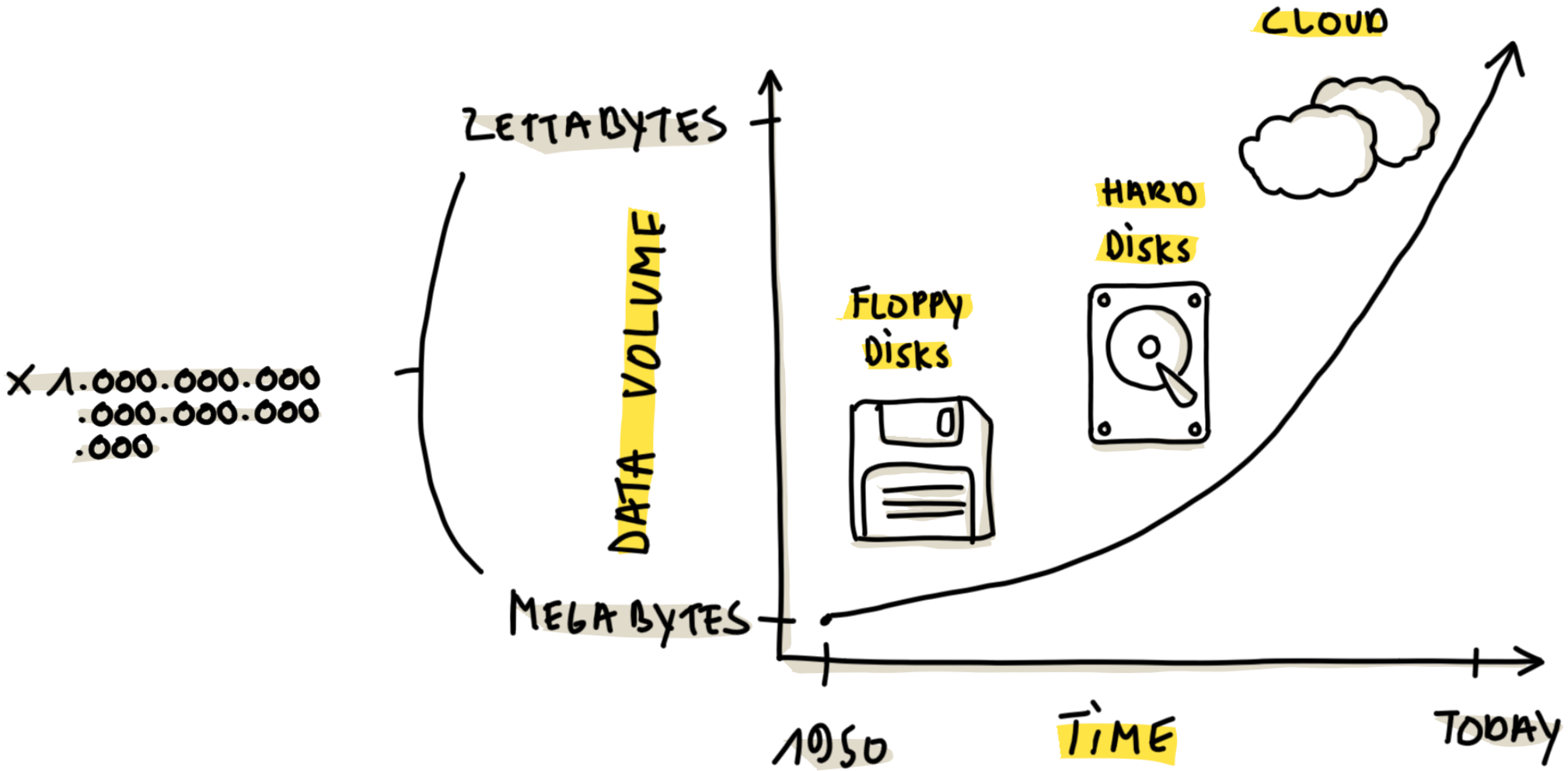
- **Defining Data**
- Data Producers
- Big Data Vs
- The Bigger Picture
- Data Management
- Data Technology

Defining **Data**

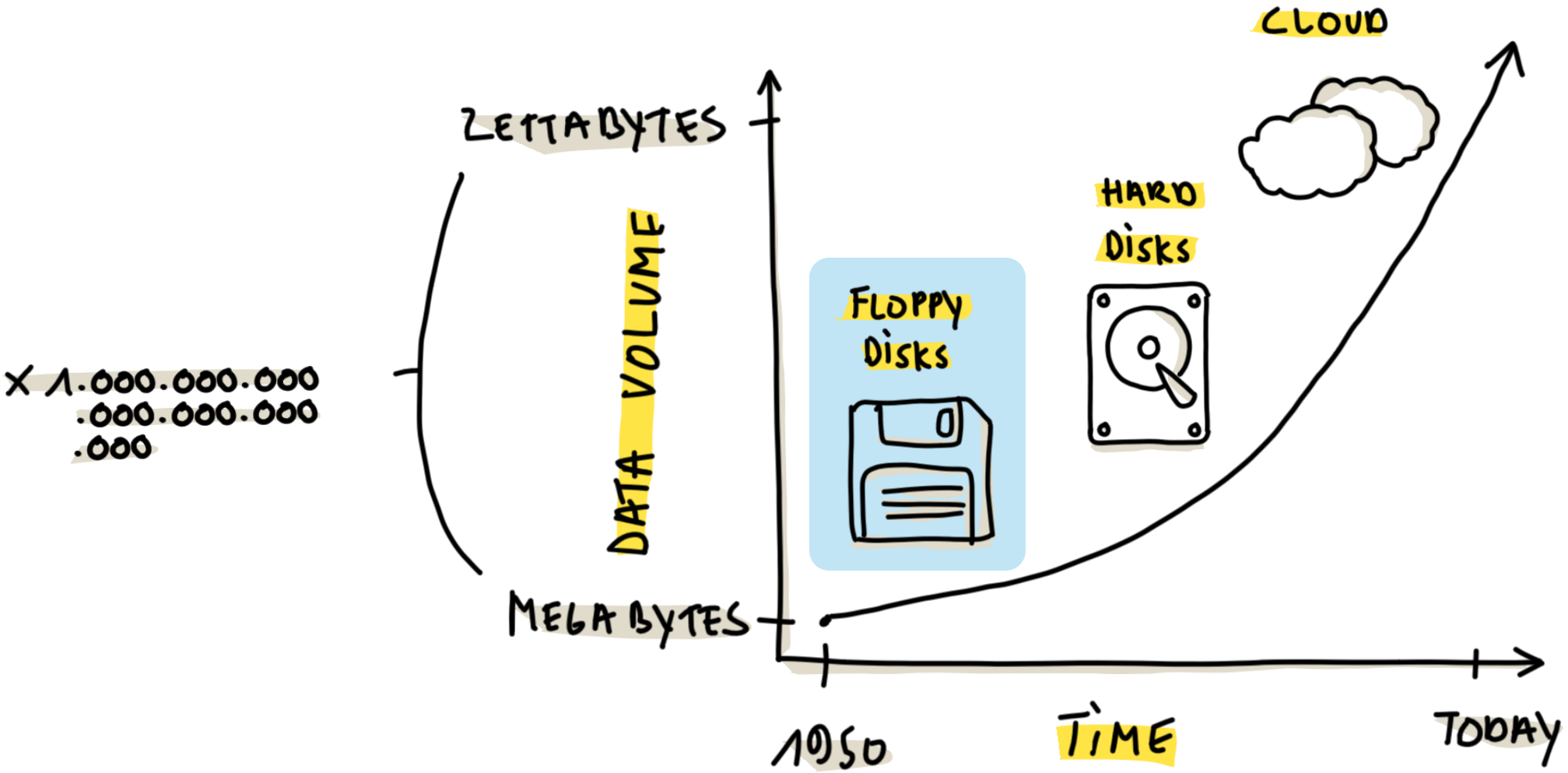
Data = Digital information that is stored, processed, and manipulated by computers and other devices.

Examples of data: text, images, figures, video, spreadsheets, word documents, internet websites, sensor measurements, …

The world of data changed a lot since the beginning (1950s) and today…


CLOUD
ZETTABYTES
HARD
DISKS
FLOPPY
DISKS
DATA VOLUME
X 1.000.000.000
.000.000.000
.000
MEGABYTES
1950
TIME
TODAY

History of Data



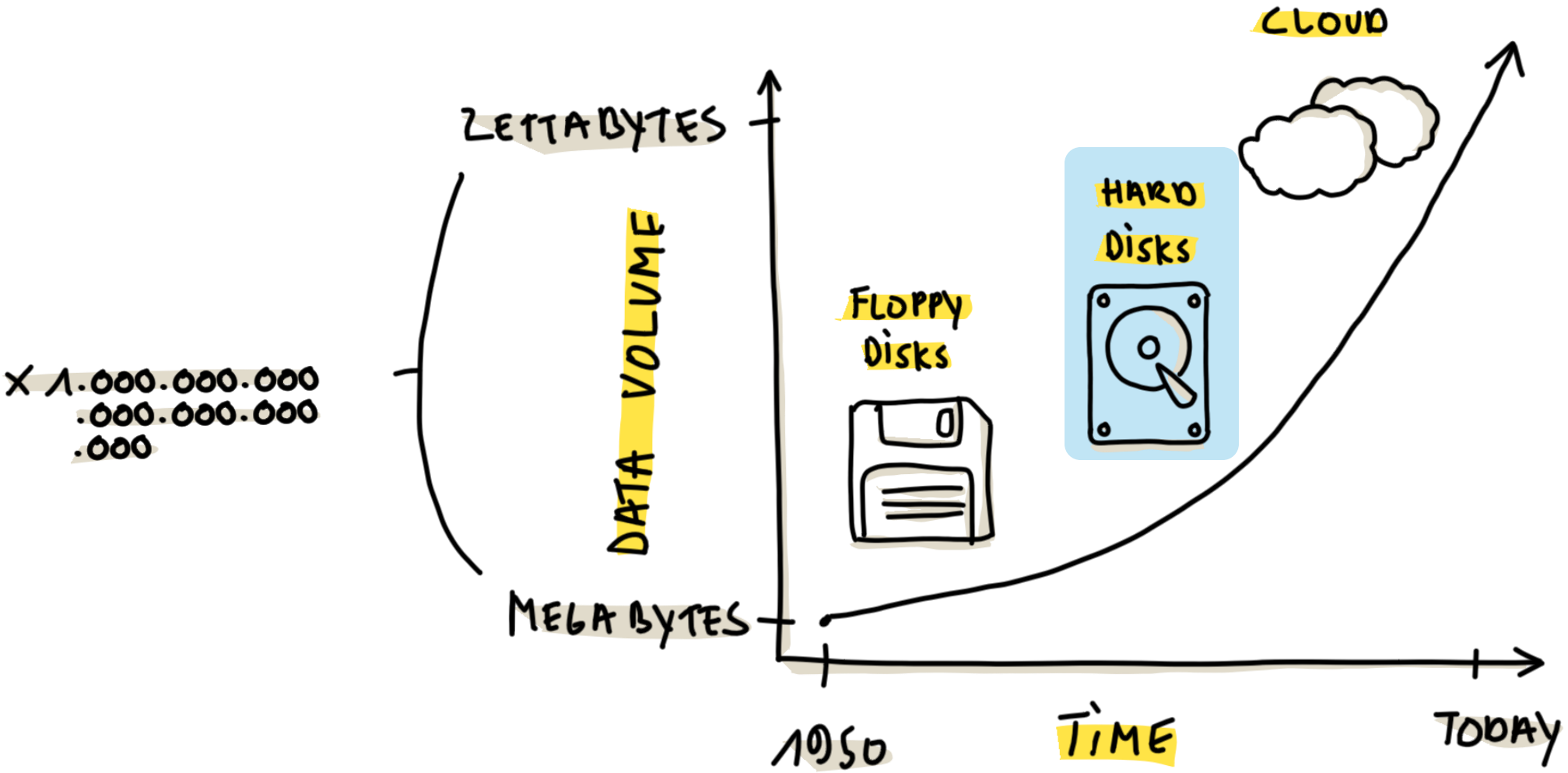

ЕС5274 DS/DD
ИЗОМ
Сделано в България
boedem
FLOPPY DISK
5,25" HD,DS

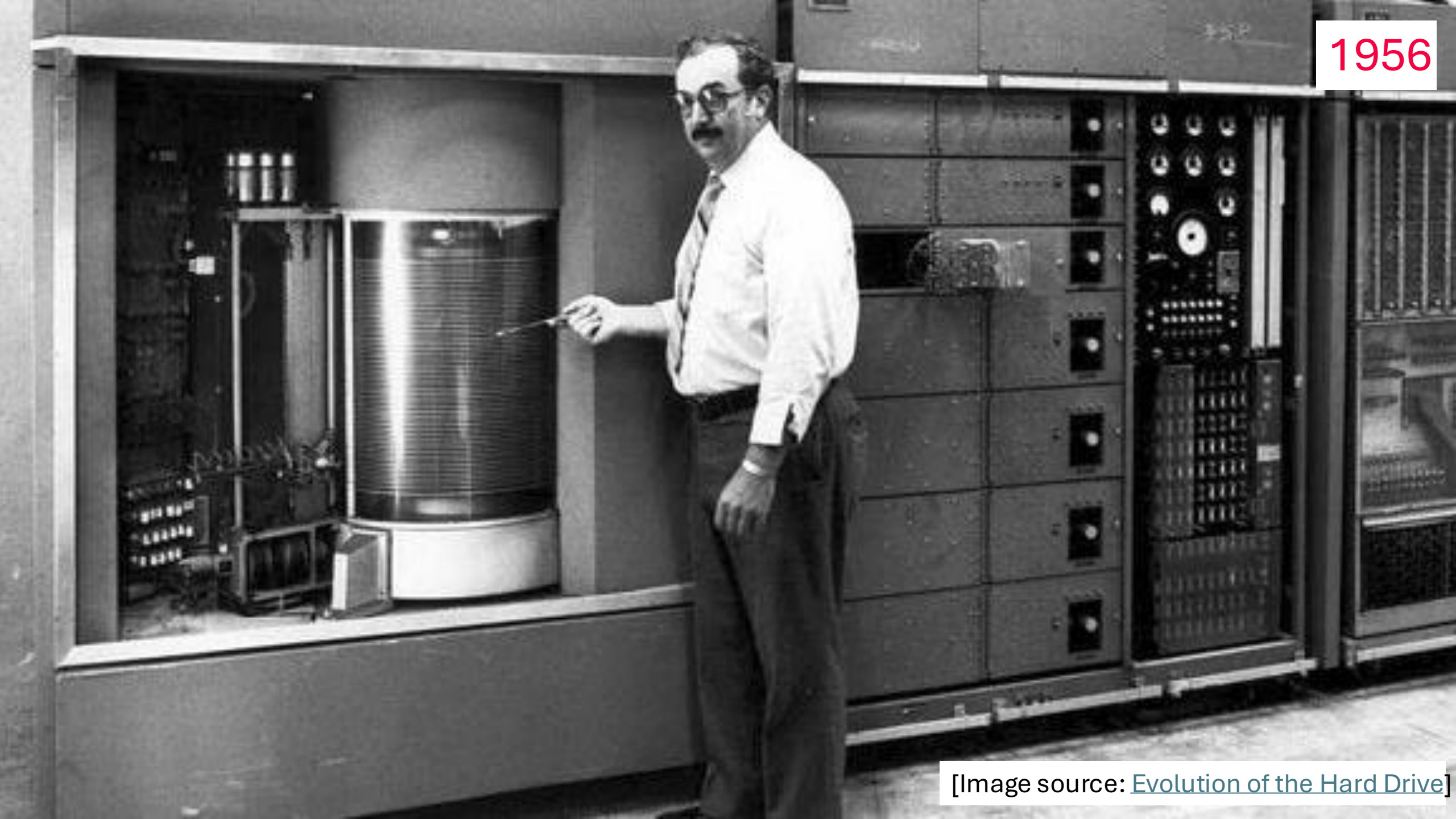

720 KB
boedem
FLOPPY DISK
5.25" HD,DS
1.4-3 MB

Floppy Disks

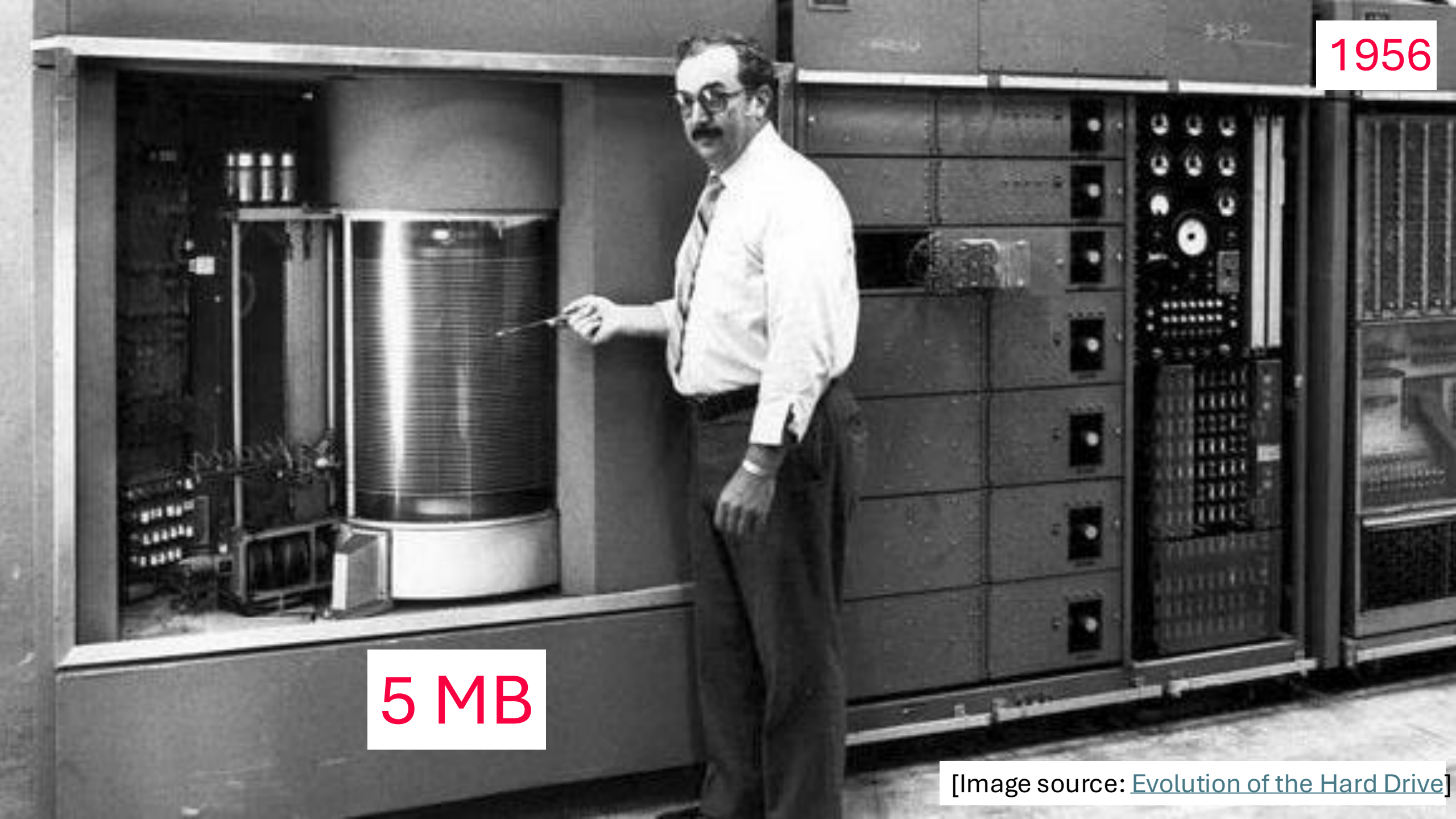
- Popular storage medium 1970s – end 1990s
- Size: kilobytes (KB) to a few megabytes (MB)
- In other words: storing a few word documents or a small image

History of Data





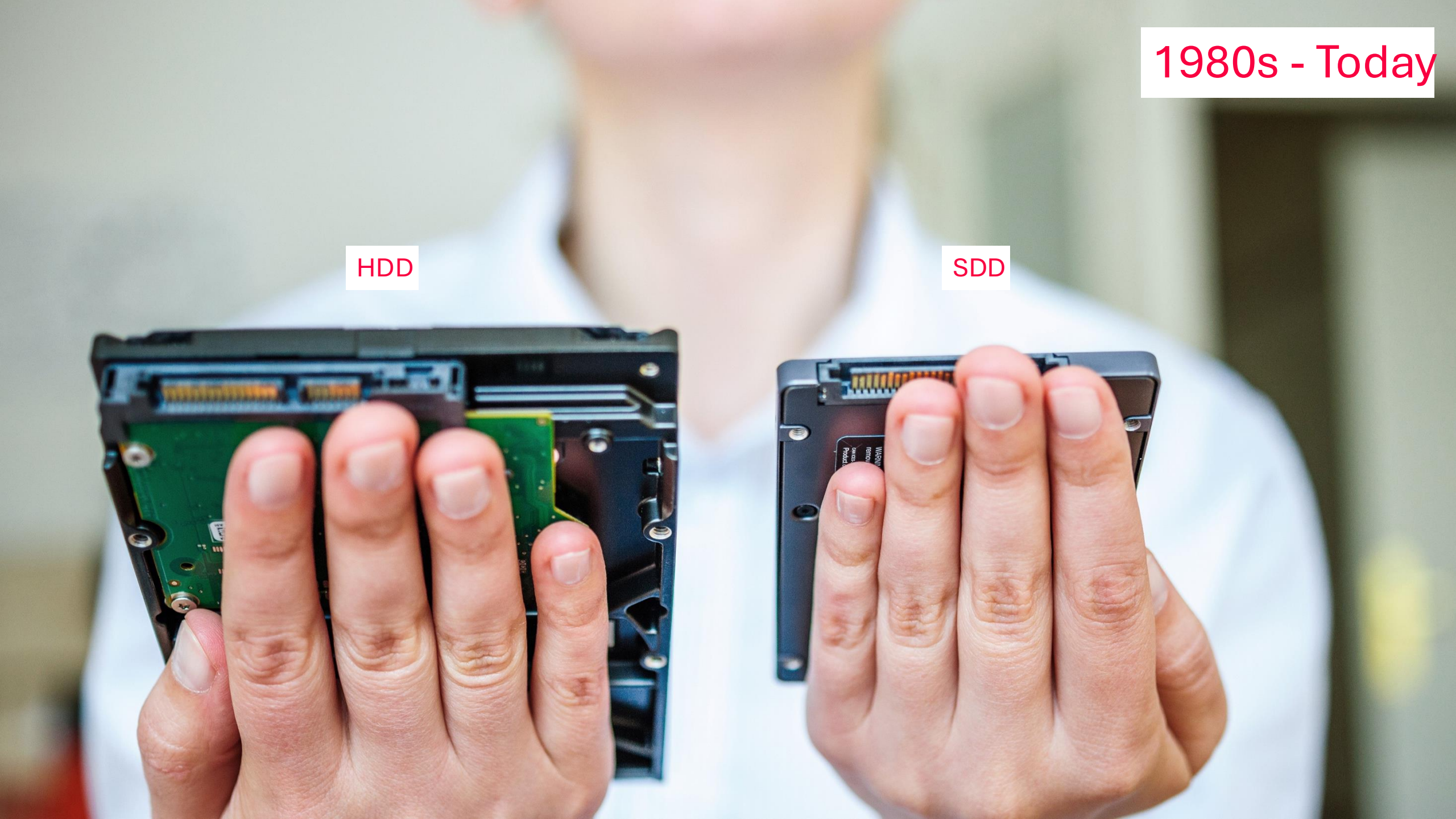
[Image source: Evolution of the Hard Drive]



[Image source: Evolution of the Hard Drive]

Removable Disk Packs

1970

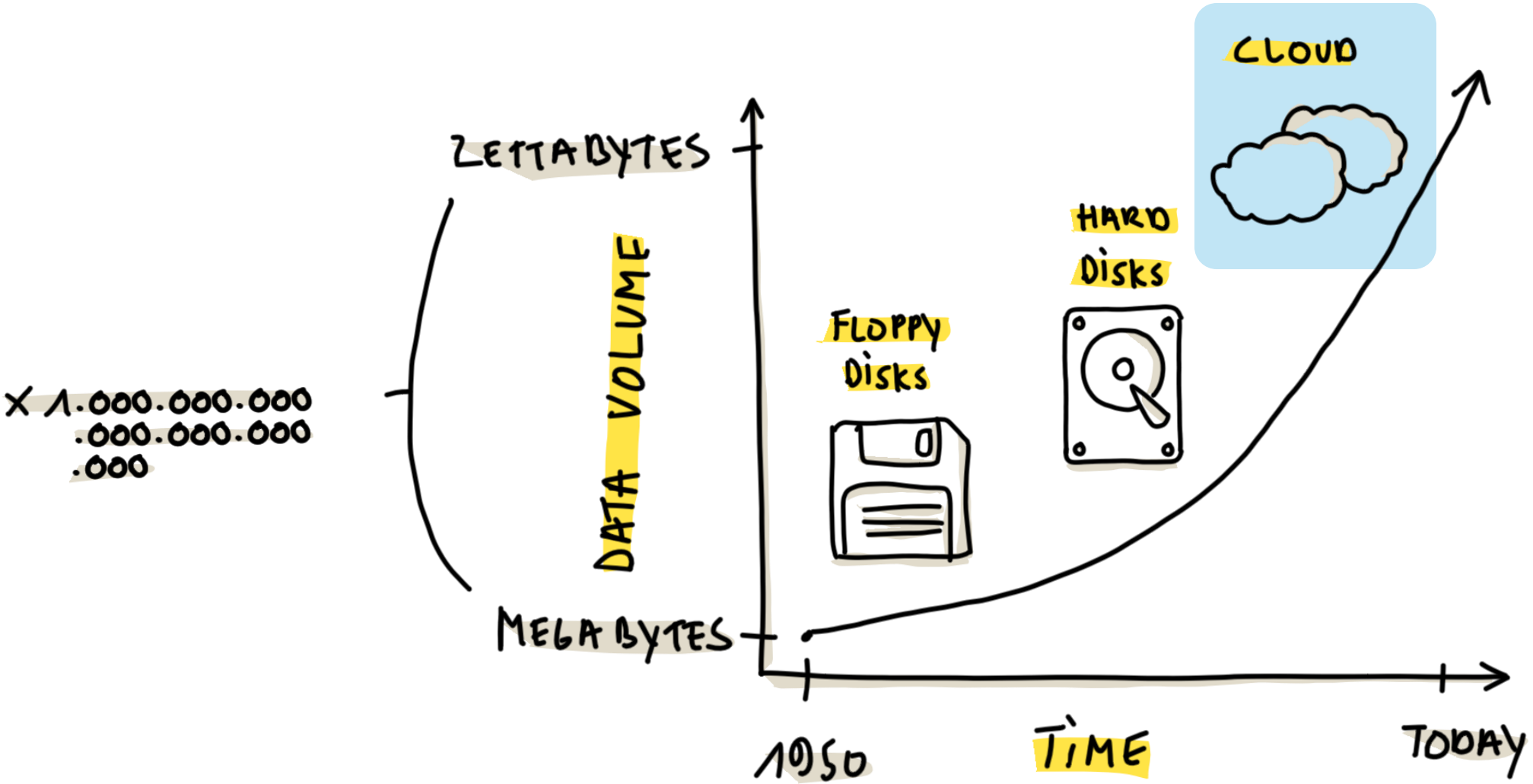

1980s - Today
HDD
SDD

Hard Disk Drives (HDD)

- Introduced in 1950s
- Remains a key component of computer storage (datacenters, personal computers)
- Capacity and speed evolved over time
- Volume can range from Gigabytes (GB) to multiple Terrabytes (TB)
- 1 TB HDD = +/- 200.000 Photos

- Note: Smartphones, tablets, modern notebooks still have hard disks (Solid State Drives – SSDs)

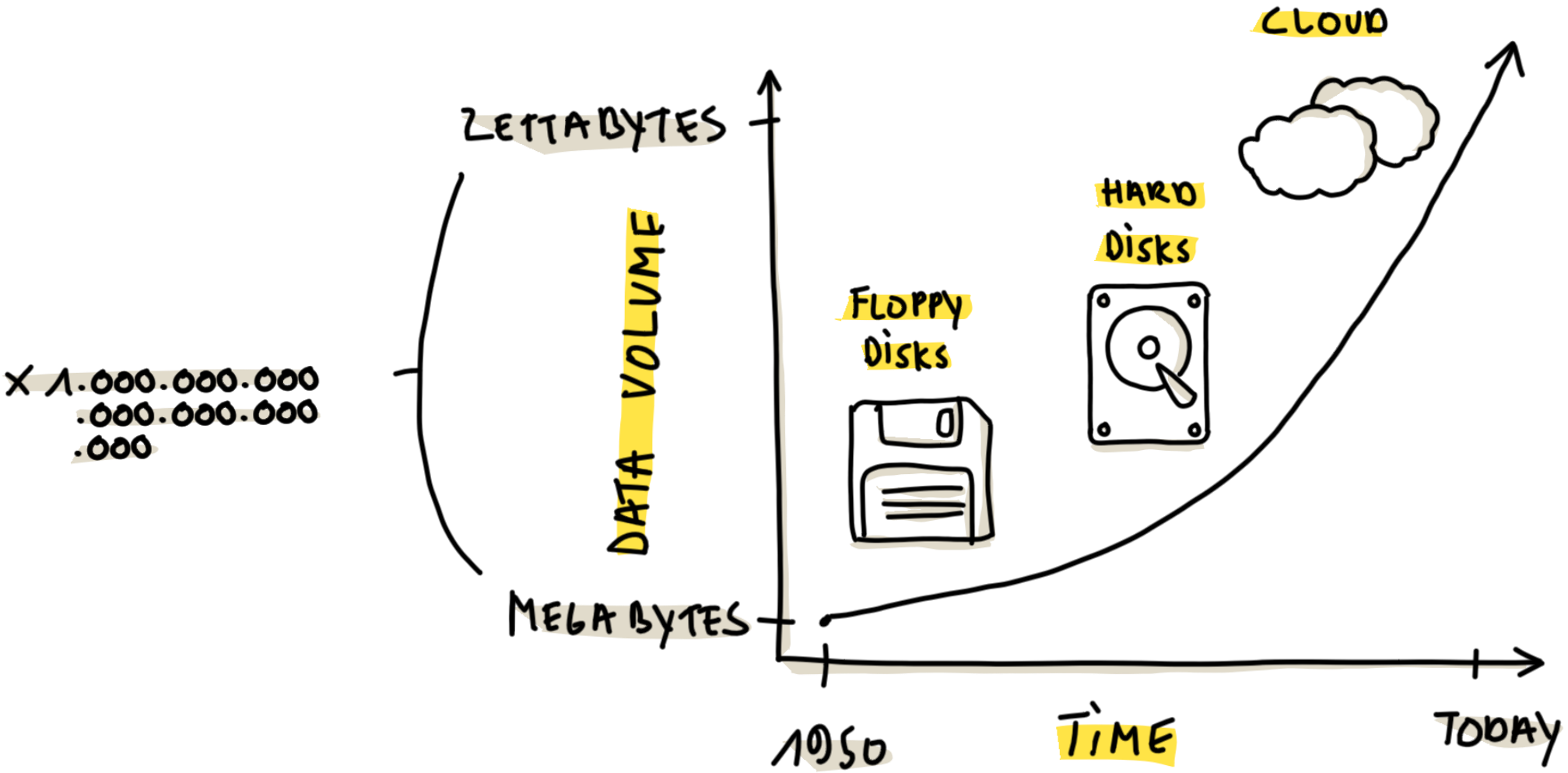
History of Data



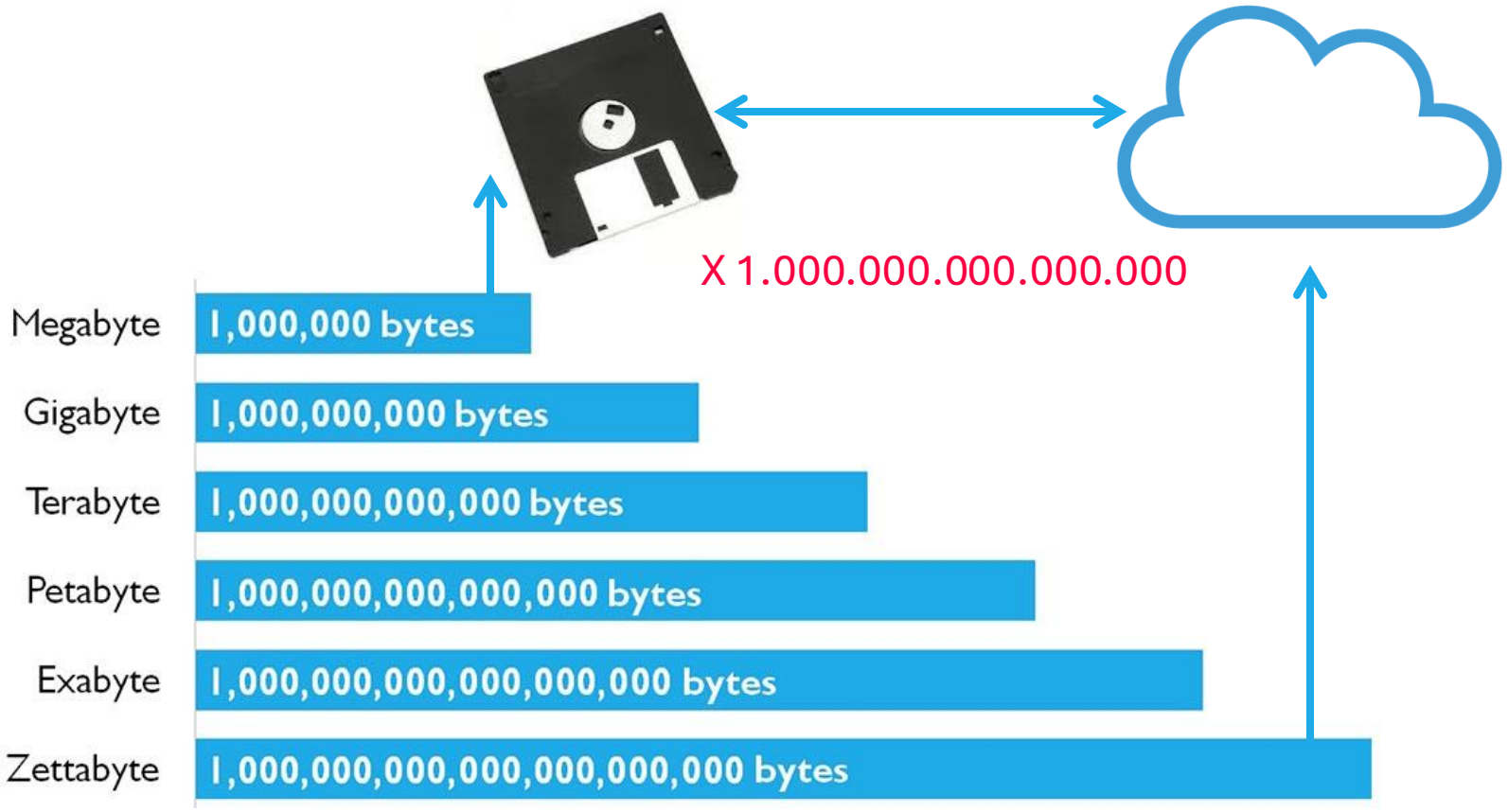
The Cloud

- Introduced in the 2000s
- Storing data on remote servers accessible via the Internet
- Cloud = service based model
- Example services Dropbox, Google Drive, iCloud, Amazon S3 or OneDrive
- Upload, store and retrieve high volumes of data from anywhere
- Cloud storage can scale up or down based on our needs
- No more physical storage limits

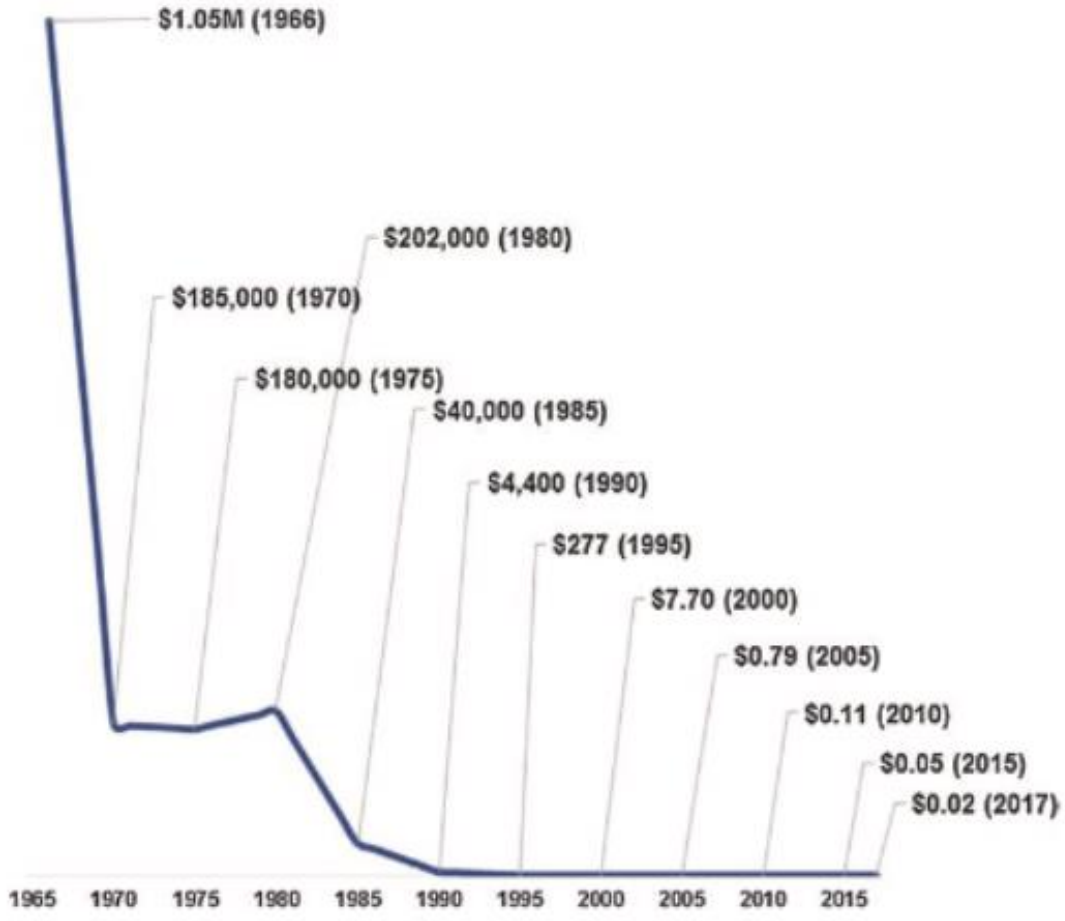
Some Observations...



1) “No more physical storage limits”

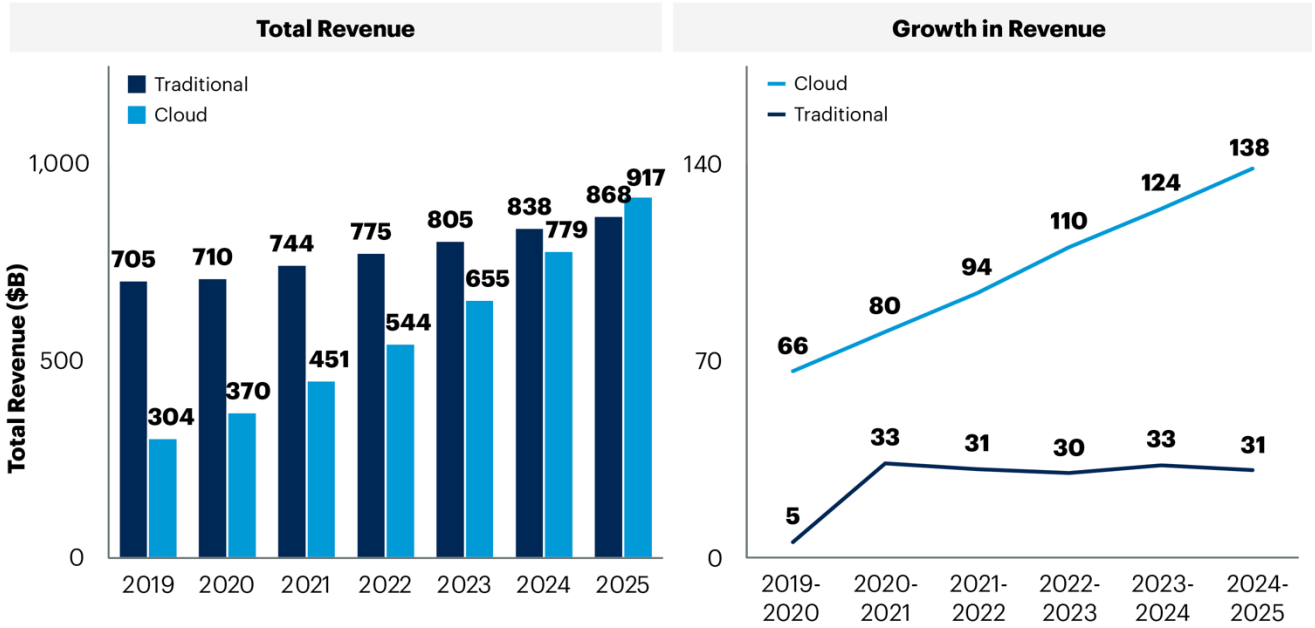


2) “Cost of storing data dropped”



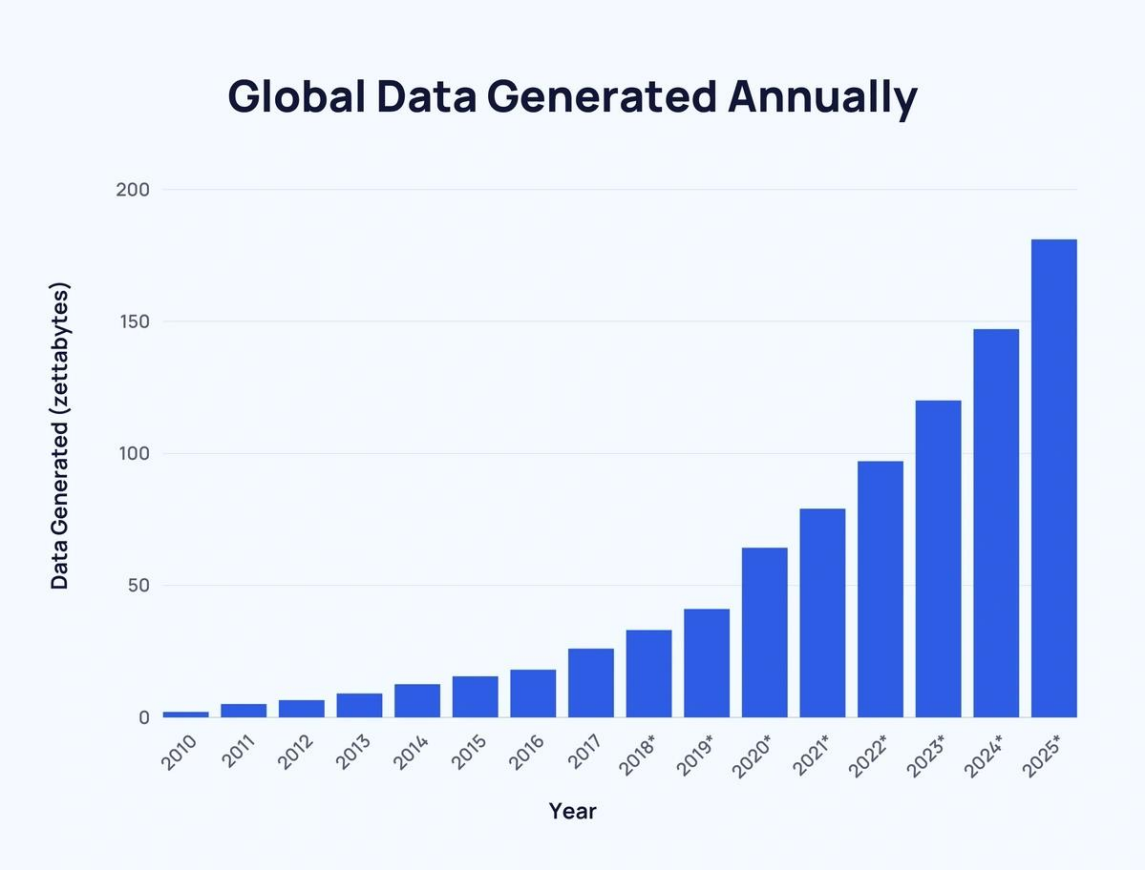
- Graph: Cost per gigabyte of storage over time
- Once extraordinarily expensive, the cost of storing large quantities of data has dropped to fractions of a cent
- Source: [The Future of Finance]

3) “More and more enterprises move to the cloud ”



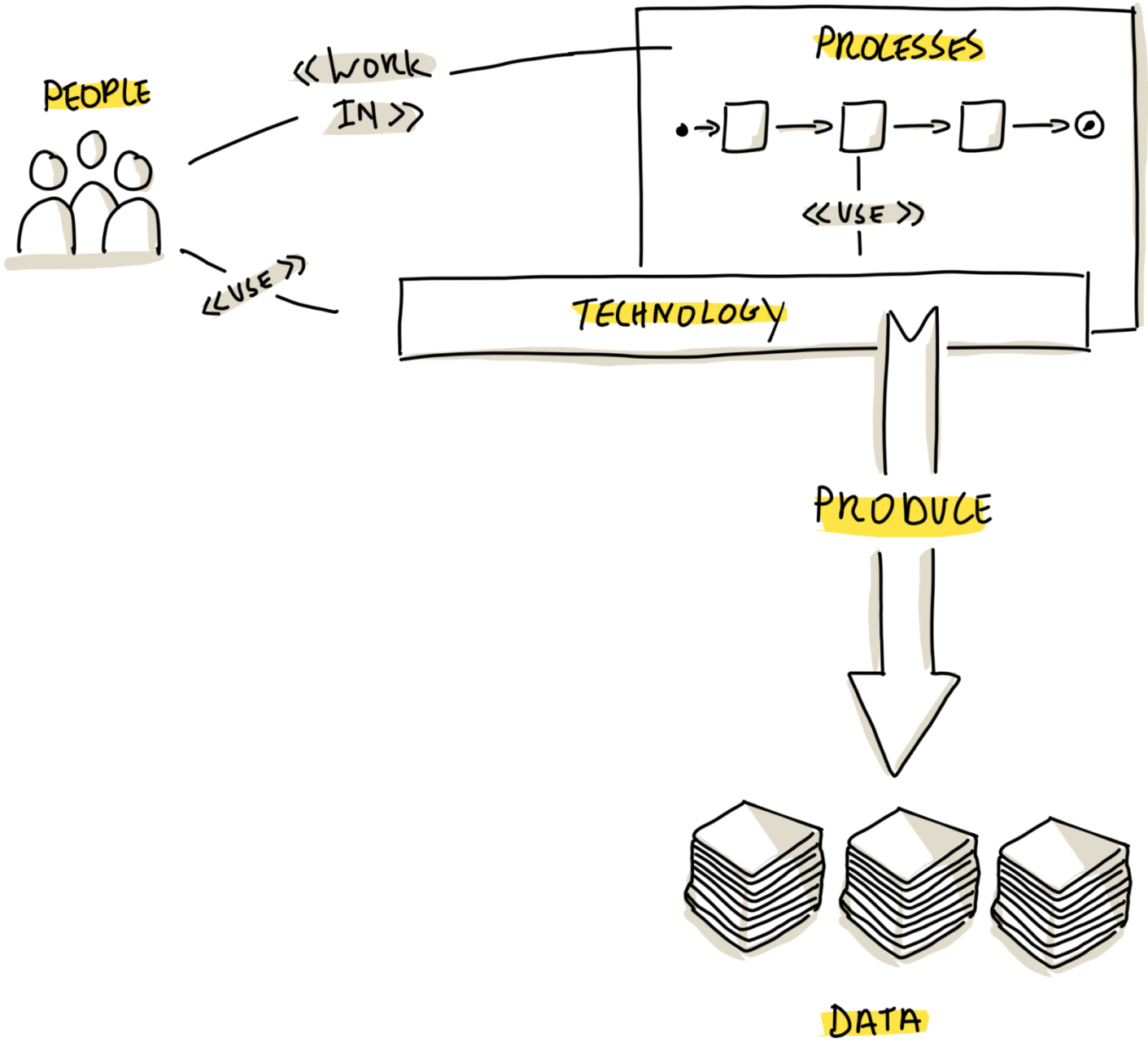
- Graph: Traditional vs Cloud computing
- By 2025, 51% of IT spending will have shifted from traditional solutions to the cloud, compared to 41% in 2022.
- Source: [Gartner Press Release]
- Cloud is an opportunity to work with higher volumes of data

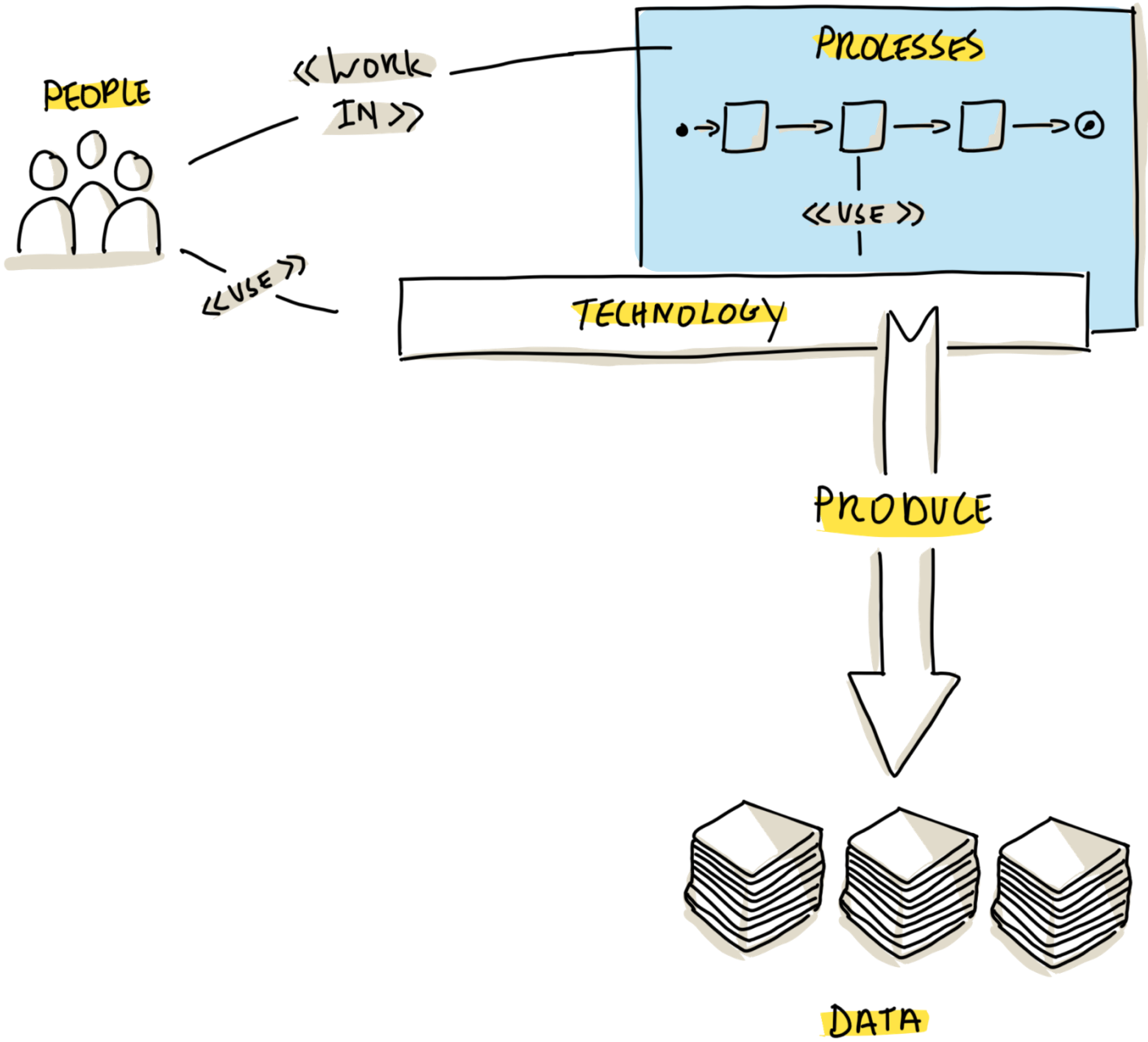
4) Data becomes “**Big Data**”



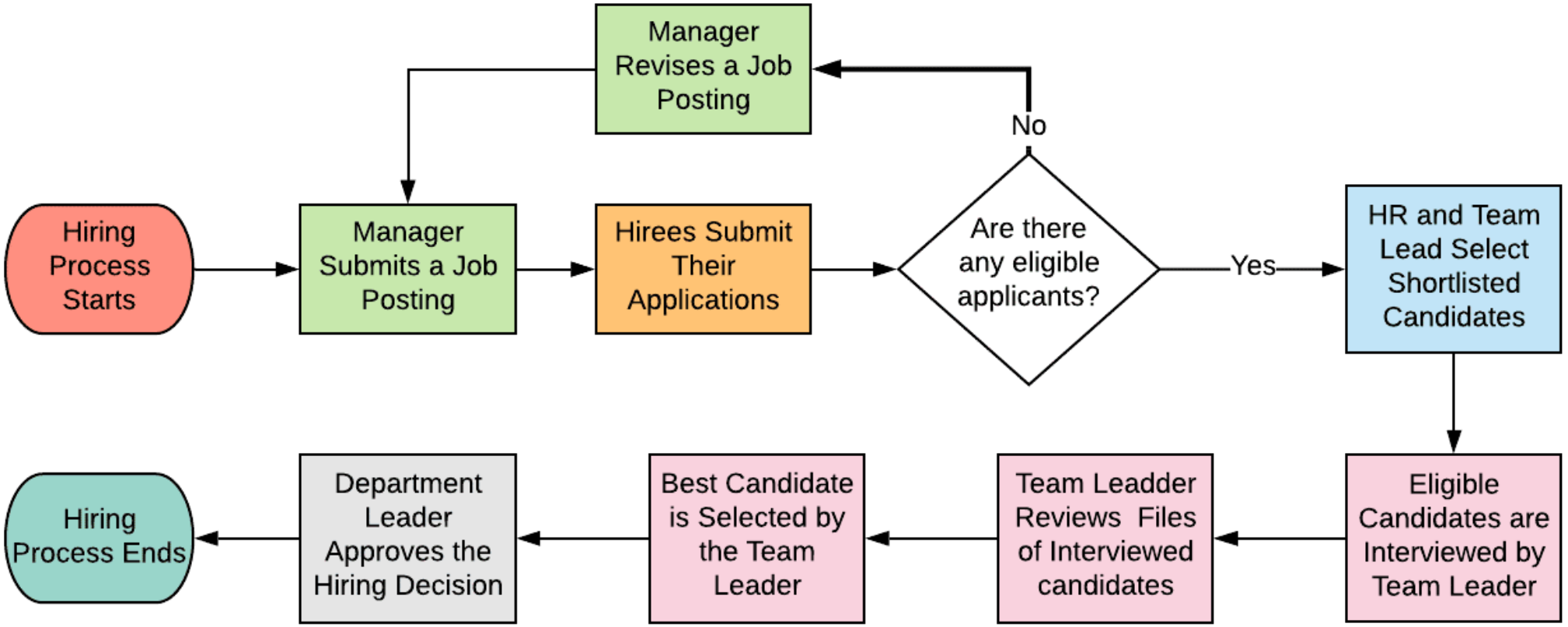
Data Management

- Defining Data
- **Data Producers**
- Big Data Vs
- The Bigger Picture
- Data Management
- Data Technology


PEOPLE
«WORK IN»
«USE»
PROCESSES
«USE»
TECHNOLOGY
PRODUCE
DATA


PEOPLE
«WORK IN»
«USE»
PROCESSES
«USE»
TECHNOLOGY
PRODUCE
DATA

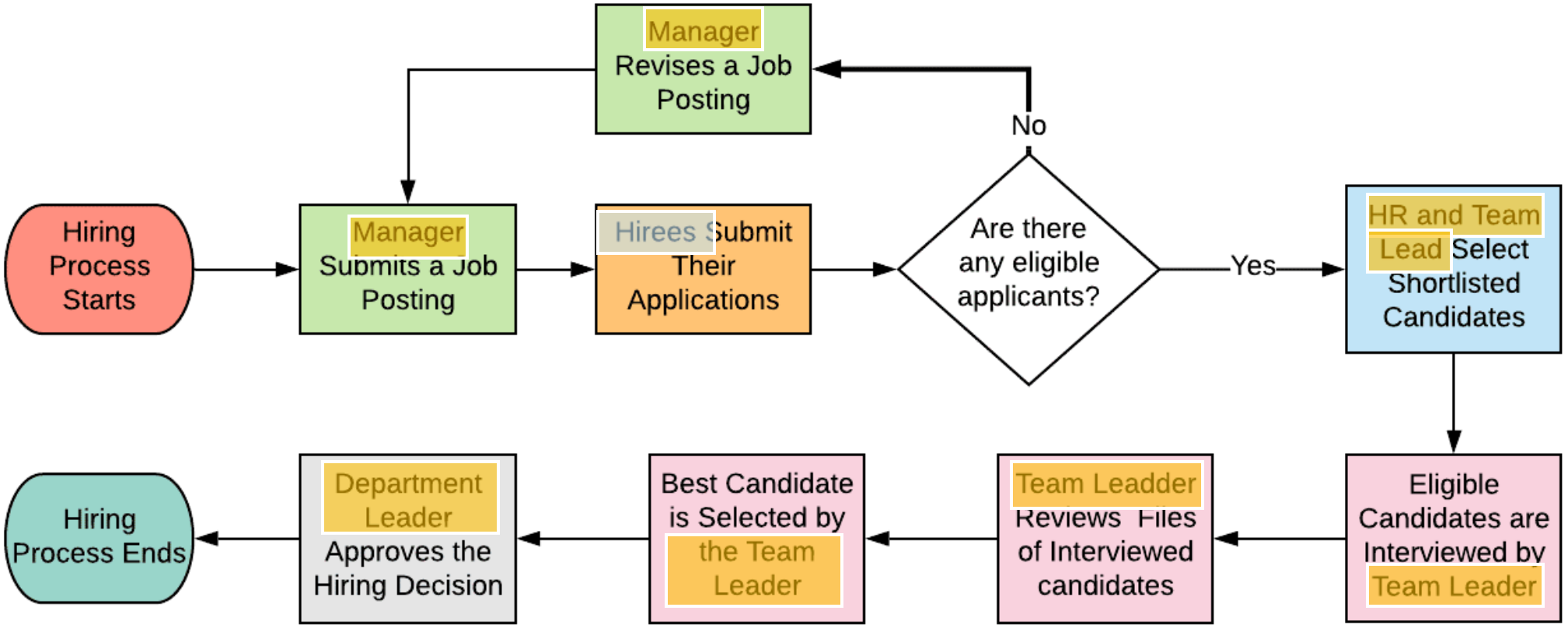
Process Example: Hiring a new Employee



[Source: Real Life Business Process Exampl

Process Example: Hiring a new Employee

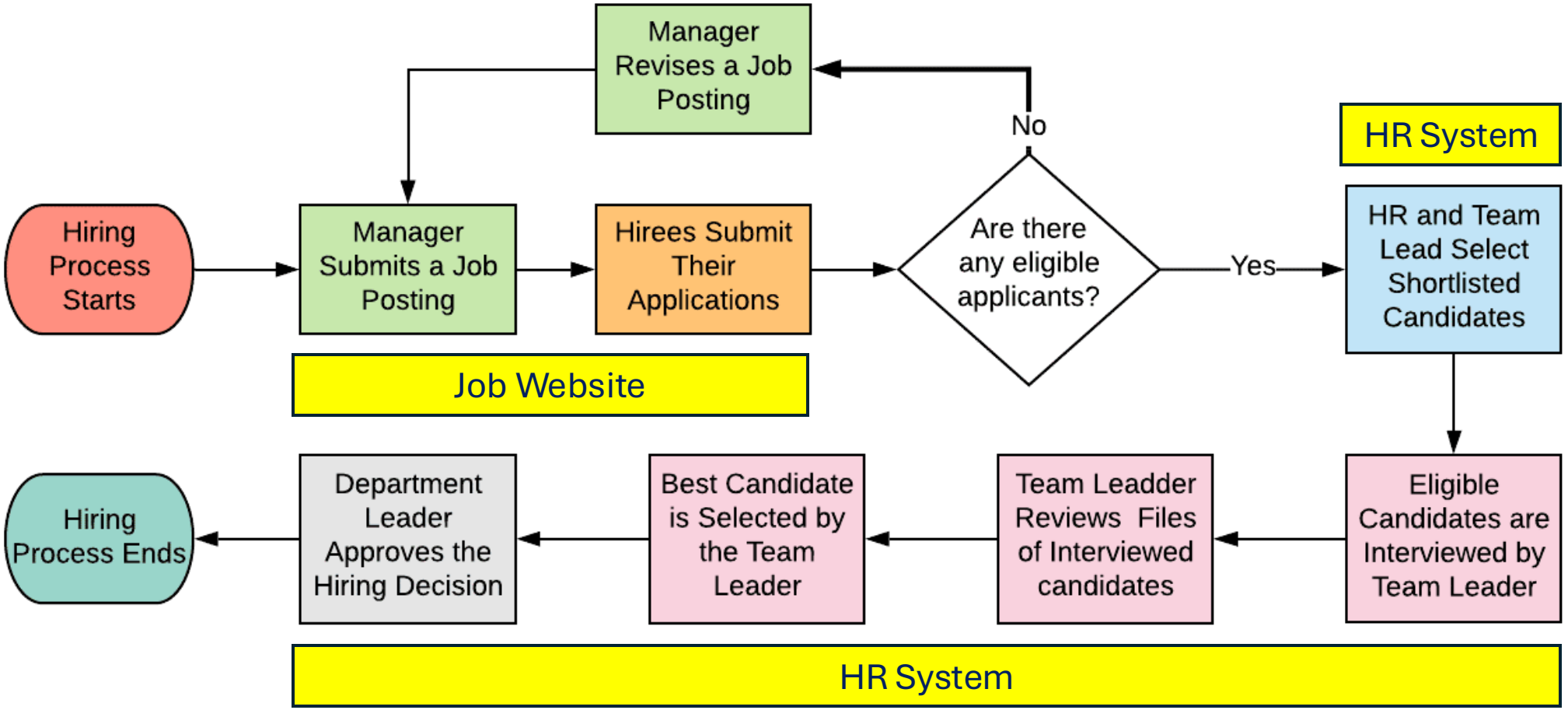
PEOPLE WORK IN THE PROCESS



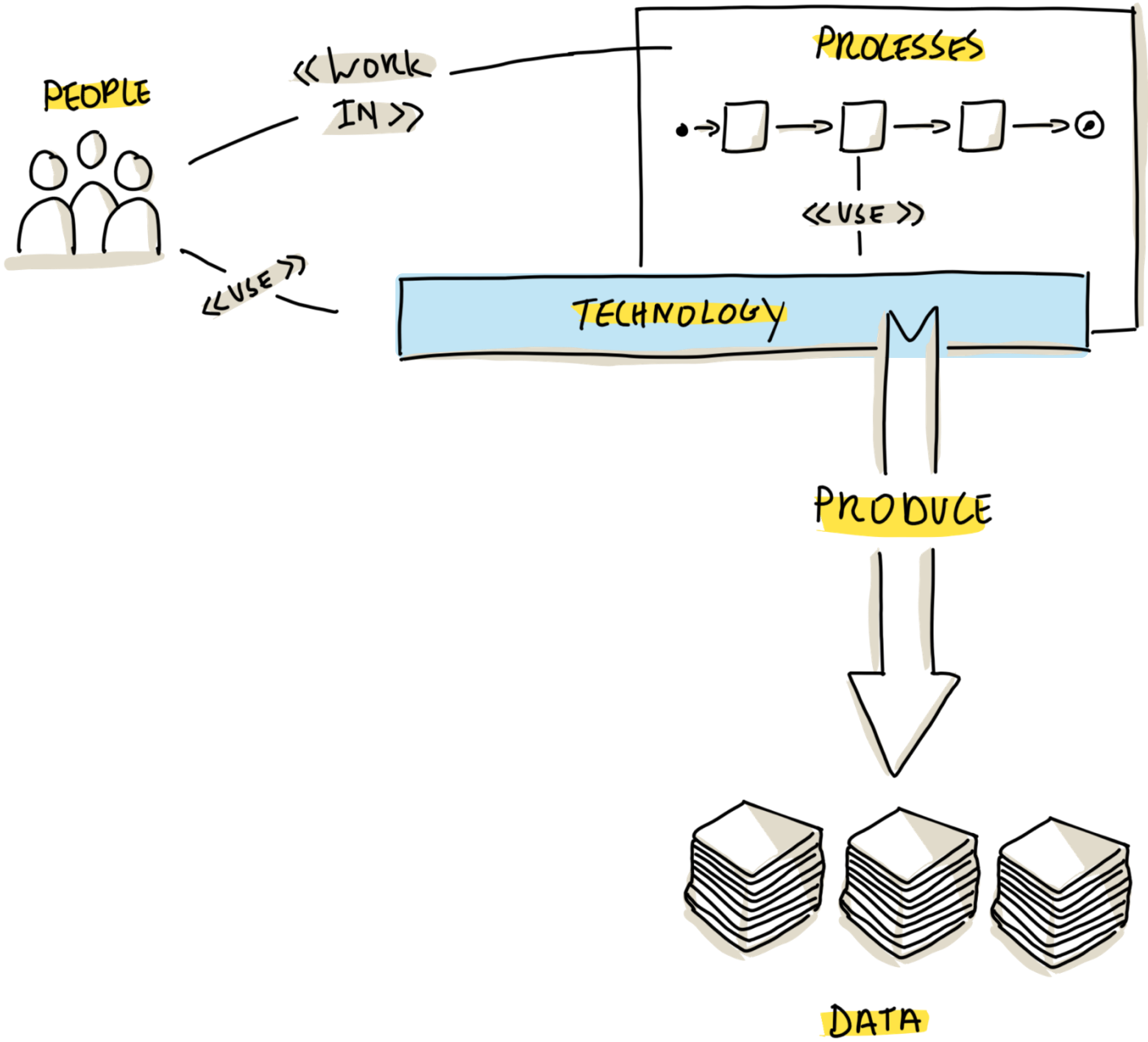
[Source: Real Life Business Process Exampl

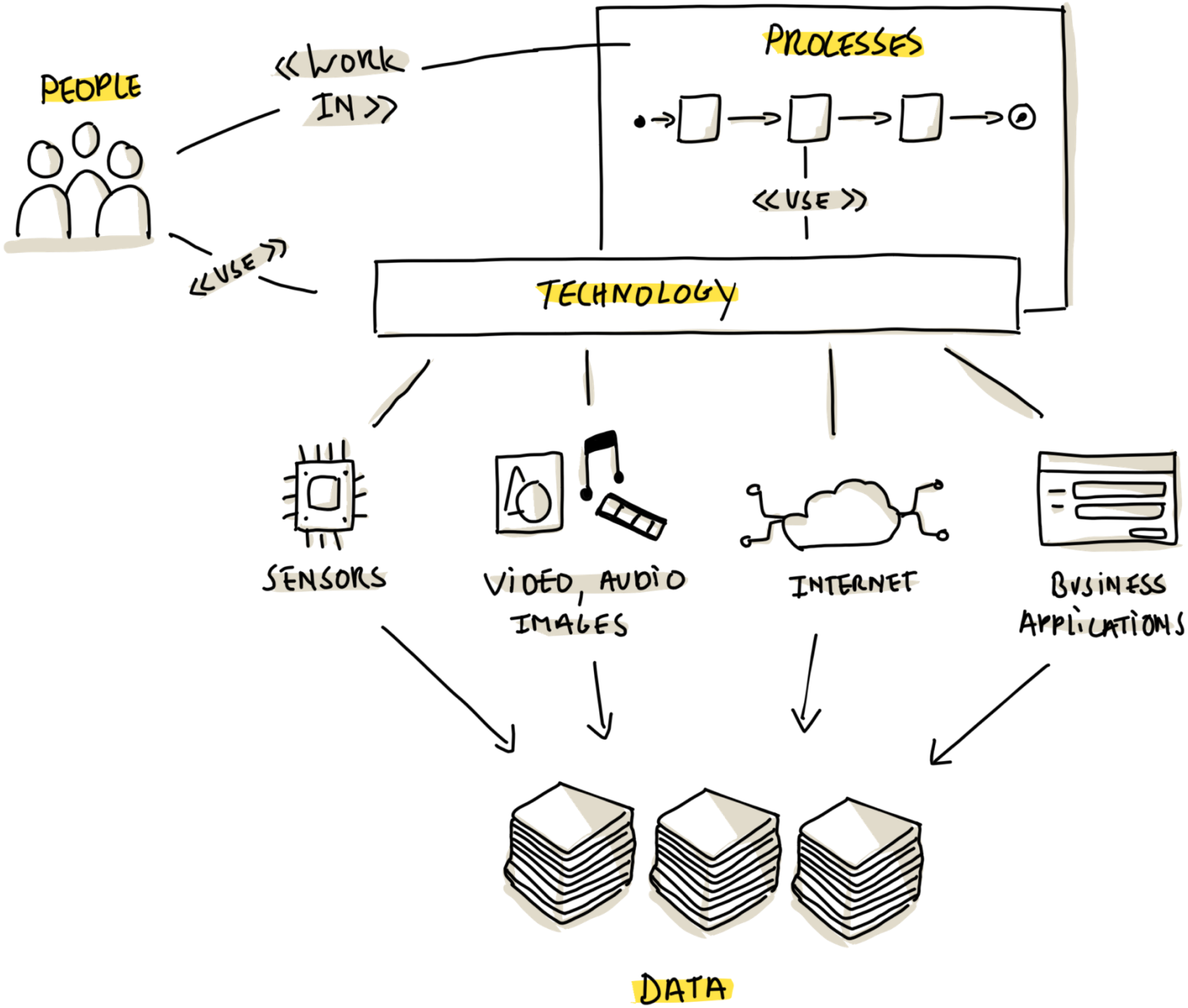
Process Example: Hiring a new Employee

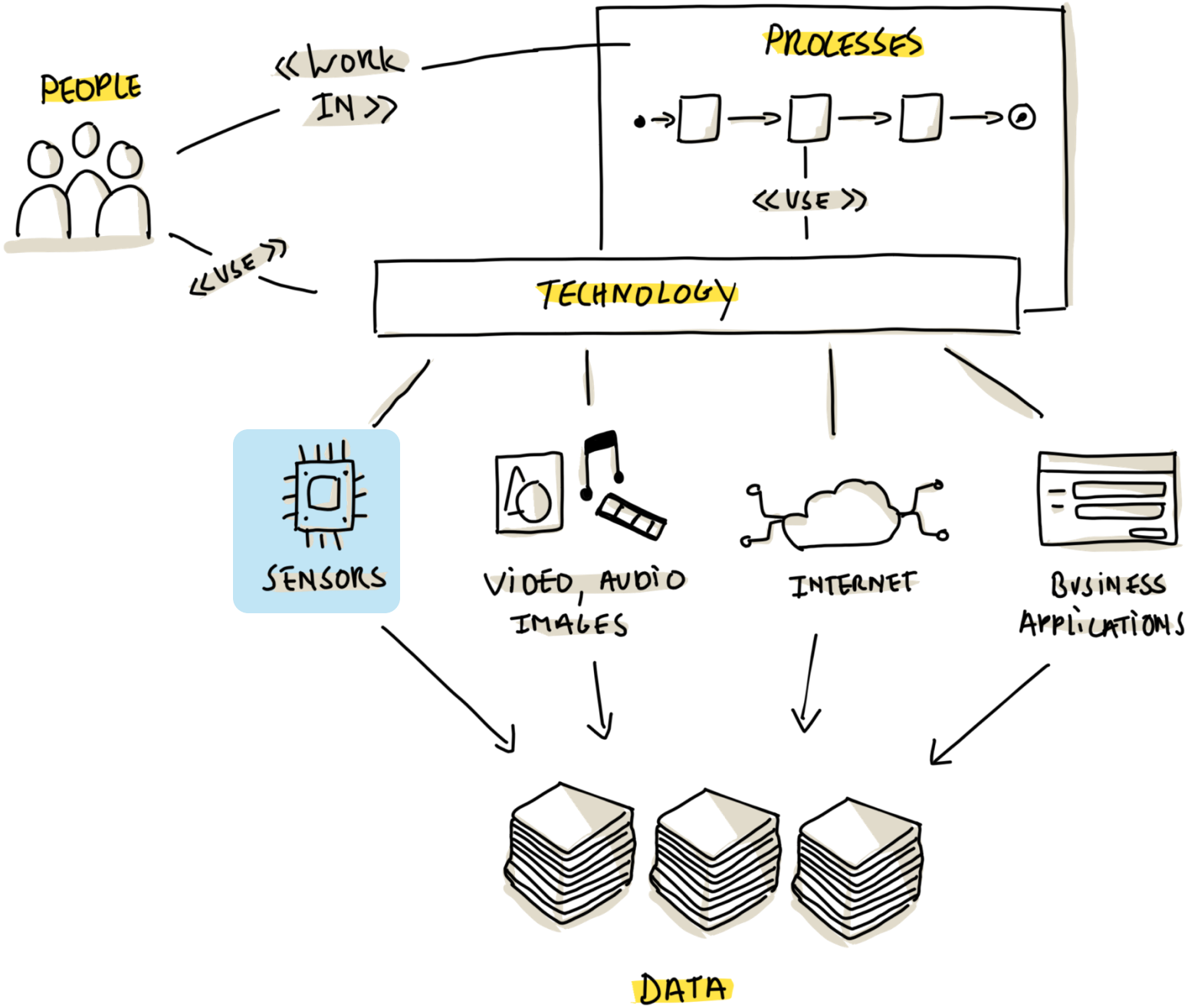
TECHNOLOGY IS BEING USED IN THE PROCESS



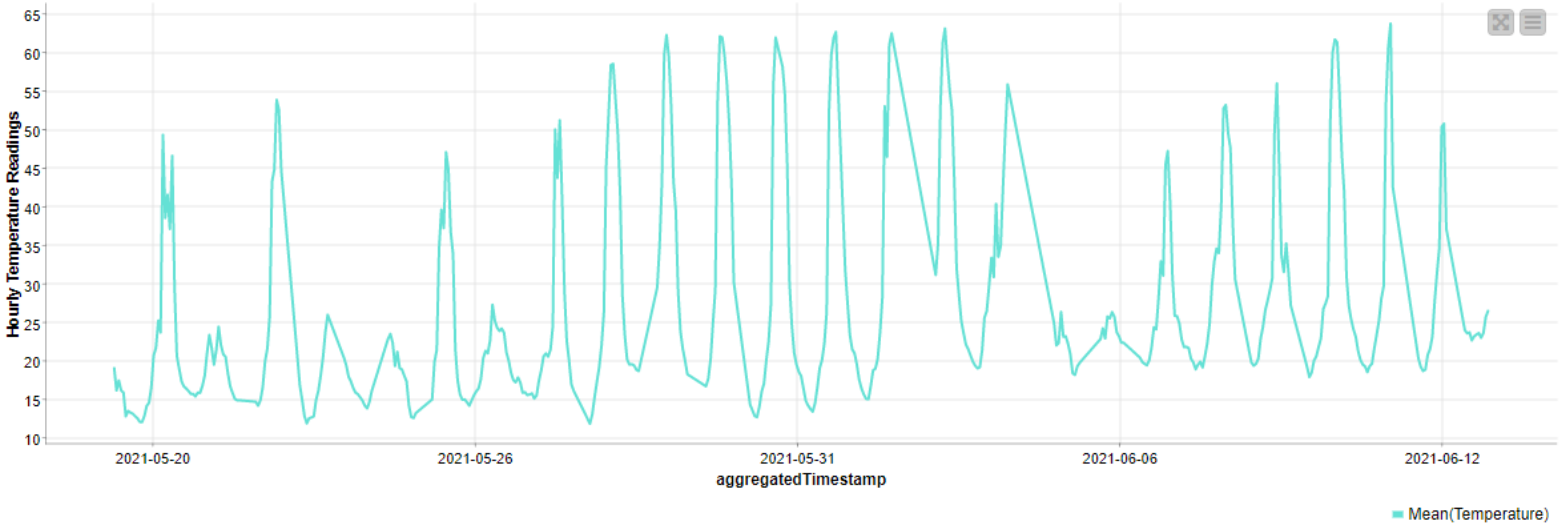
[Source: Real Life Business Process Exampl


PEOPLE
«WORK IN»
«USE»
PROCESSES
«USE»
TECHNOLOGY
PRODUCE
DATA


PEOPLE
«WORK IN»
PROCESSES
«USE»
«USE»
TECHNOLOGY
SENSORS
VIDEO, AUDIO IMAGES
INTERNET
BUSINESS APPLICATIONS
DATA


PEOPLE
«WORK IN»
«USE»
PROCESSES
«USE»
TECHNOLOGY
SENSORS
VIDEO, AUDIO IMAGES
INTERNET
BUSINESS APPLICATIONS
DATA

Temperature Sensors

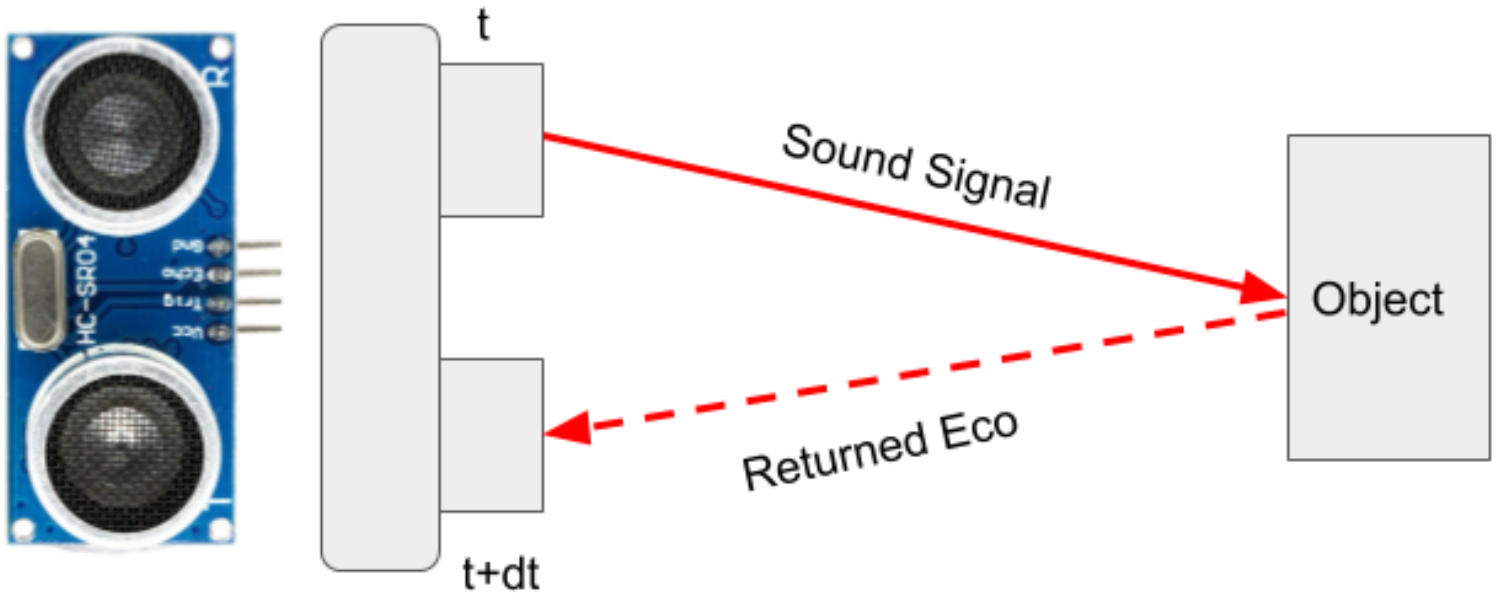


Every 10 miliseconds, a temperature sensor will measure temperature.
This gives **8.640.000** temperature values per day.

Temperature Sensors: Usage

- Controlling the temparature throughout process steps
- Example applications:
  - Food production
  - Chemical processes
  - Agriculture
  - Heating control
  - Transport Refrigeration
  - …

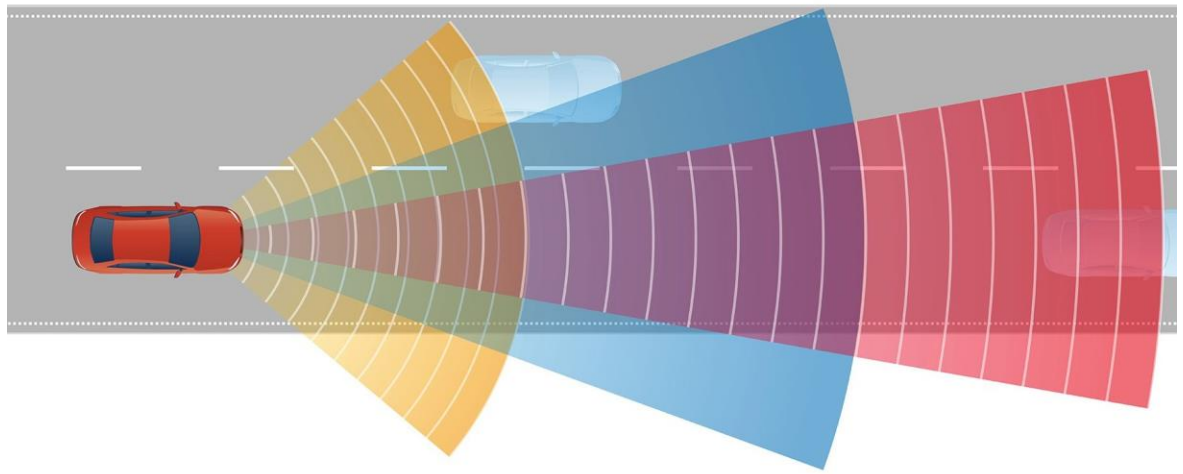
Proximity Sensors



Proximity sensors detect the **presence or absence of a nearby object** or material.

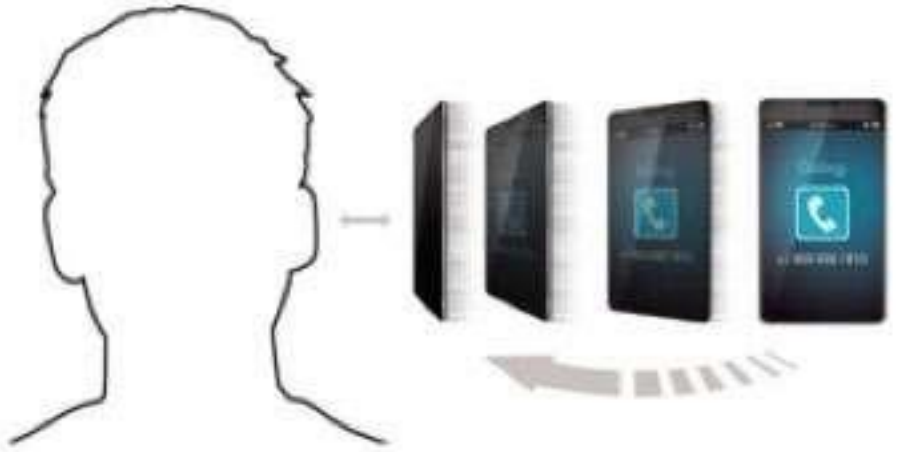
Proximity Sensors: Cars & Smartphones

CARS



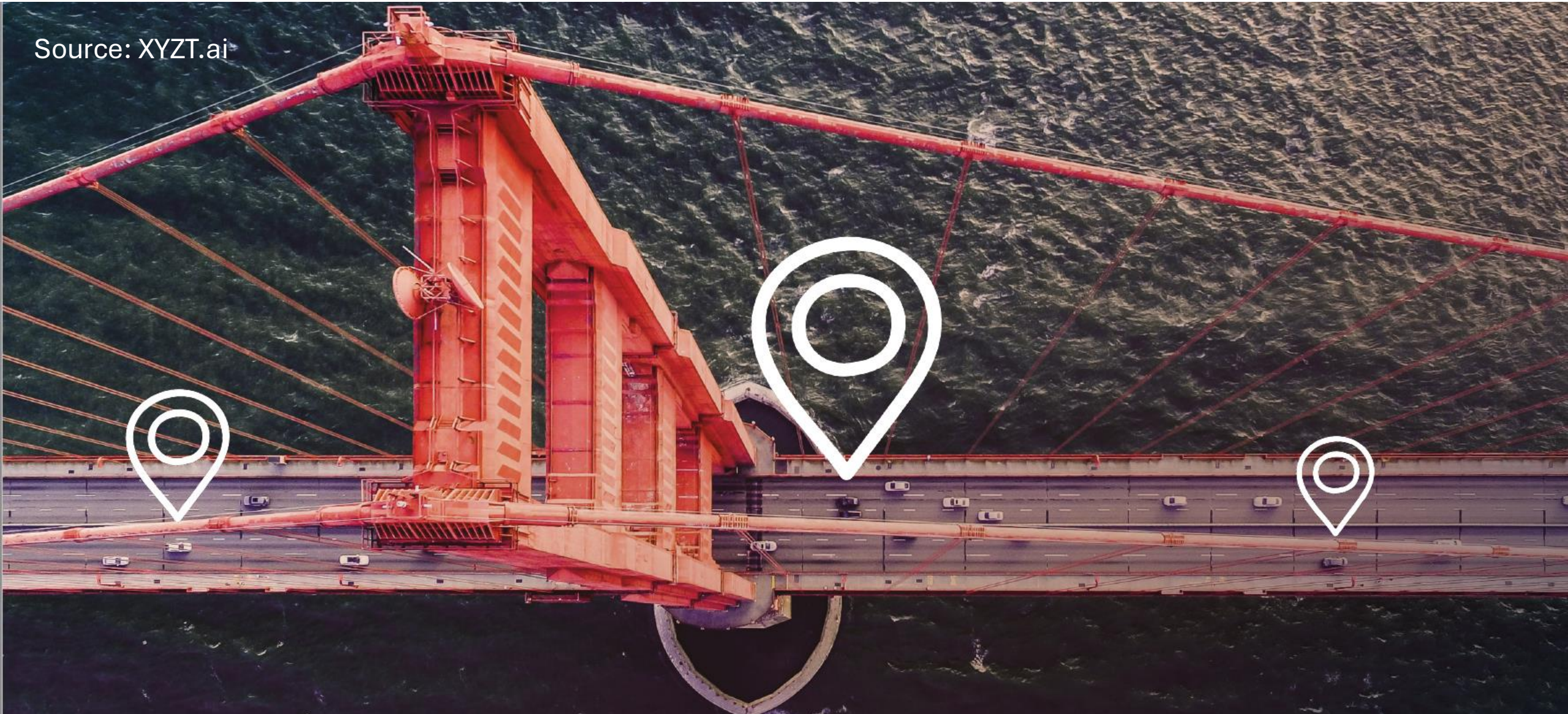
Proximity sensors detect nearby objects to facilitate driving assistance.

SMARTPHONES

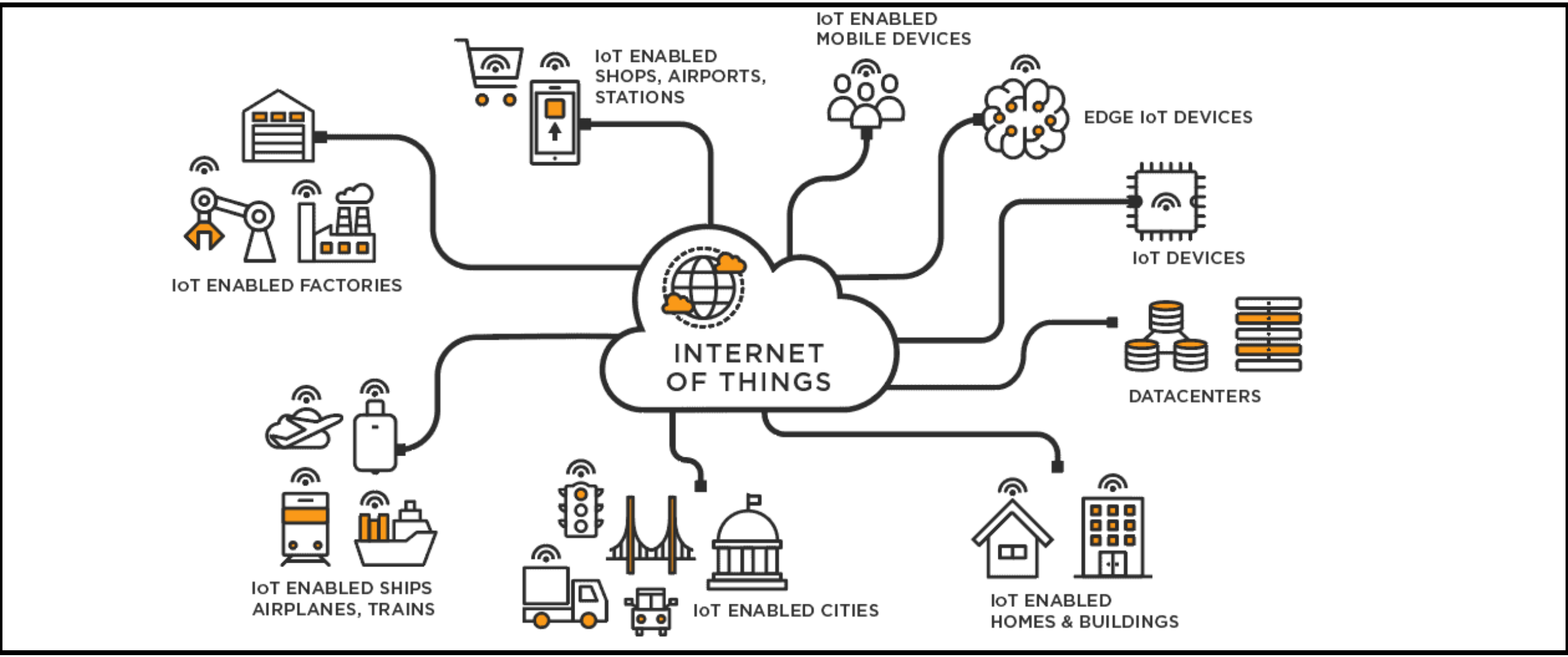


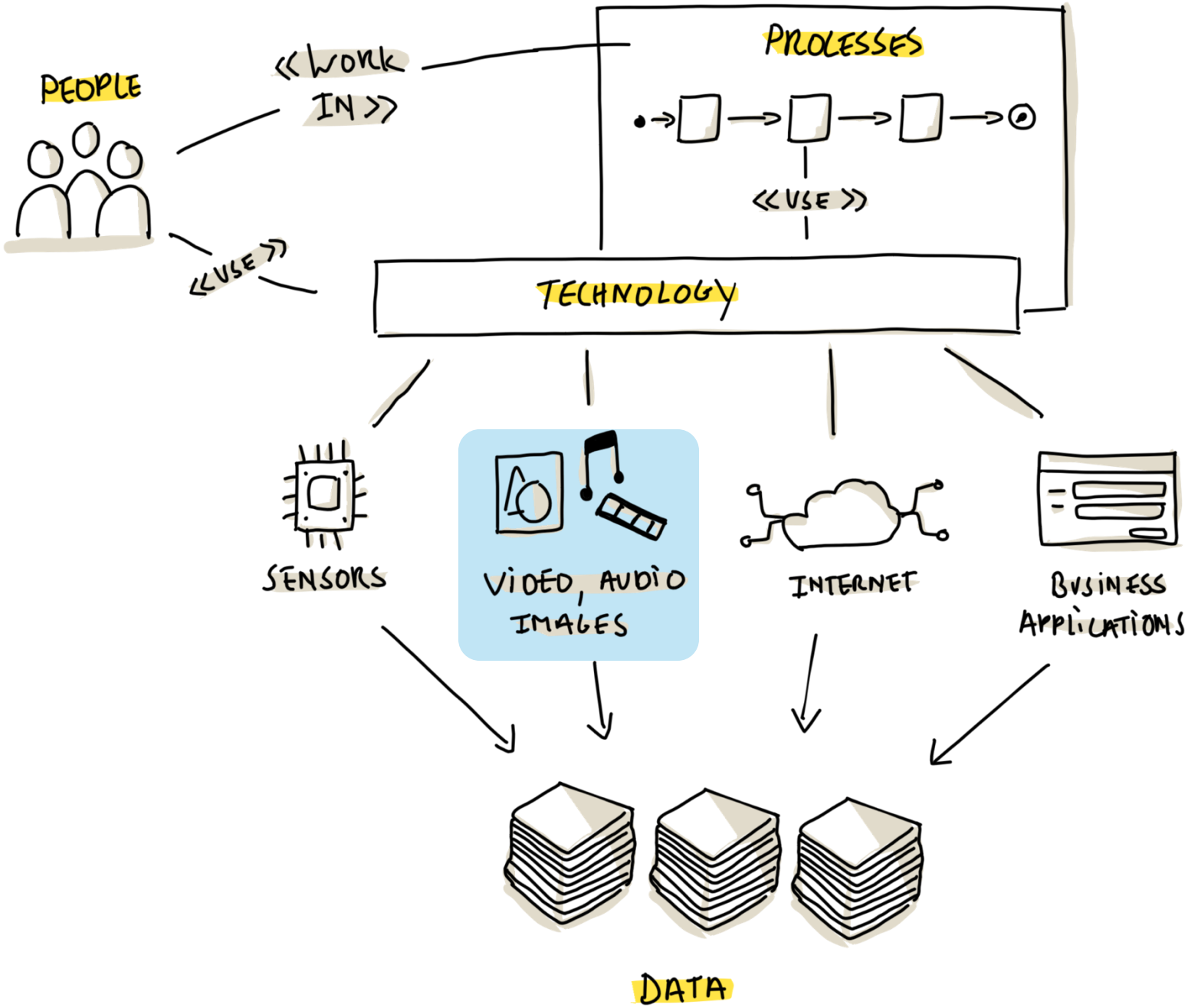
When a proximity sensor detects a face/ear, the smartphone screen turns of (energy saving).

GPS Location Sensors



Big Data Producers: **Internet Of Things (IOT)**



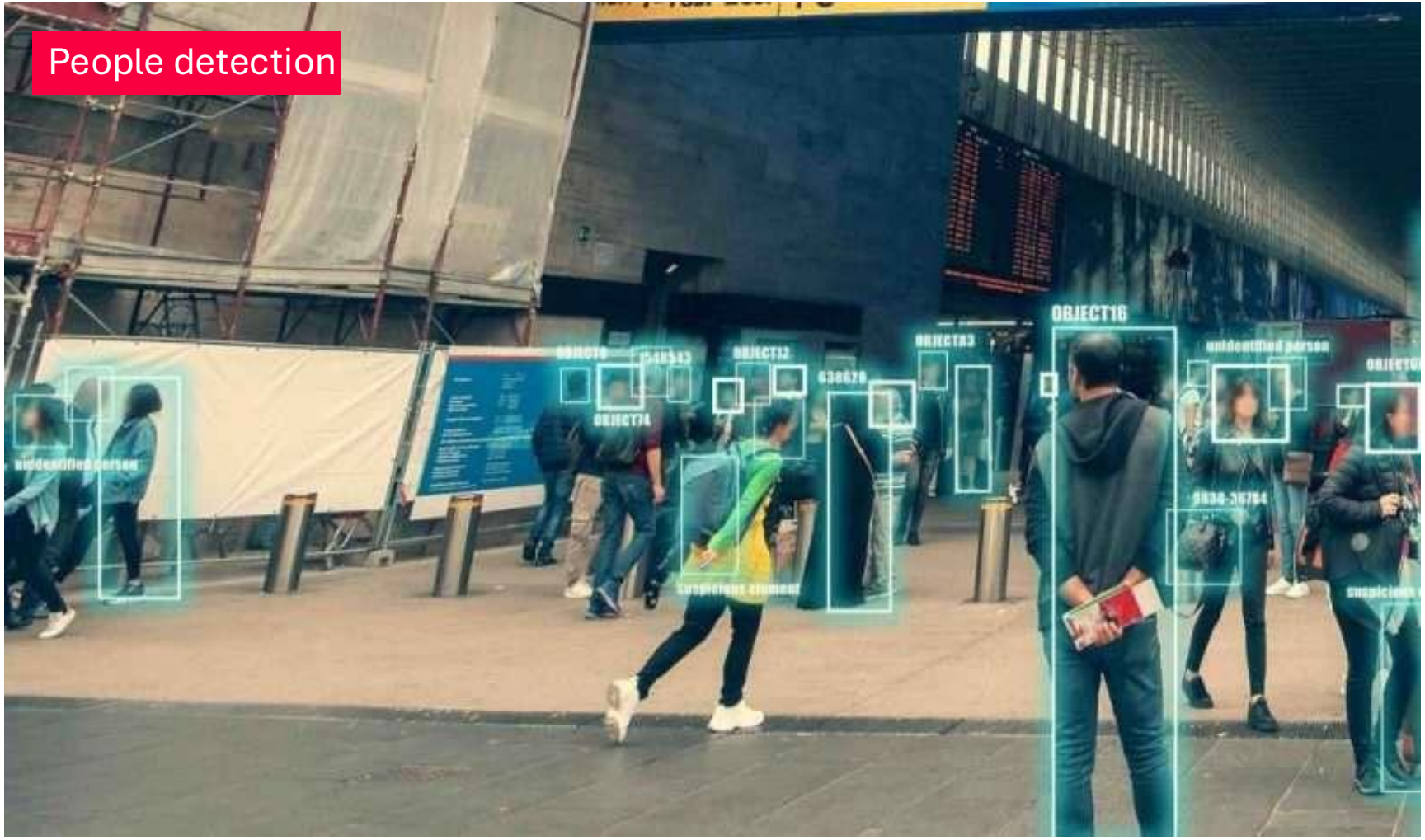

PEOPLE
«WORK IN»
«USE»
PROCESSES
«USE»
TECHNOLOGY
SENSORS
VIDEO, AUDIO IMAGES
INTERNET
BUSINESS APPLICATIONS
DATA

Video

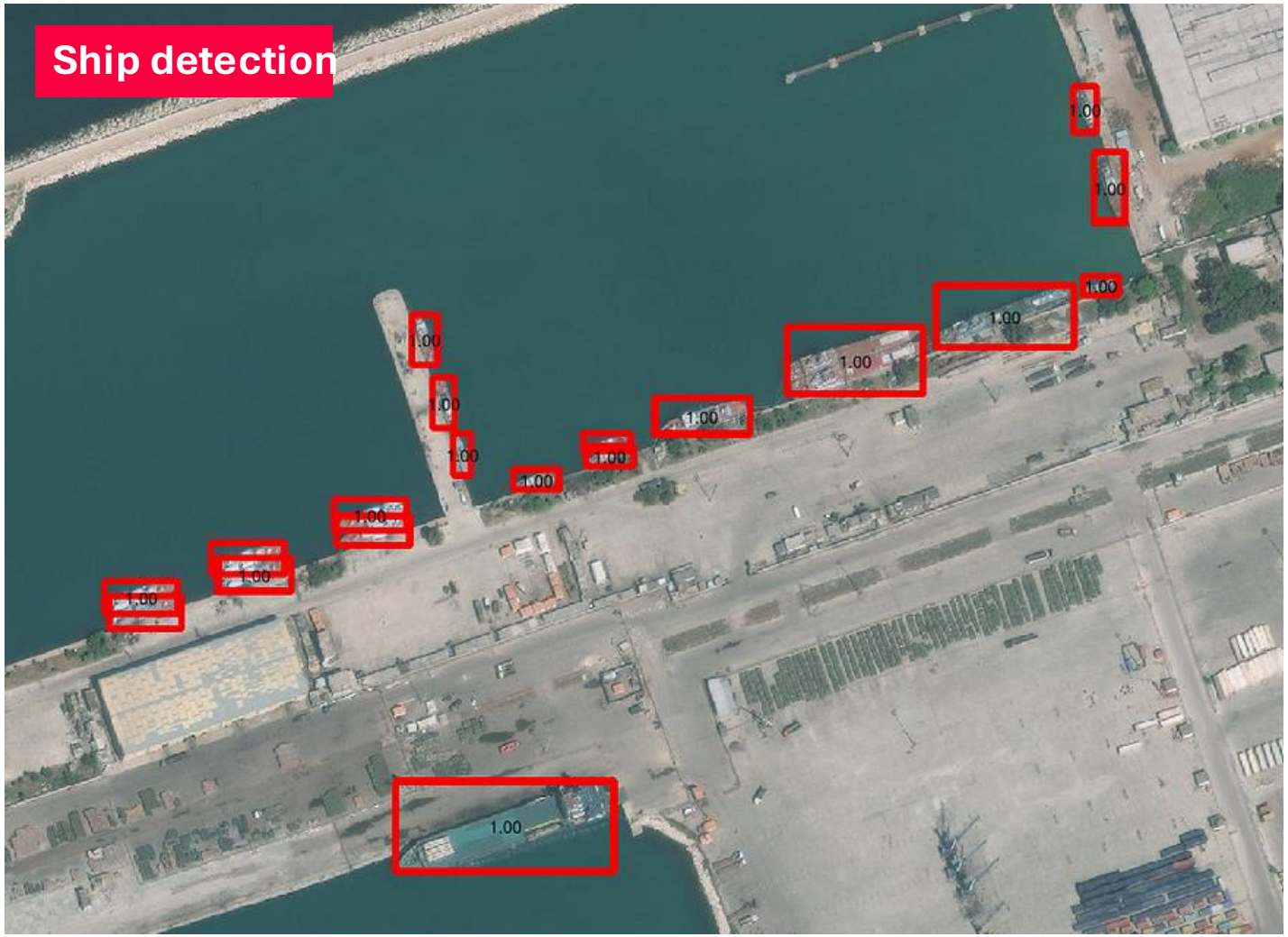
- Data generated by cameras
- Various types:
  - **Regular cameras**, just recording video
  - **Smart cameras**, recording video + object/precense recognition
  - **Sattelite** cameras
  - **Drone** cameras

Smart Camera’s with Object Detection

Smart Camera’s with Object Detection



Sattelite Camera’s with Object Detection



Volume

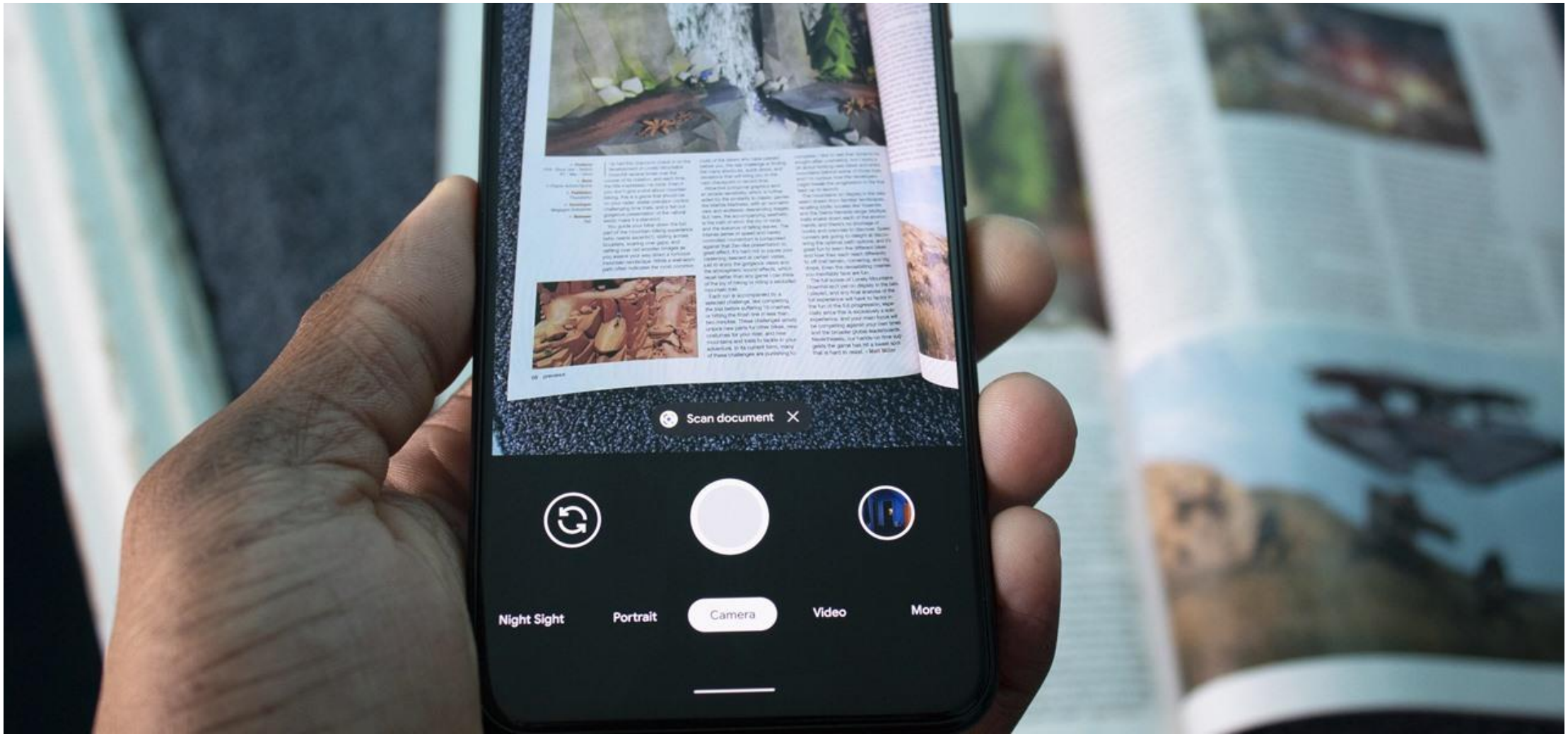
LOW ← VIDEO QUALITY → HIGH

DATA VOLUME

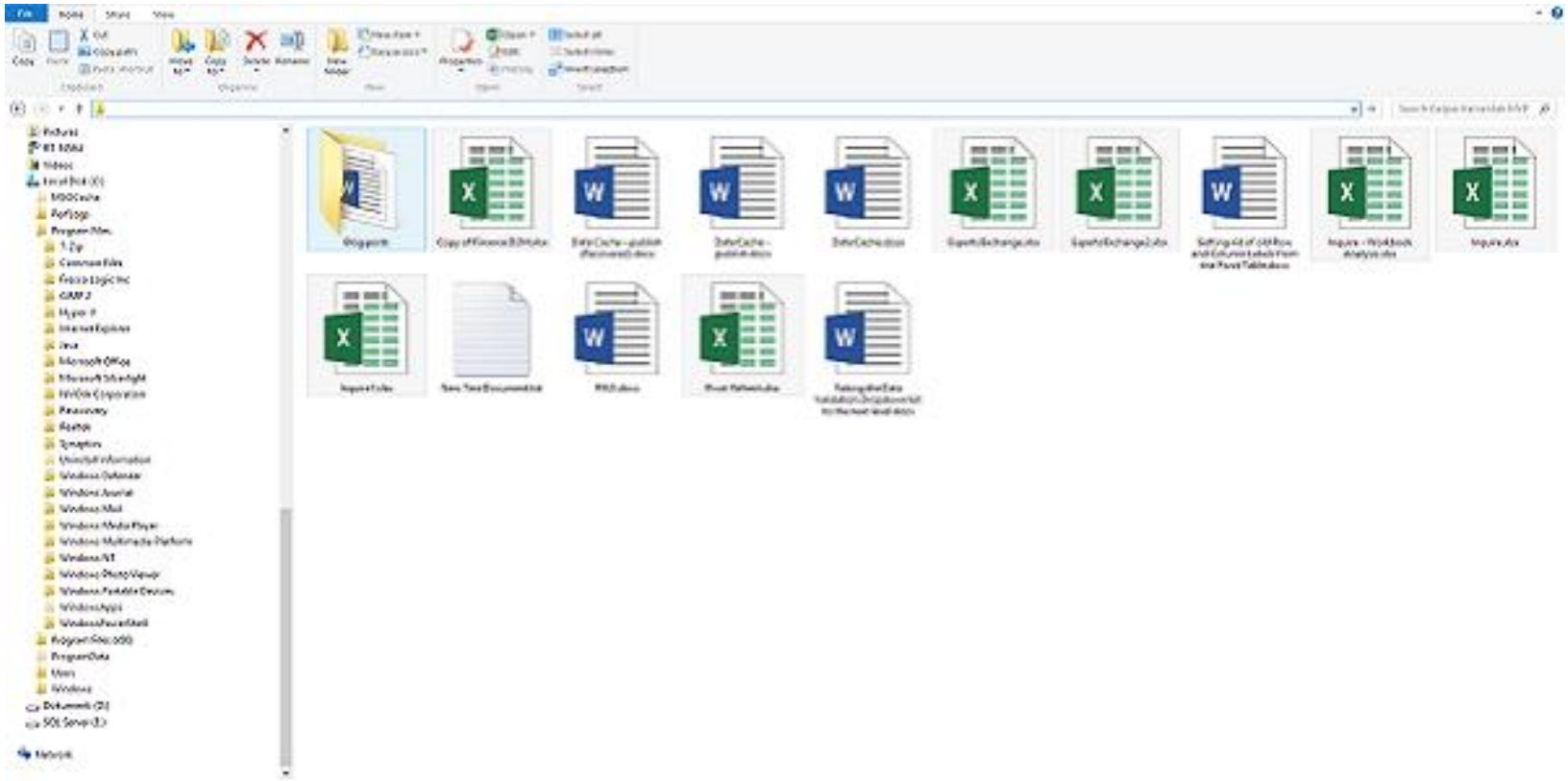
	144p	240p	360p	480p	720p	1080p
Per minute	1.3MB	3.3MB	5MB	8.3MB	25MB	50MB
Per hour	80MB	200MB	300MB	500MB	1.5GB	3GB

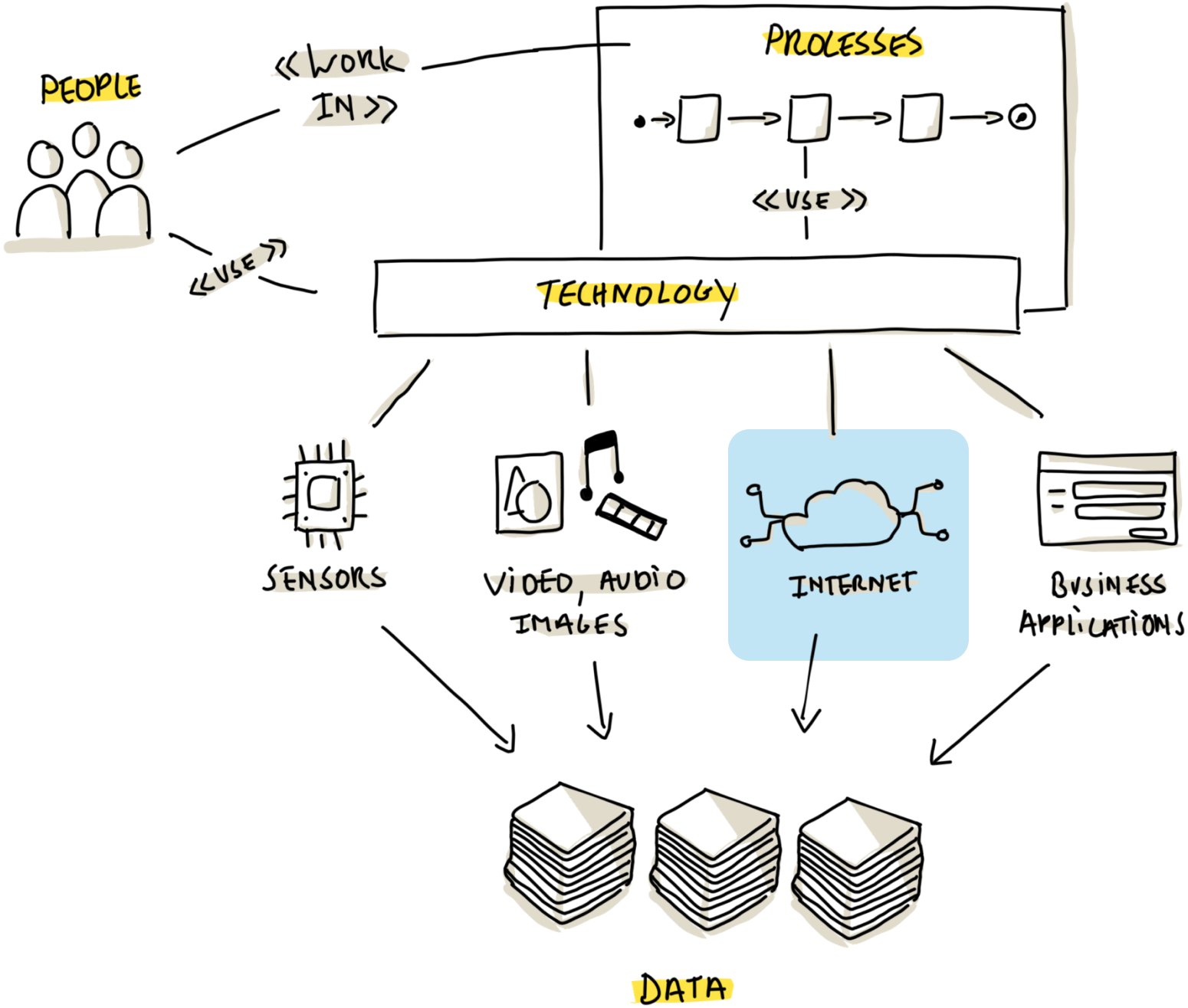
1 minute/hour of video data is **way larger** than 1 minute/hour of sensor data.

Document Scanning Apps

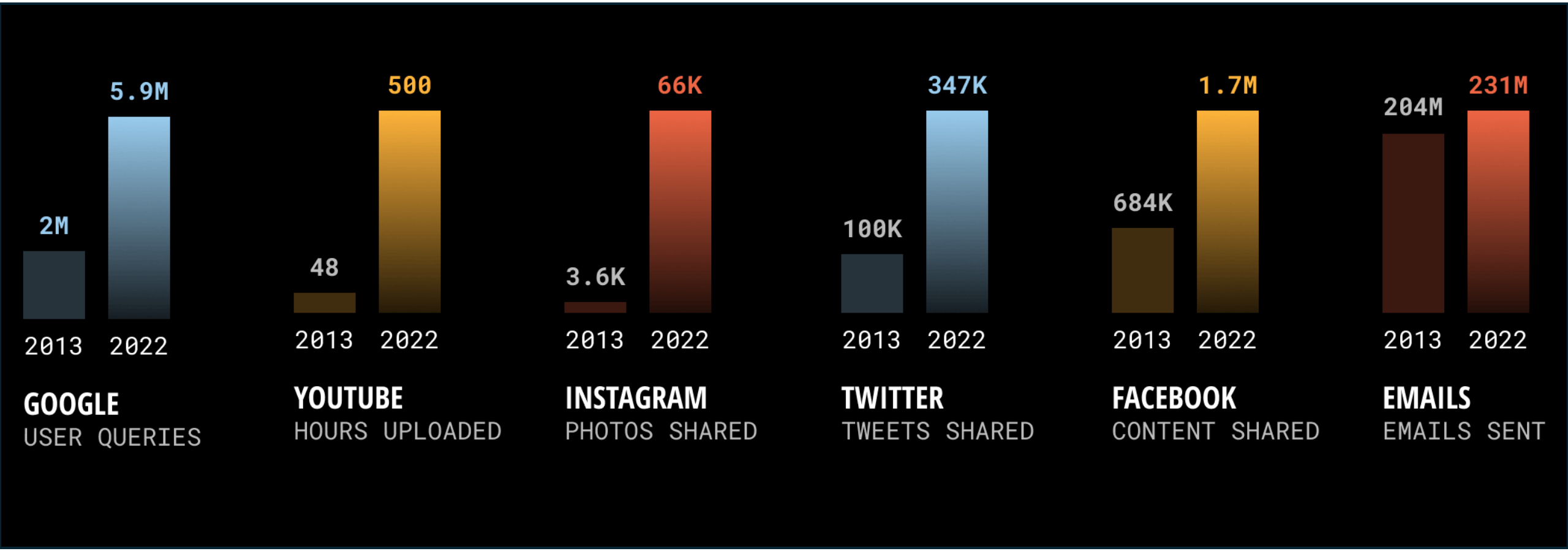


Digital Files




PEOPLE
«WORK IN»
PROCESSES
«USE»
«USE»
TECHNOLOGY
SENSORS
VIDEO, AUDIO IMAGES
INTERNET
BUSINESS APPLICATIONS
DATA

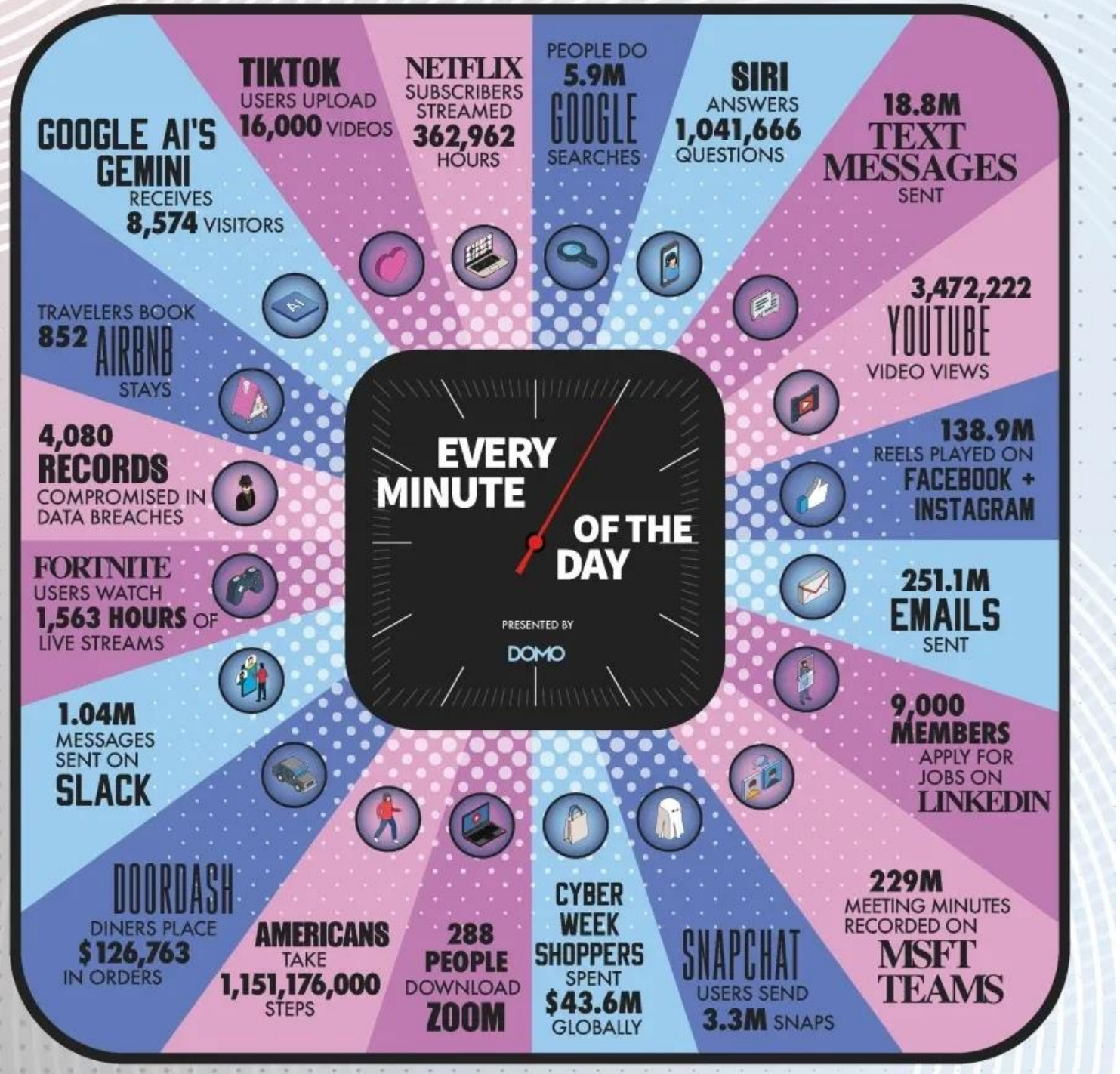
Big Data Producers: **The Internet**

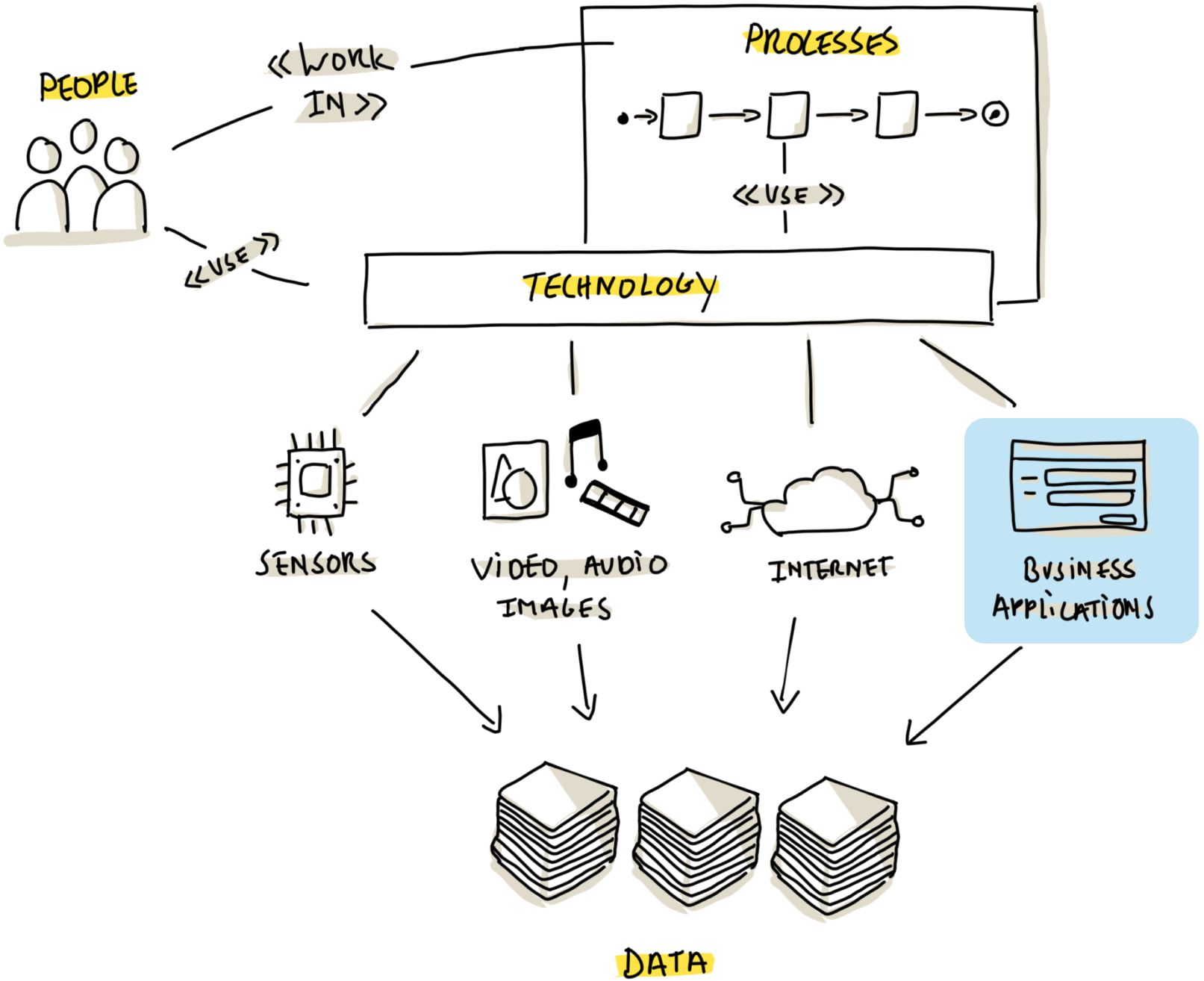


[Data Never Sleeps (Domo)]

1 MINUTE ON THE INTERNET

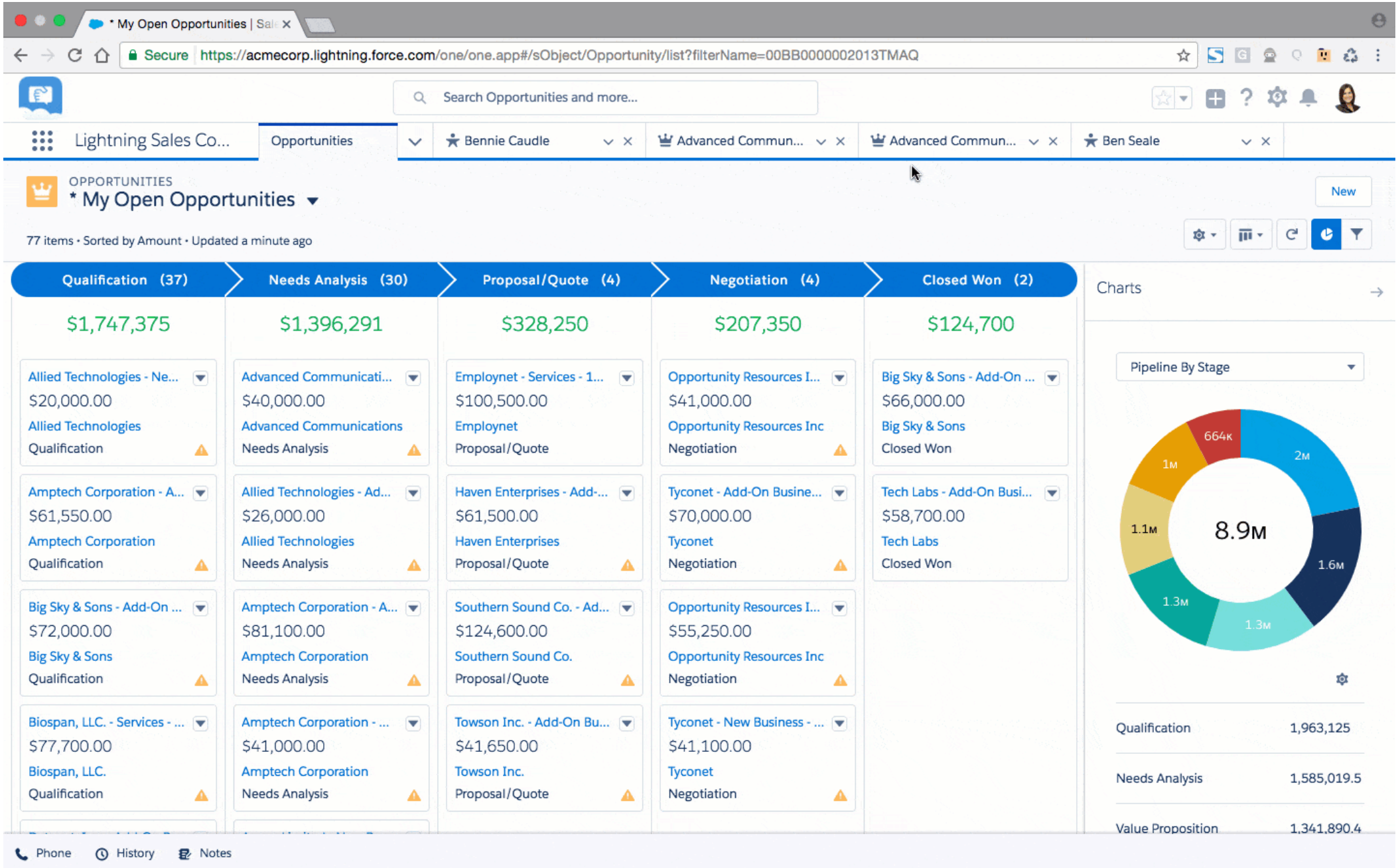
WWW.DCPWEB.CO.UK / @dcpWebDesigners


EVERY MINUTE OF THE DAY
PRESENTED BY DOMO
GOOGLE AI'S GEMINI RECEIVES 8,574 VISITORS
TIKTOK USERS UPLOAD 16,000 VIDEOS
NETFLIX SUBSCRIBERS STREAMED 362,962 HOURS
PEOPLE DO 5.9M GOOGLE SEARCHES
SIRI ANSWERS 1,041,666 QUESTIONS
18.8M TEXT MESSAGES SENT
3,472,222 YOUTUBE VIDEO VIEWS
138.9M REELS PLAYED ON FACEBOOK + INSTAGRAM
251.1M EMAILS SENT
9,000 MEMBERS APPLY FOR JOBS ON LINKEDIN
229M MEETING MINUTES RECORDED ON MSFT TEAMS
SNAPCHAT USERS SEND 3.3M SNAPS
CYBER WEEK SHOPPERS SPENT $43.6M GLOBALLY
288 PEOPLE DOWNLOAD ZOOM
AMERICANS TAKE 1,151,176,000 STEPS
DOORDASH DINERS PLACE $126,763 IN ORDERS
1.04M MESSAGES SENT ON SLACK
FORTNITE USERS WATCH 1,563 HOURS OF LIVE STREAMS
4,080 RECORDS COMPROMISED IN DATA BREACHES
TRAVELERS BOOK 852 AIRBNB STAYS


PEOPLE
«WORK IN»
PROCESSES
«USE»
«USE»
TECHNOLOGY
SENSORS
VIDEO, AUDIO IMAGES
INTERNET
BUSINESS APPLICATIONS
DATA

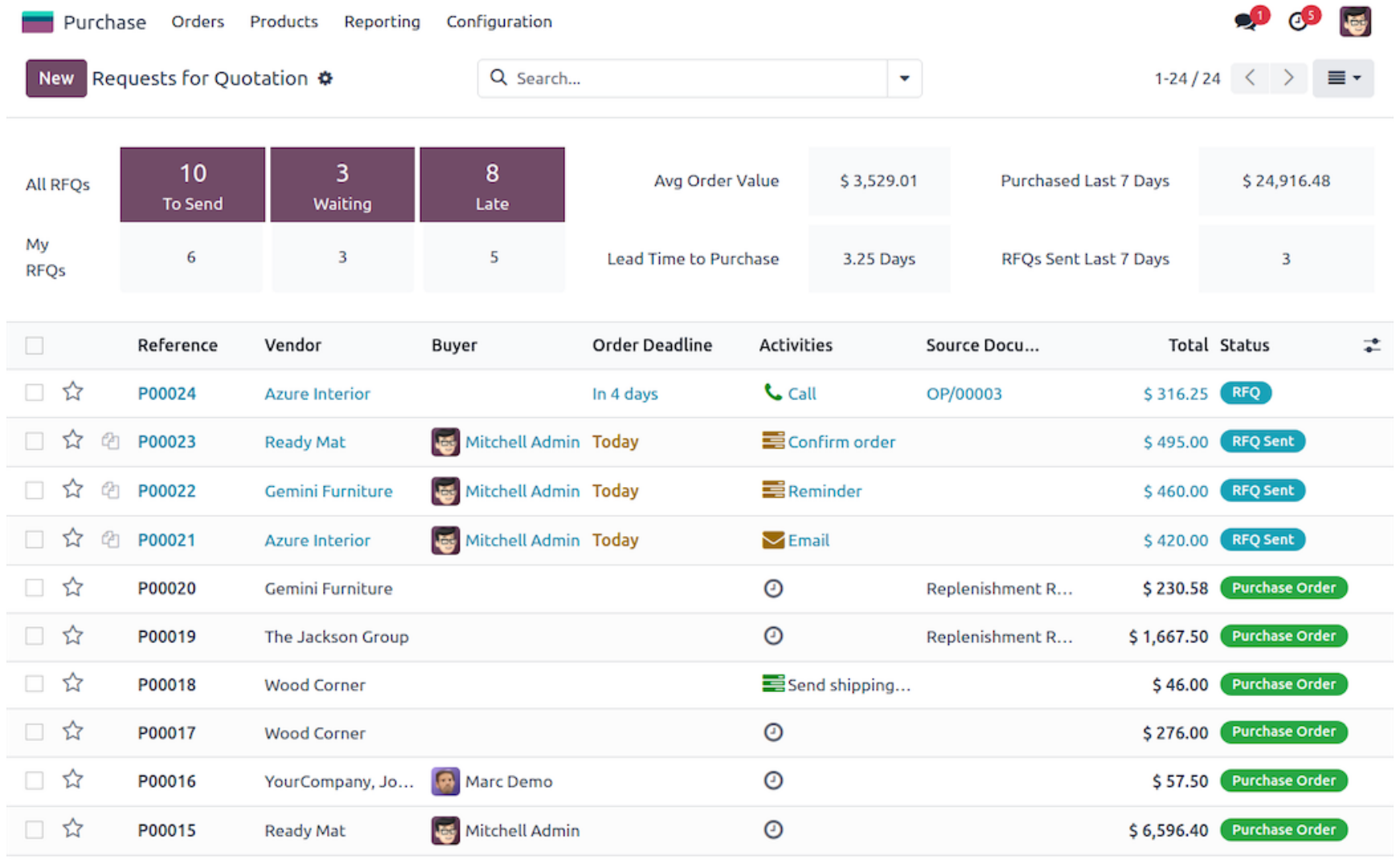
Business Applications

- Business data is collected by **Business Applications**
- Examples:
  - **CRM systems**
  - **ERP systems**
  - **Service Desk Management system**
  - **…**

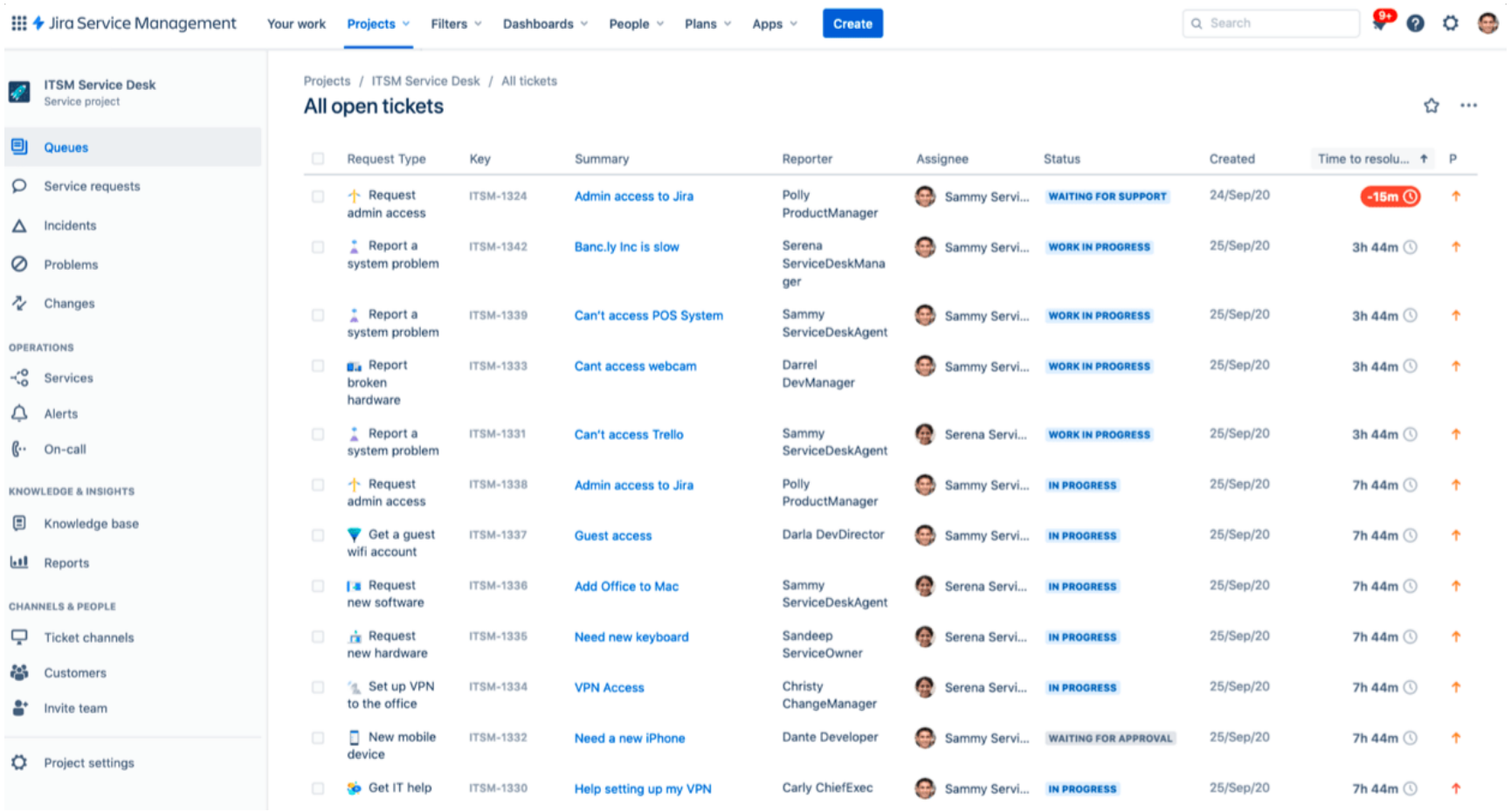


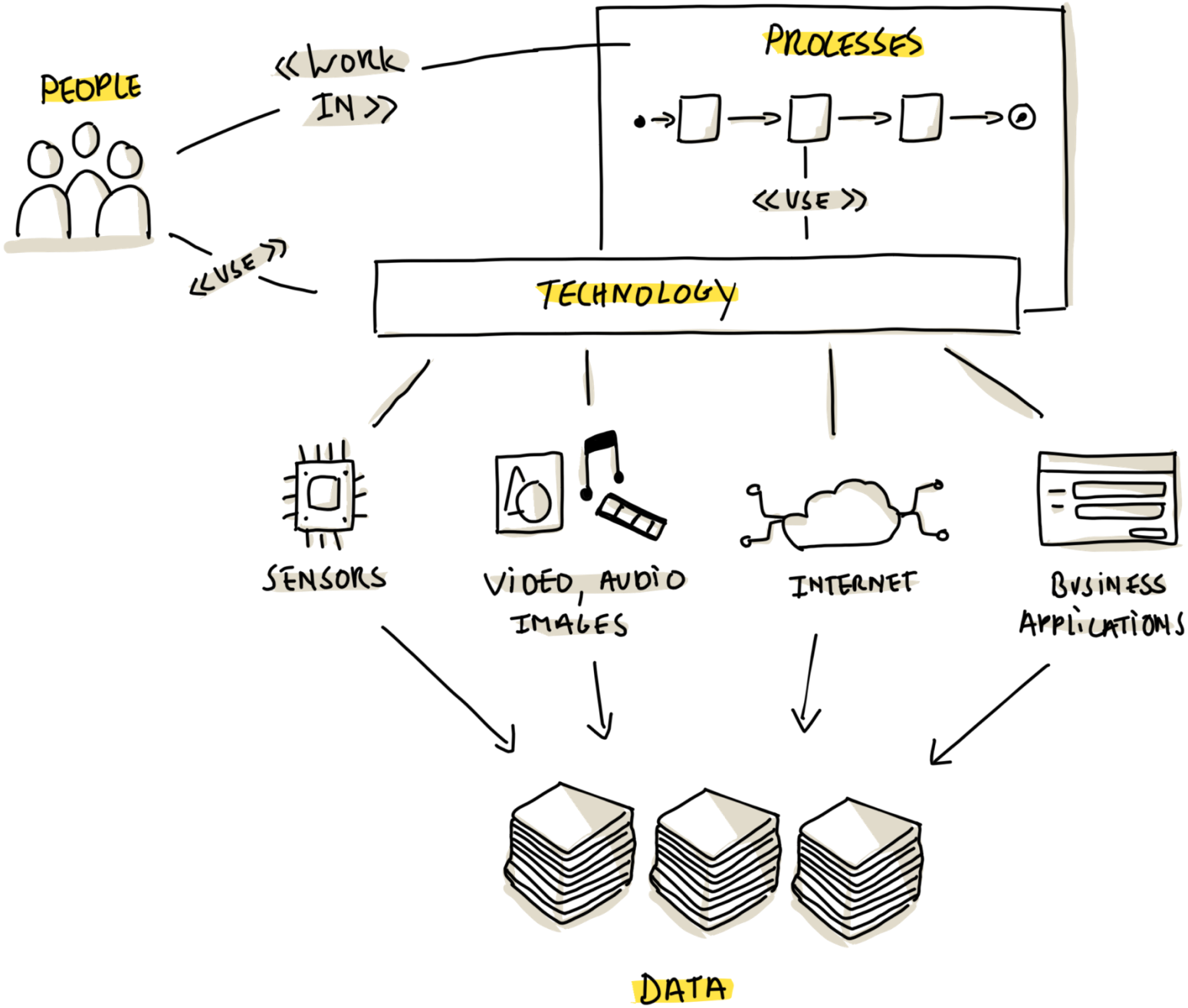
[Source: Salesforce Lightning]

ERP: Purchasing Example in ODOO



Service Desk Management: Example




PEOPLE
«WORK IN»
PROCESSES
«USE»
«USE»
TECHNOLOGY
SENSORS
VIDEO, AUDIO IMAGES
INTERNET
BUSINESS APPLICATIONS
DATA

A global port in the heart of Europe

One port
Two sites

Port of Antwerp Bruges

2nd largest port in **Europe**

Number one **export** port in Europe

147 mio tons/year

Largest **car handling** port in Europe

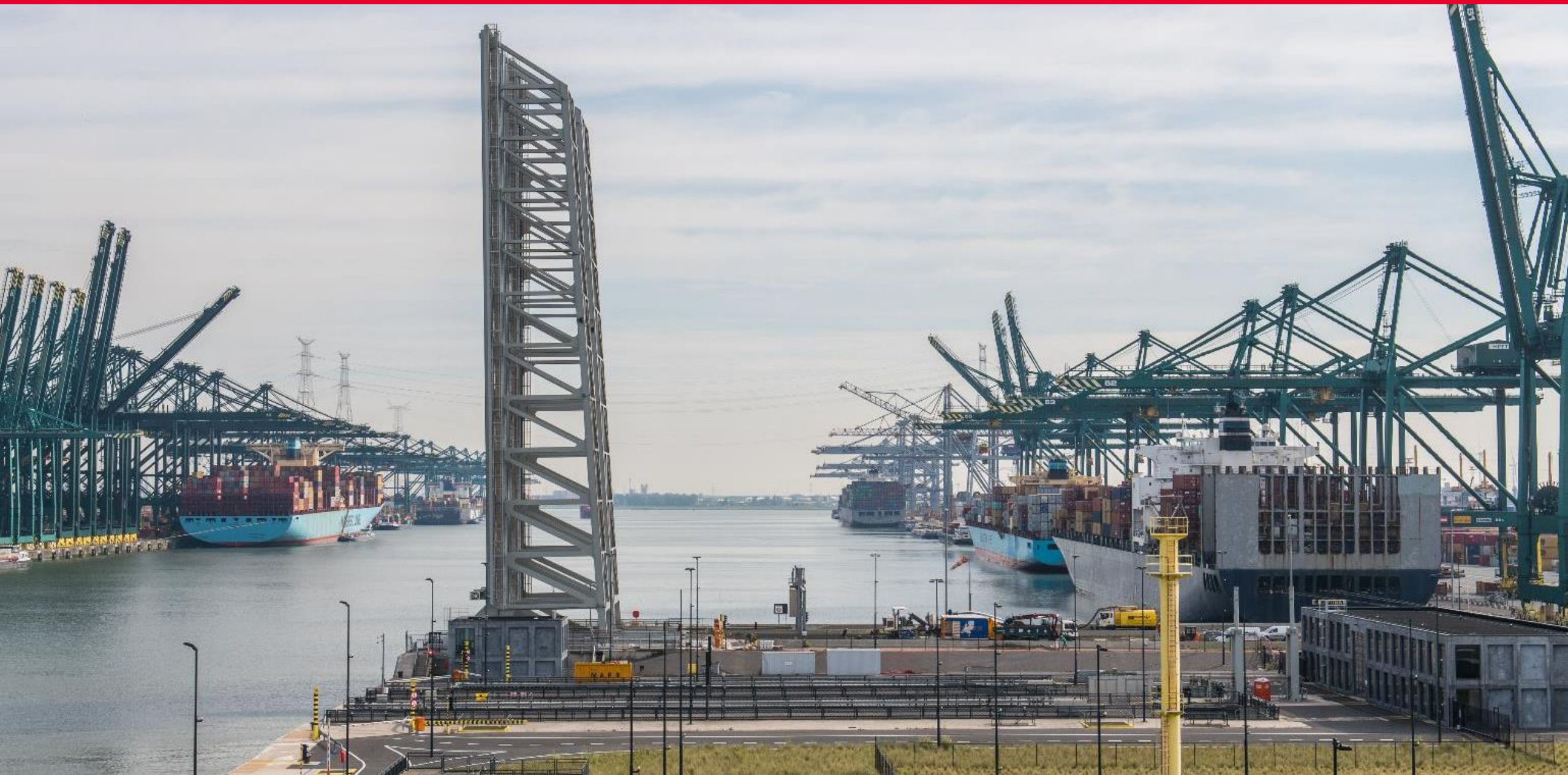
3,160,000 cars/year

21,400 **Seagoing ships**/year

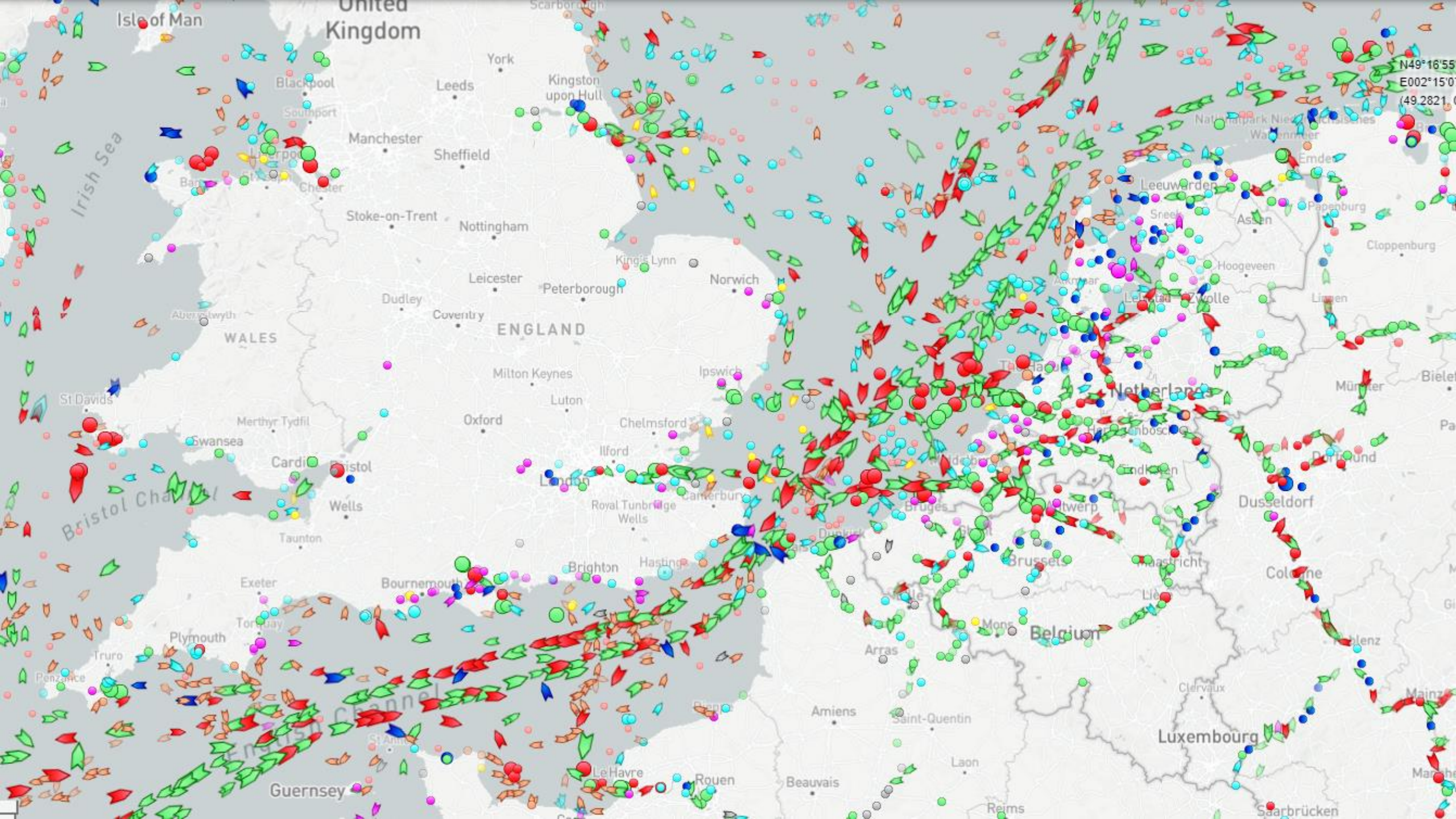
15% of EU **gas** market

Which Data is Available at Port Of Antwerp Bruges?

Data Is Everywhere at The Port



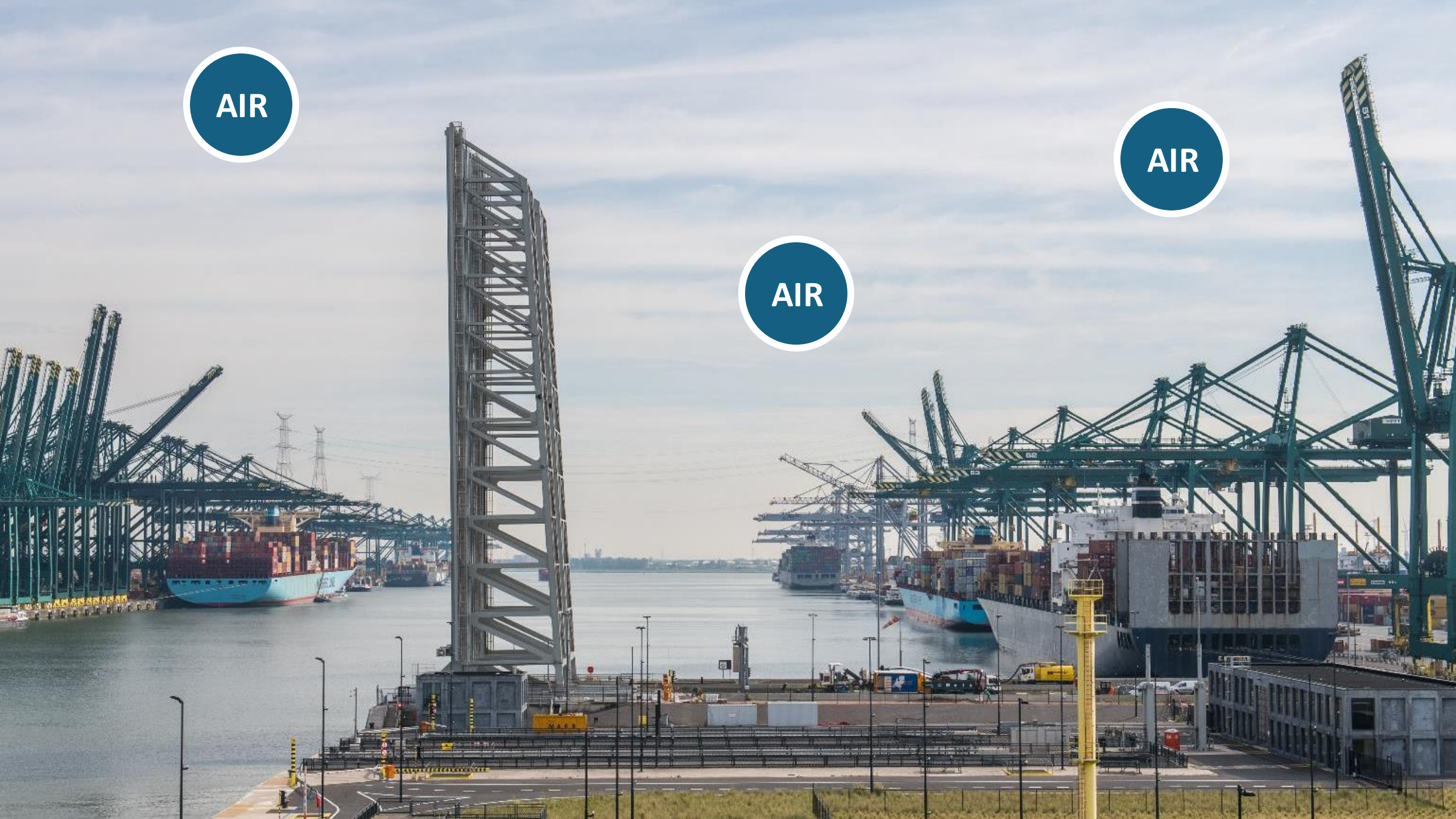

pos
pos
pos
pos
pos


Isle of Man
United Kingdom
York
Blackpool
Leeds
Kingston upon Hull
Manchester
Sheffield
Irish Sea
Stoke-on-Trent
Nottingham
Leicester
Peterborough
Norwich
King's Lynn
Dudley
Coventry
ENGLAND
WALES
Milton Keynes
Ipswich
Luton
Oxford
Chelmsford
St Davids
Swansea
Ilford
London
Bristol
Wells
Taunton
Royal Tunbridge Wells
Canterbury
Brighton
Bournemouth
Exeter
Plymouth
Truro
Penzance
Guernsey
Le Havre
Rouen
Amiens
Beauvais
Arras
Saint-Quentin
Laon
Reims
Brussels
Belgium
Netherlands
Luxembourg
Dusseldorf
Cologne
Leeuwarden
Hoogeveen
Zwolle
Assen
Emden
Cloppenburg
Eindhoven
Antwerp
Bruges
Maastricht
Mons
Saarbrücken
Mainz
N49°16'55
E002°15'0
(49.2821,

Automatic Identification System (AIS)

“Public” Data

Port of Antwerp Bruges

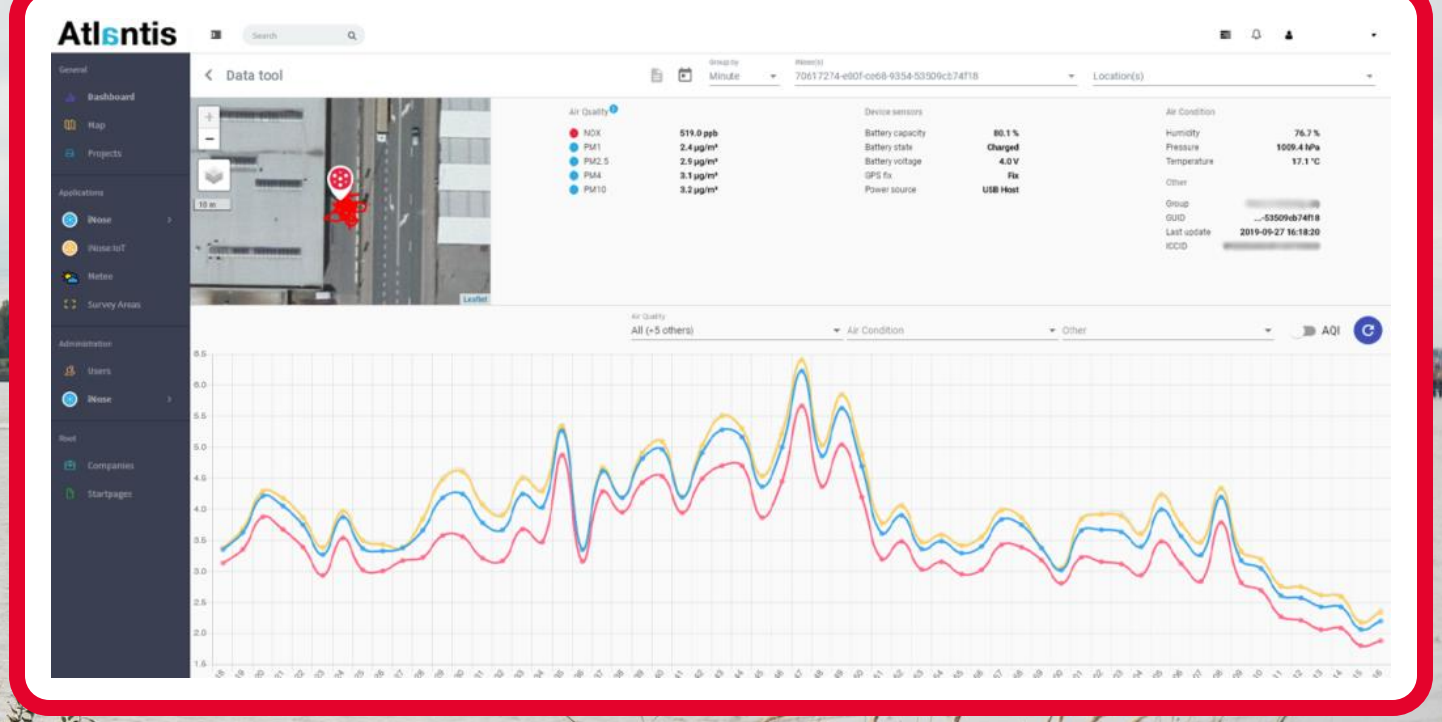

AIR
AIR
AIR

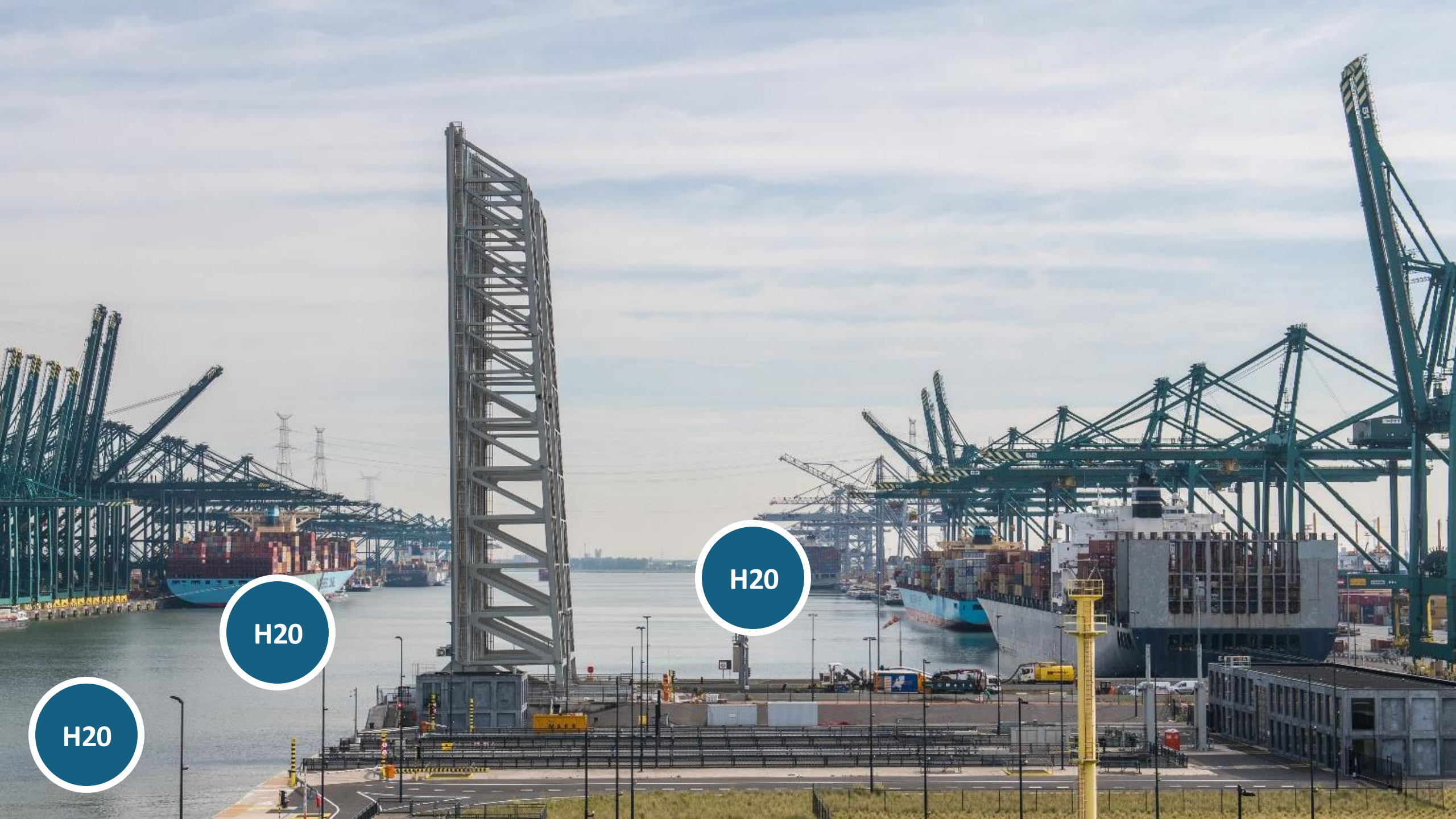
Air Quality Sensors

Port of Antwerp Bruges

VLAAMSE MILIEUMAATSCHAPPIJ

Vlaanderen is milieu




H20
H20
H20

Water Level Sensors

Vlaanderen
is wetenschap

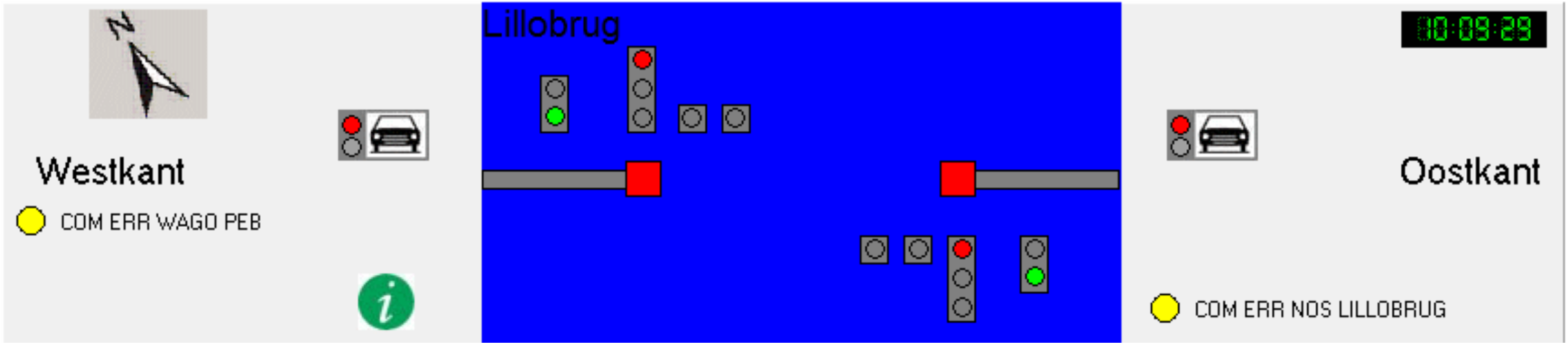
Floating Waste

Port of Antwerp Bruges


BRIDGE

Bridges

Port of Antwerp Bruges

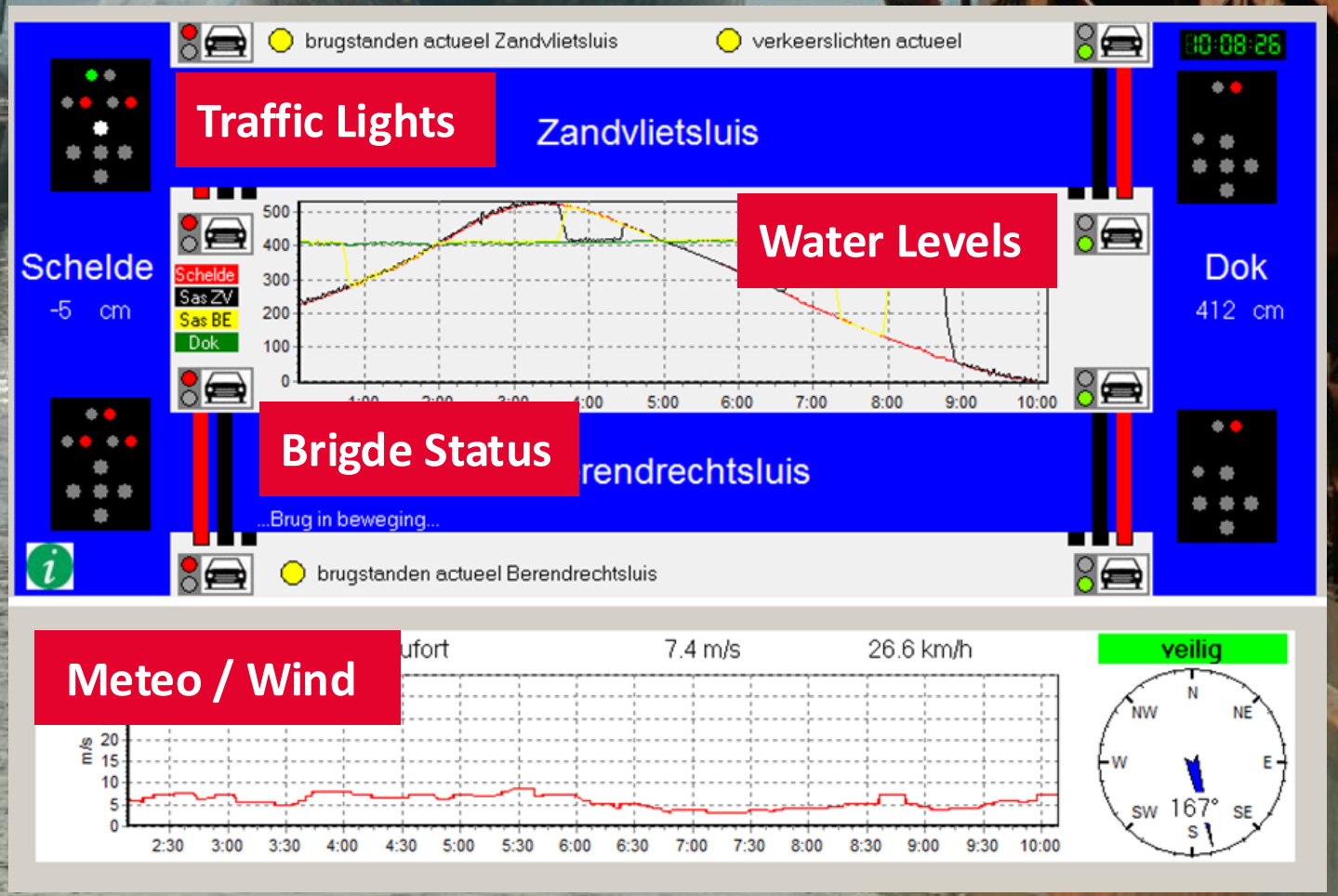


Locks

Port of Antwerp Bruges

Locks

Port of Antwerp Bruges




ENERGY

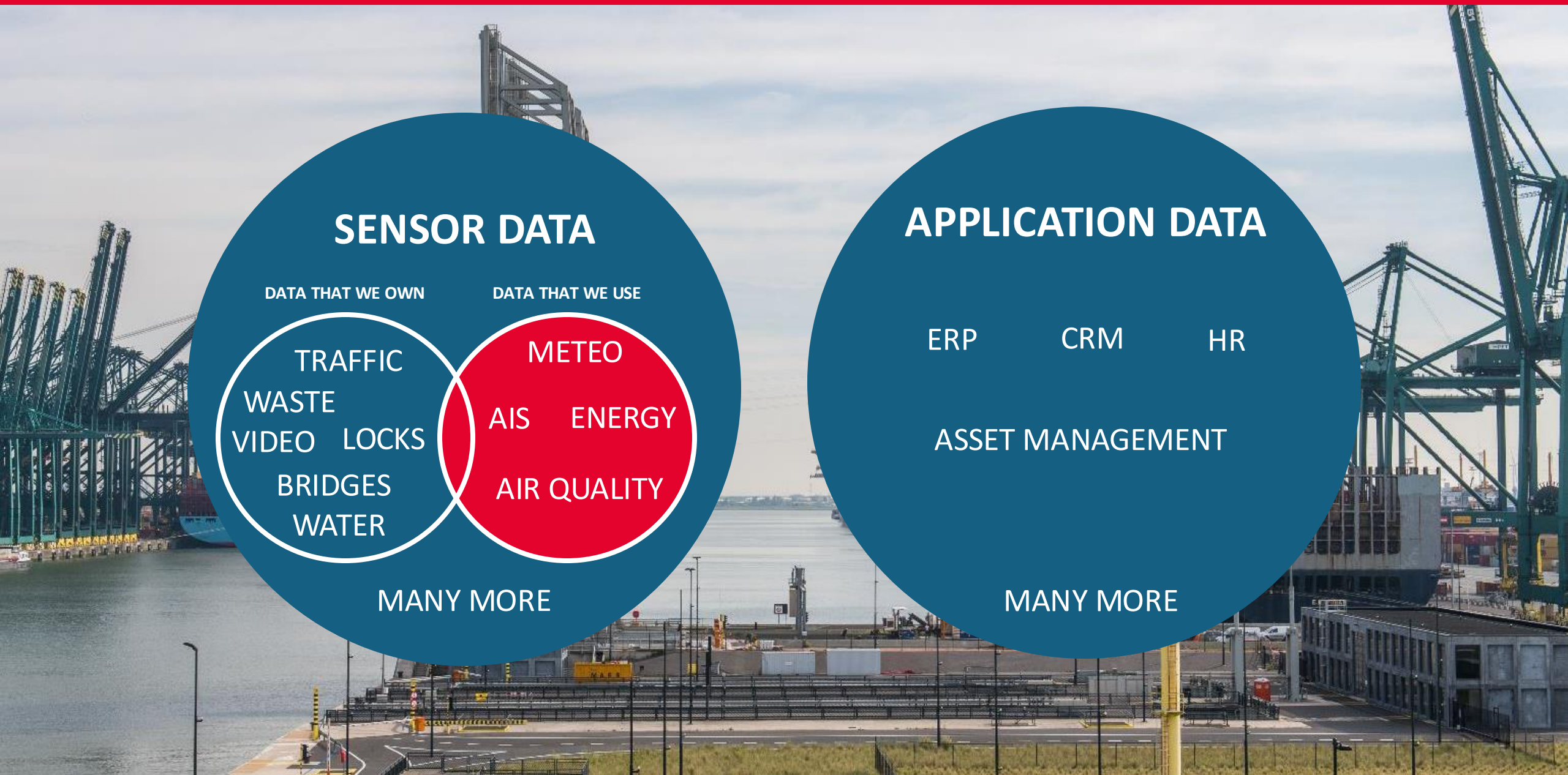
Energy

Port of Antwerp Bruges

fluvius.

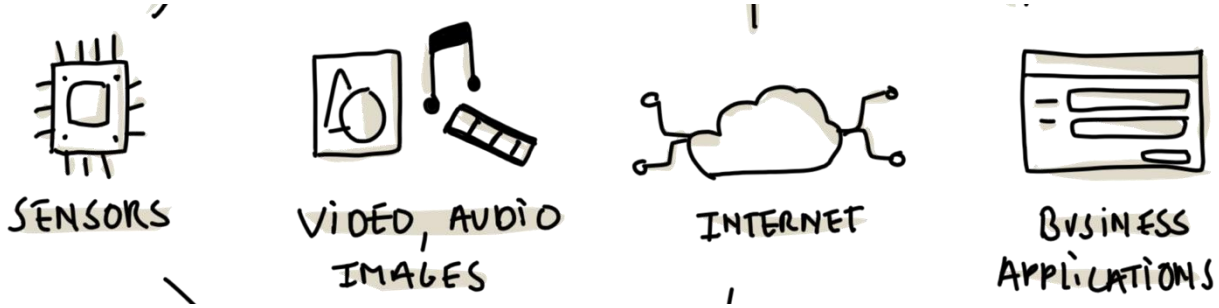
Data Is Everywhere at The Port

Data Is Everywhere at The Port



OEFENING: Welke Data Providers leveren data voor jullie use case?

- Welke data providers?
- Welk type providers?

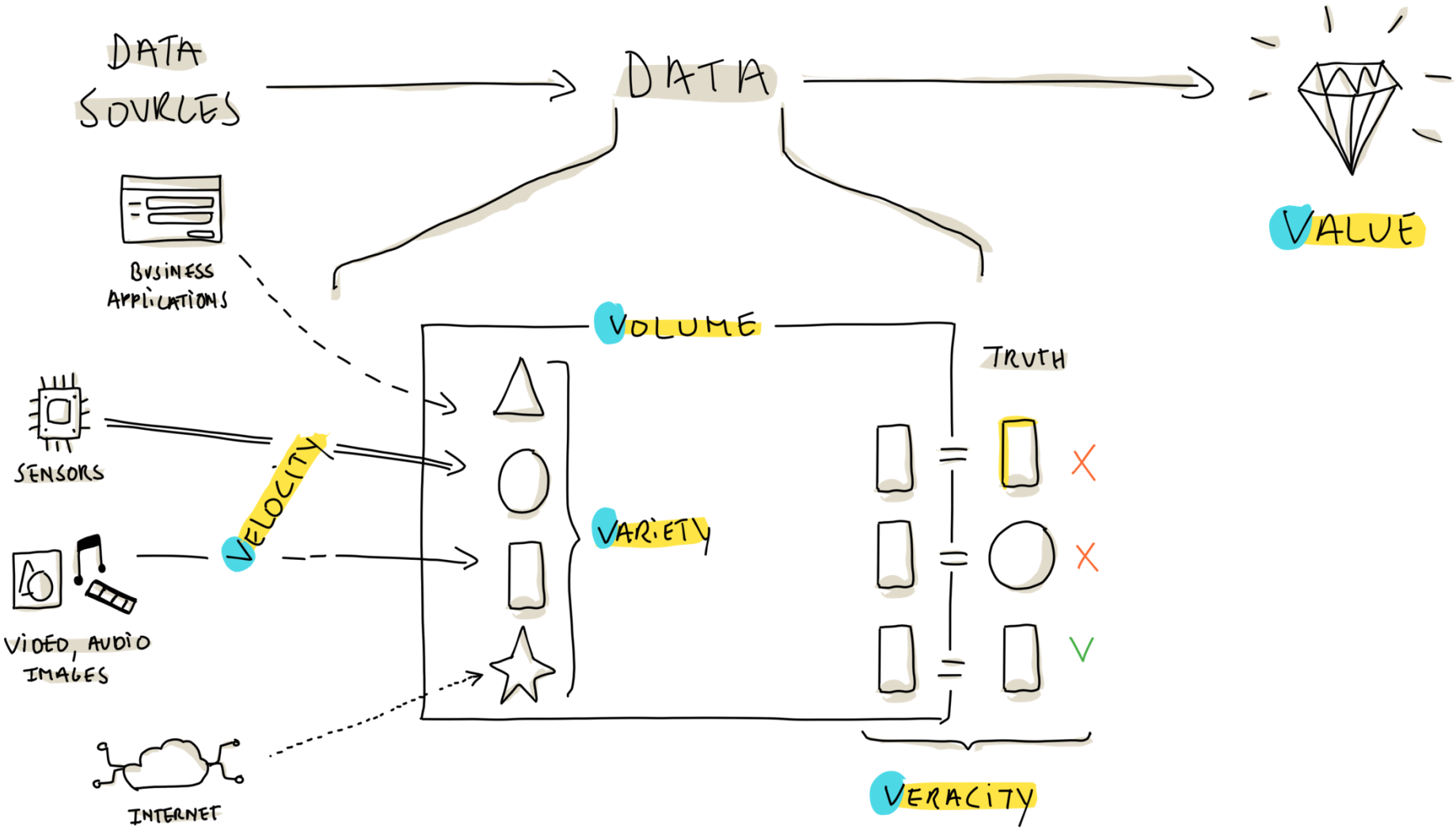


- Bespreek kort in groep – delen achteraf

Data Management

- Defining Data
- Data Producers
- **Big Data Vs**
- The Bigger Picture
- Data Management
- Data Technology

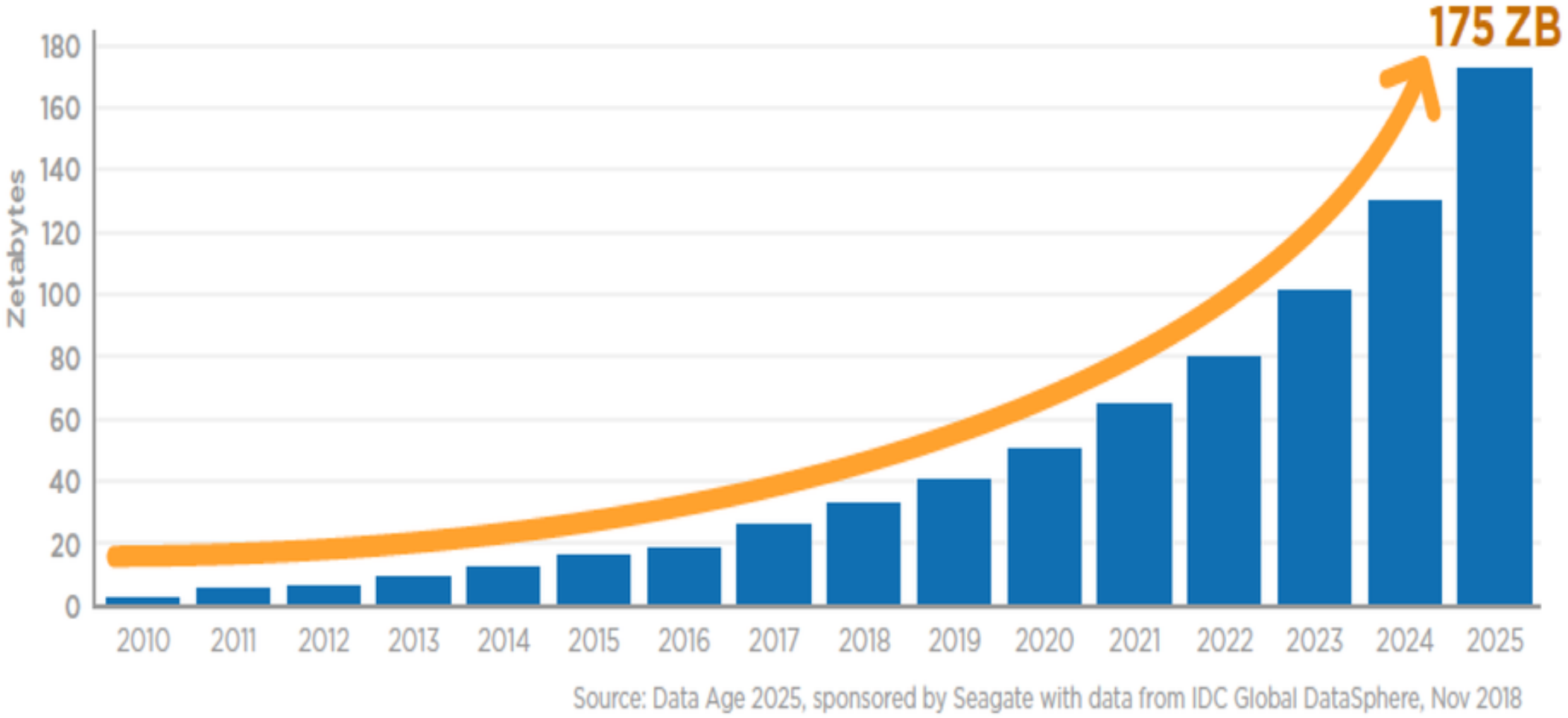
5 Data V's


DATA SOURCES
DATA
VALUE
BUSINESS APPLICATIONS
VOLUME
TRUTH
SENSORS
VELOCITY
VARIETY
X
X
VIDEO, AUDIO IMAGES
V
INTERNET
VERACITY

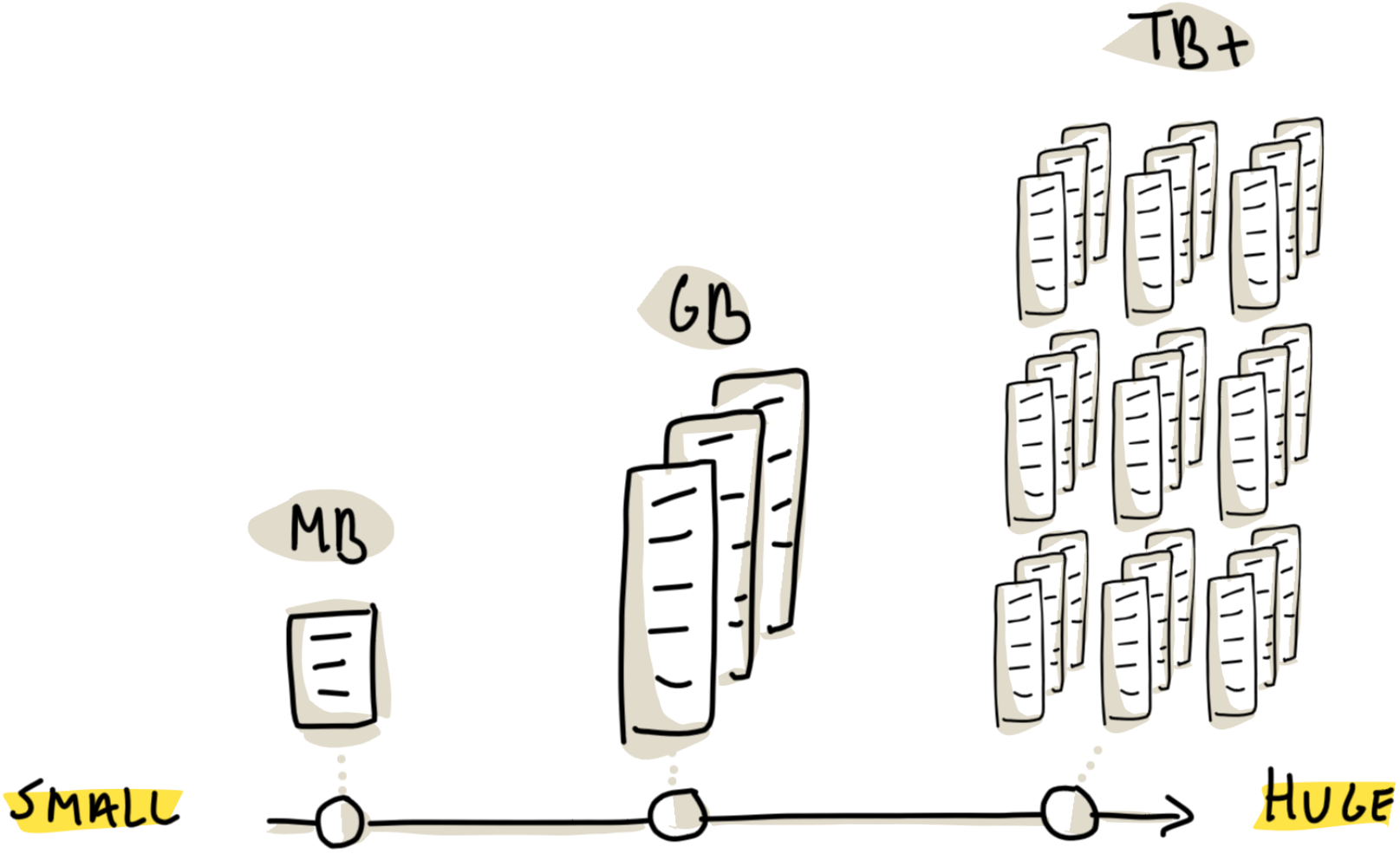
Volume - the **size** of Big Data

- Big data analytics entails handling and analyzing **vast amounts of data**.
- To effectively work with such massive datasets, **specialized tools and infrastructure are necessary** for capturing, storing, managing, cleaning, transforming, analyzing, and reporting the data.

Volume

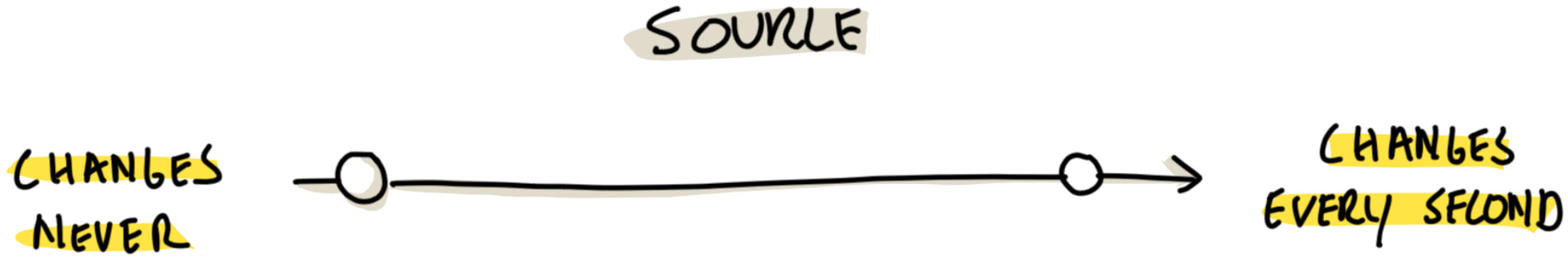

175 ZB
Zetabytes
180
160
140
120
100
80
60
40
20
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Volume

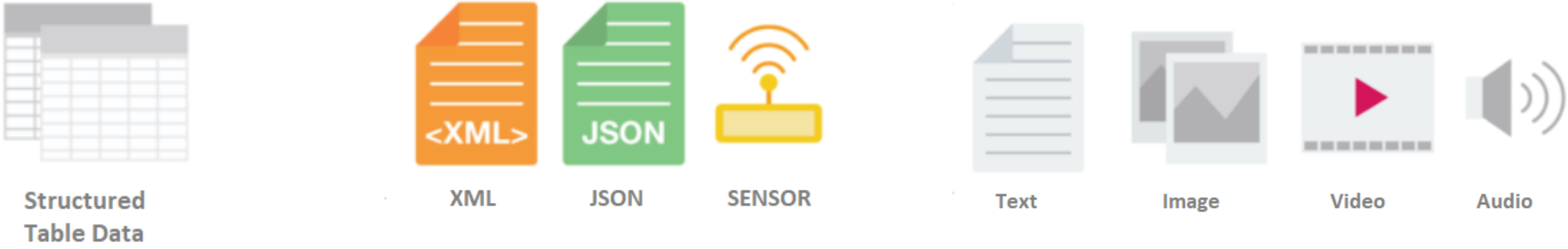


Velocity - the **speed** at which data is growing

- Velocity denotes the speed at which data is generated.
- To keep up with the rapid generation of data, systems for processing and analyzing data must possess **sufficient capacity to handle the influx of data and deliver timely, actionable insights**.



Variety - the **different** structures of data



STRUCTURED **SEMI STRUCTURED** **UNSTRUCTURED**

Variety – Structured Data

- Constituted of elements that are readily addressable to facilitate effective analysis.
- This category typically arranges data into tabular structures with rows and columns.
- Structured data presents a straightforward starting point for data analysis.
- Examples encompass relational data, CSV files, and spreadsheets.

Tabular Data

ID	TOTAL ACTIONS	ACTION 1	ACTION 2	TOTAL TIME
10	120	80	40	0:50:05
11	255	130	125	1:40:03
12	180	100	80	1:20:19
13	305	205	100	1:58:58
14	71	50	21	0:35:41
15	418	310	108	2:08:18
16	222	150	72	1:32:58

Variety – Semi-structured Data

- This data type lacks the rigid tabular format
- Retains some inherent organizational traits that render it more analyzable compared to unstructured data.
- Illustrations include XML data and JSON data.

Variety – Semi-structured Data

XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<endereco>
    <cep>31270901</cep>
    <city>Belo Horizonte</city>
    <neighborhood>Pampulha</neighborhood>
    <service>correios</service>
    <state>MG</state>
    <street>Av. Presidente Antônio Carlos, 6627</street>
</endereco>
```

vs.

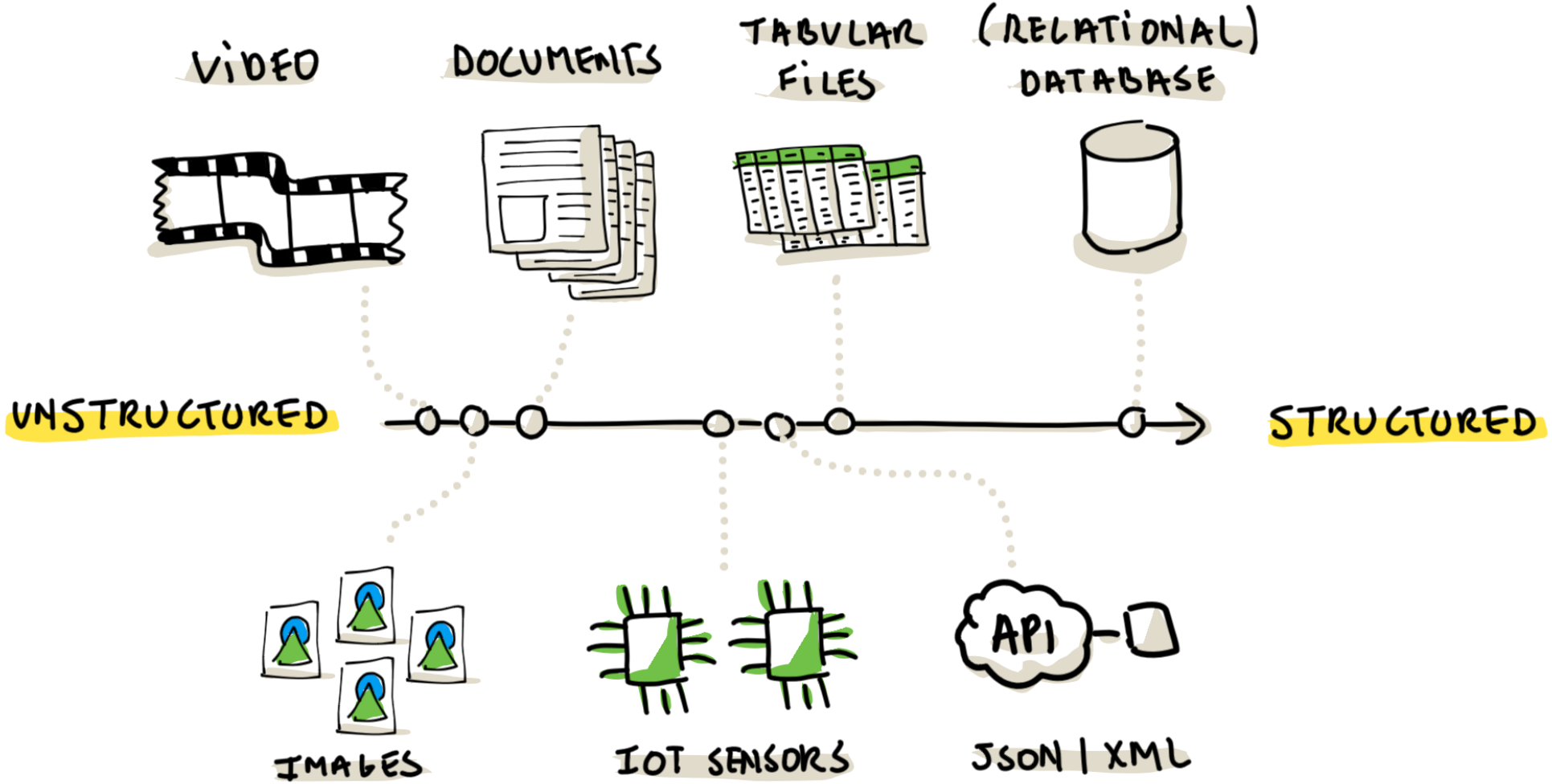
JSON

```json
{
    "endereco": {
        "cep": "31270901",
        "city": "Belo Horizonte",
        "neighborhood": "Pampulha",
        "service": "correios",
        "state": "MG",
        "street": "Av. Presidente Antônio Carlos, 6627"
    }
}
```

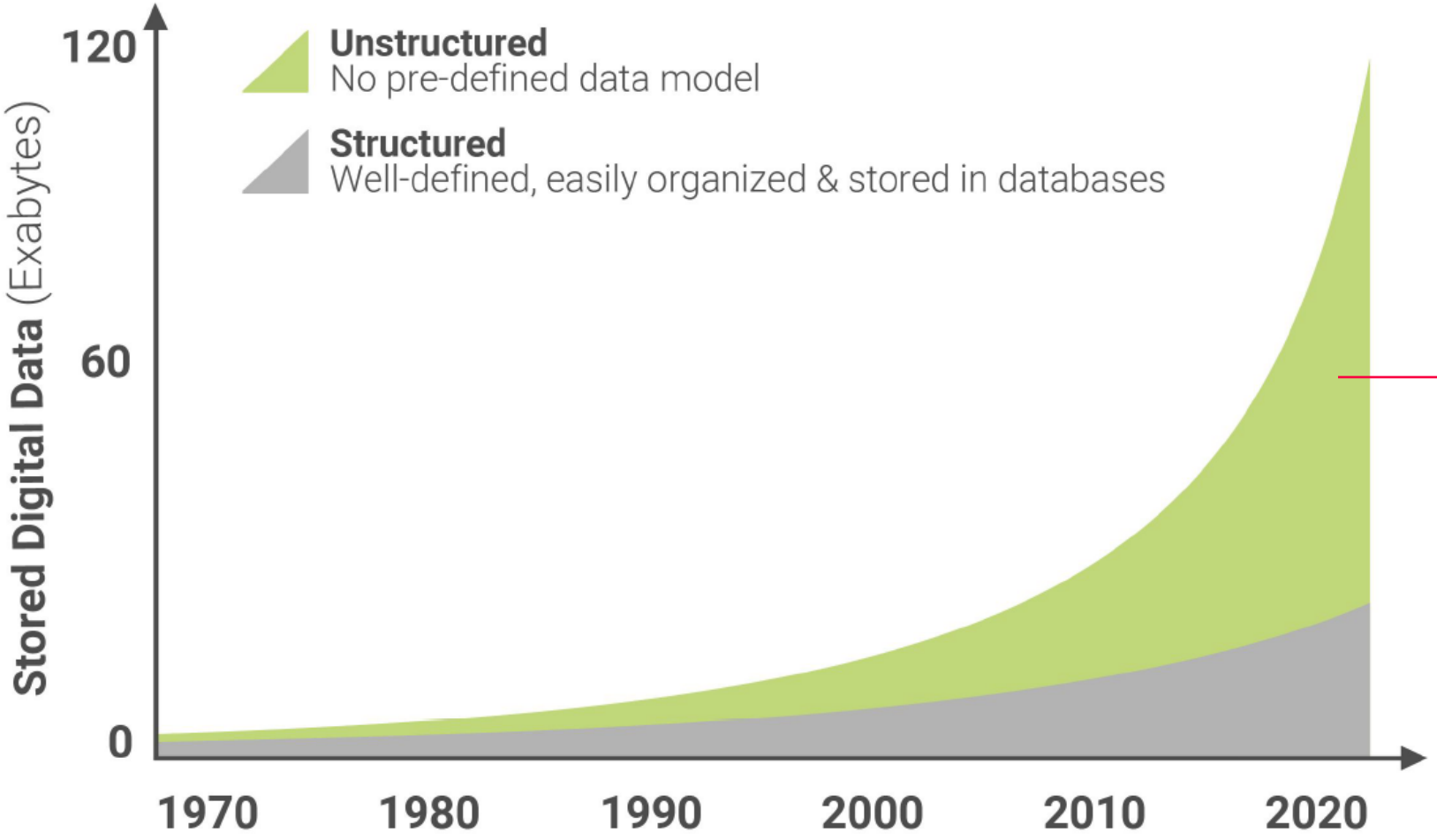
Variety – Unstructured Data

- Data lacking predefined organization or structure.
- The absence of clear formatting renders this data type intricate to analyze.
- Examples entail documents like Word files, PDFs, textual content, images, videos, and audio recordings.

Variety

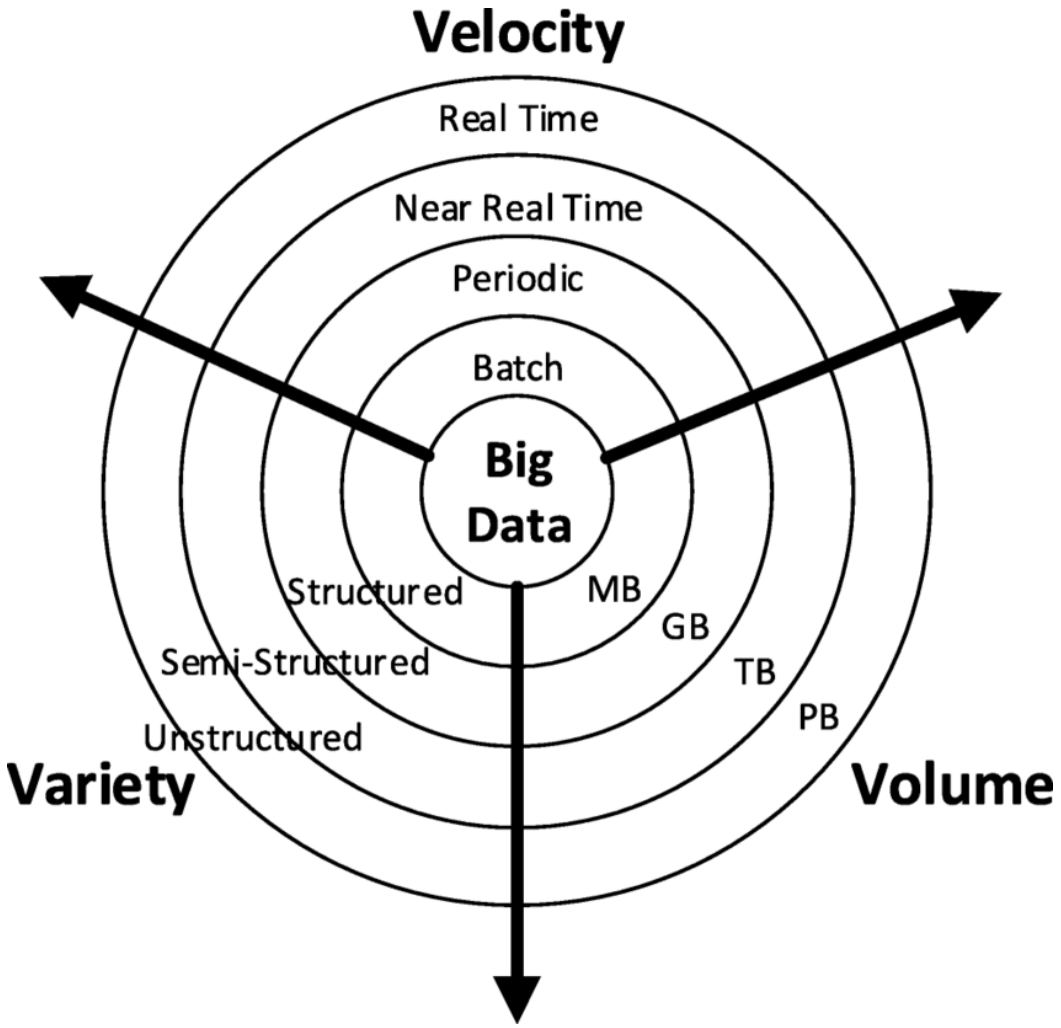

VIDEO
DOCUMENTS
TABULAR FILES
(RELATIONAL) DATABASE
UNSTRUCTURED
STRUCTURED
IMAGES
IOT SENSORS
API
JSON | XML

Variety / Velocity / Volume



Exponential growth of unstructured data!

Variety / Velocity / Volume



Veracity – Accuracy or truthfulness of data?

- Veracity pertains to the accuracy and authenticity of the data.
- Data must undergo validation to ensure that it accurately represents essential business functions.
- Any data manipulation does not compromise the data's accuracy.

Value – How useful is the data?

Value – How useful is the data?

- Big data must generate value.
- The insights derived from an analysis should for example provide meaningful guidance for improving operations, enhancing customer service, or creating other forms of value.

OEFENING: De 5 V’s

- Wat zijn de belangrijkste eigenschappen (5 V’s) van de data voor jullie use cases?

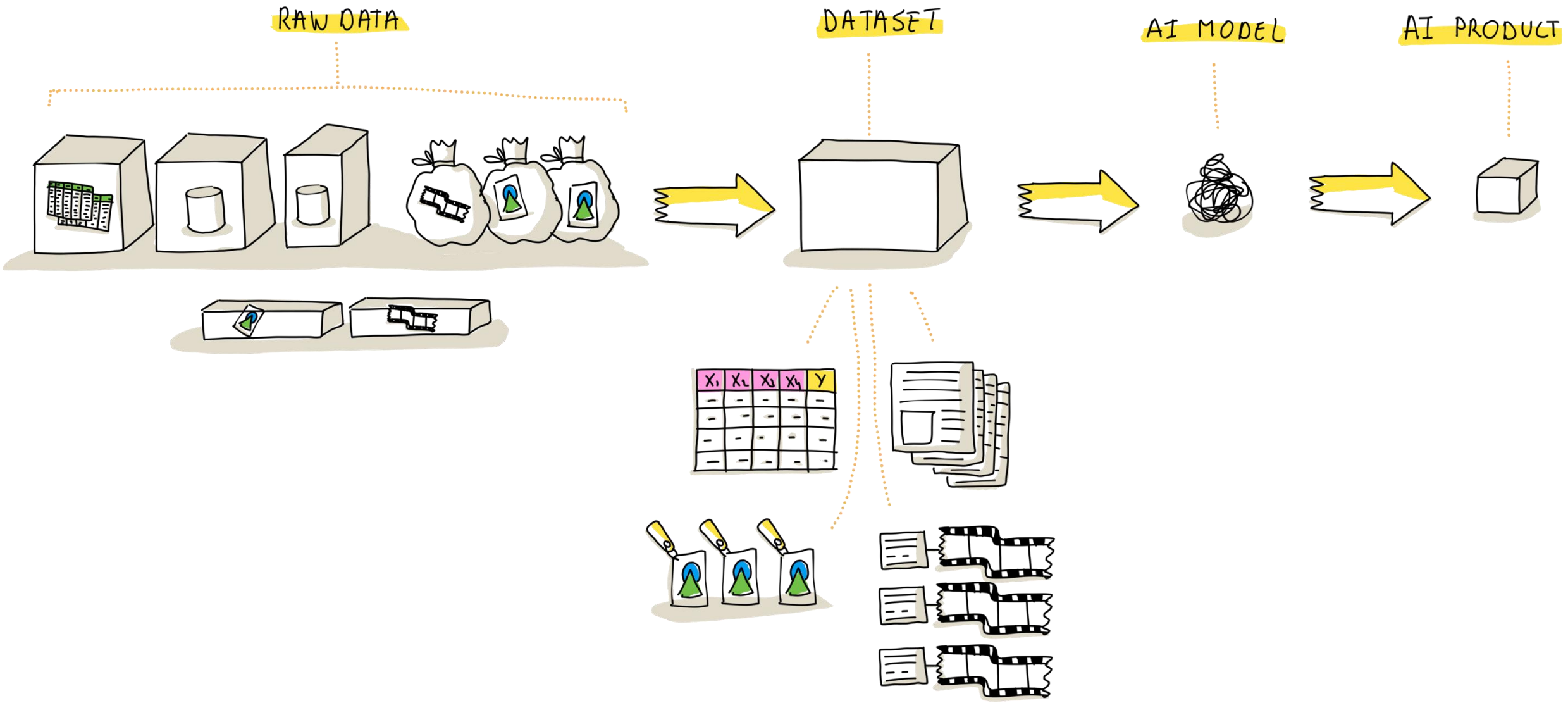
    1. Velocity
    2. Volume
    3. Variety
    4. Veracity
    5. Value

- Overleg kort in groep

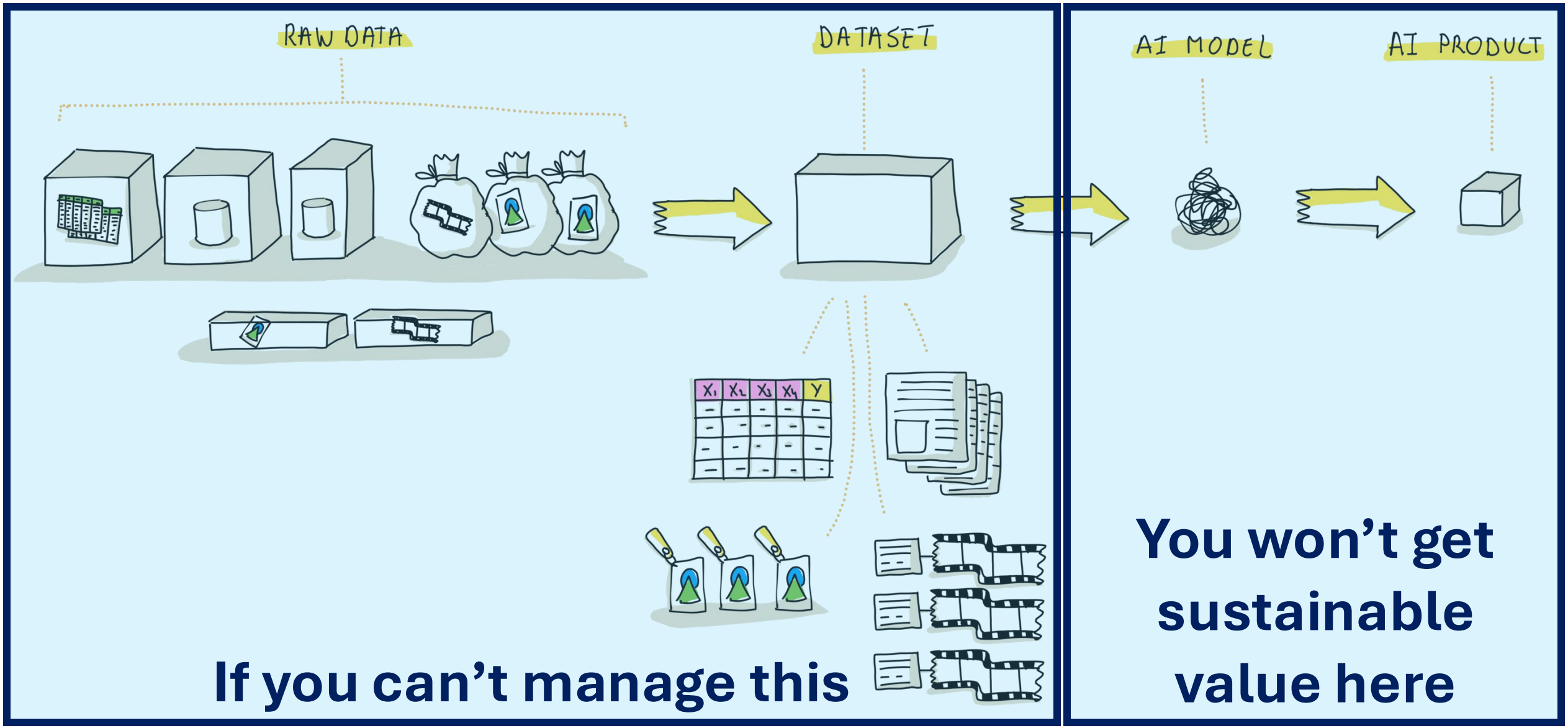
Data Management

- Defining Data
- Data Producers
- Big Data Vs
- **The Bigger Picture**
- Data Management
- Data Technology

The Bigger Picture



The Bigger Picture



THE WALL STREET JOURNAL.

CIO JOURNAL

Rush to Use Generative AI Pushes Companies to Get Data in Order

Data management is under the spotlight again as companies seek to out-innovate competitors with large language models

By *Belle Lin* Follow

June 8, 2023 7:00 am ET

Corporate technology chiefs are under pressure to ensure companies' data is stored, filtered, and protected for use with AI. PHOTO: I-HWA CHENG/BLOOMBERG NEWS

Barriers to Preventing Organizations From Delivering More Value From Data, Analytics, and AI

Q11: Please rank all barriers preventing you from delivering more value from data, analytics, and AI, where 1 is the biggest/most present barrier (Note: Other (n=41) not shown):

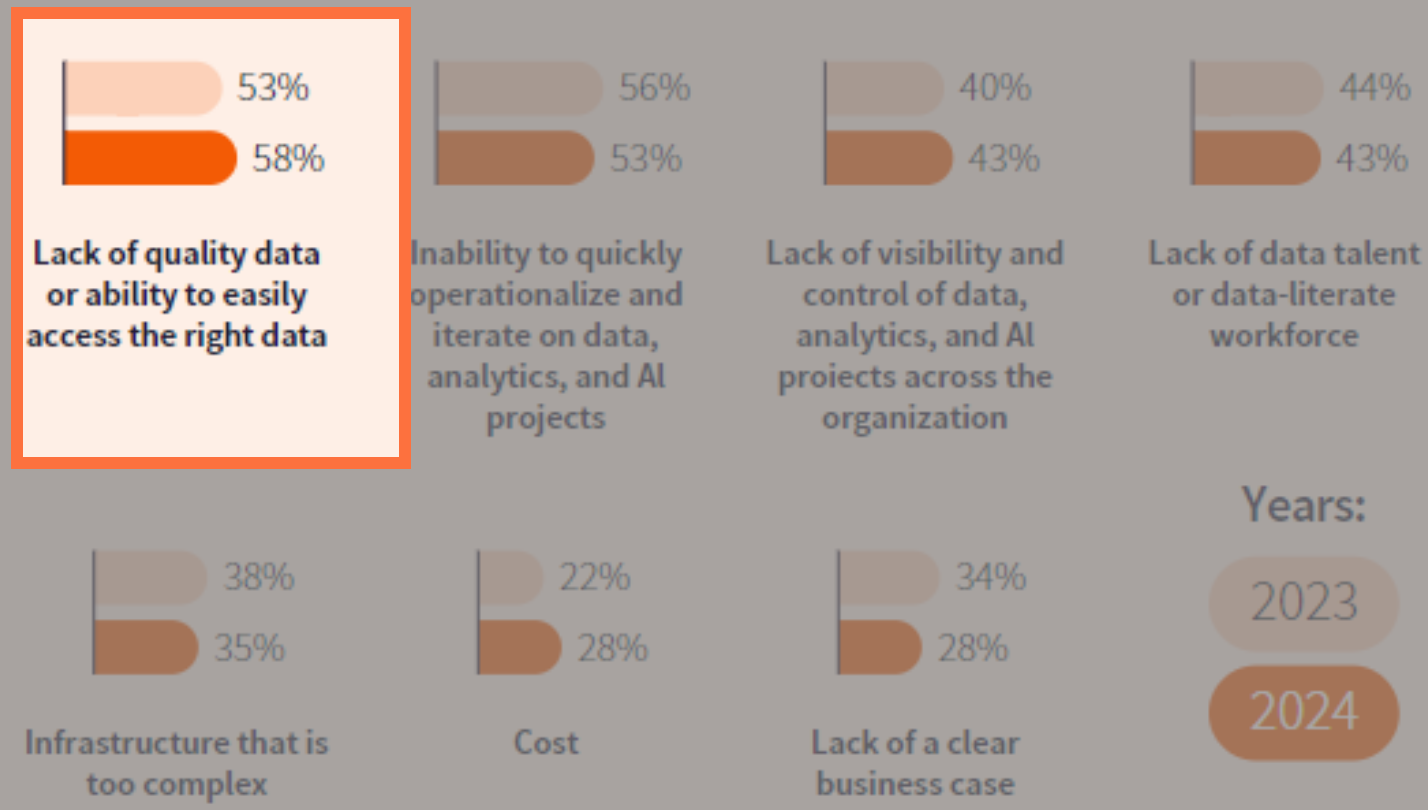
% Ranking Each 'Rank 1 - Biggest/Most Present Barrier,' 'Rank 2,' or 'Rank 3'

[Dataiku: AI, Today (400 respondents – 2024)]

Barriers to Preventing Organizations From Delivering More Value From Data, Analytics, and AI

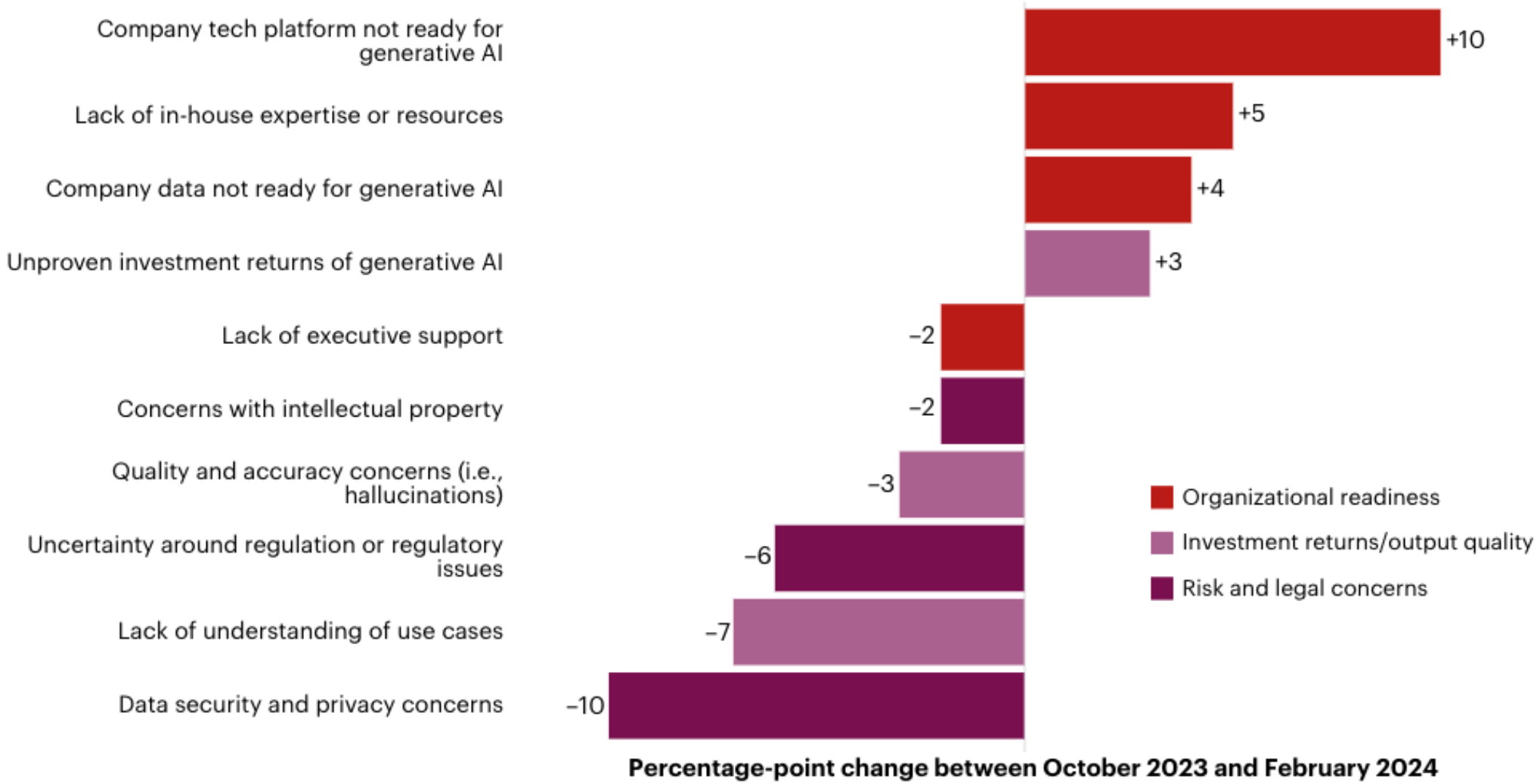
Q11: Please rank all barriers preventing you from delivering more value from data, analytics, and AI, where 1 is the biggest/most present barrier (Note: Other (n=41) not shown):

% Ranking Each 'Rank 1 - Biggest/Most Present Barrier,' 'Rank 2,' or 'Rank 3'



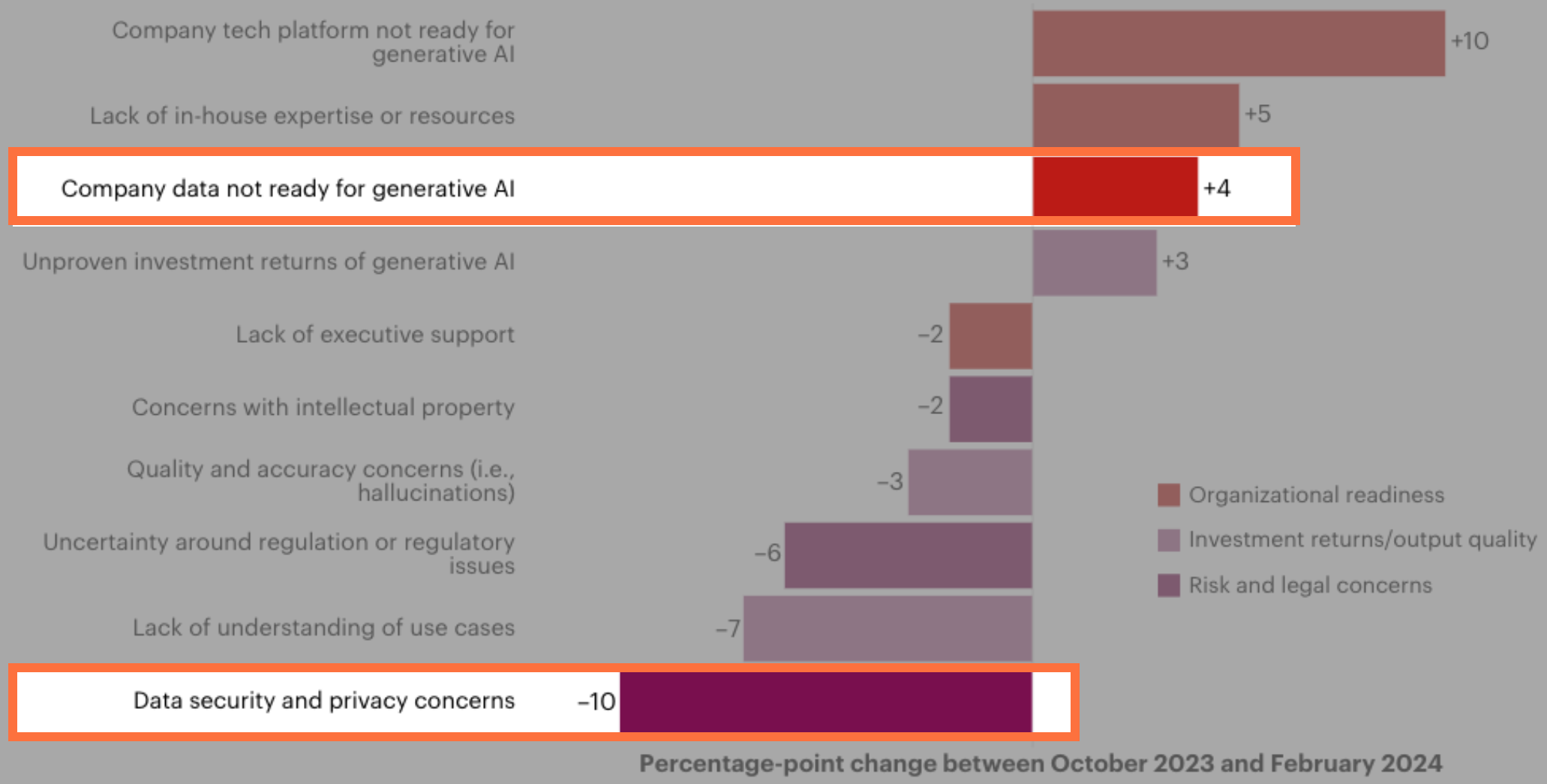
[Dataiku: AI, Today (400 respondents – 2024)]

What are the top reasons preventing your company from moving faster with generative AI?



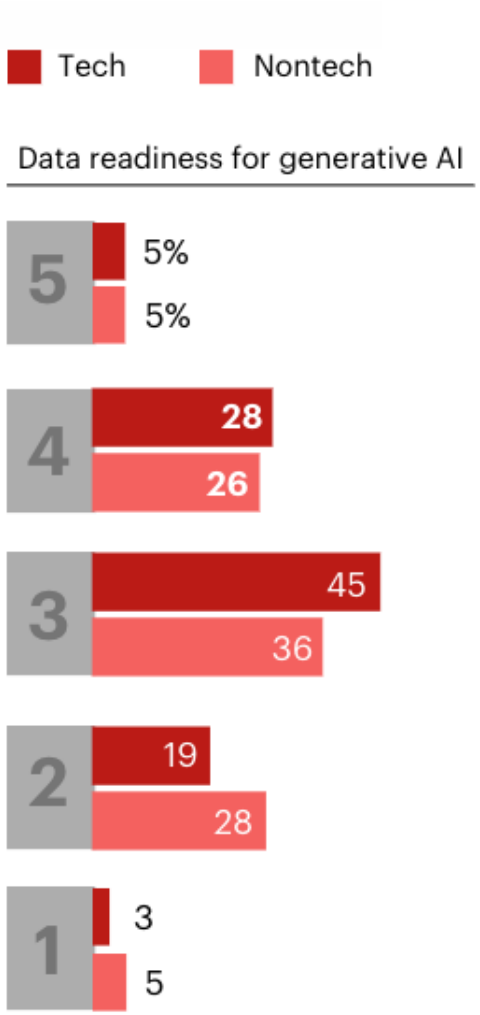
Sources: Bain Generative AI Surveys, October 2023 (N=198) and February 2024 (N=200)

What are the top reasons preventing your company from moving faster with generative AI?



Sources: Bain Generative AI Surveys, October 2023 (N=198) and February 2024 (N=200)

Data Readiness for Generative AI



Structured and unstructured data are centralized, quality controlled and readily available

Structured and unstructured data are available, but both need to be quality controlled and centralized

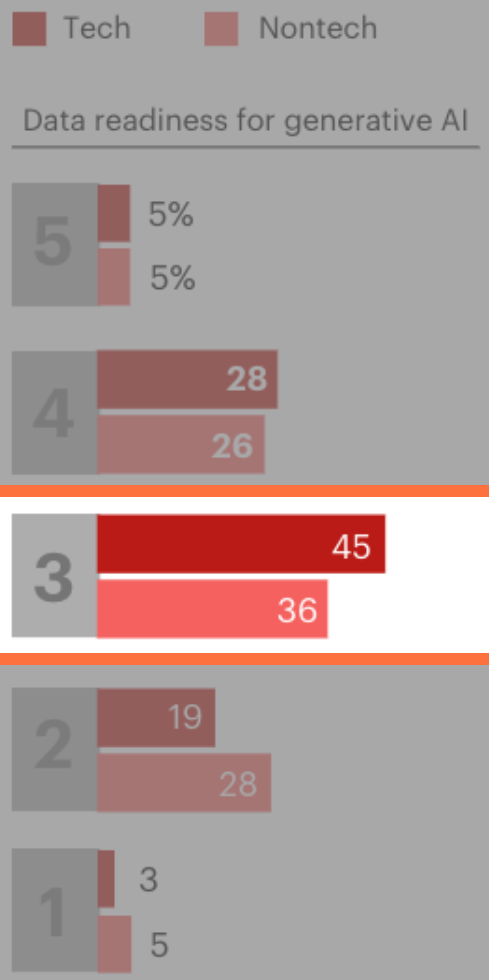
Structured data is readily available, unstructured data takes some time to collect, process and check quality

Structured and unstructured data can be made available, but doing so requires significant effort and enhancements

Structured and unstructured data are not readily available.

Source: Bain Generative Artificial Intelligence Survey, February 2024 (N=200)

Data Readiness for Generative AI



Structured and unstructured data are centralized, quality controlled and readily available

Structured and unstructured data are available, but both need to be quality controlled and centralized

Structured data is readily available, unstructured data takes some time to collect, process and check quality

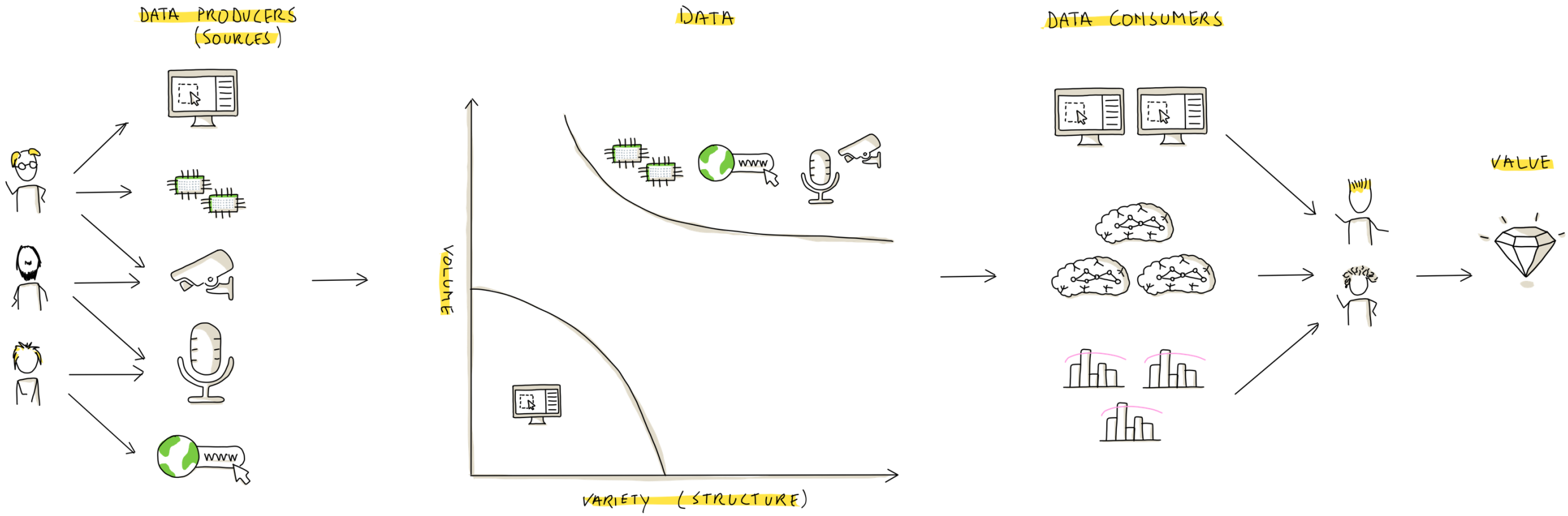
Structured and unstructured data can be made available, but doing so requires significant effort and enhancements

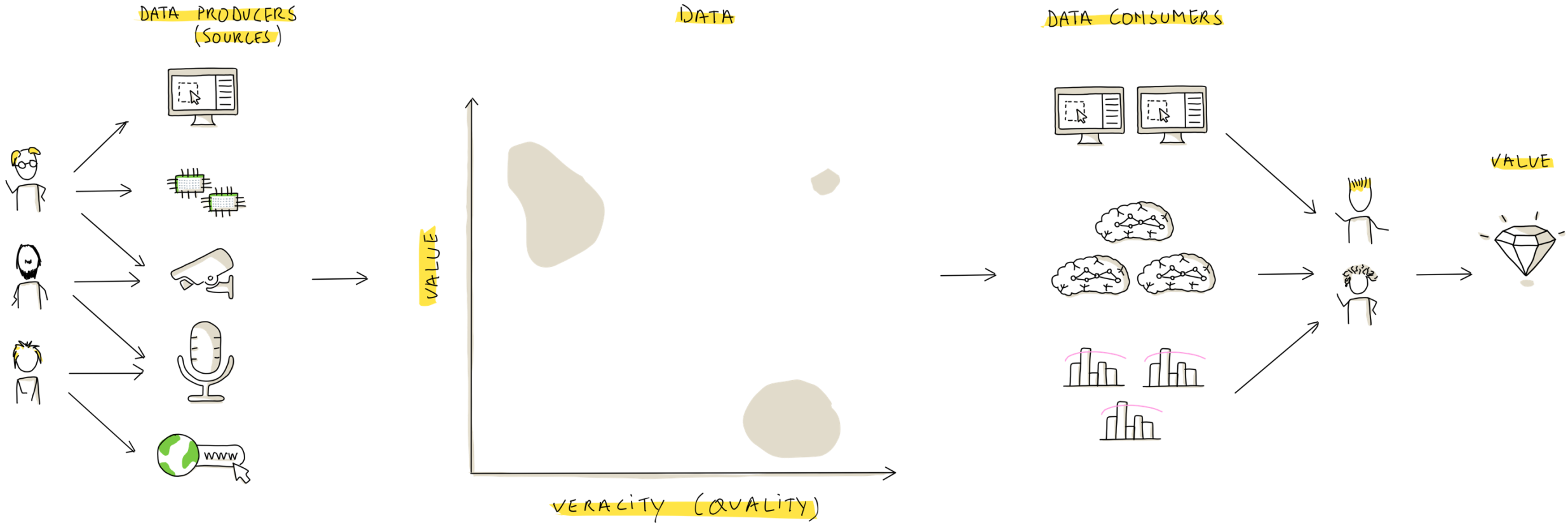
Structured and unstructured data are not readily available.

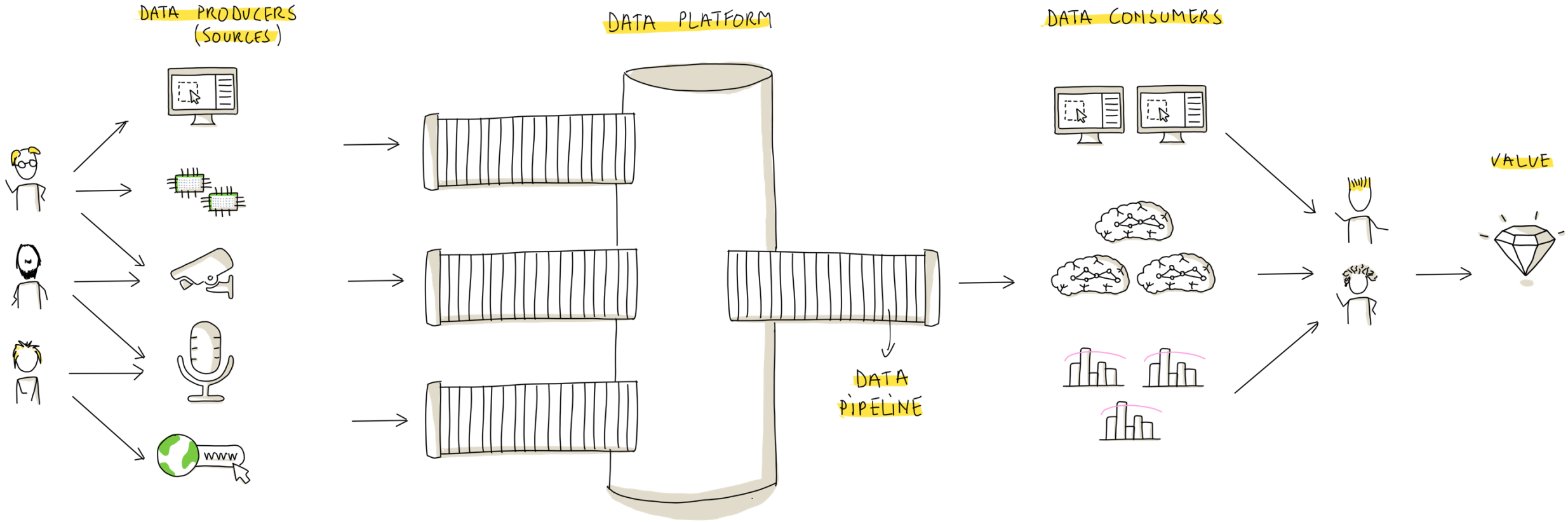
Source: Bain Generative Artificial Intelligence Survey, February 2024 (N=200)

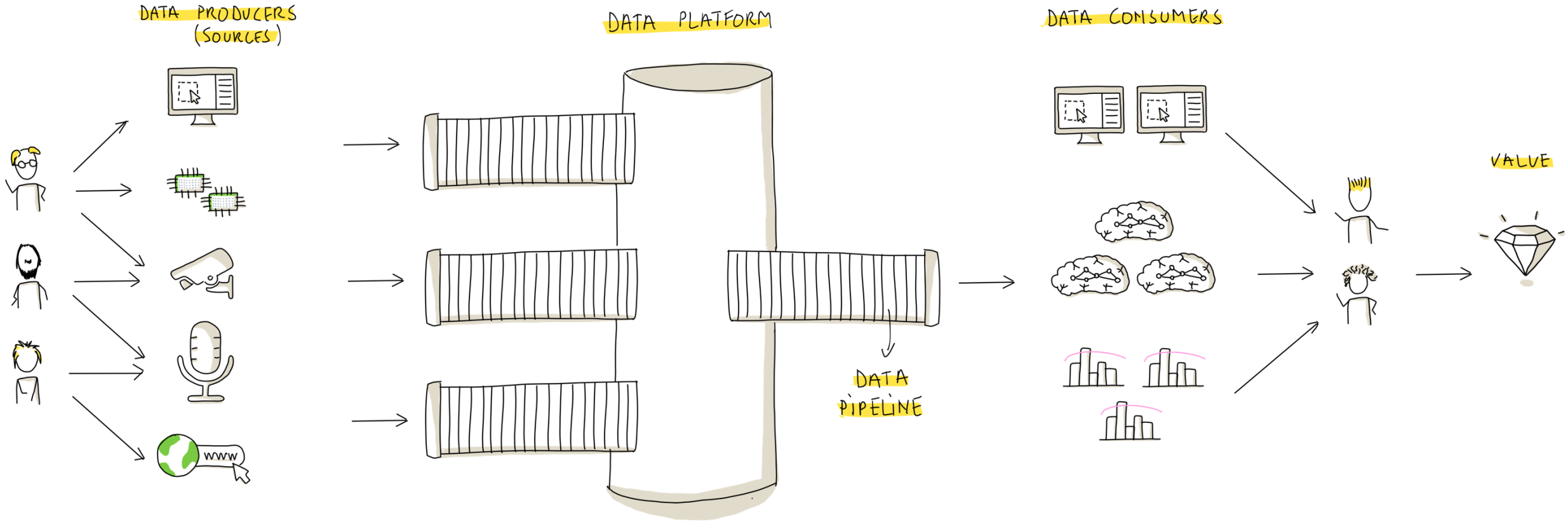
Data Management

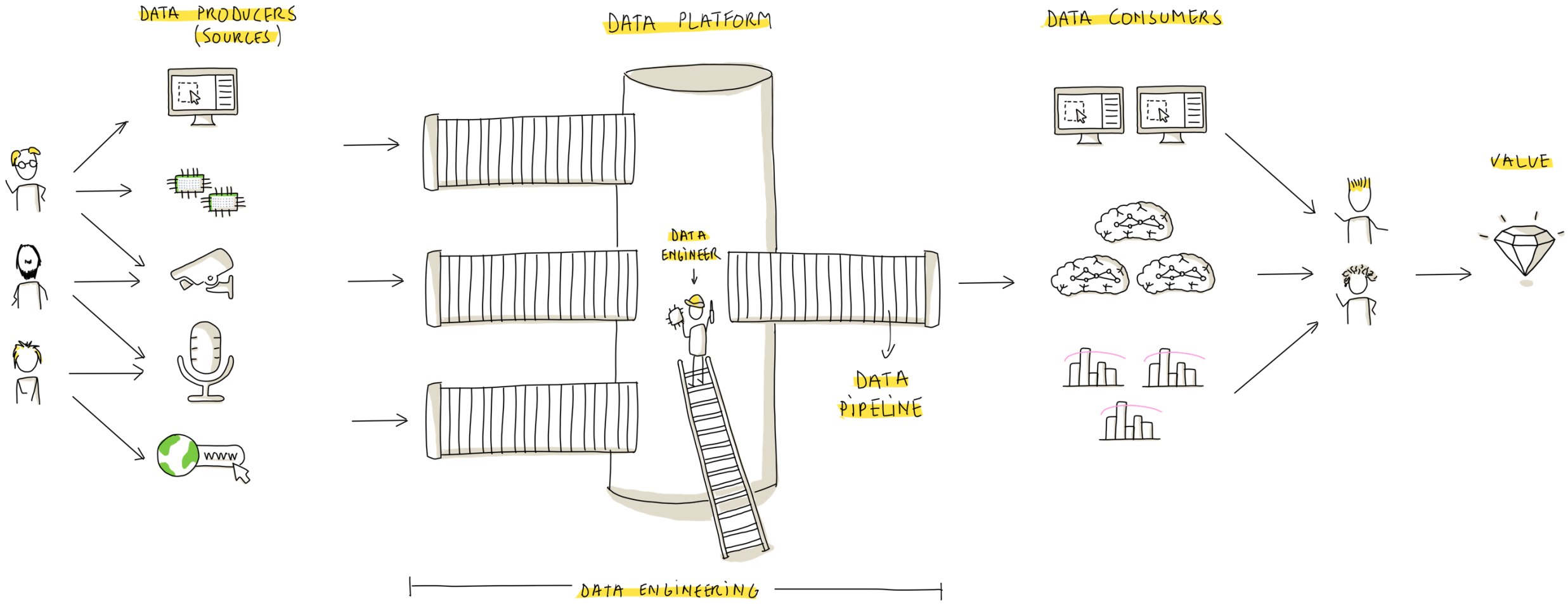
- Defining Data
- Data Producers
- Big Data Vs
- The Bigger Picture
- **Data Management**
- Data Technology

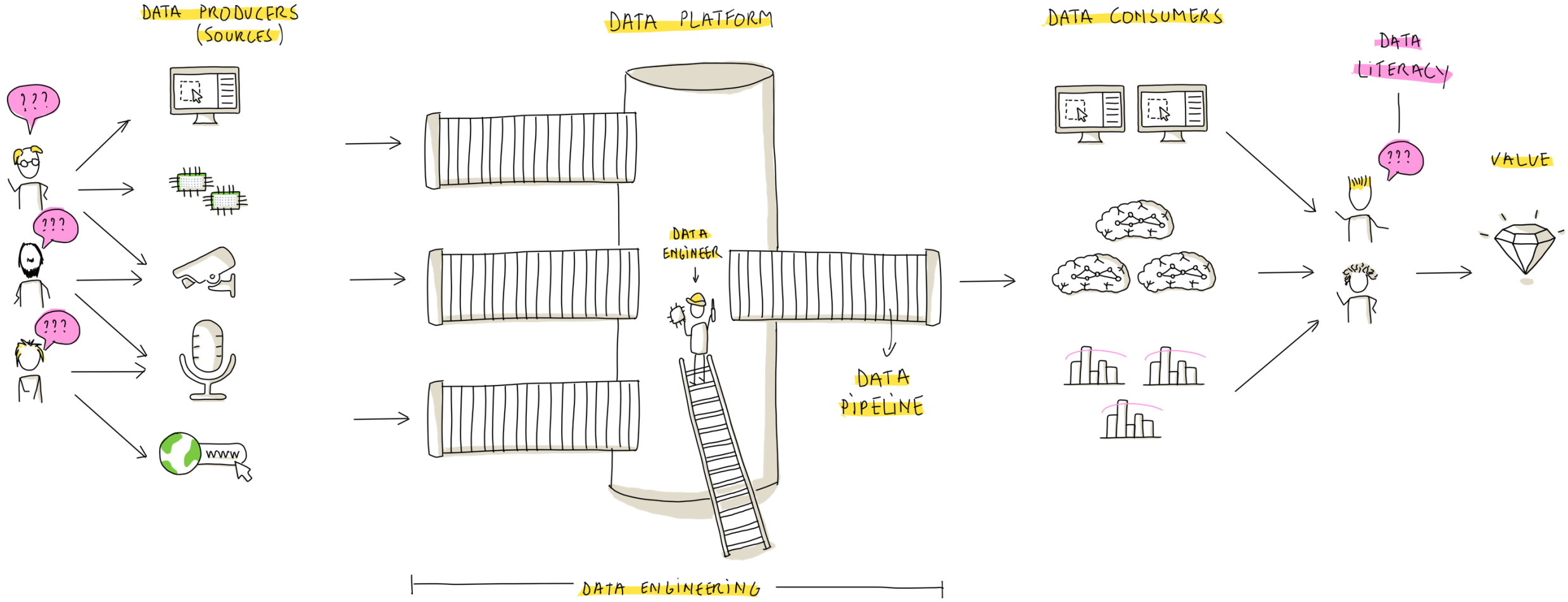

DATA PRODUCERS
(SOURCES)
DATA
VOLUME
VARIETY (STRUCTURE)
DATA CONSUMERS
VALUE
www
www

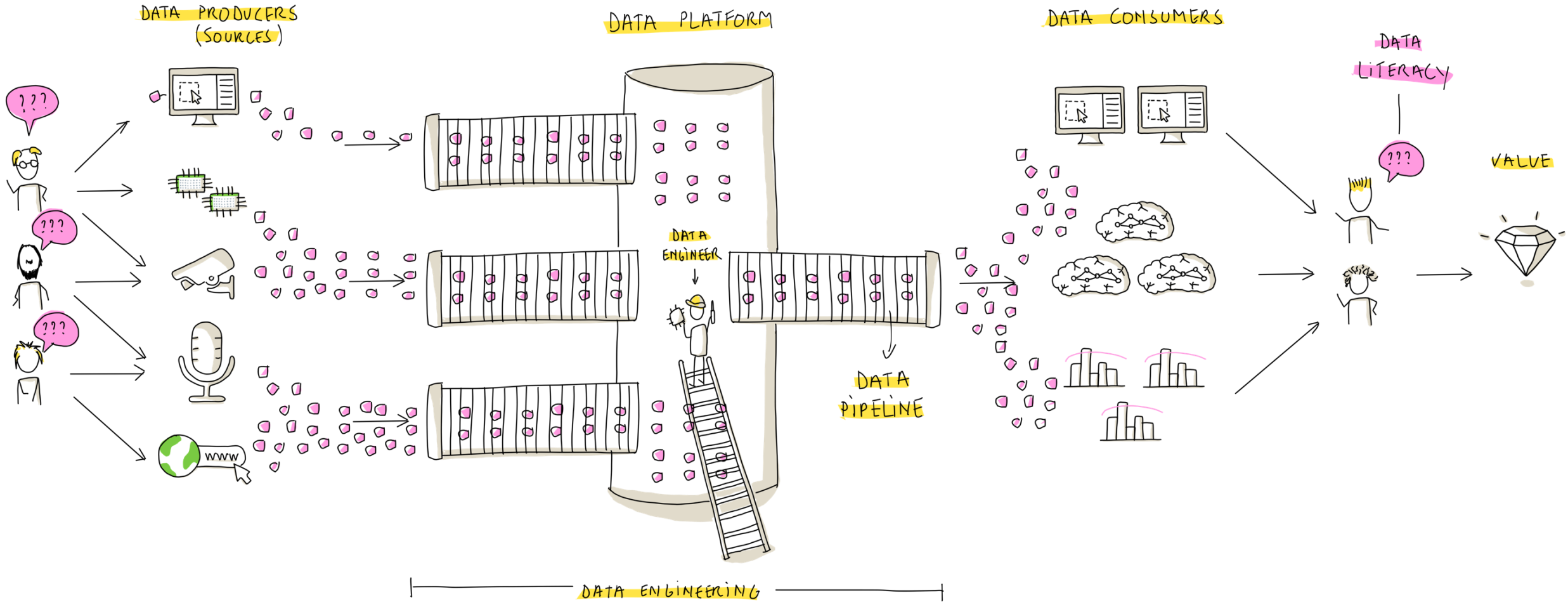

DATA PRODUCERS
(SOURCES)
DATA
VALUE
VERACITY (QUALITY)
DATA CONSUMERS
VALUE
www

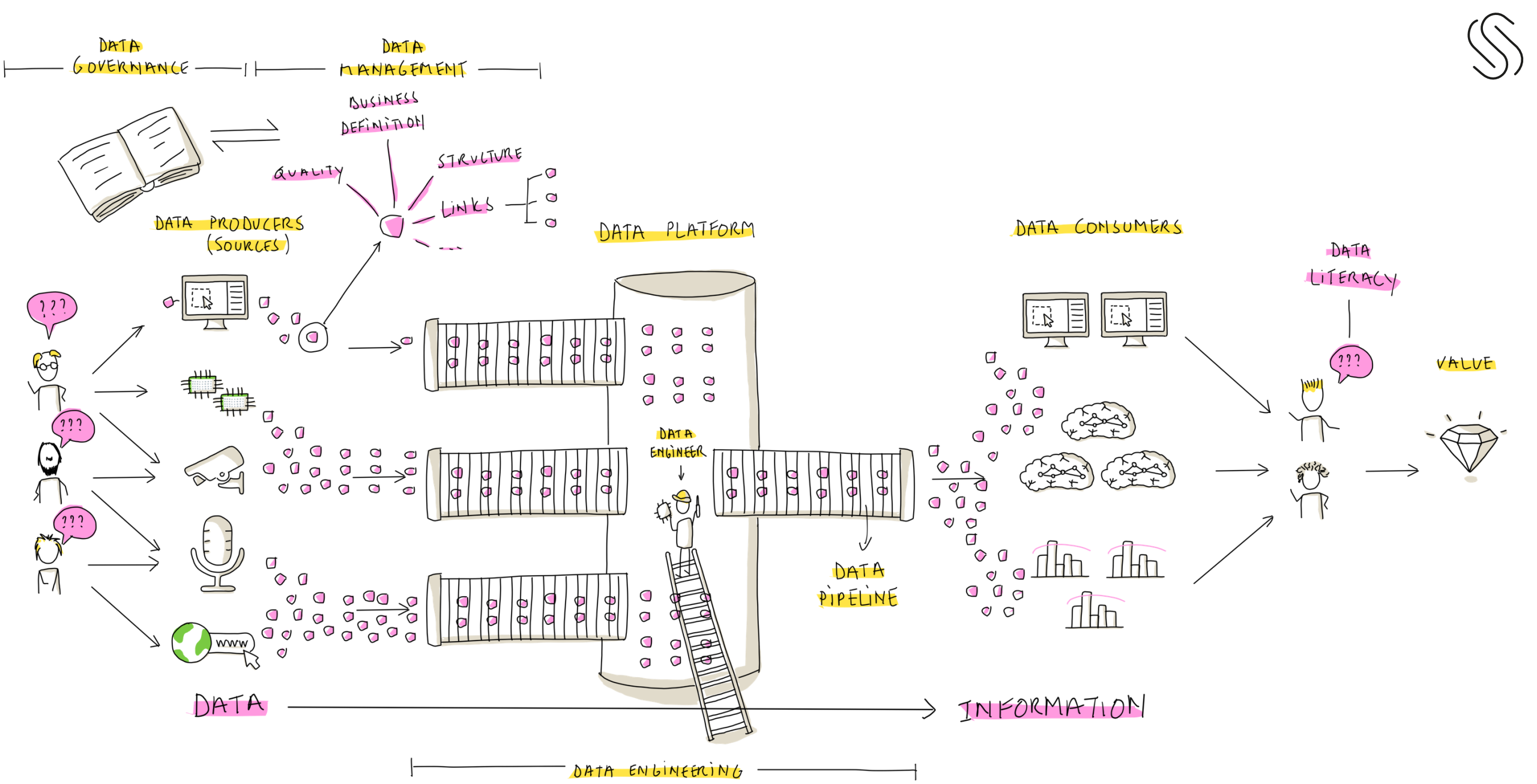

DATA PRODUCERS
(SOURCES)
DATA PLATFORM
DATA CONSUMERS
VALUE
DATA
PIPELINE
www


DATA PRODUCERS
(SOURCES)
DATA PLATFORM
DATA CONSUMERS
VALUE
DATA
PIPELINE
www


DATA PRODUCERS
(SOURCES)
DATA PLATFORM
DATA CONSUMERS
VALUE
DATA ENGINEER
DATA PIPELINE
www
DATA ENGINEERING


DATA PRODUCERS
(SOURCES)
DATA PLATFORM
DATA CONSUMERS
DATA LITERACY
VALUE
DATA ENGINEER
DATA PIPELINE
DATA ENGINEERING
???
???
???
???
www


DATA PRODUCERS
(SOURCES)
DATA PLATFORM
DATA CONSUMERS
DATA LITERACY
???
???
???
???
VALUE
DATA ENGINEER
DATA PIPELINE
www
DATA ENGINEERING


DATA GOVERNANCE
DATA MANAGEMENT
BUSINESS DEFINITION
QUALITY
STRUCTURE
LINKS
...
DATA PRODUCERS (SOURCES)
DATA PLATFORM
DATA CONSUMERS
DATA LITERACY
???
VALUE
DATA ENGINEER
DATA PIPELINE
www
DATA
INFORMATION
DATA ENGINEERING

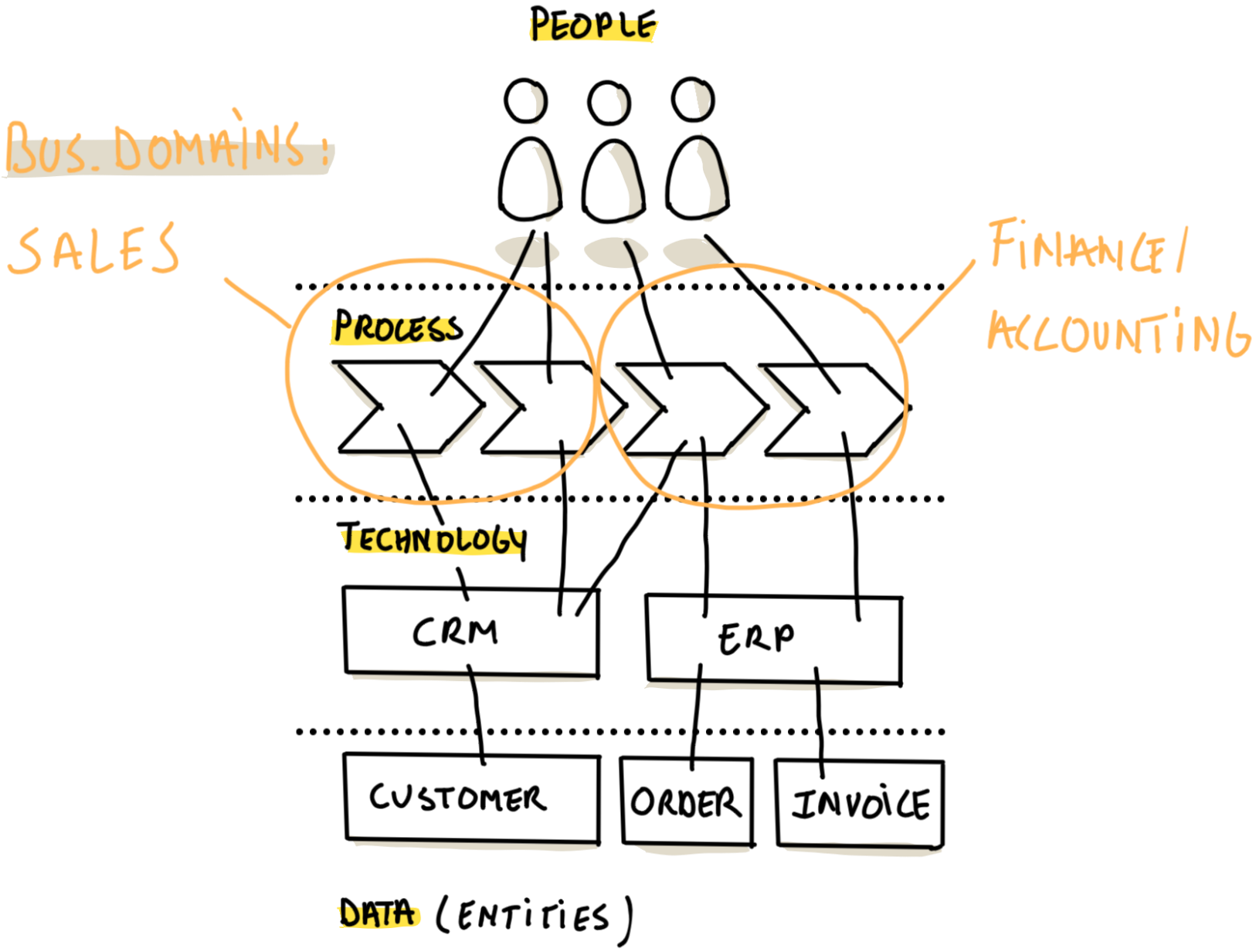
There is **Room for Improvement...**

DMBOK – DAta MAnagement WHEEL

Data Governance defines the rules of play, Data Management executes these rules

Data Architecture

Which data do we have and how does it flow through the organization?


PEOPLE
BUS. DOMAINS:
SALES
FINANCE / ACCOUNTING
PROCESS
TECHNOLOGY
CRM
ERP
CUSTOMER
ORDER
INVOICE
DATA (ENTITIES)

Business Domain

A specific **area of expertise or knowledge within a company** that focuses on particular business functions or industries. For example, "Finance" or "Human Resources" are business domains. It defines the scope of responsibilities, rules, and activities in a specific field.

Business Process

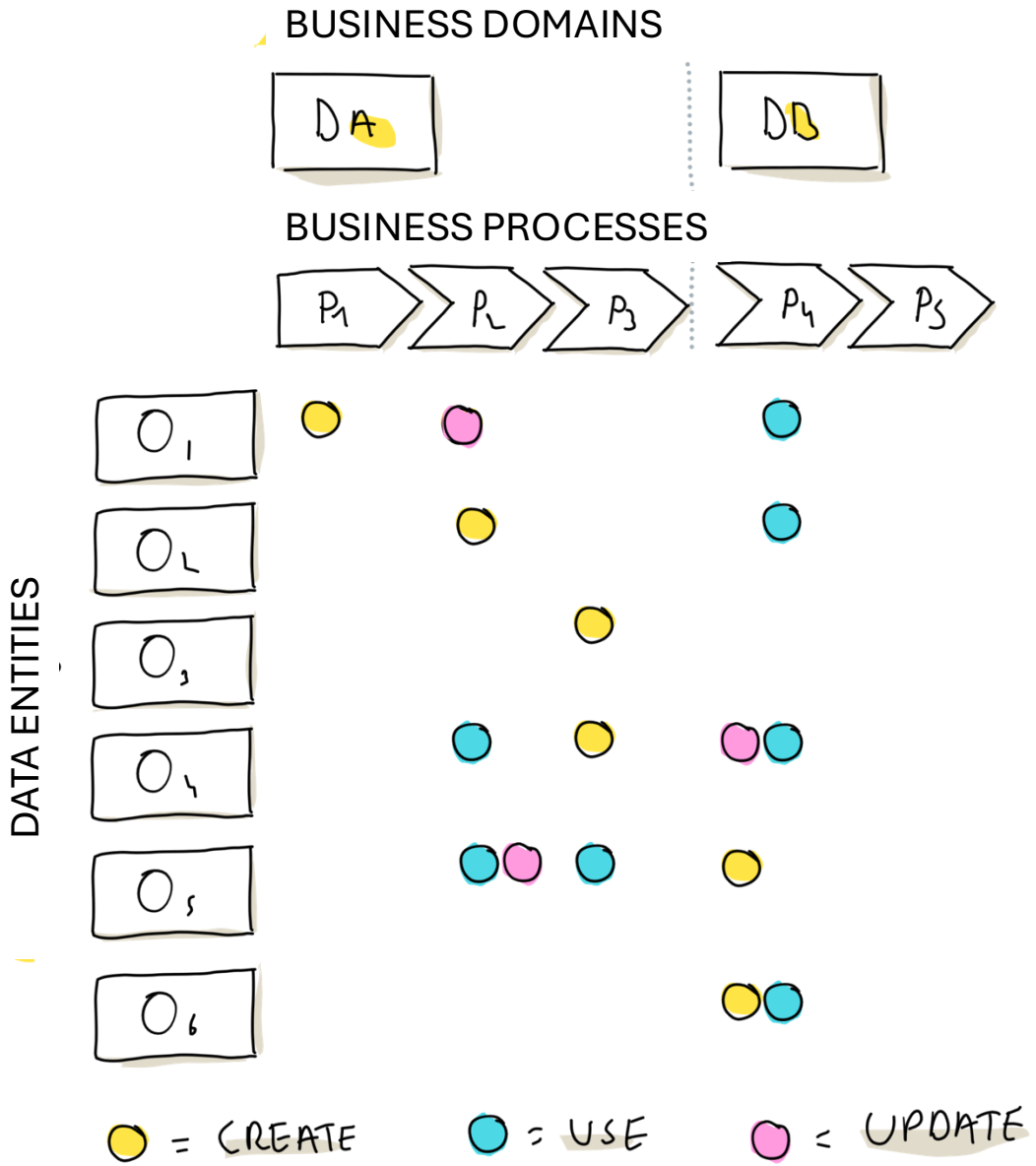
A series of tasks or steps carried out to achieve a specific business goal. **It's how work gets done within an organization**. For example, the **order-to-cash process** includes all the steps from receiving an order to collecting payment.

Entities

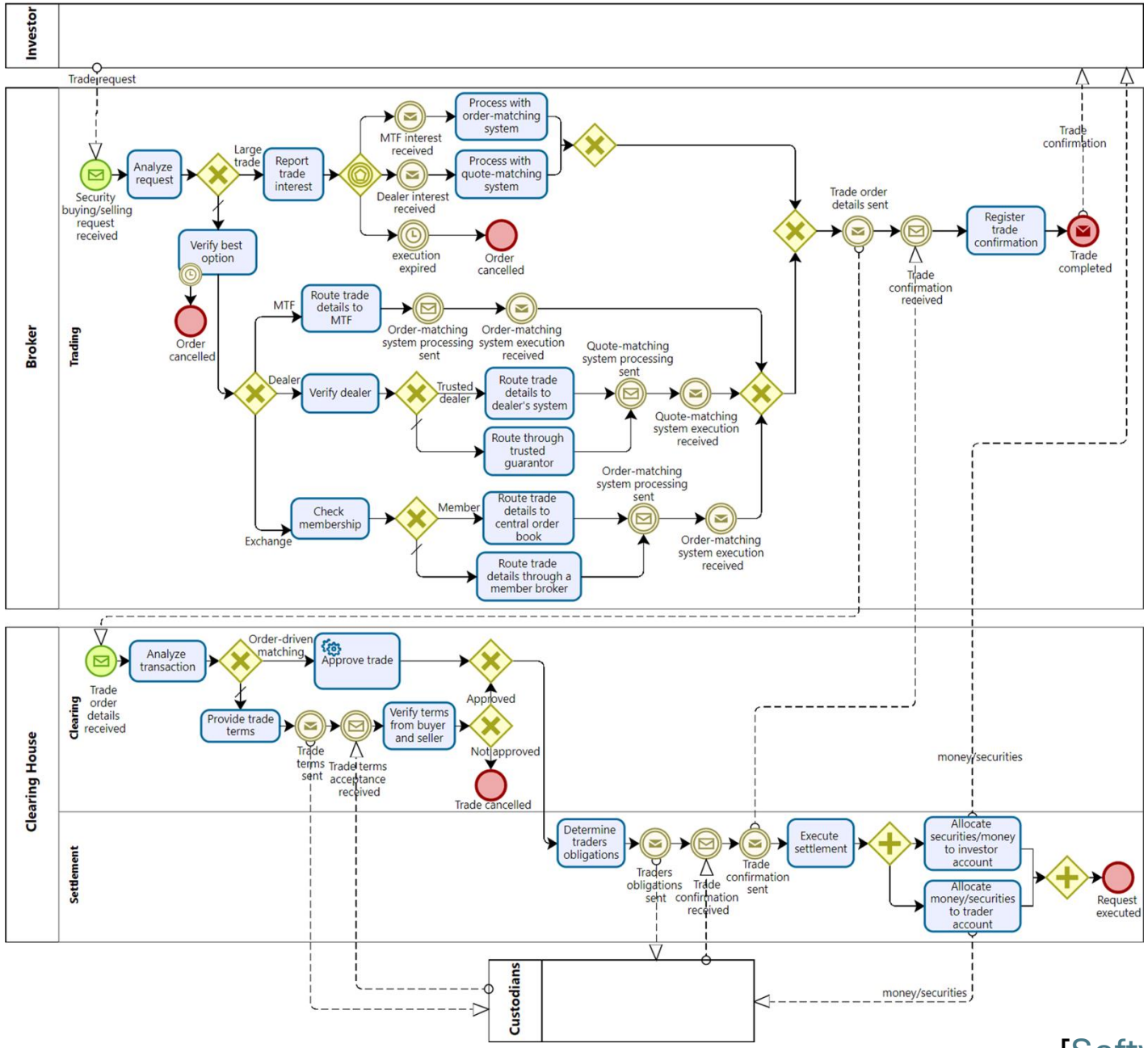
An entity is **a conceptual representation of something relevant to a business**, usually a tangible or abstract entity, **that groups data**. Examples:

- **Customer**, grouping data like name, address, and purchase history
- **Product**, grouping data like product ID, price, and stock level.

Domains, Processes and Entities



"Process P1 (part of Domain **DA**) creates Entity O1. This entity is updated in Process P2 and used in Process P4."

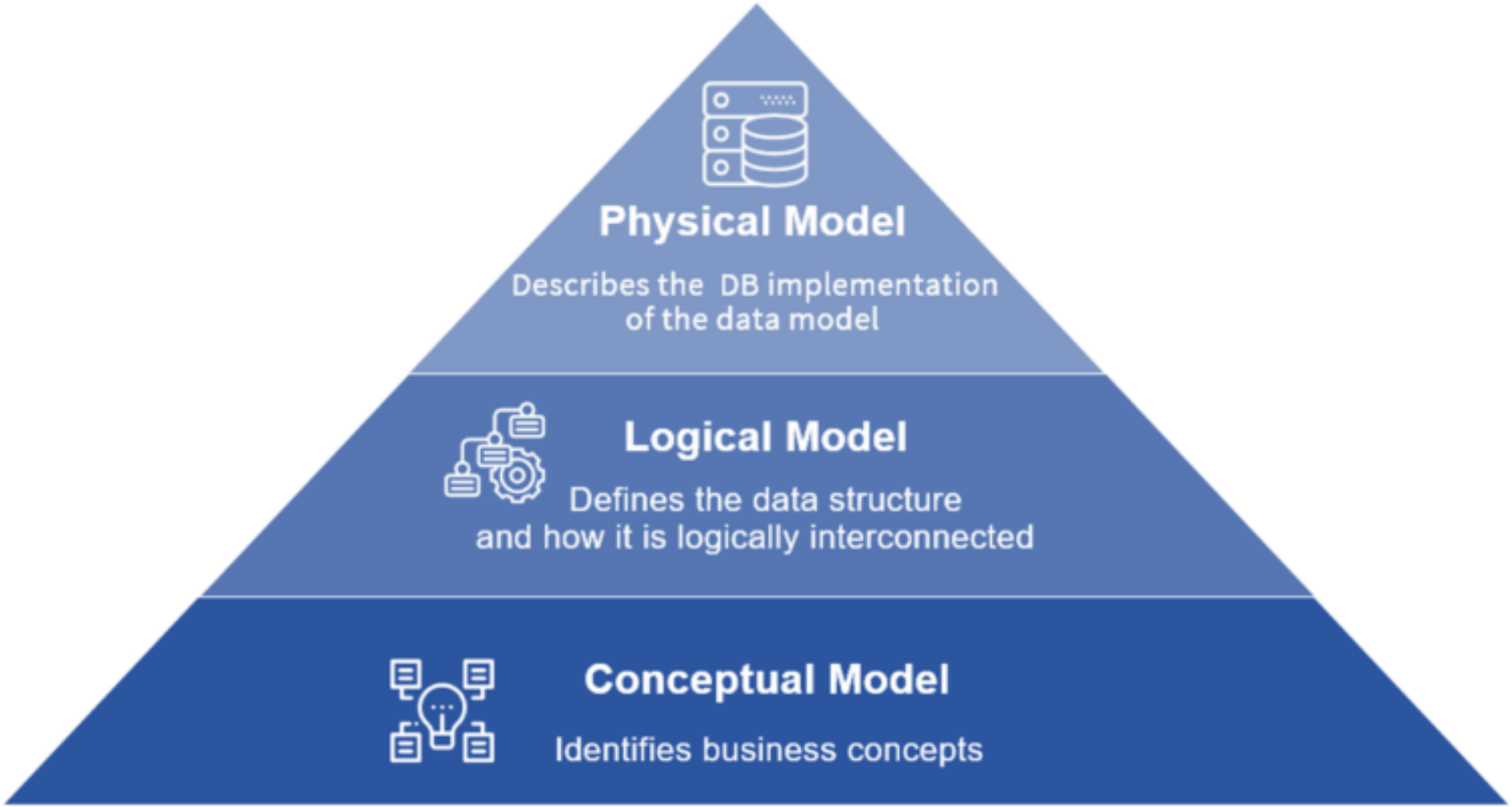


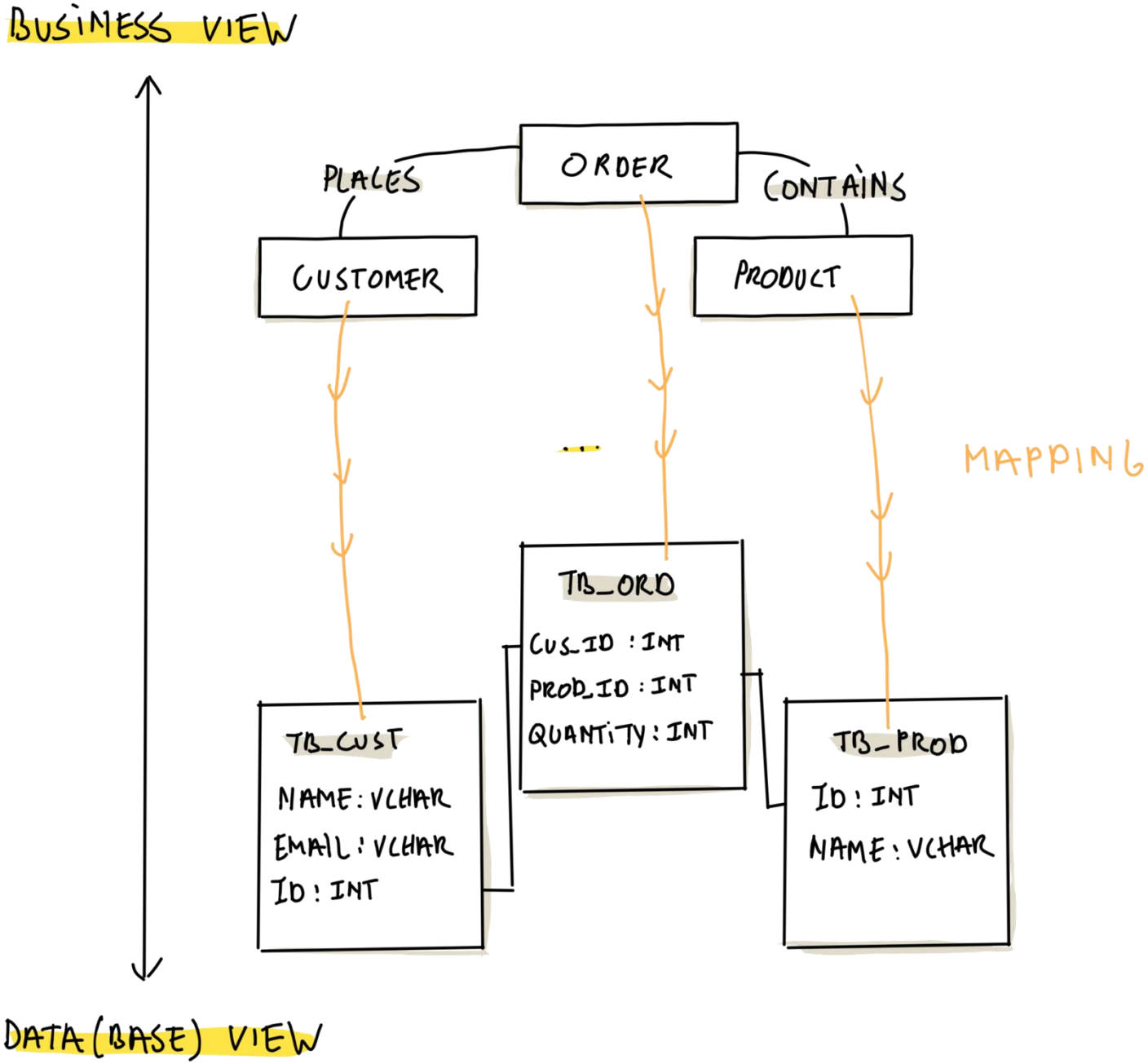
[Software & Systemd Modeling]

Data Modeling & Design

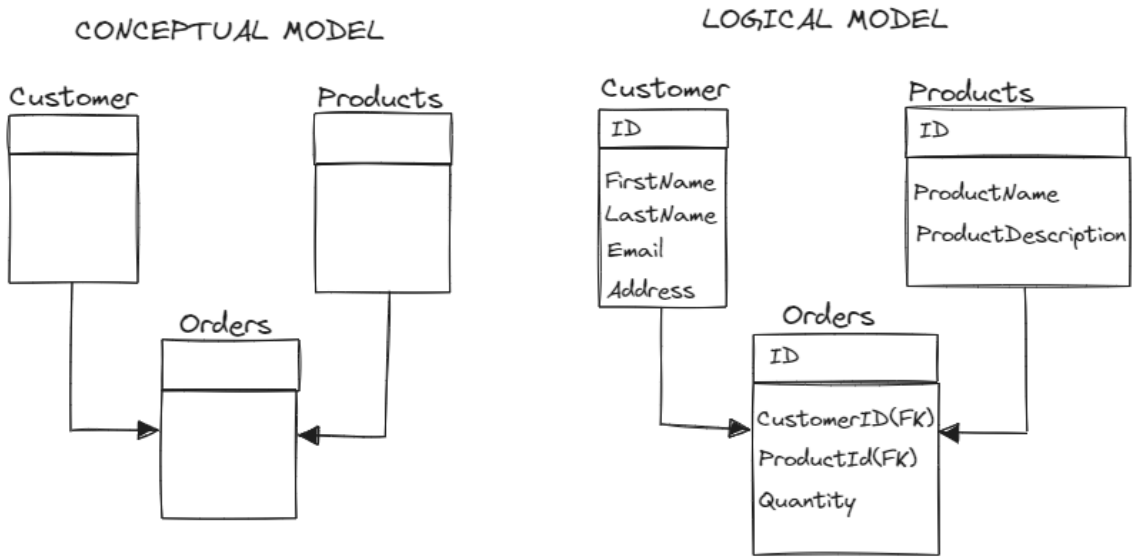
Providing a Common Vocabulary around Data – Often communicated in a ‘Data Model’

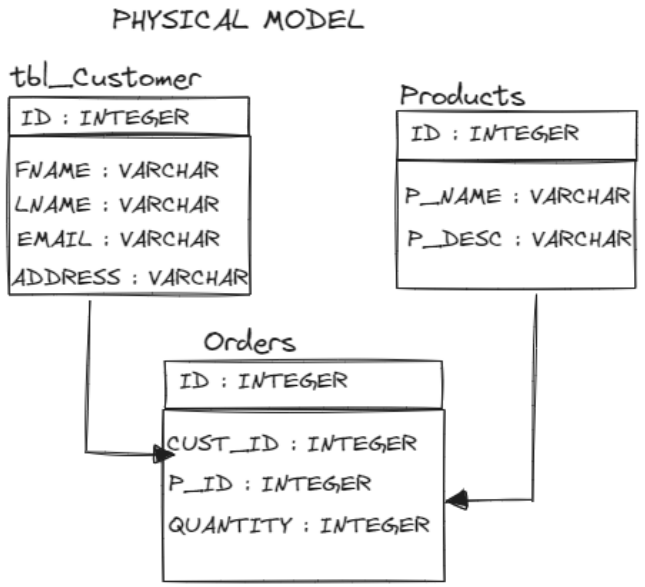
Physical, Logical & Conceptual Model

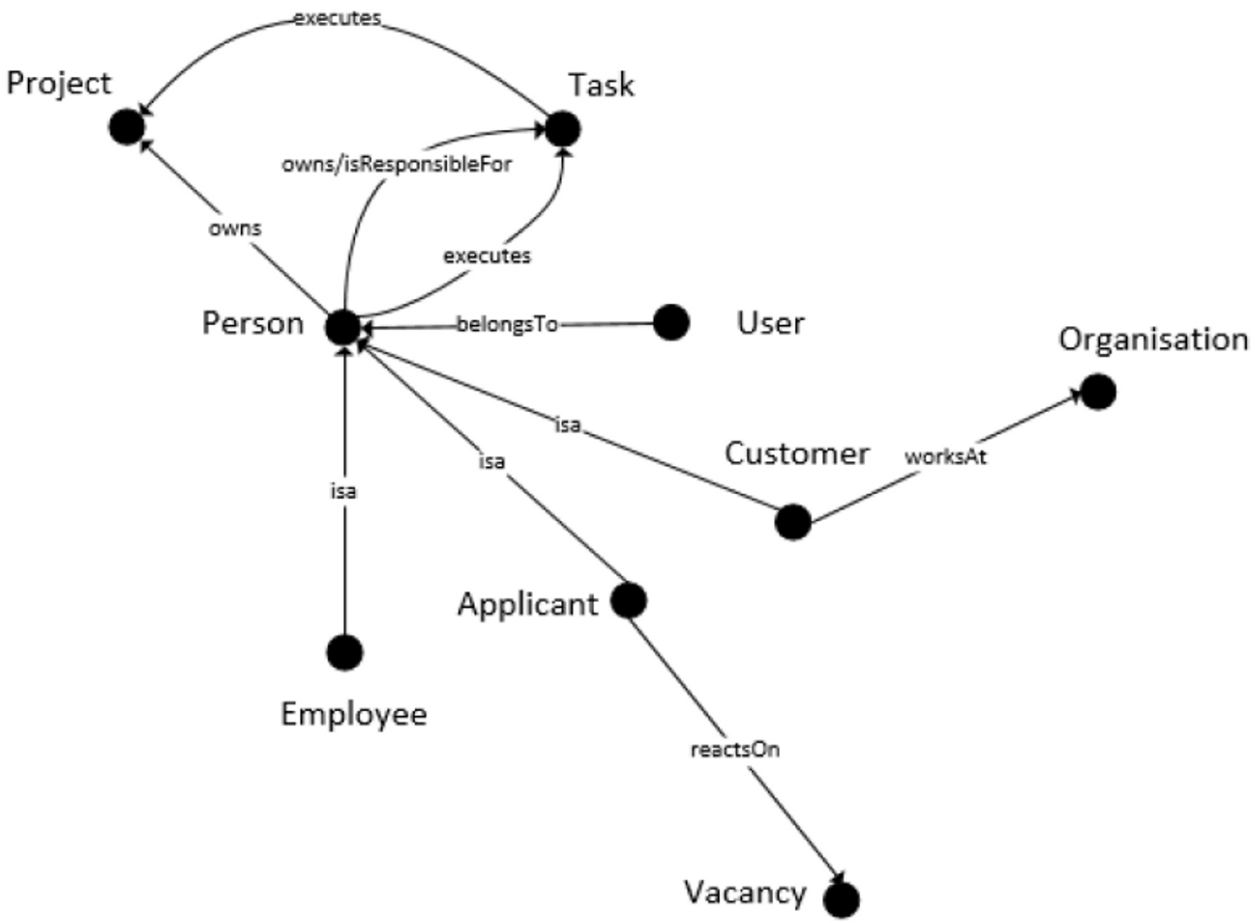



BUSINESS VIEW
ORDER
PLACES
CONTAINS
CUSTOMER
PRODUCT
MAPPING
TB_ORD
CUS_ID : INT
PROD_ID : INT
QUANTITY : INT
TB_CUST
NAME : VCHAR
EMAIL : VCHAR
ID : INT
TB_PROD
ID : INT
NAME : VCHAR
DATA (BASE) VIEW

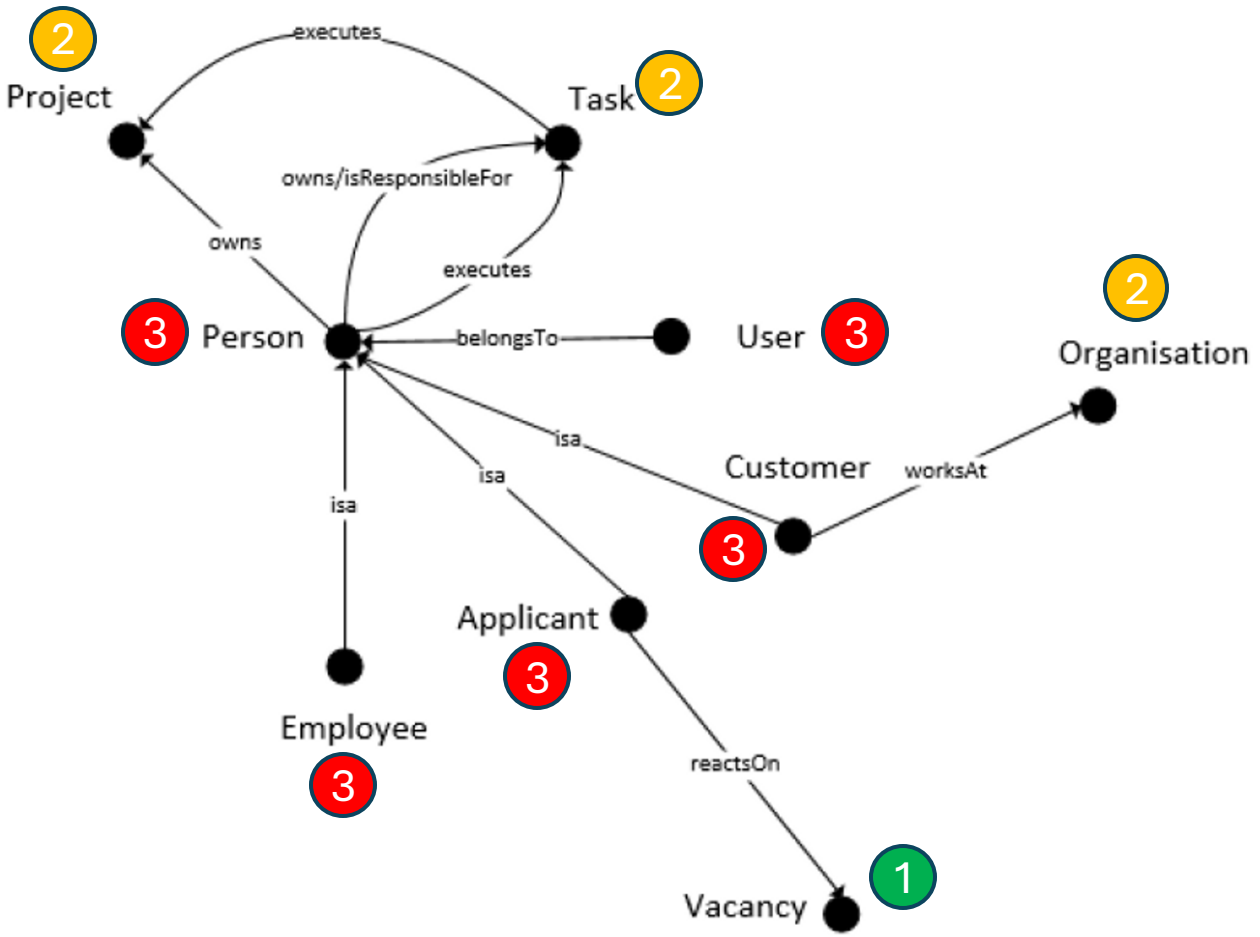
Physical, Logical & Conceptual Model







[What is a Semantic Data Model?]



1 Public data

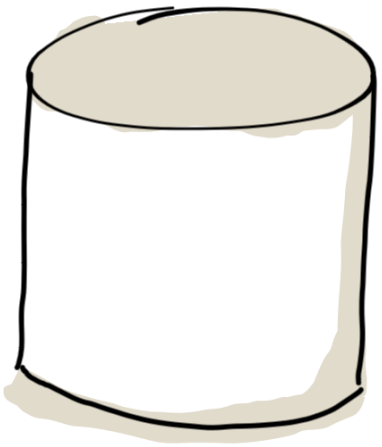
2 Internal

3 Sensitive

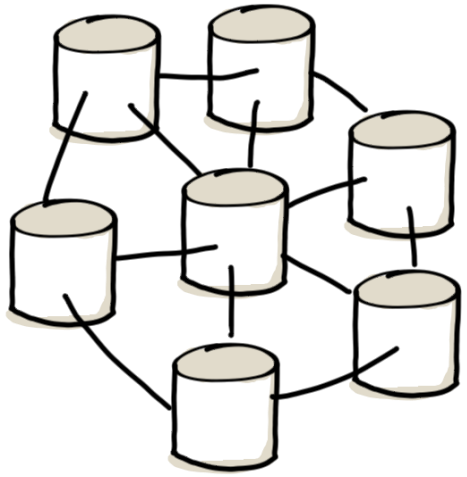
Classification: Adding Sensitivity Information!

Data Storage & Operations

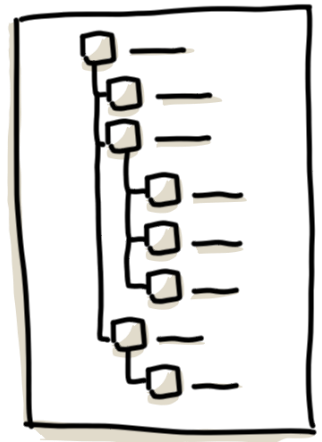
How do we store data for analytical use?



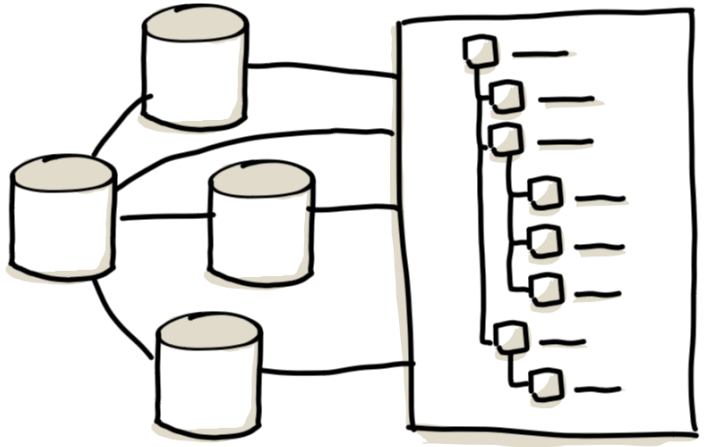
DATABASE



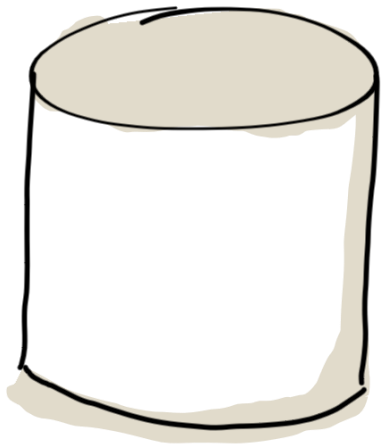
ANALYTICAL DATABASE



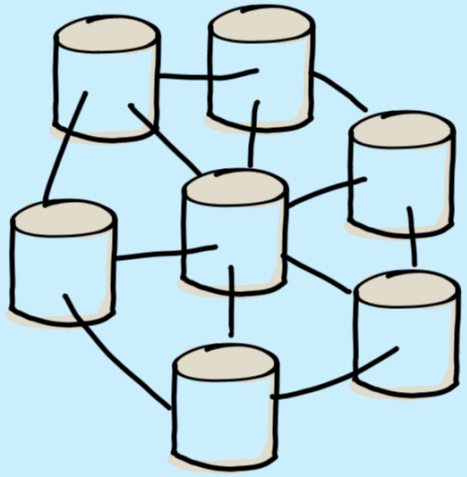
DATA LAKE



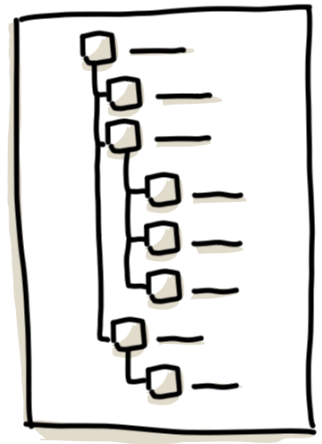
DATA LAKEHOUSE



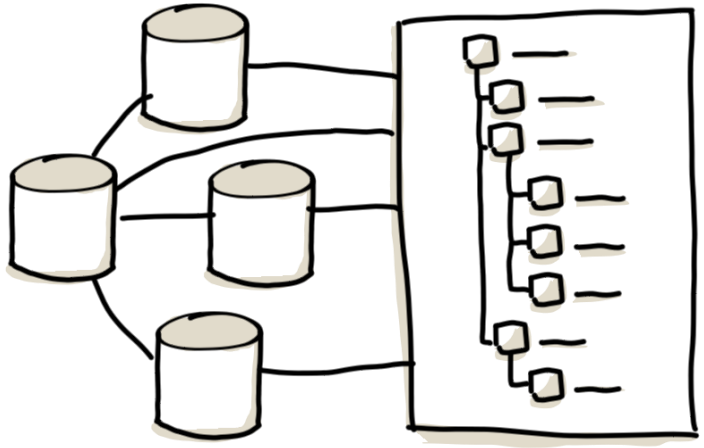
DATABASE



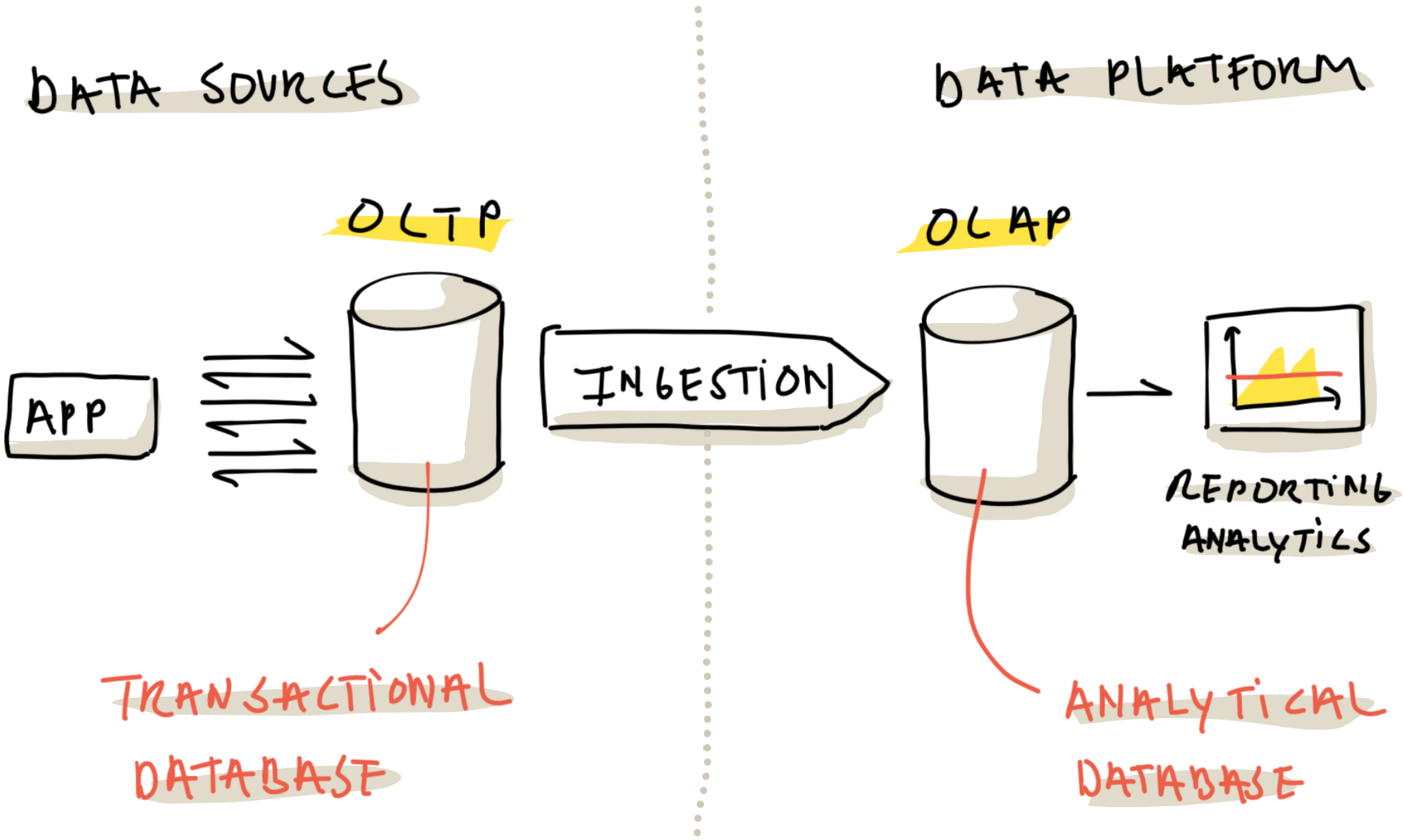
ANALYTICAL DATABASE



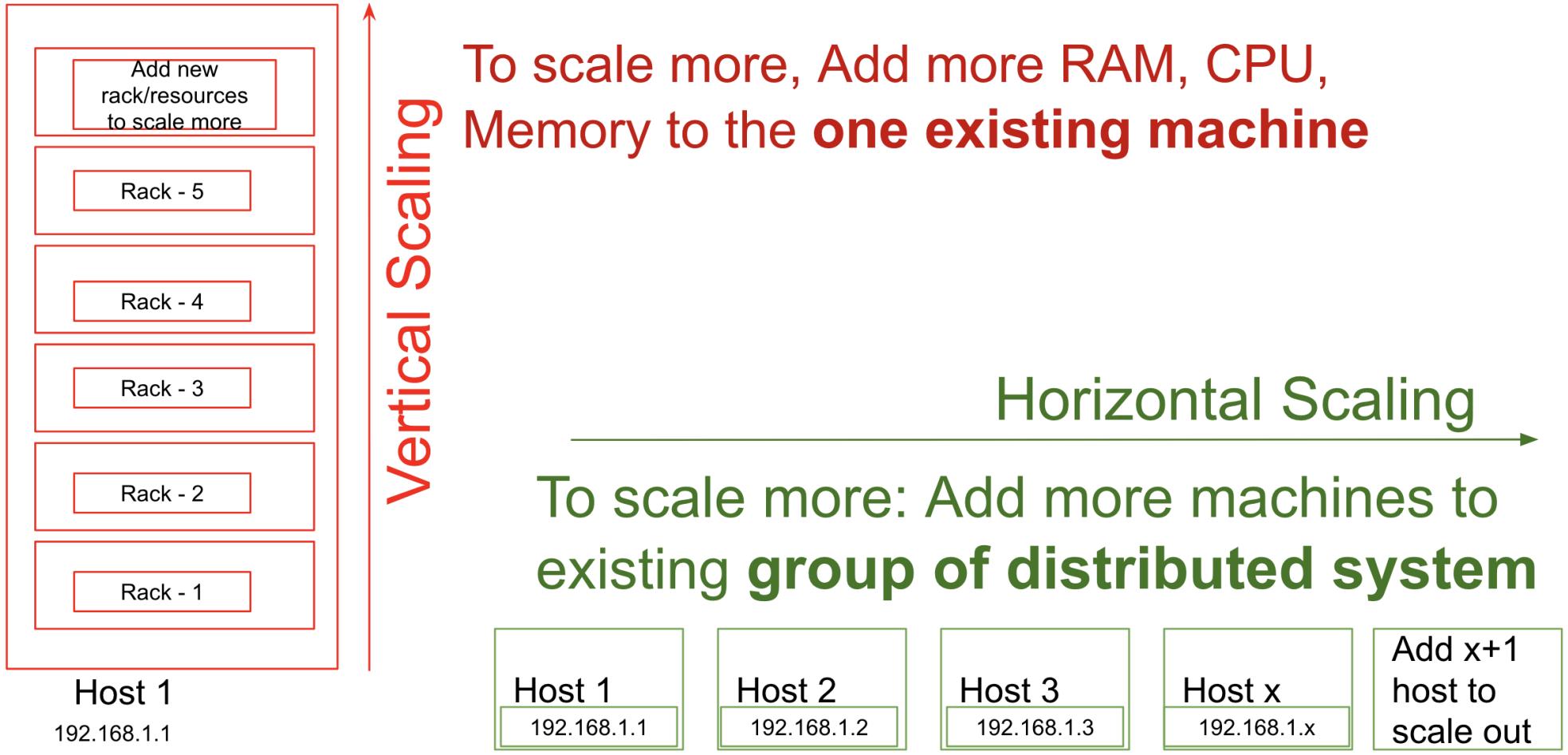
DATA LAKE



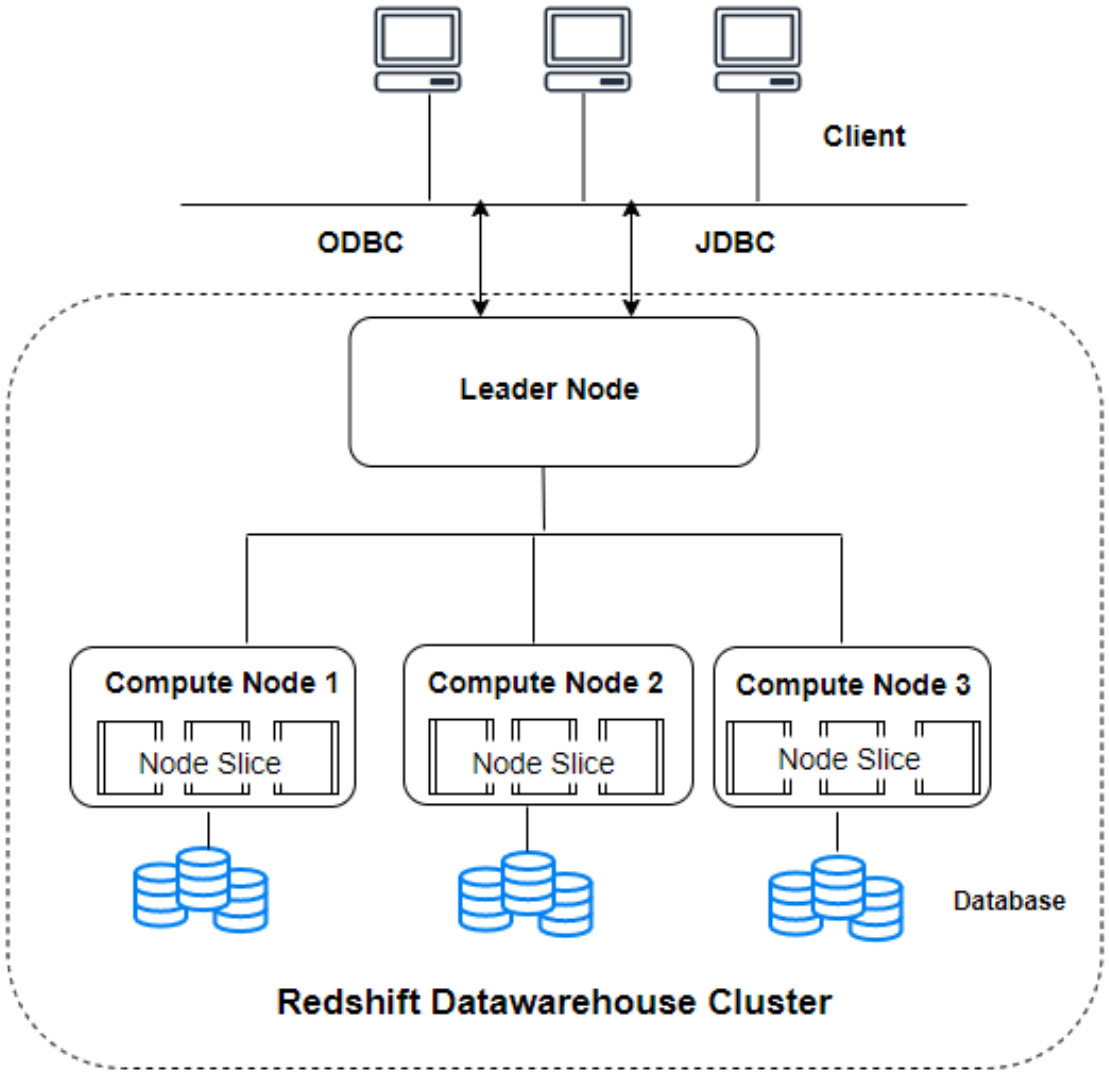
DATA LAKEHOUSE


DATA SOURCES
DATA PLATFORM
OLTP
OLAP
APP
INGESTION
REPORTING
ANALYTICS
TRANSACTIONAL
DATABASE
ANALYTICAL
DATABASE

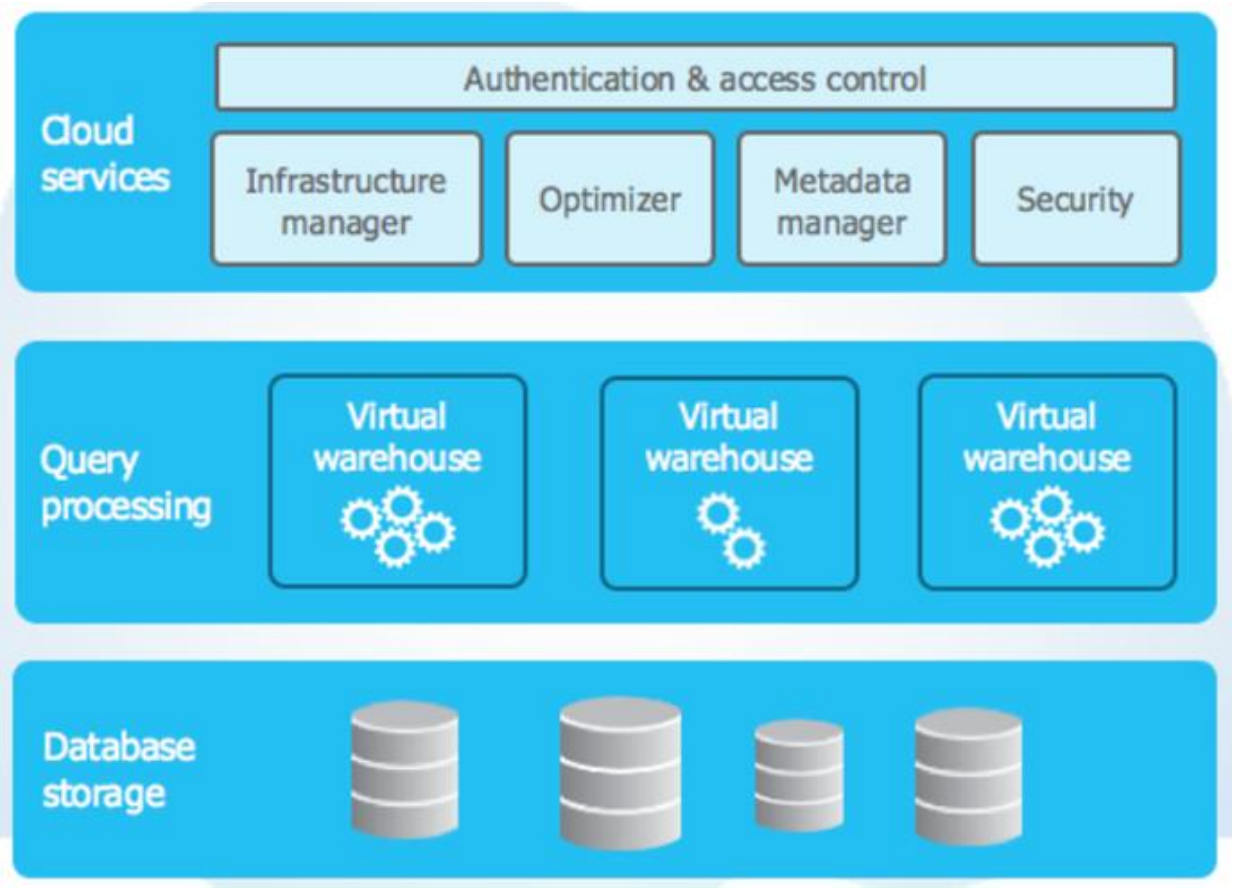
A DWH Database (often called ‘Cloud DWH’) is tuned for **horizontal scaling**.



Example: Amazon Redshift



Example: Snowflake

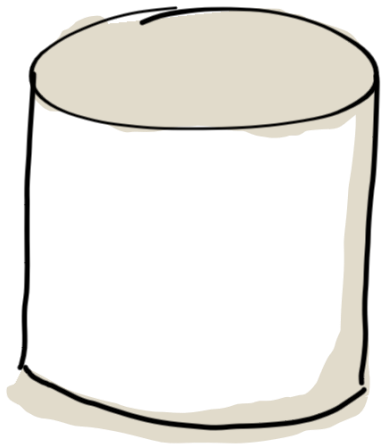



snowflake

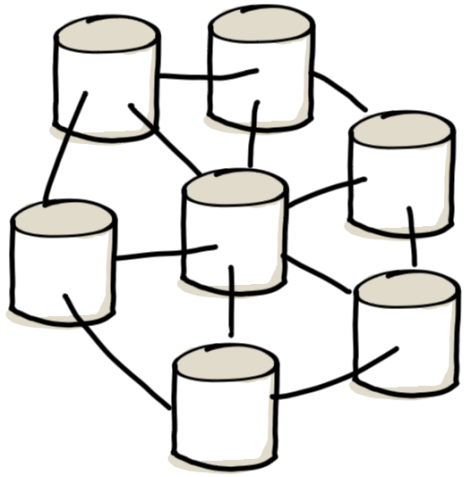

amazon
REDSHIFT


databricks

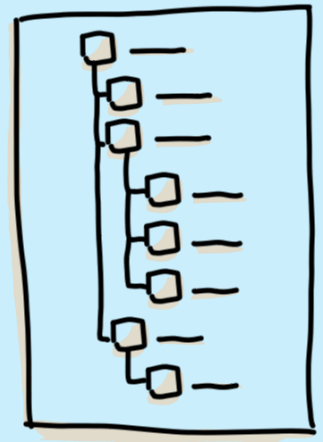

BigQuery



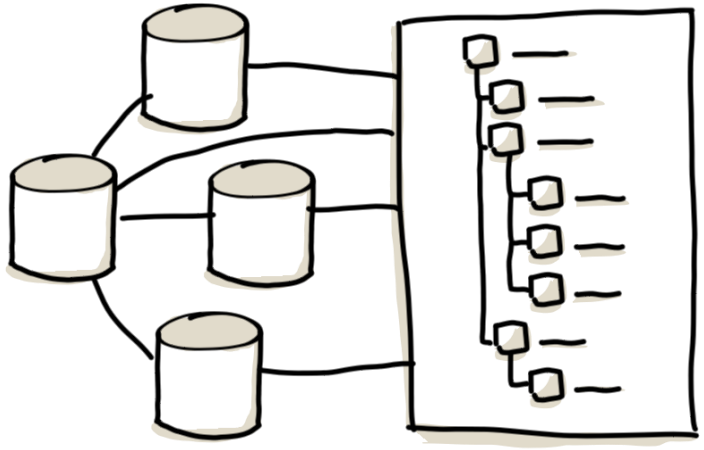
DATABASE



ANALYTICAL DATABASE

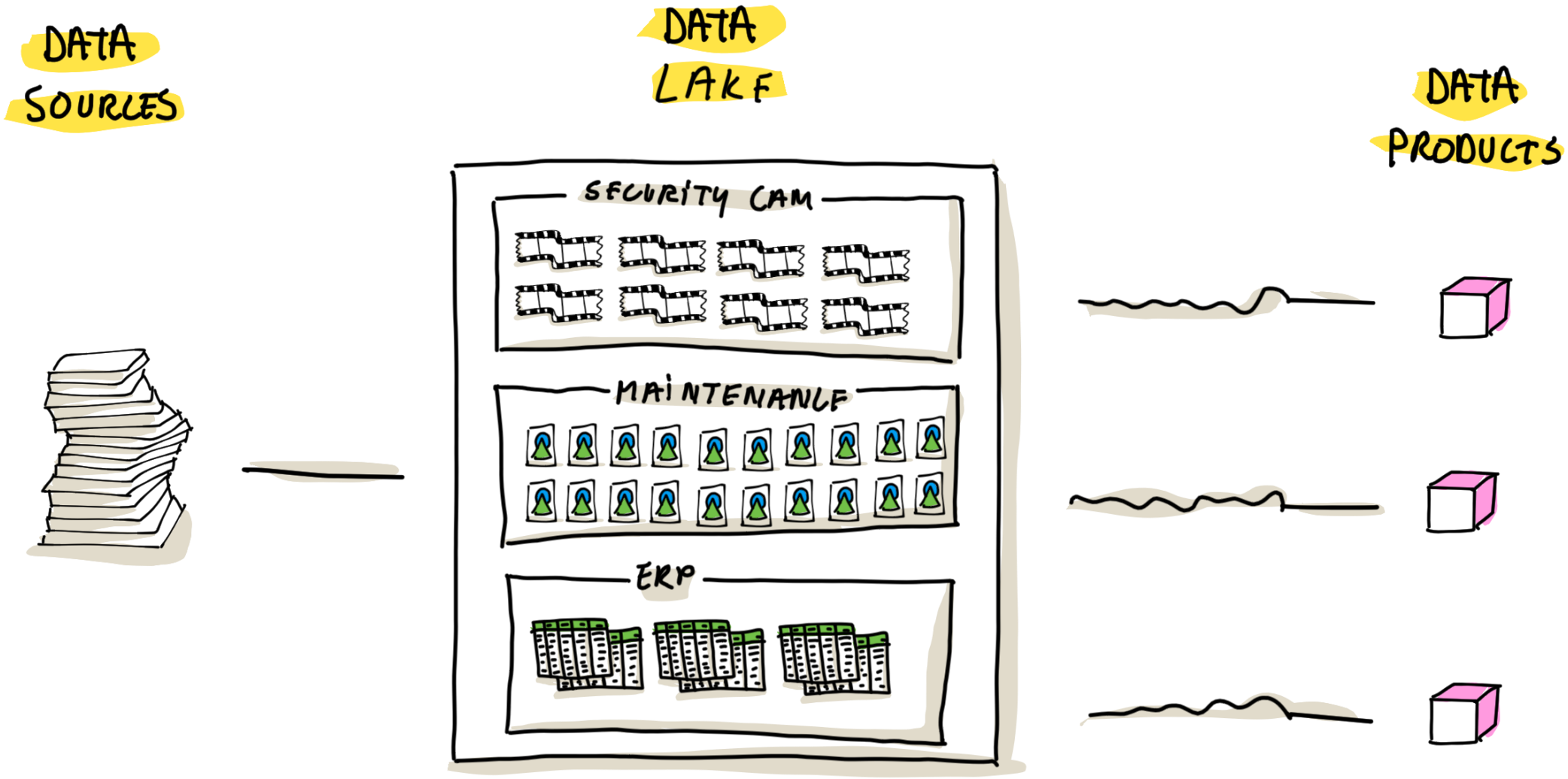


DATA LAKE

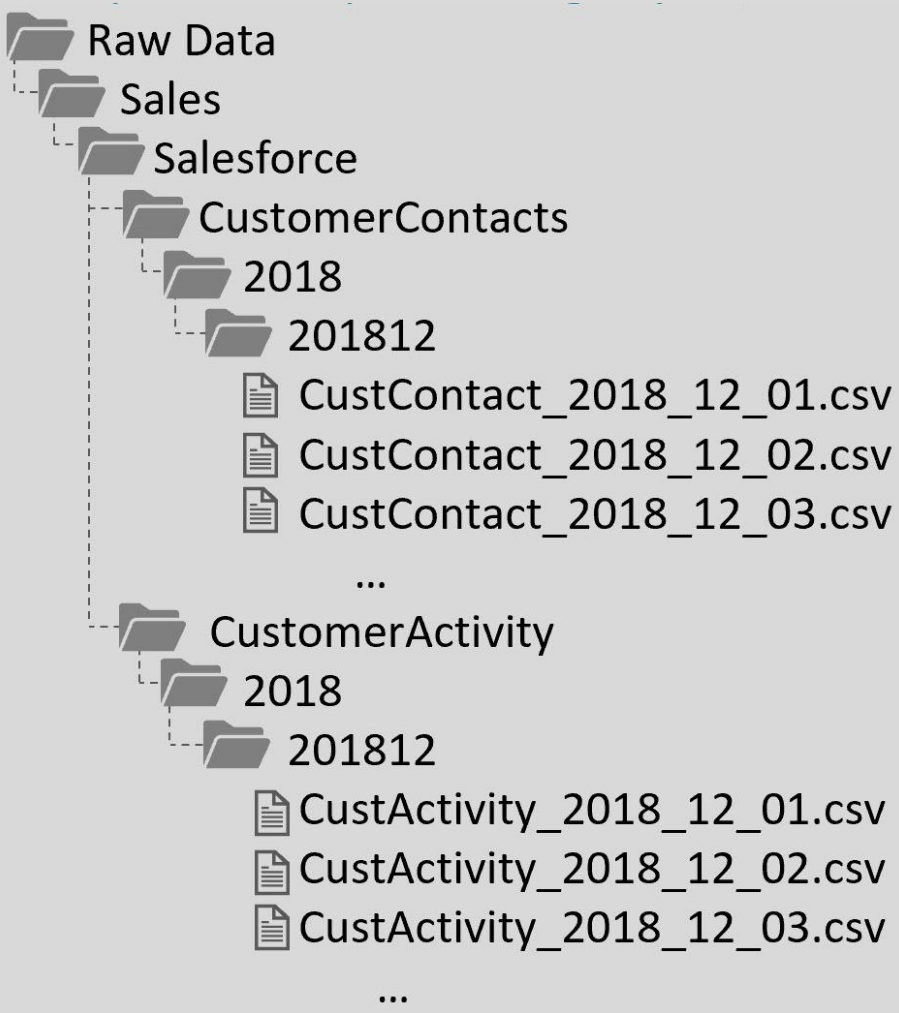


DATA LAKEHOUSE

Data Lake

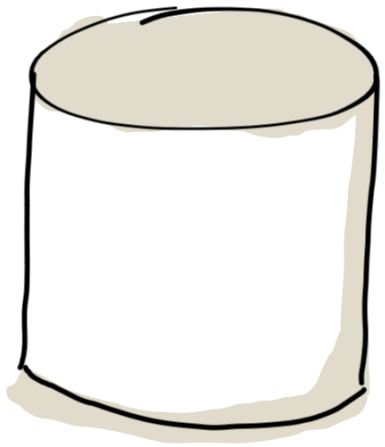

DATA SOURCES
DATA LAKE
DATA PRODUCTS
SECURITY CAM
MAINTENANCE
ERP

Data Lake: A **Structured** File Repository

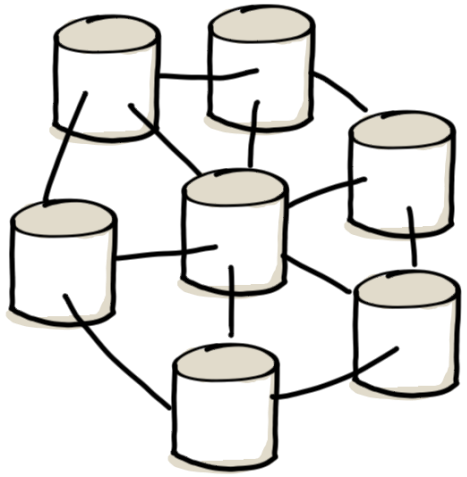


DELTA LAKE

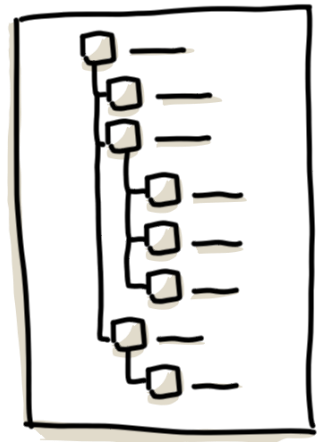
Parquet



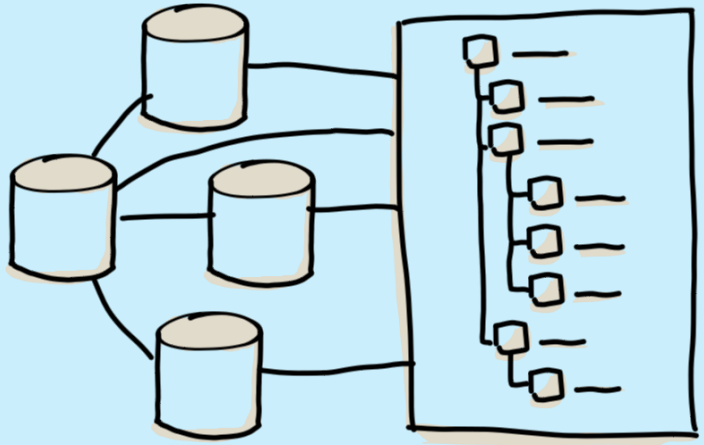
DATABASE



ANALYTICAL DATABASE

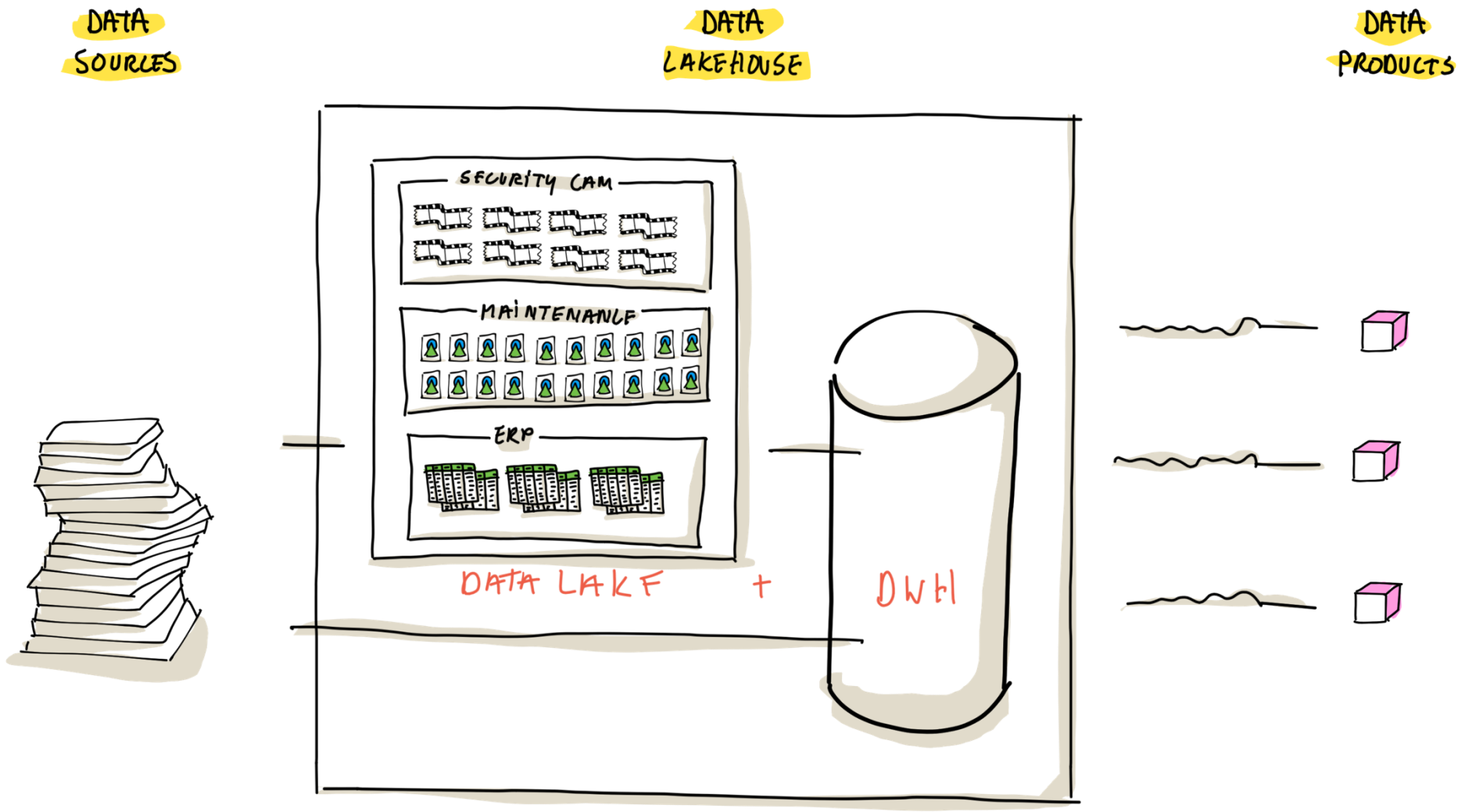


DATA LAKE



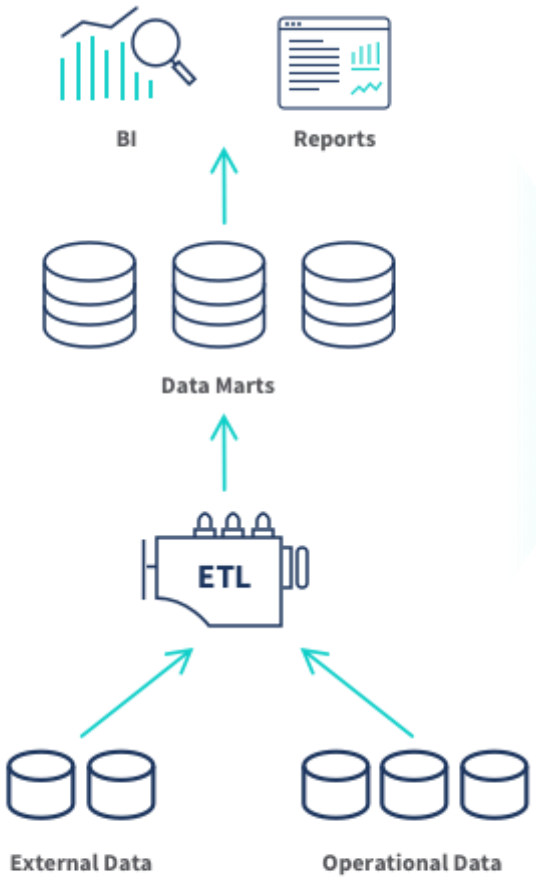
DATA LAKEHOUSE

Data Lakehouse



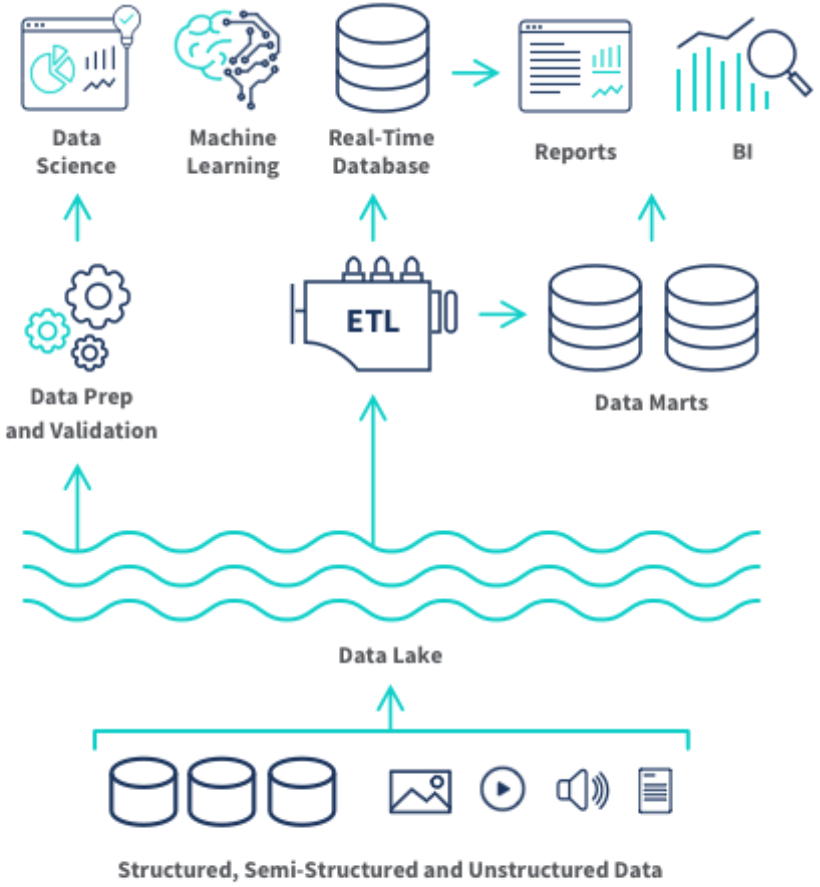
LATE 1980'S

Data Warehouse



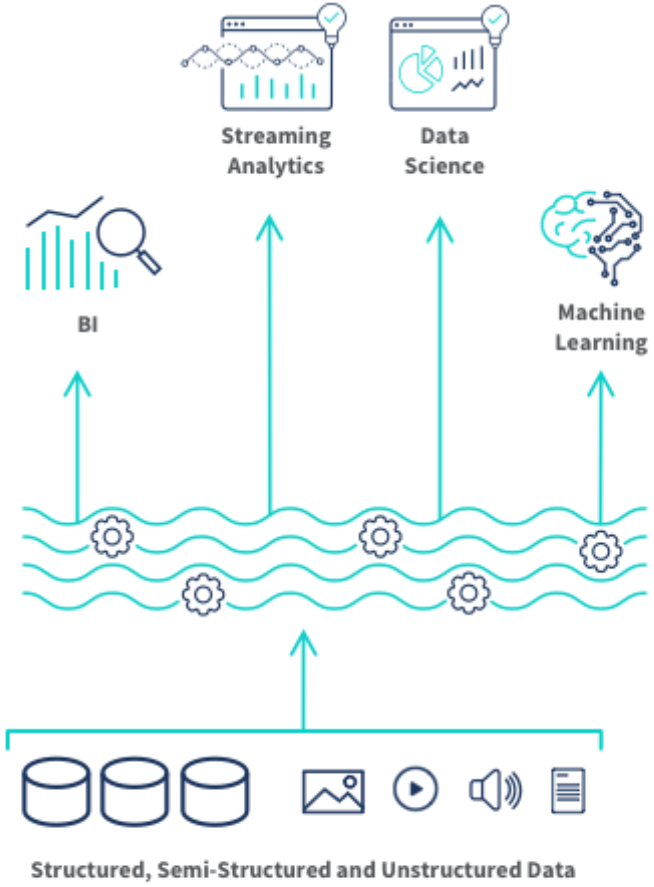
2011

Data Lake



2020

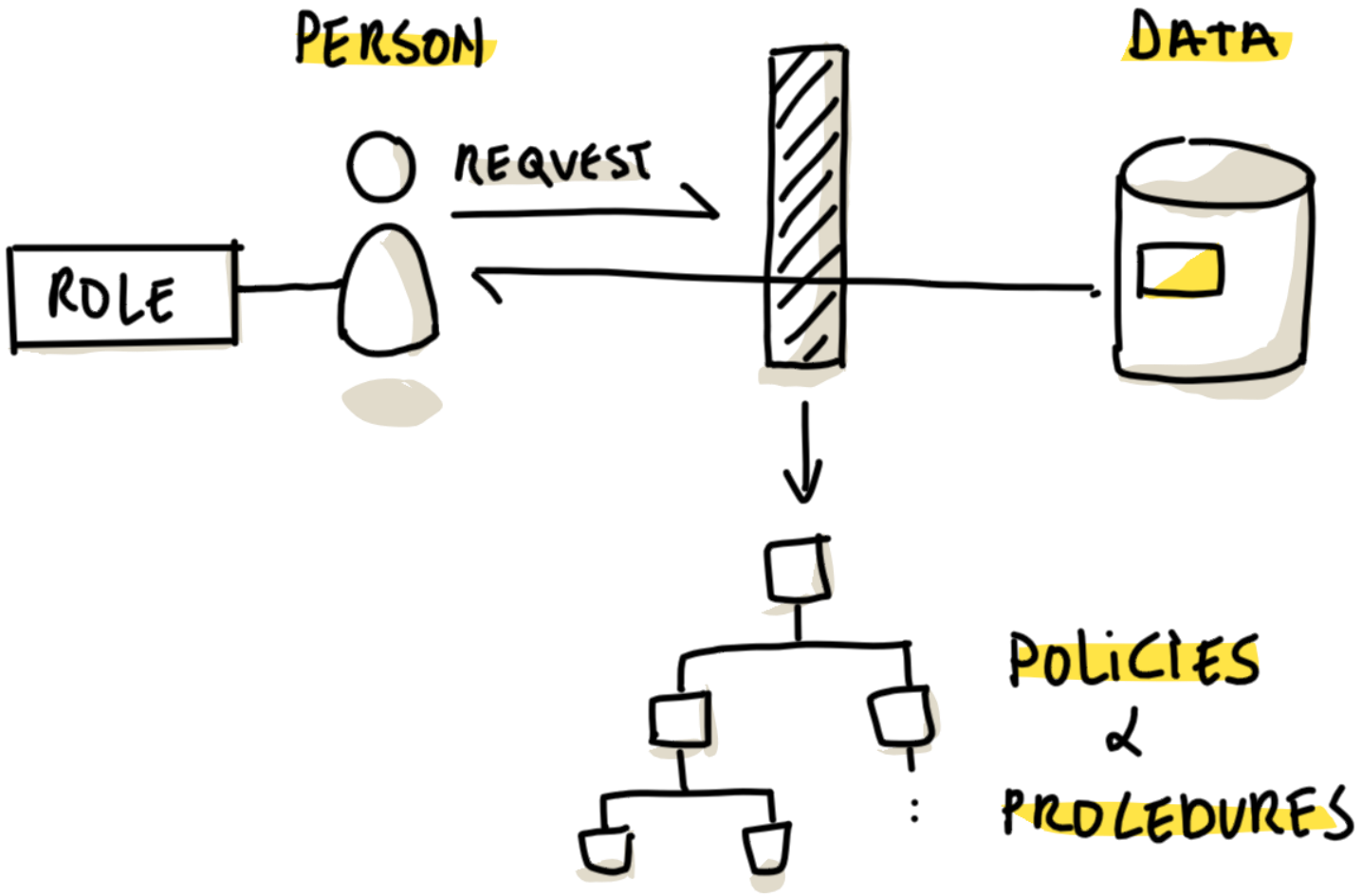
Lakehouse

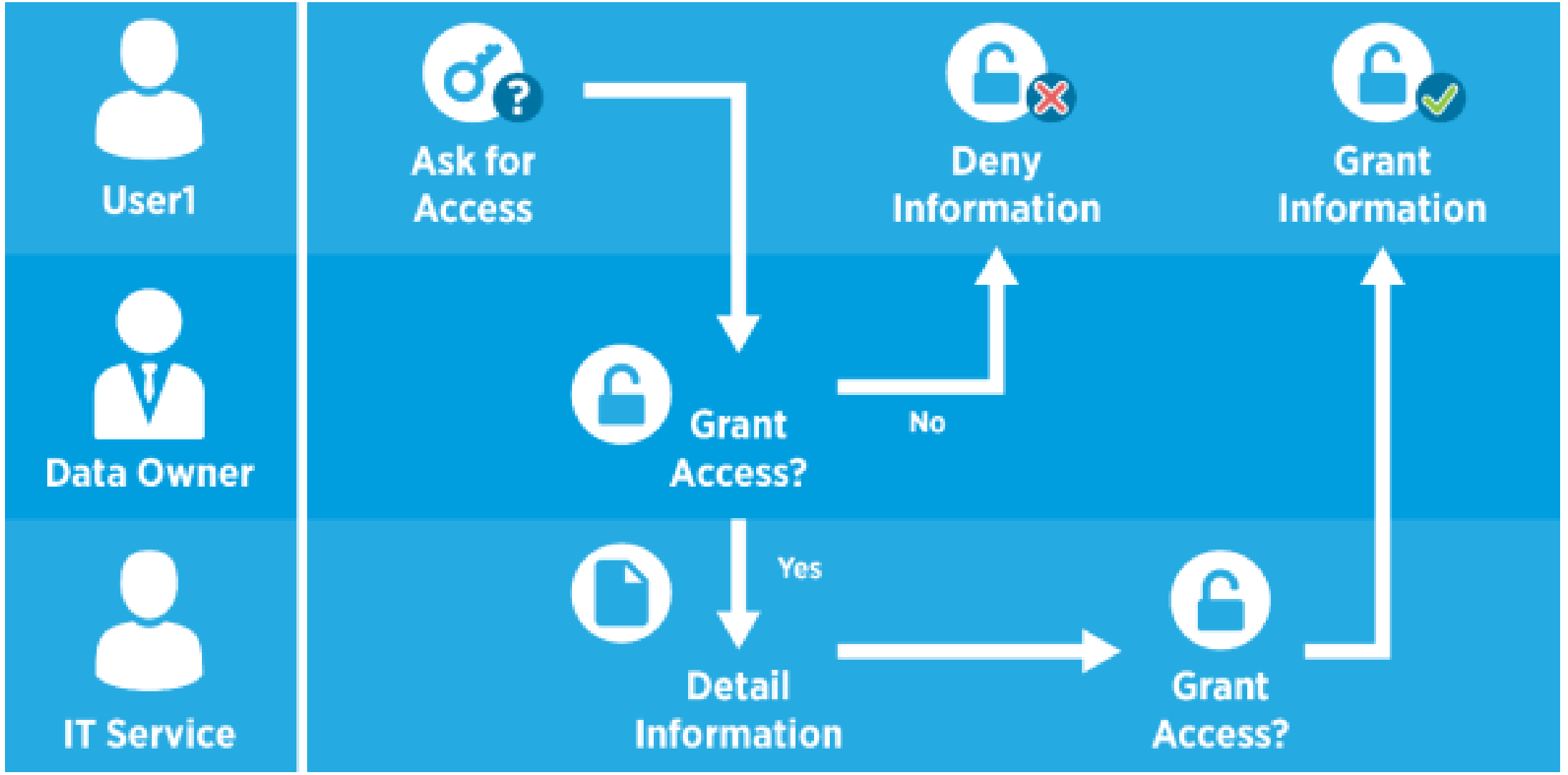


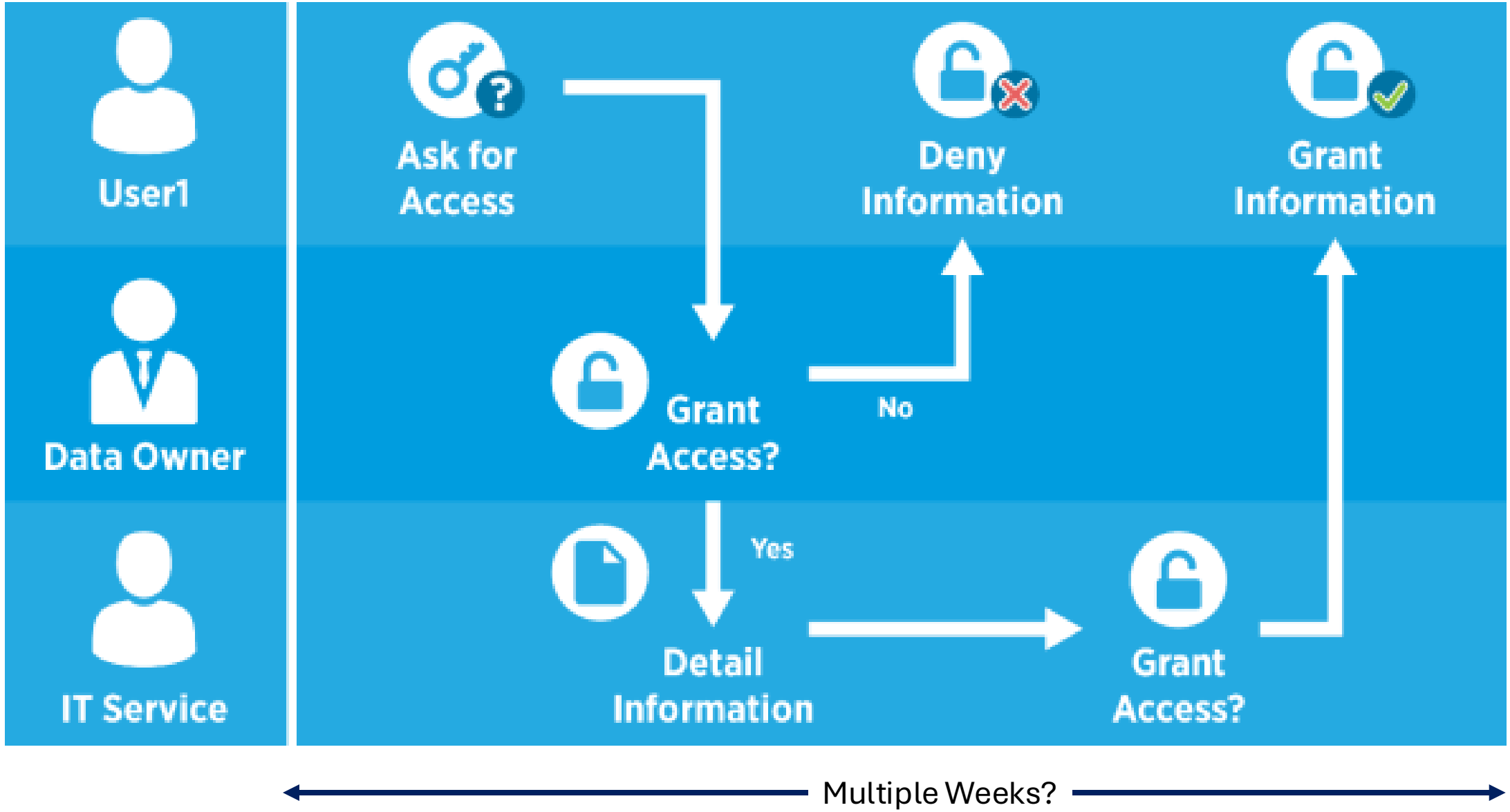
	Database	DWH Database	Data Lake	Data Lakehouse
Cost	+++	+++	+	++
Agility	+	+	++++	+++
Users	Anyone	IT / Business Users	Data Scientists	Anyone
Scaling	Vertical (expensive)	Horizontal (cheaper)		Horizontal (cheaper)
Volume	++	+++	+++++	+++++
Type of Data	Structured	Structured & Semi-Structured	Structured, Semi-Structured & Unstructured	Structured, Semi-Structured & Unstructured
Read Performance	++++ (Depending on the type)	++++	++	+++(+)

Data Security

Defining Security Policies and Procedures to provide Data Access


PERSON
DATA
REQUEST
ROLE
POLICIES
&
PROCEDURES


User1
Data Owner
IT Service
Ask for Access
Grant Access?
No
Deny Information
Yes
Detail Information
Grant Access?
Grant Information


User1
Data Owner
IT Service
Ask for Access
Grant Access?
No
Deny Information
Grant Information
Yes
Detail Information
Grant Access?
Multiple Weeks?

Step 3

What are the data objects you need access to?

Data object	Permissions
SALES Schema	Read ×
PURCHASING Schema	Read ×

Request received — Implementation — Closed

Notifications

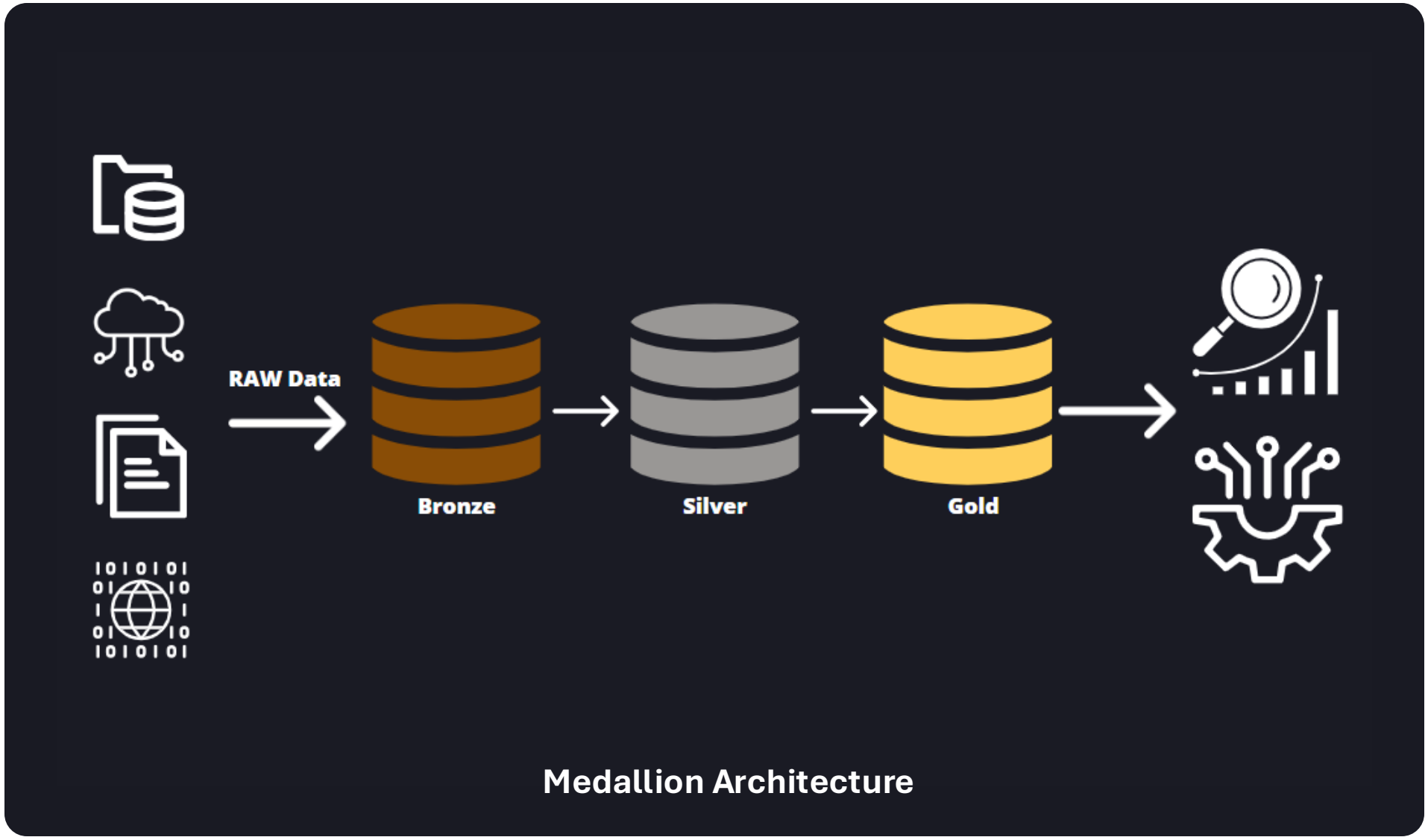
Nick Nguyen sent a request that requires your approval #109

3 hours ago

Access control	Owners	Action
DATA_ANALYST 32 data objects	NP TT	Implement
FINANCE 16 data objects	Jin Doe jin@raito.io	Implement

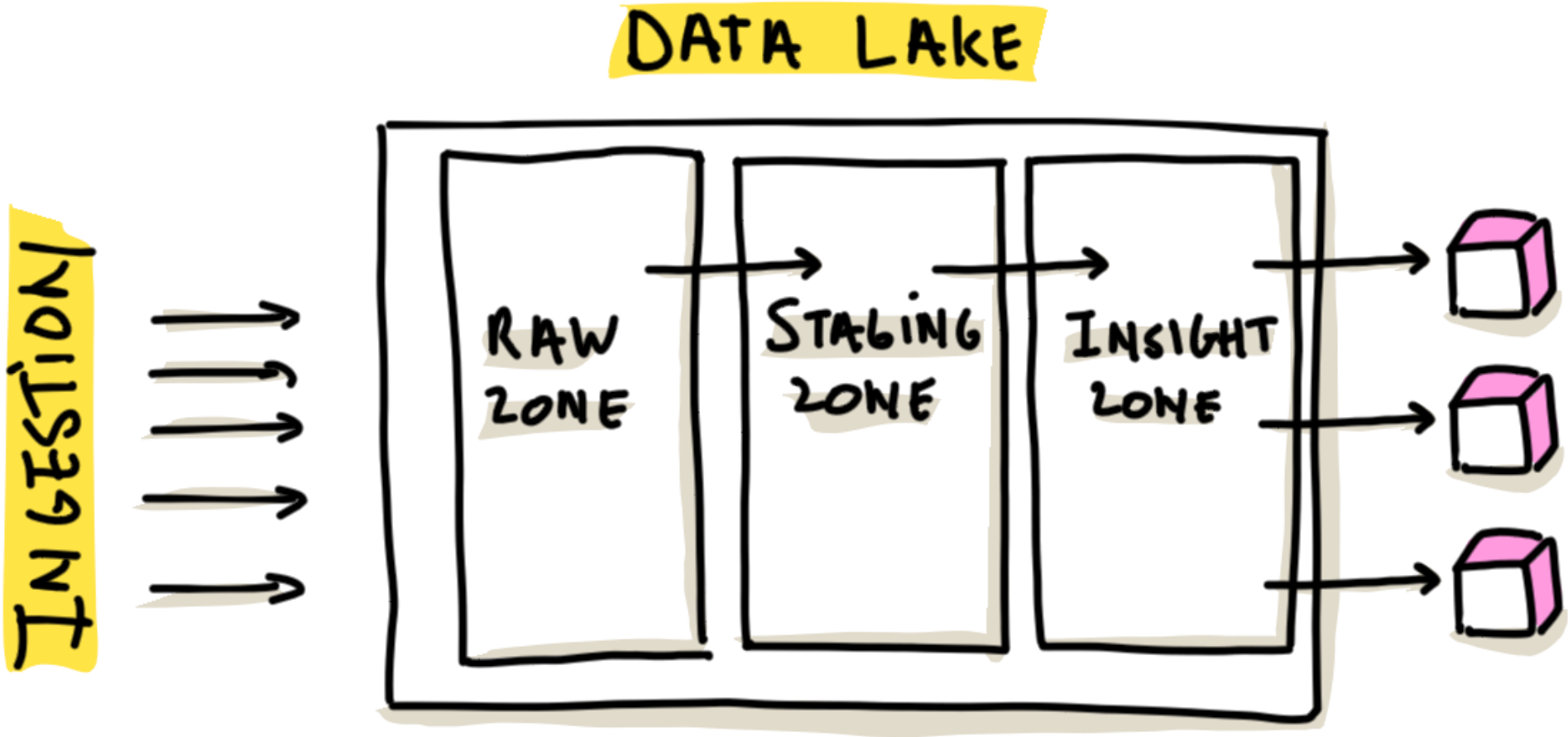
Policy in a Data Lake

Example Policy



Policy in a Data Lake

Example Policy



Medallion Architecture

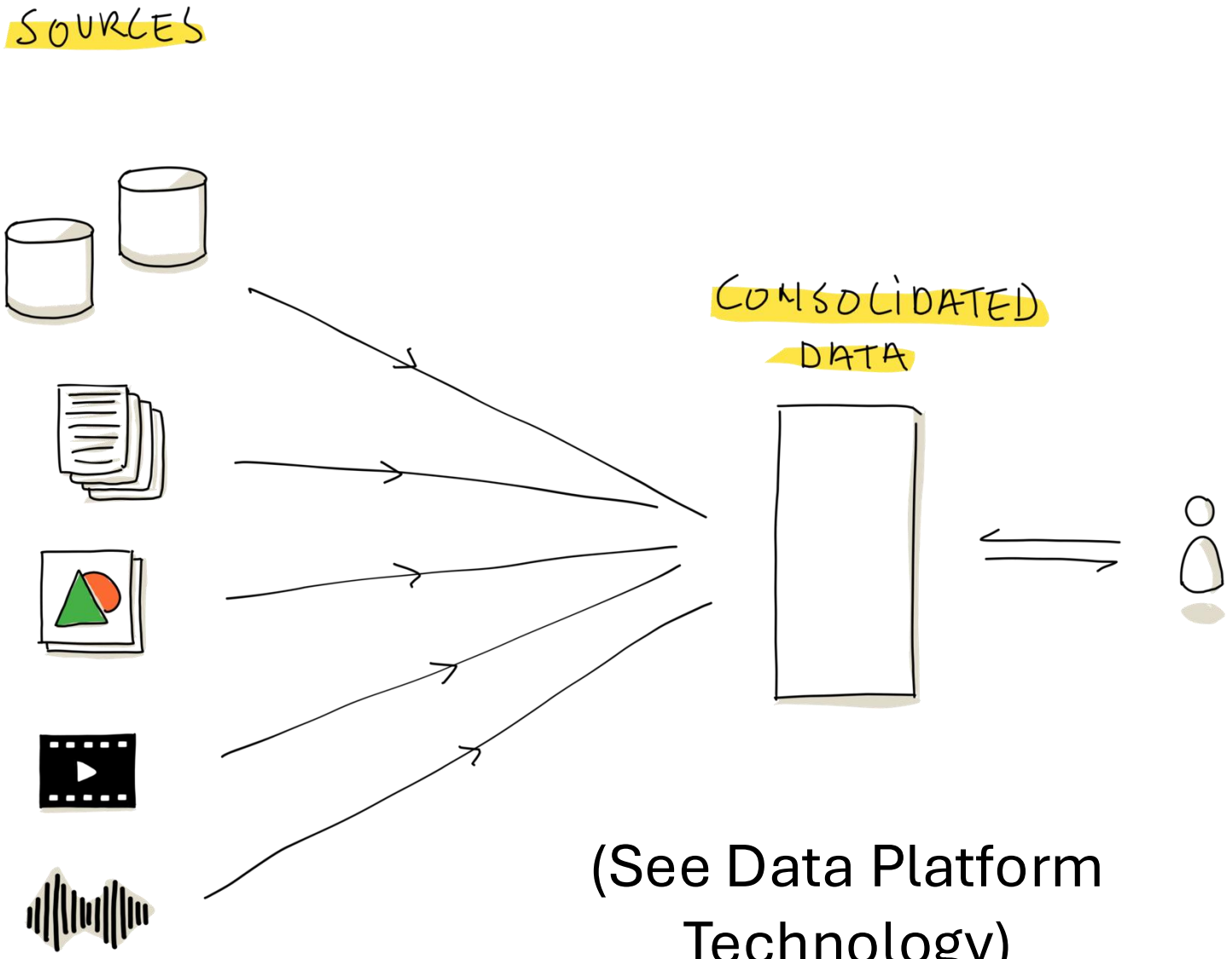
Example Data Policy

Example Policy

	RAW ZONE	STAGING ZONE	INSIGHT ZONE
TYPES OF DATA	• Any kind of data including unstructured data • Examples: videos, tekst files, .csv files, …	• Known and structured data • Data from multiple sources is likely to be joined here • Data engineers prep and cleanse data	• Known, enriched, integrated and cleaned data • Privacy controles like removing personal data
ACCESS	• Very restricted access • Likely a handful of people or just an admin	• More access • Mostly data engineers	• Highest level of access • Most if not all data analysts/scientists

Data Integration & Interoperability

The Movement and Consolidation of Data Within and Between Data Stores, Applications and Organizations

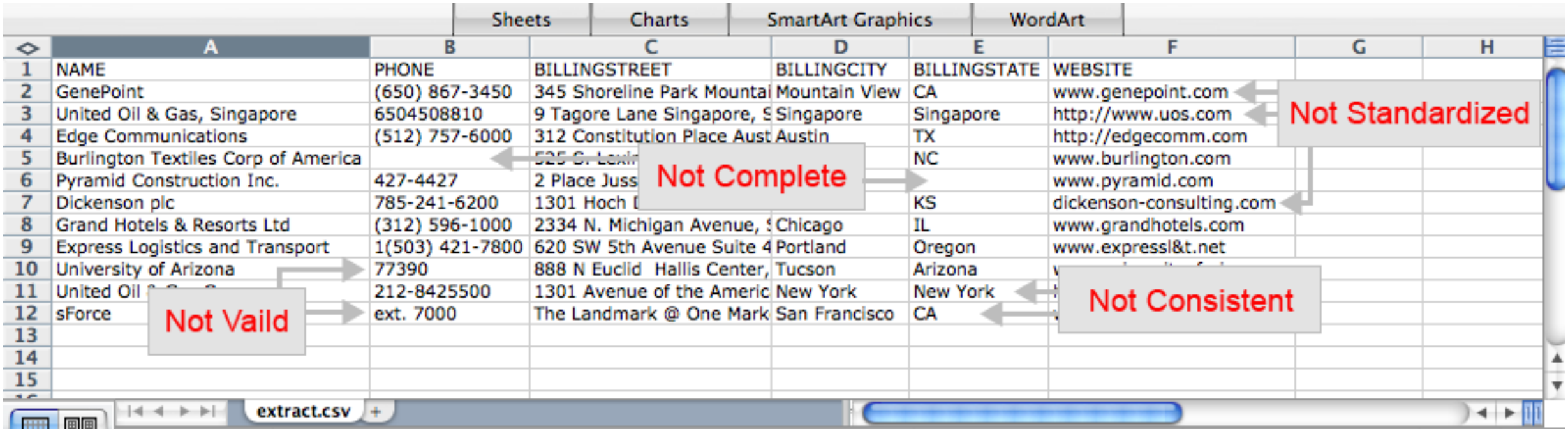


(See Data Platform Technology)

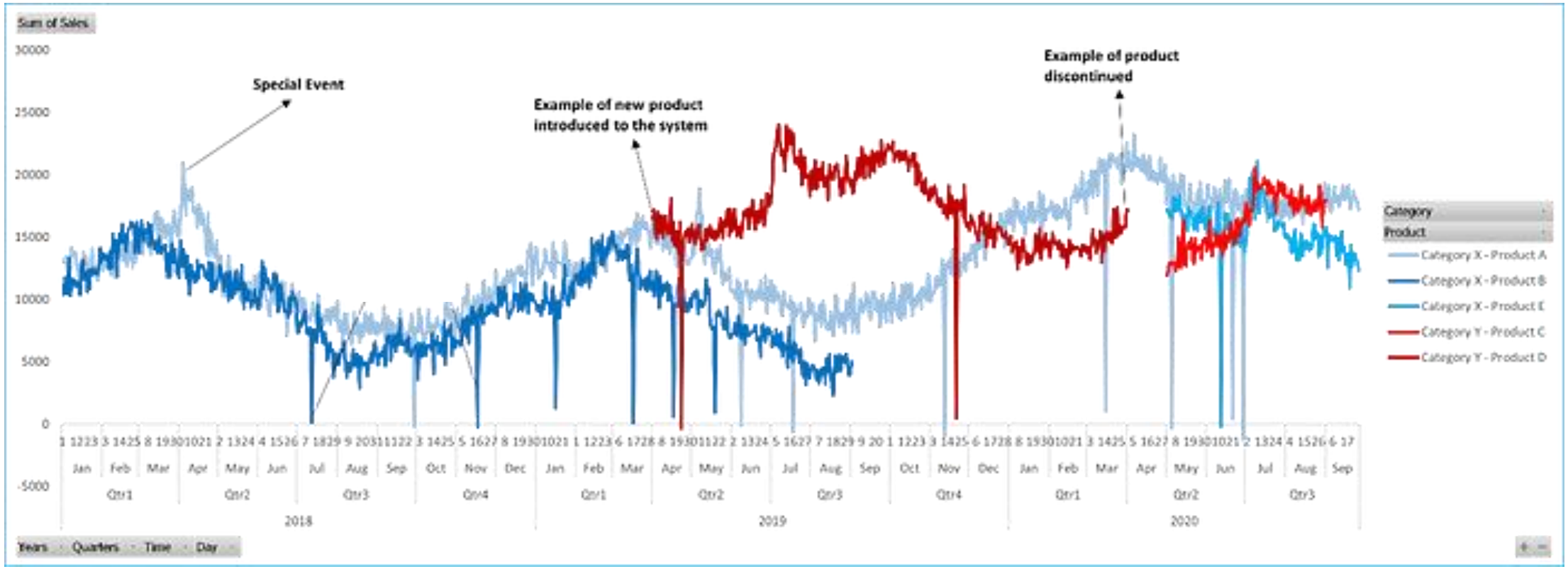
Data Quality

Defining standards and Data Quality controls. Implementing processes to manage and improve DQ

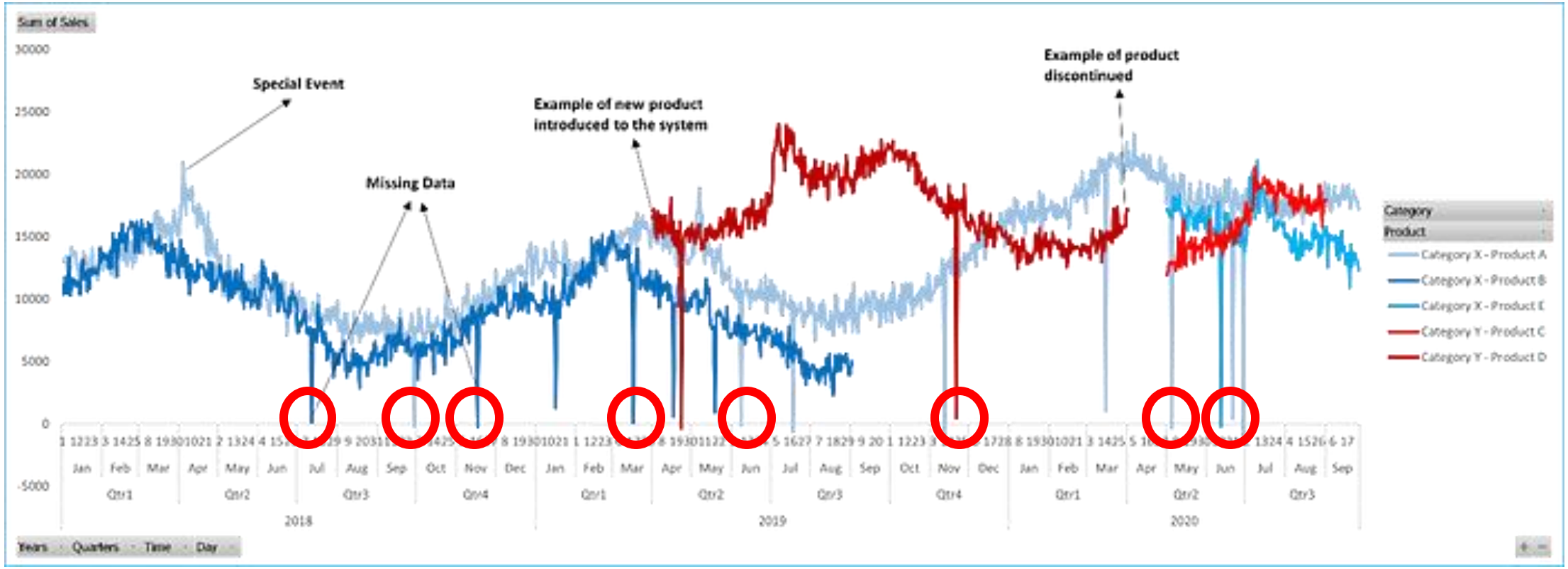
DQ : Example



	A	B	C	D	E	F	G	H
1	NAME	PHONE	BILLINGSTREET	BILLINGCITY	BILLINGSTATE	WEBSITE		
2	GenePoint	(650) 867-3450	345 Shoreline Park Mountai	Mountain View	CA	www.genepoint.com		
3	United Oil & Gas, Singapore	6504508810	9 Tagore Lane Singapore, S	Singapore	Singapore	http://www.uos.com		
4	Edge Communications	(512) 757-6000	312 Constitution Place Aust	Austin	TX	http://edgecomm.com		
5	Burlington Textiles Corp of America		525 S. Lexi		NC	www.burlington.com		
6	Pyramid Construction Inc.	427-4427	2 Place Juss			www.pyramid.com		
7	Dickenson plc	785-241-6200	1301 Hoch D		KS	dickenson-consulting.com		
8	Grand Hotels & Resorts Ltd	(312) 596-1000	2334 N. Michigan Avenue, S	Chicago	IL	www.grandhotels.com		
9	Express Logistics and Transport	1(503) 421-7800	620 SW 5th Avenue Suite 4	Portland	Oregon	www.expressl&t.net		
10	University of Arizona	77390	888 N Euclid Hallis Center,	Tucson	Arizona			
11	United Oil	212-8425500	1301 Avenue of the Americ	New York	New York			
12	sForce	ext. 7000	The Landmark @ One Mark	San Francisco	CA			
13								
14								
15								

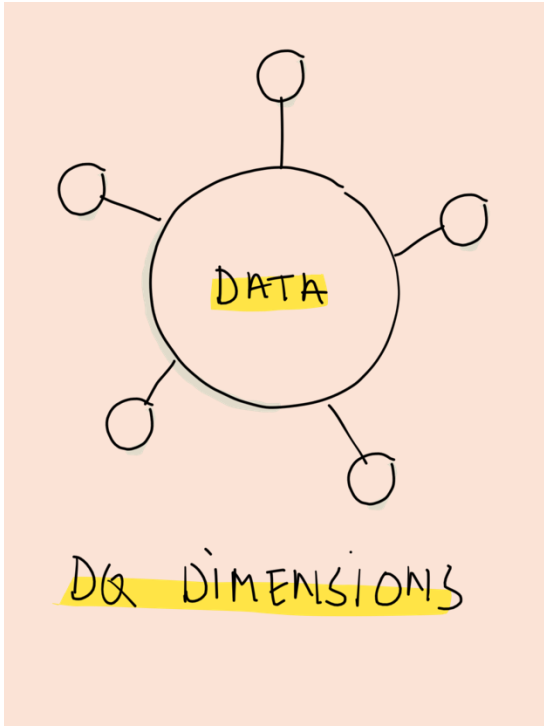


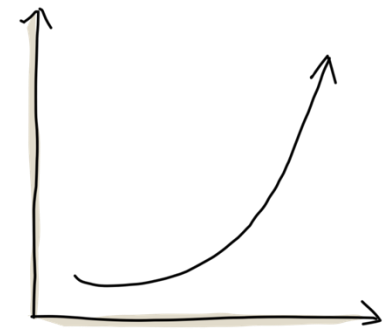
[Time Series Forecasting]



[Time Series Forecasting]

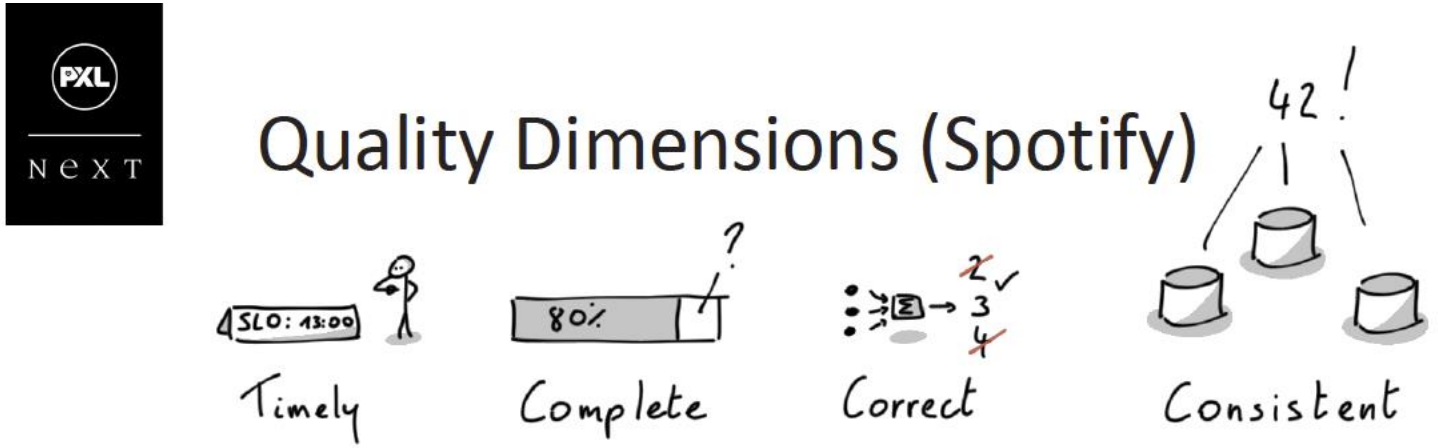
Data Quality





DQ PROCESS
&TOOLS

Introduced in the Previous Lesson



- **Timely**: data is on time (SLO)
- **Complete**: all required data is available
- **Correct**: correct w.r.t. specs, input produces output
- **Consistent**: same meaning across systems, data in sync

What are Data Quality Dimensions?

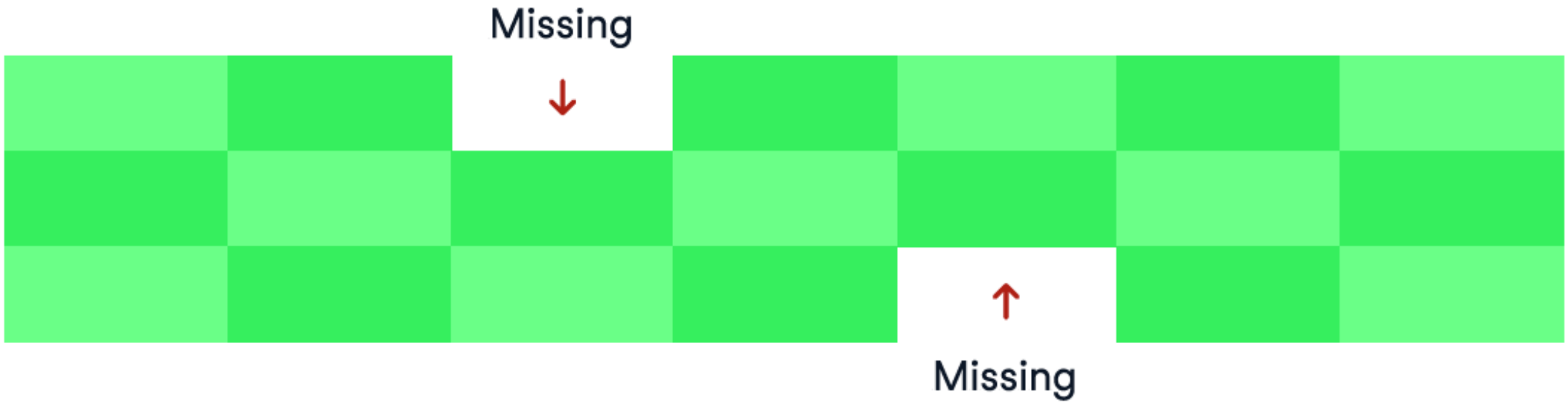
Data Quality is a measurement of the degree to which data is fit for purpose. Good data quality generates trust in data. Data Quality Dimensions are a measurement of a specific attribute of a data's quality.

Completeness – Validity – Accuracy – Uniqueness
Timeliness – Consistency

[Source: Datacamp]

> Completeness

Completeness measures the degree to which all expected records in a dataset are present. At a data element level, completeness is the degree to which all records have data populated when expected.



[Source: Datacamp]

Completeness Example

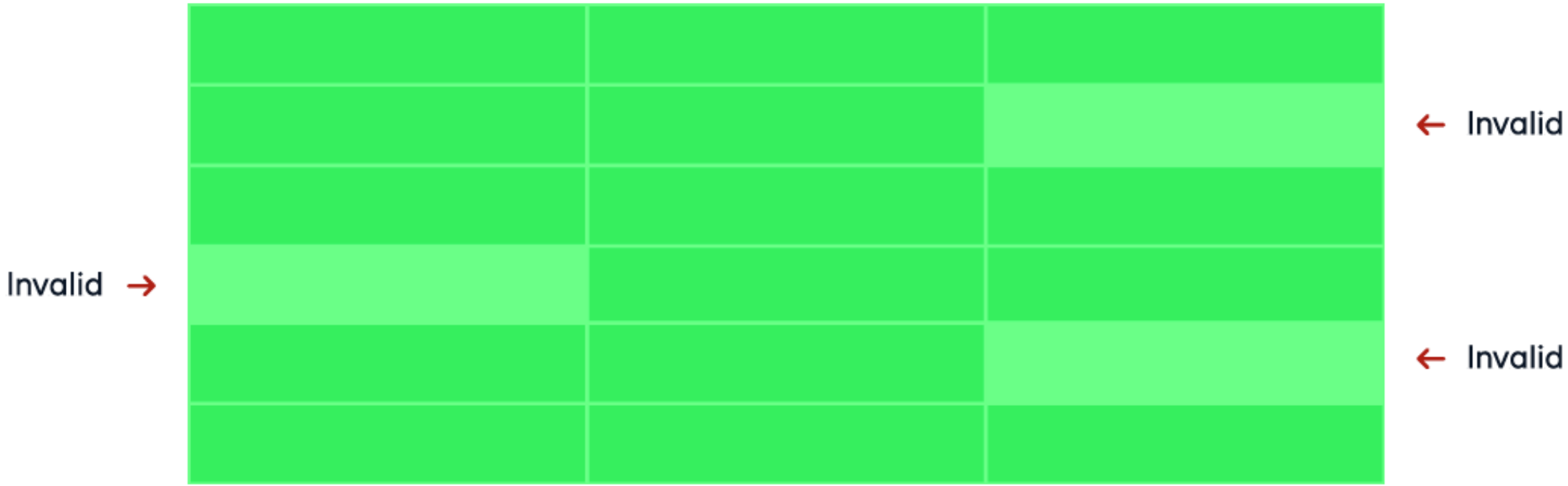
All records must have a value populated in the CustomerName field.

CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

[Source: Datacamp]

> Validity

Validity measures the degree to which the values in a data element are valid.



[Source: Datacamp]

Validity Example

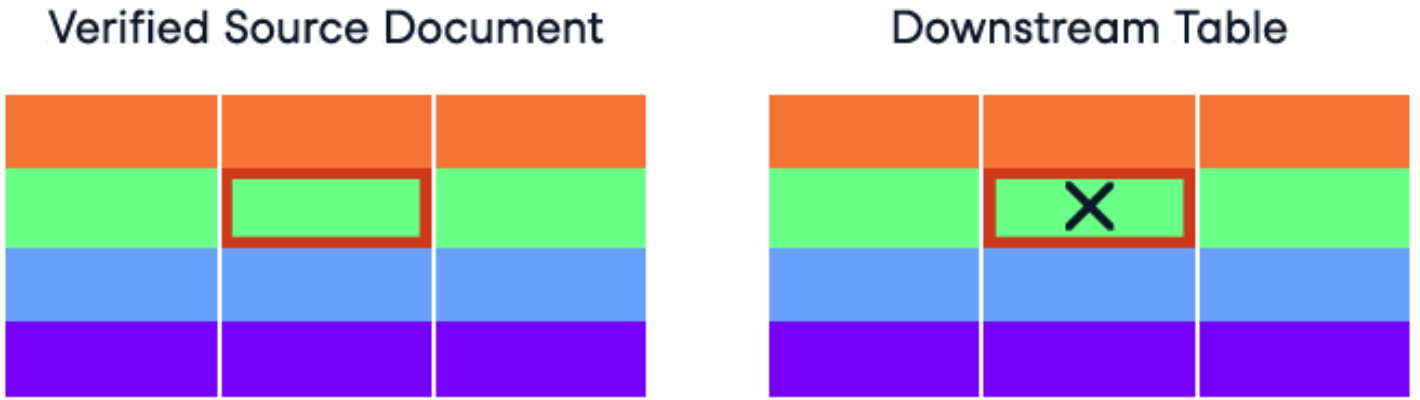
- CustomerBirthDate value must be a date in the past.
- CustomerAccountType value must be either Loan or Deposit.
- LatestAccountOpenDate value must be a date in the past.

CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

[Source: Datacamp]

> Accuracy

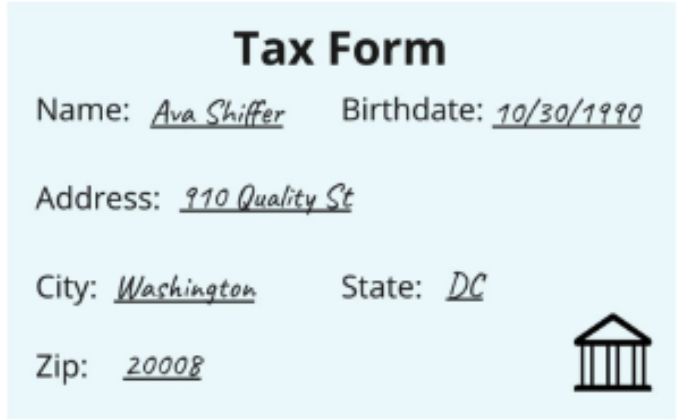
Accuracy measures the degree to which data is correct and represents the truth.



[Source: Datacamp]

Accuracy Example

All records in the Customer Table must have accurate Customer Name, Customer Birthdate, and Customer Address fields when compared to the Tax Form.



CustomerName	CustomerBirthDate	CustomerAddress	CustomerCity	CustomerState	CustomerZip
Ava Shiffer	10/31/1990	910 Quality St	Washington	WA	20008

[Source: Datacamp]

After I sent **a late notice about an outstanding invoice** to a third-party firm I sub-contract for, we discovered that while the check was indeed in the mail, unfortunately it was mailed to the wrong address—a valid but inaccurate address.

> Uniqueness

Uniqueness measures the degree to which the records in a dataset are not duplicated.

[Source: Datacamp]

Uniqueness Example

All records must have a unique CustomerID and CustomerName.

CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

[Source: Datacamp]

> Timeliness

Timeliness is the degree to which a dataset is available when expected and depends on service level agreements being set up between technical and business resources.

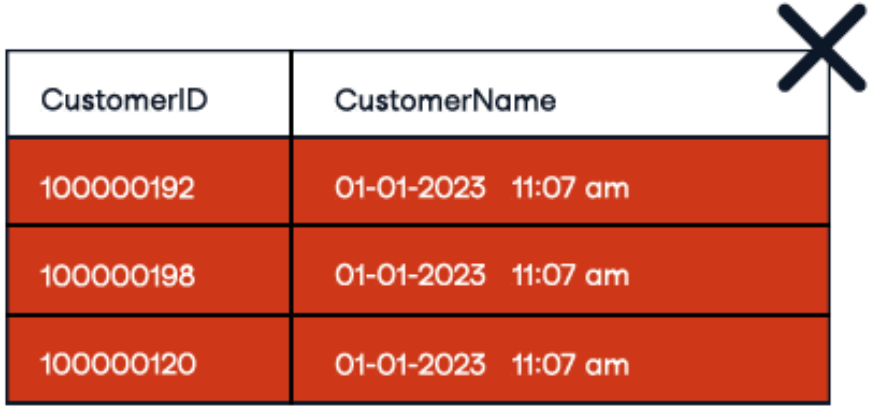
SLA	Table Load Time
08:00 am	07:59 am
10:00 am	09:59 am
11:00 am	11:01 am

← Missed the SLA

[Source: Datacamp]

Timeliness Example

All records in the customer dataset must be loaded by the 9:00 am.



CustomerID	CustomerName
100000192	01-01-2023 11:07 am
100000198	01-01-2023 11:07 am
100000120	01-01-2023 11:07 am

[Source: Datacamp]

> Consistency

Consistency is a data quality dimension that measures the degree to which data is the same across all instances of the data. Consistency can be measured by setting a threshold for how much difference there can be between two datasets.

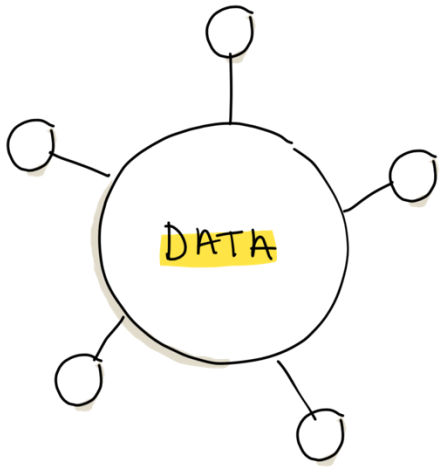
[Source: Datacamp]

Consistency Example

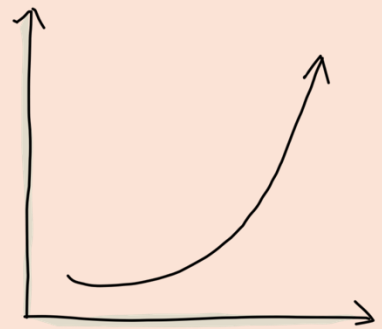
AccountTableCustomerID	CustomerTableCustomerID
108394858	108394858
192039482	192039482
203475849	NULL ✗
2930485953	NULL ✗
102832748	102832748

[Source: Datacamp]

Data Quality

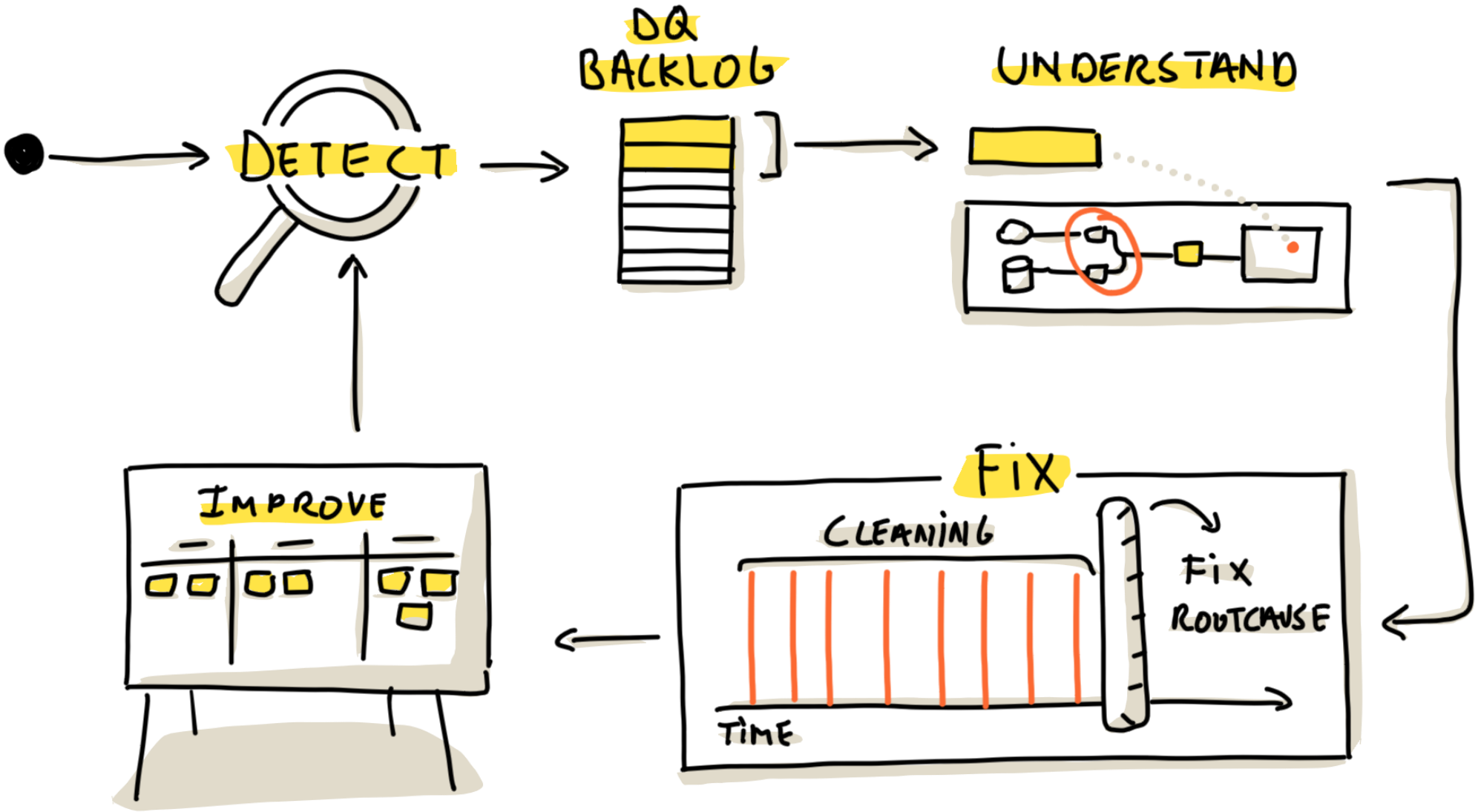


DQ DIMENSIONS

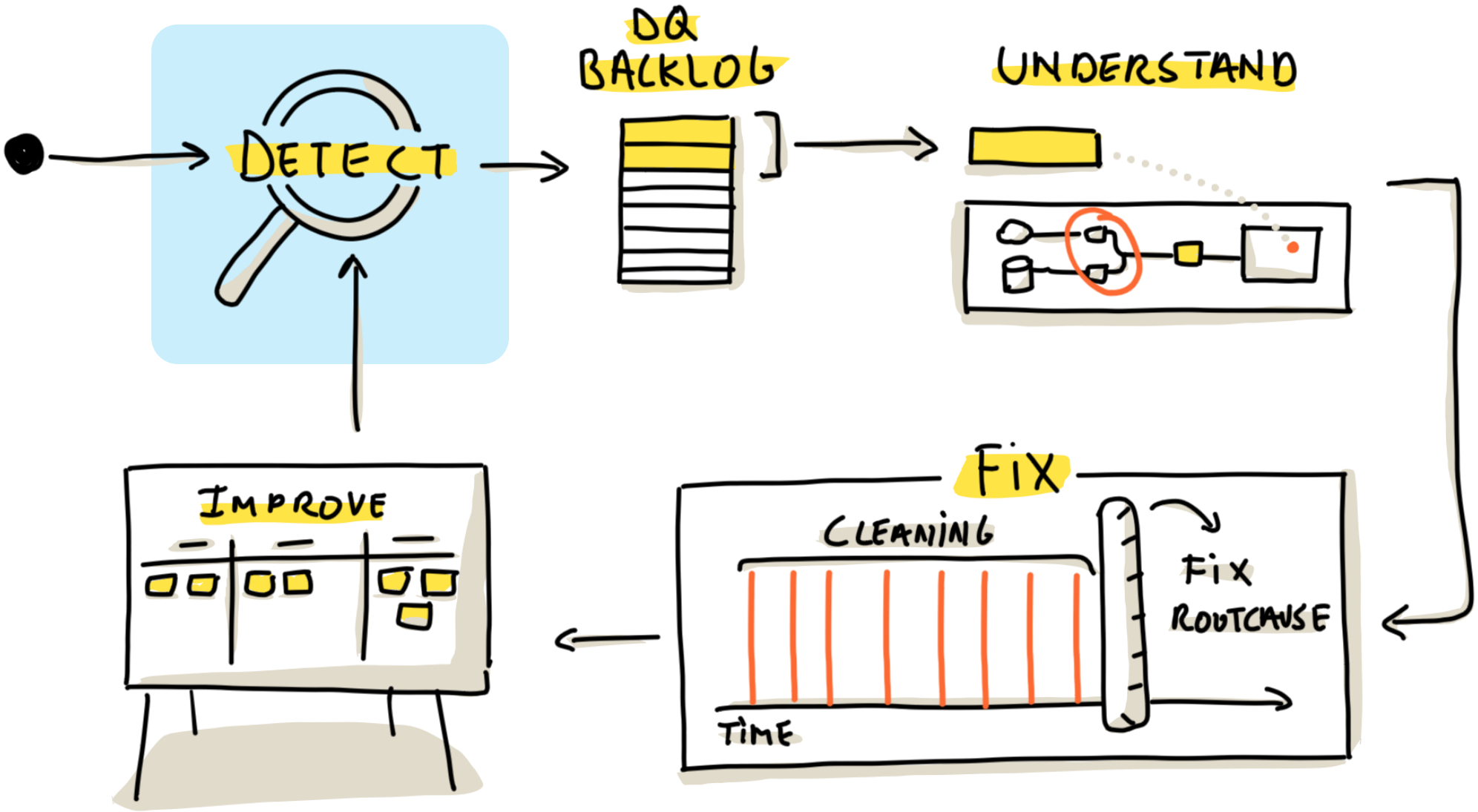


DQ PROCESS
&TOOLS

Data Quality Process



Data Quality Process



DQ Tests

SQL customer_id.assert.sql

```sql
-- check if customers contains null values
SELECT customer_id as customer_id,
       customer_name as customer_name
FROM customers
WHERE customer_id IS NULL
```

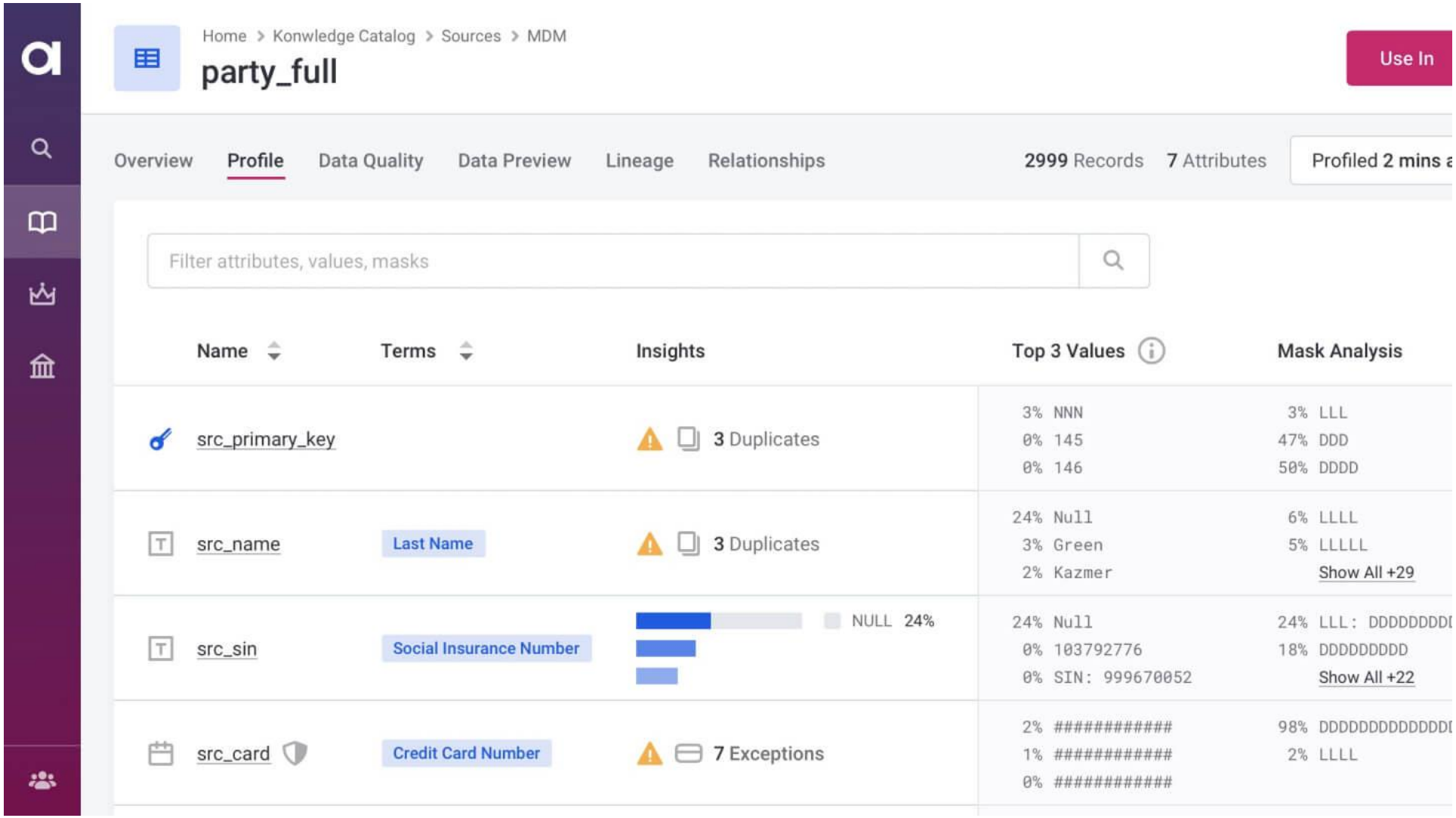
if 0 row returned:

Assertion passed ✅

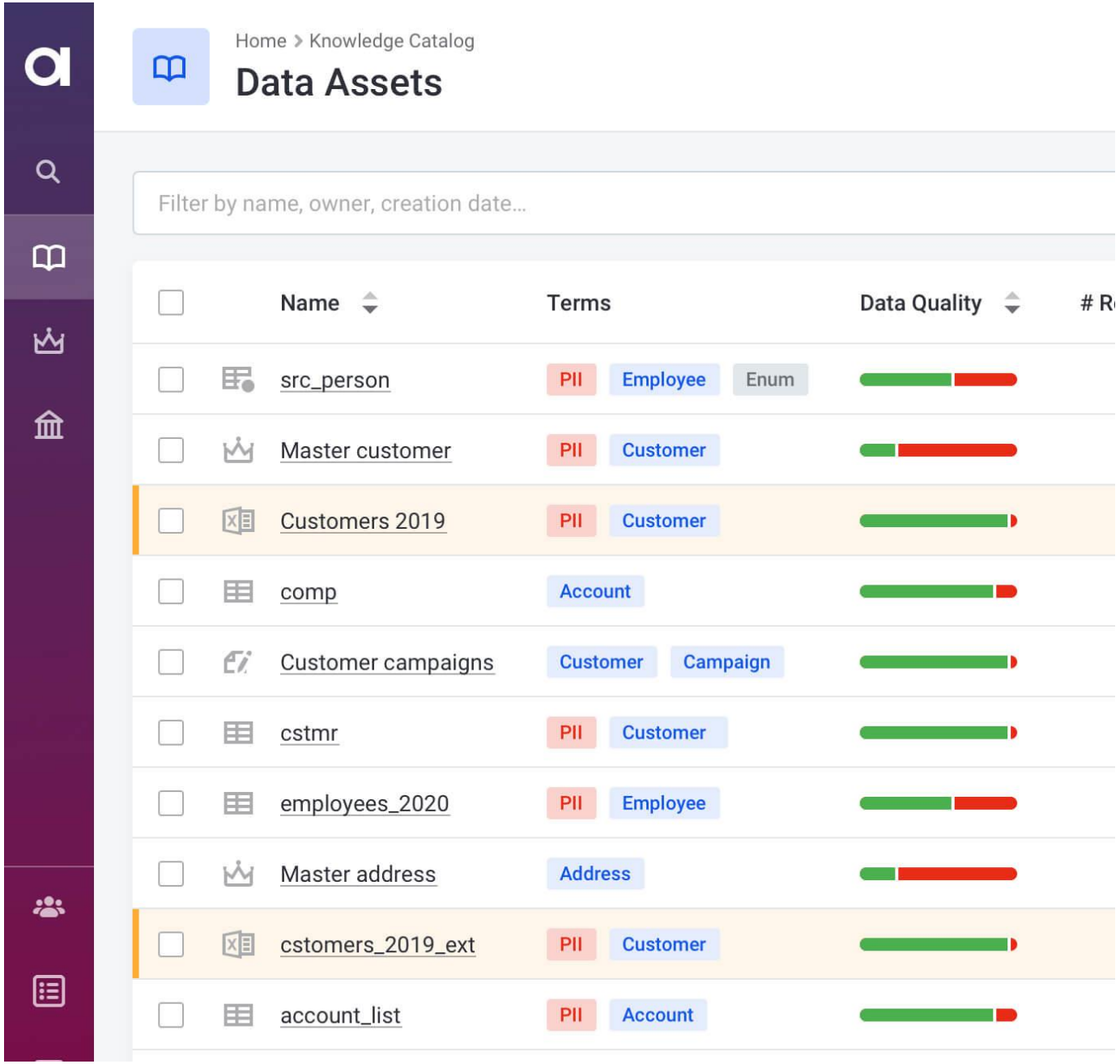
if >=1 row(s) returned:

Assertion failed ❌

DQ Monitoring

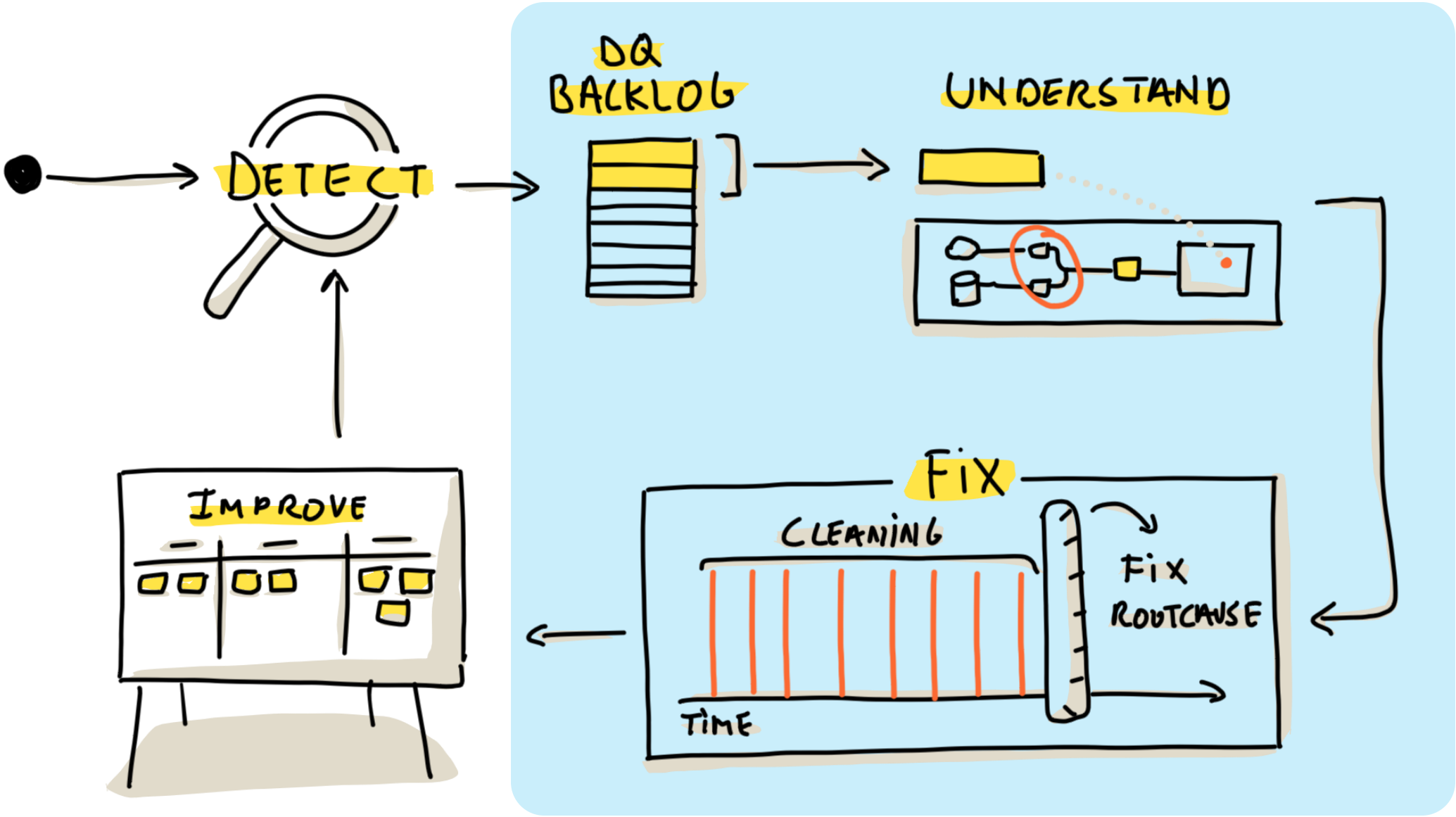


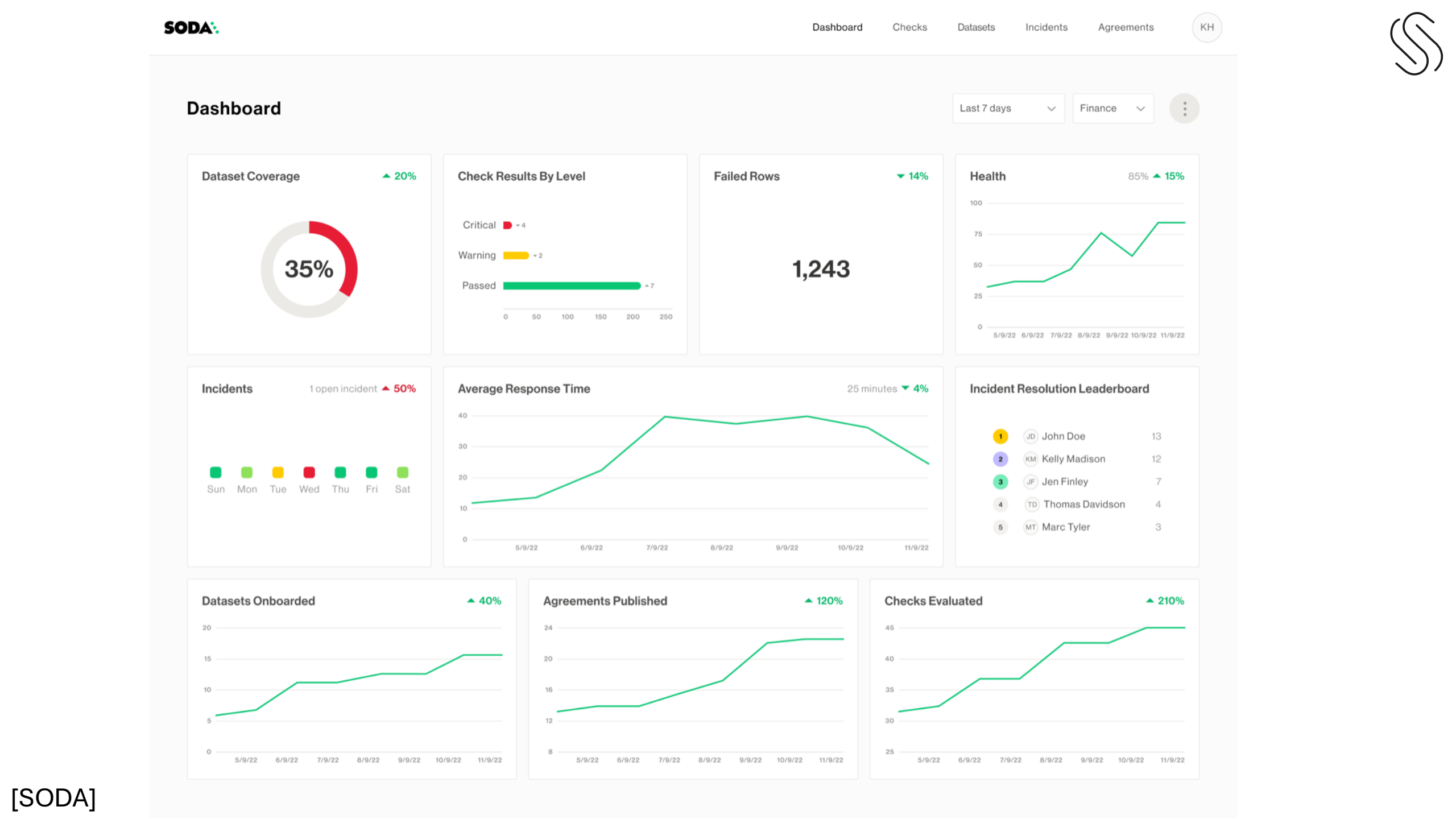
[Alation]


Home > Knowledge Catalog
Data Assets
Filter by name, owner, creation date...
Name
Terms
Data Quality
R
src_person
PII
Employee
Enum
Master customer
PII
Customer
Customers 2019
PII
Customer
comp
Account
Customer campaigns
Customer
Campaign
cstmr
PII
Customer
employees_2020
PII
Employee
Master address
Address
cstomers_2019_ext
PII
Customer
account_list
PII
Account

[Alation]

Data Quality Process



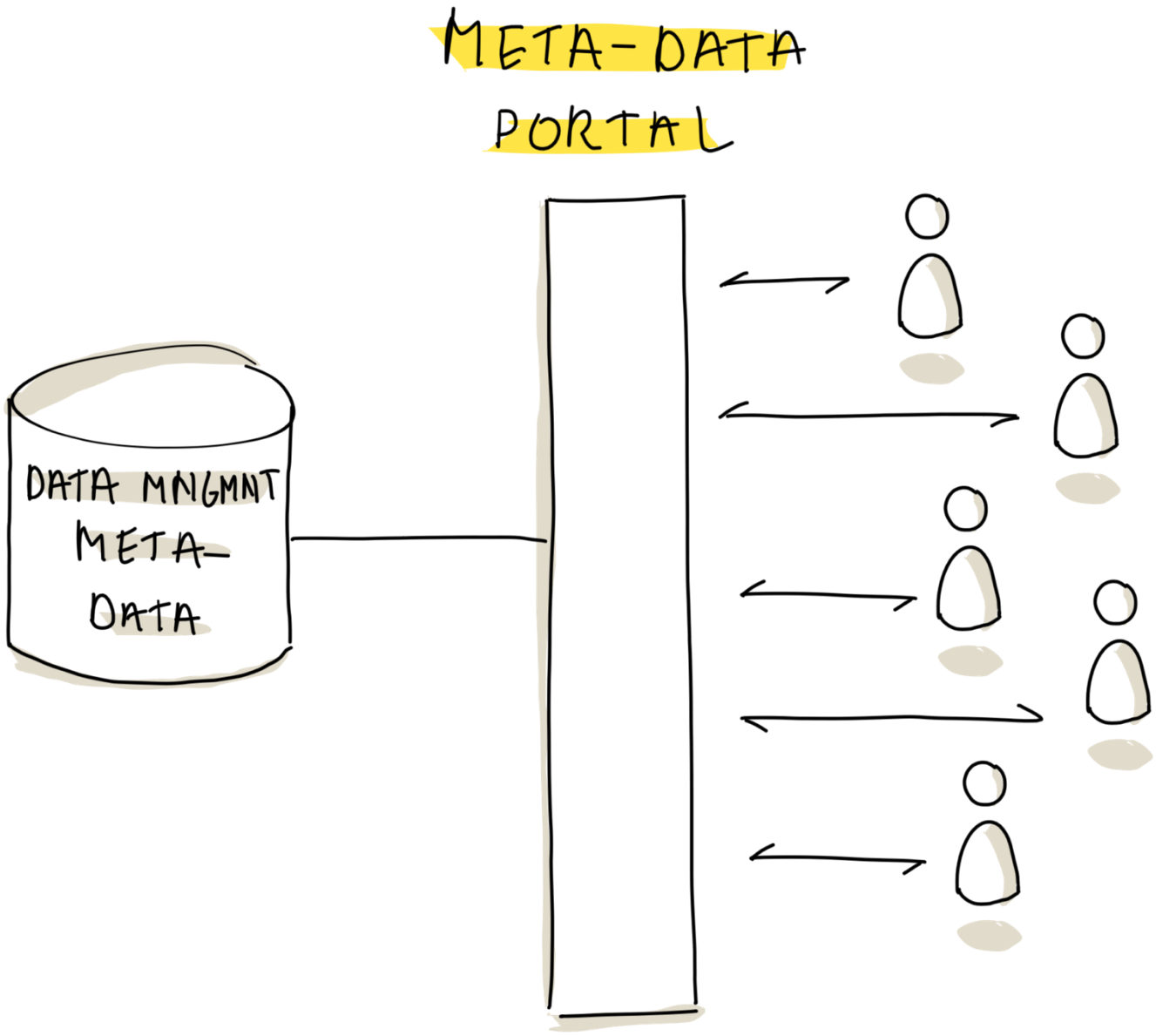

SODA
Dashboard
Checks
Datasets
Incidents
Agreements
KH
Dashboard
Last 7 days
Finance
Dataset Coverage
20%
35%
Check Results By Level
Critical
4
Warning
2
Passed
7
0
50
100
150
200
250
Failed Rows
14%
1,243
Health
85%
15%
100
75
50
25
0
5/9/22
6/9/22
7/9/22
8/9/22
9/9/22
10/9/22
11/9/22
Incidents
1 open incident
50%
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Average Response Time
25 minutes
4%
40
30
20
10
0
5/9/22
6/9/22
7/9/22
8/9/22
9/9/22
10/9/22
11/9/22
Incident Resolution Leaderboard
1
JD
John Doe
13
2
KM
Kelly Madison
12
3
JF
Jen Finley
7
4
TD
Thomas Davidson
4
5
MT
Marc Tyler
3
Datasets Onboarded
40%
20
15
10
5
0
5/9/22
6/9/22
7/9/22
8/9/22
9/9/22
10/9/22
11/9/22
Agreements Published
120%
24
20
16
12
8
5/9/22
6/9/22
7/9/22
8/9/22
9/9/22
10/9/22
11/9/22
Checks Evaluated
210%
45
40
35
30
25
5/9/22
6/9/22
7/9/22
8/9/22
9/9/22
10/9/22
11/9/22

[SODA]

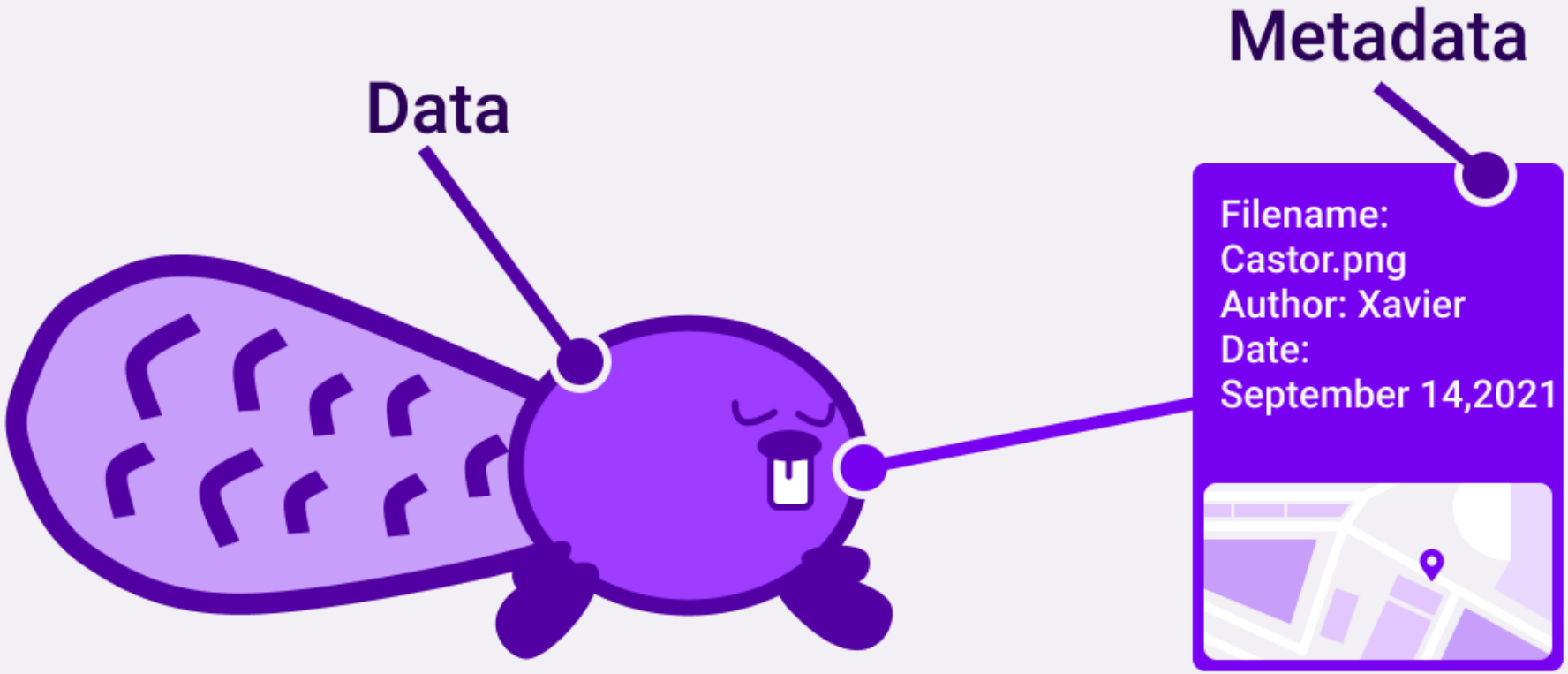
Meta Data


Data Quality
Data Architecture
Meta Data
Data Modeling & Design
7
Data Warehousing & Business Intelligence
Data Governance
Data Storage & Operations
Reference & Master Data
Data Security
Documents & Content
Data Integration & Inter-operability

...


META-DATA
PORTAL
DATA MNGMNT
META-
DATA

Metadata = “Data about Data”




Data
(What you can use)
Metadata
(Describing the data format, the actions leading to this data, documentation, business context, ...)

Types of Metadata

Technical

Definitional

Schemas, data types, models, etc.

Operational

Descriptive

How is data being produced? (Source / Transformations / …)

Business

Descriptive

Data tags, classifications, mappings to business relationships, etc.

Social

Descriptive

Metadata about user-generated content, business knowledge, etc.

Types of Metadata

Technical

Definitional

Schemas, data types, models, etc.

Operational

Descriptive

How is data being produced? (Source / Transformations / …)

Business

Descriptive

Data tags, classifications, mappings to business relationships, etc.

Social

Descriptive

Metadata about user-generated content, business knowledge, etc.

Table Structure

emlployee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Data

Metadata

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
department_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee sti

Types of Metadata

Technical

Definitional

Schemas, data types, models, etc.

Operational

Descriptive

How is data being produced? (Source / Transformations / …)

Business

Descriptive

Data tags, classifications, mappings to business relationships, etc.

Social

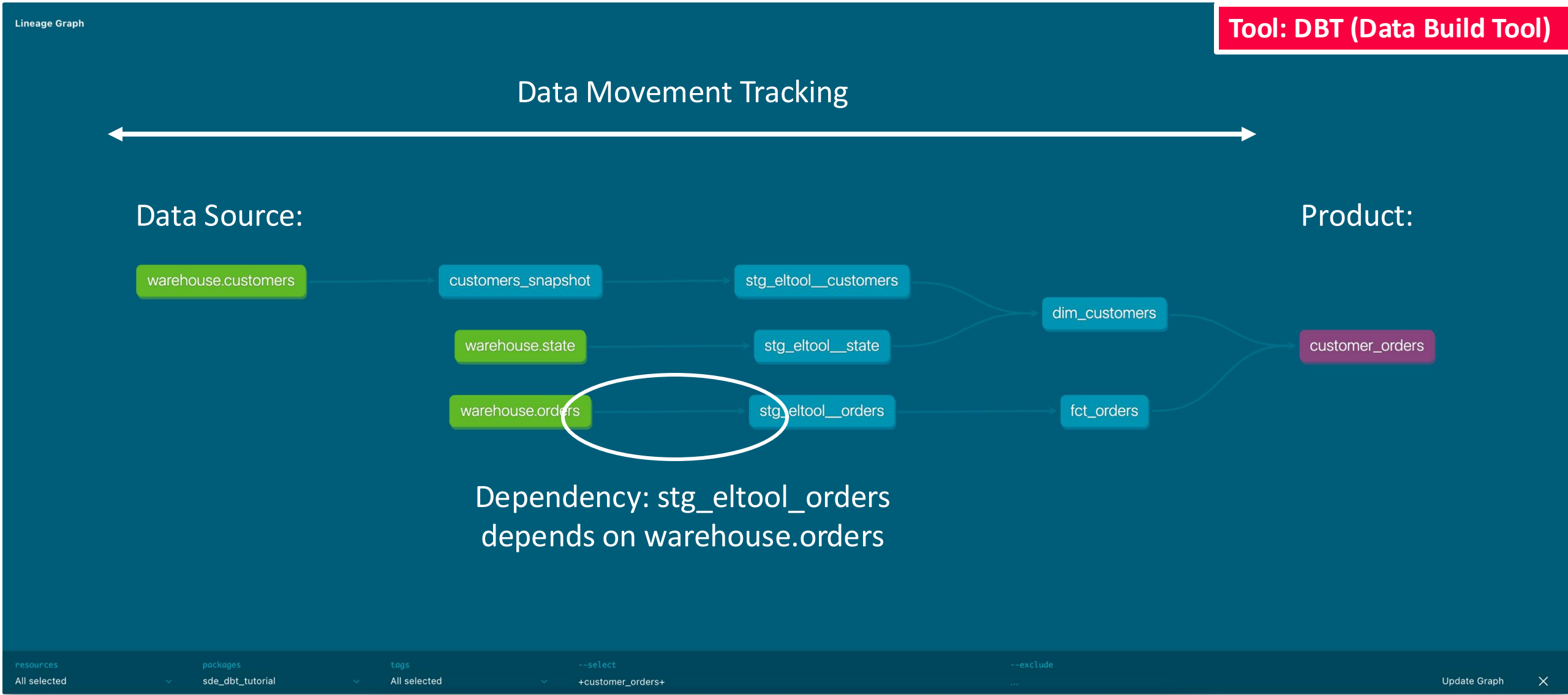
Descriptive

Metadata about user-generated content, business knowledge, etc.

Data Lineage

- Documentation of the flow and transformation of data as it moves from source to product.
- It is crucial for understanding a data product's origins, transformations
- Four key points to define data lineage:
  1. **Data Source Identification**: identifying the original sources of data
  2. **Data Movement Tracking**: It traces how data is transferred and transformed
  3. **Dependency Mapping**: Data lineage maps dependencies between different data elements
  4. **End-to-End Visibility**: It provides a comprehensive view of data's journey

Example: E2E Visible Data Lineage



Types of Metadata

Technical

Definitional

Schemas, data types, models, etc.

Operational

Descriptive

How is data being produced? (Source / Transformations / …)

Business

Descriptive

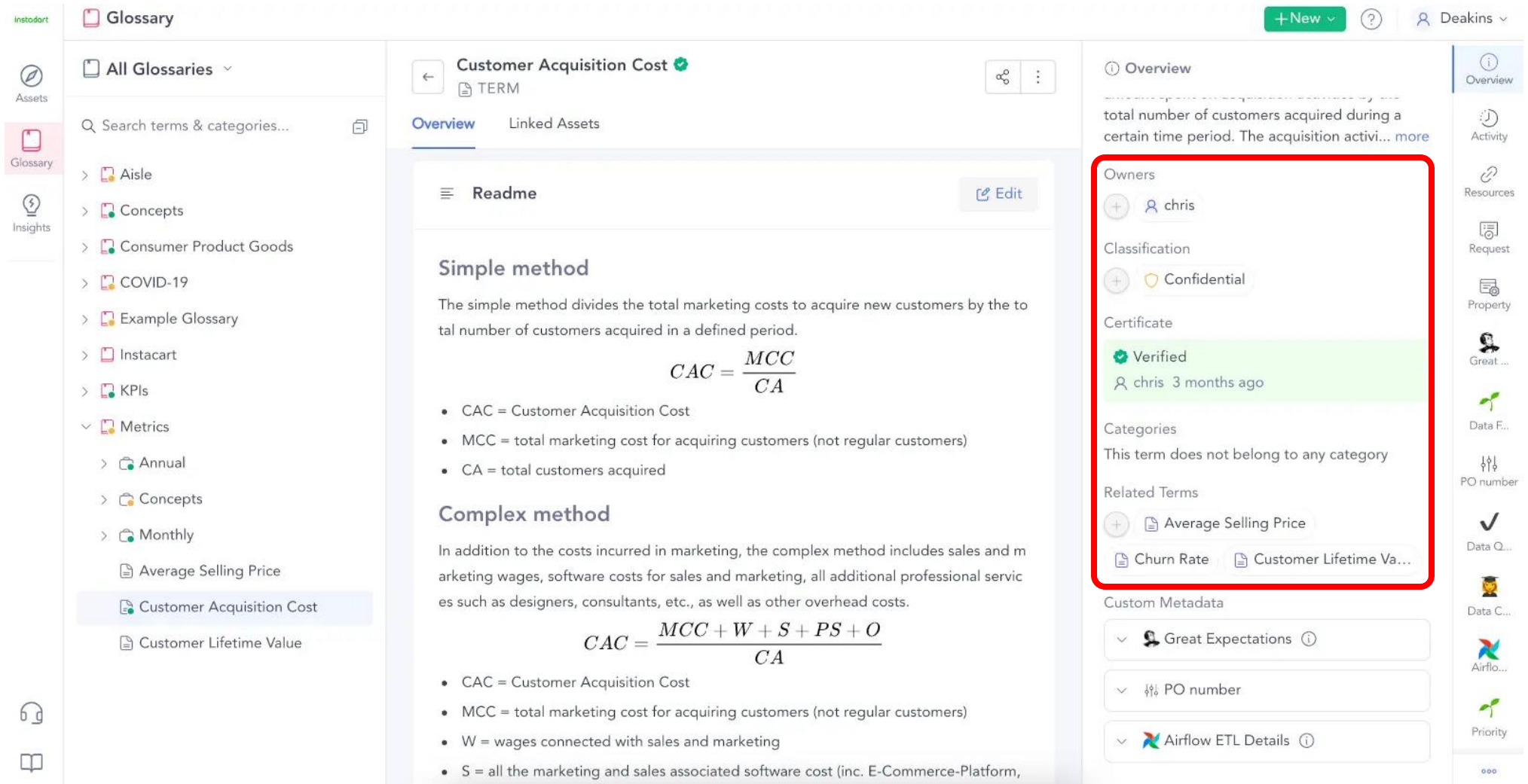
Data tags, classifications, mappings to business relationships, etc.

Social

Descriptive

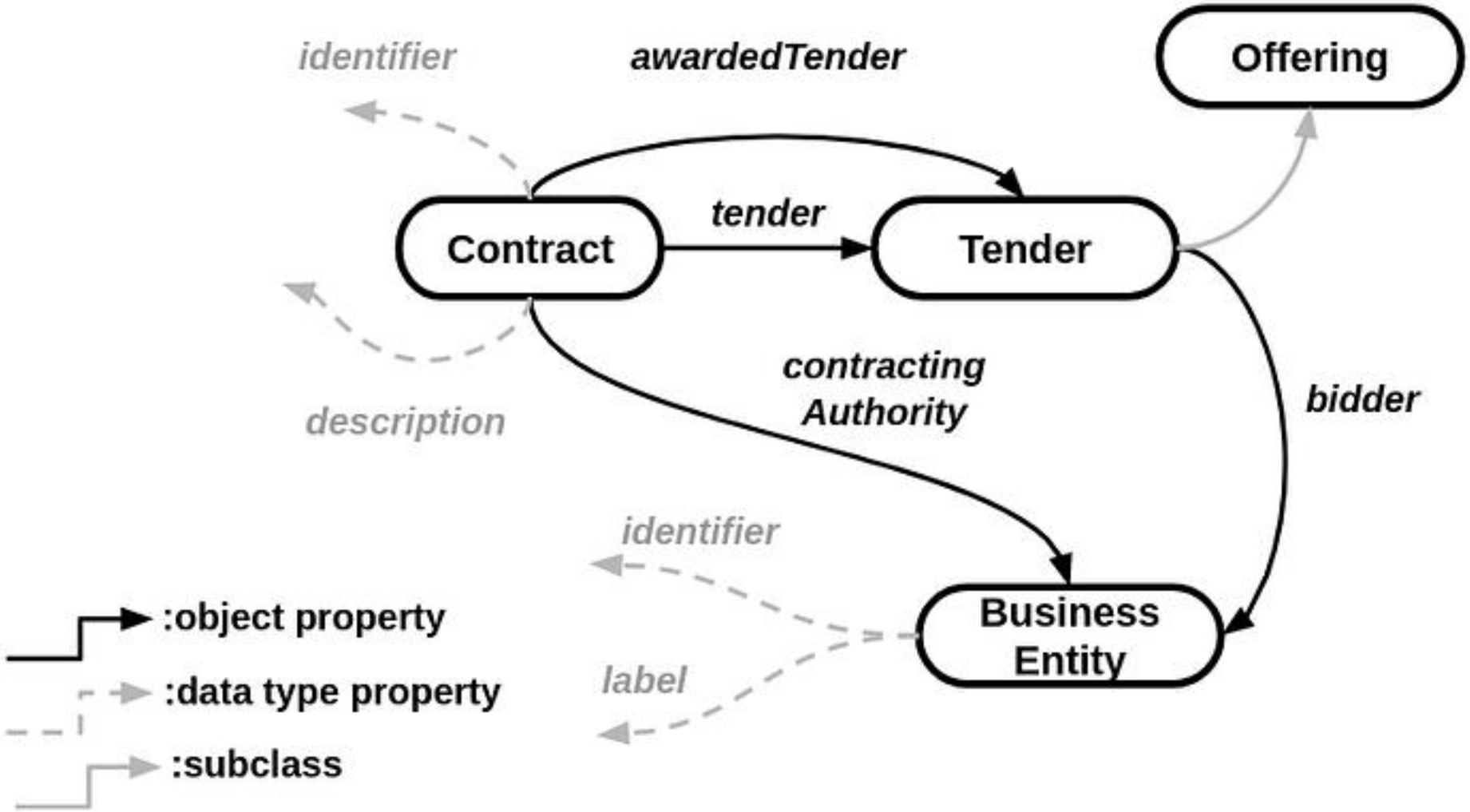
Metadata about user-generated content, business knowledge, etc.

Business Glossary



[Atlan – What is a Business Glossary]

Data Models


identifier
awardedTender
Offering
Contract
tender
Tender
description
contracting
Authority
bidder
identifier
Business
Entity
label
:object property
:data type property
:subclass

Types of Metadata

Technical

Definitional

Schemas, data types, models, etc.

Operational

Descriptive

How is data being produced? (Source / Transformations / …)

Business

Descriptive

Data tags, classifications, mappings to business relationships, etc.

Social

Descriptive

Metadata about user-generated content, business knowledge, etc.

Data models

Conceptual		Logical		Physical		Canonical	
Entity	Relationship	Entity	Attribute	File	Table	Type	Namespace
Glossary	Identifier	Model	Relationship	Column	Document	Schema	Attribute

Business capabilities		
Process	Organization	Application function

Semantics		
Business rule	Code definition	Attribute

Data objects	
Attributes	
Topic	Database
API	File

Dataset	
Elements	
Data domain	Data owner
Relationships	Definitions

Collaboration		
Comment	Tag	Like
Rating	Label	Note

Data quality	
Metric	Dimension
Score	Attribute

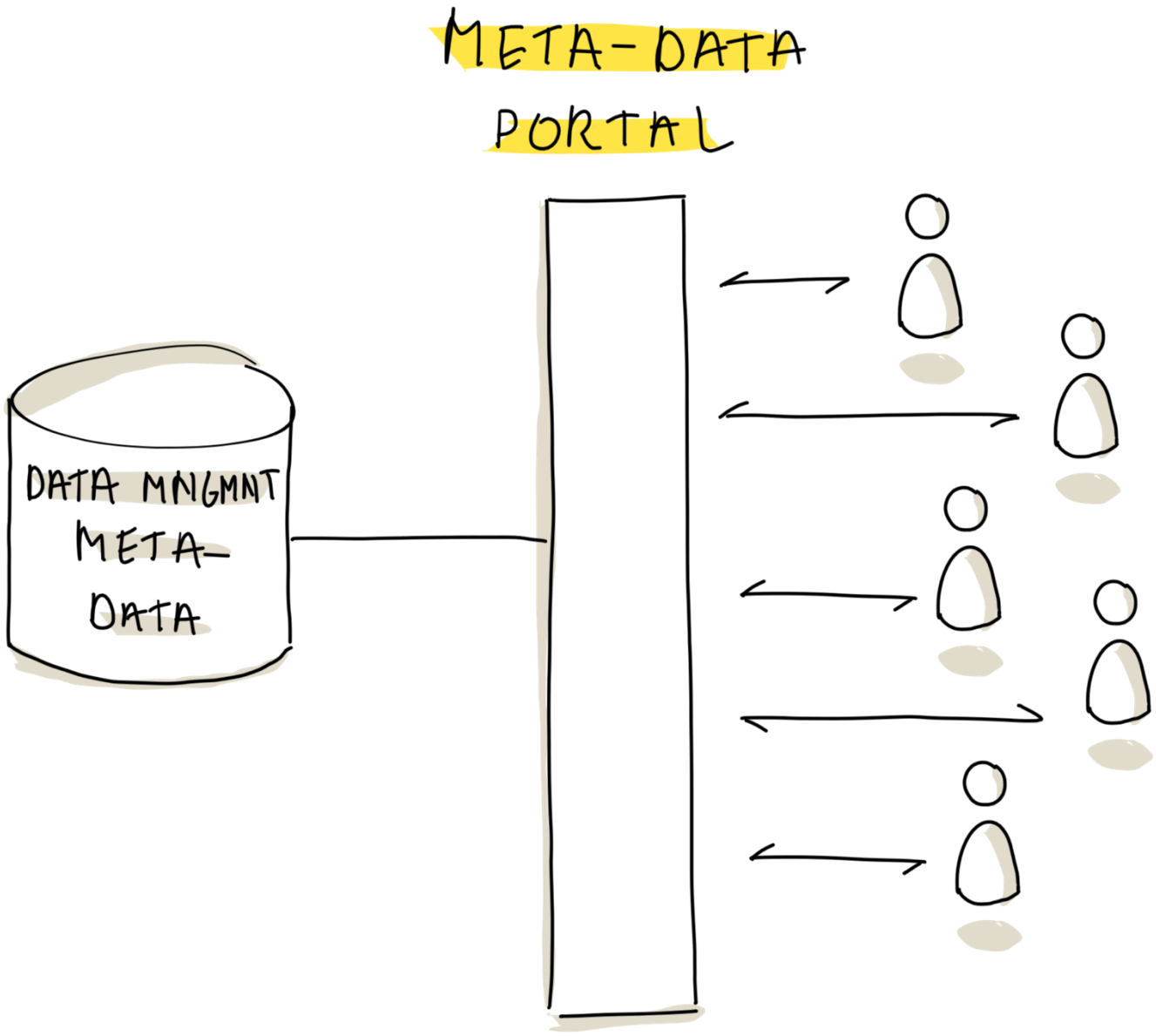
Lineage		
Process	Dataset	Attribute
Source:target		

Technology	
Platform	Datastore
Application	Data flow

Organization								
Process	Roles	Agreement	Owner	Purpose	Sensitivity	Confidentiality	Organization	SLA

Figure 10-2. An overview showing all the major metadata management subject areas for the enterprise. These subject areas are nonexhaustive. Each organization uses the areas differently.

[Datamanagement at Scale, Piethein Strengholt]


META-DATA
PORTAL
DATA MNGMNT
META-
DATA

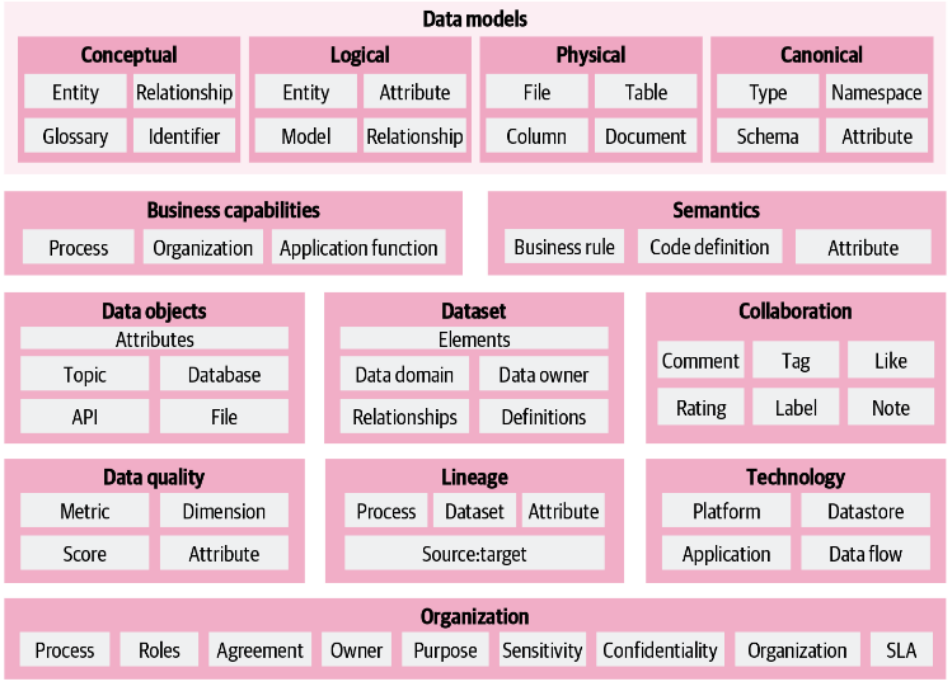
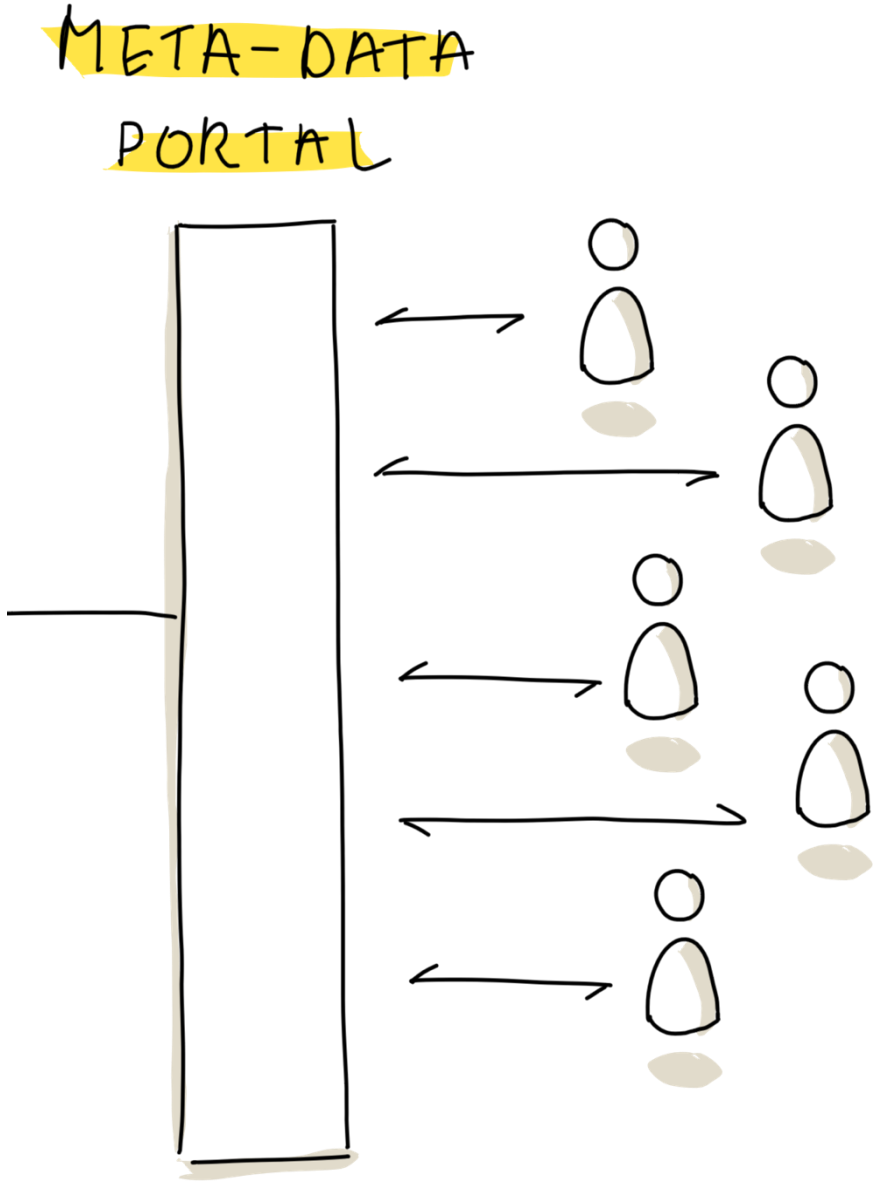
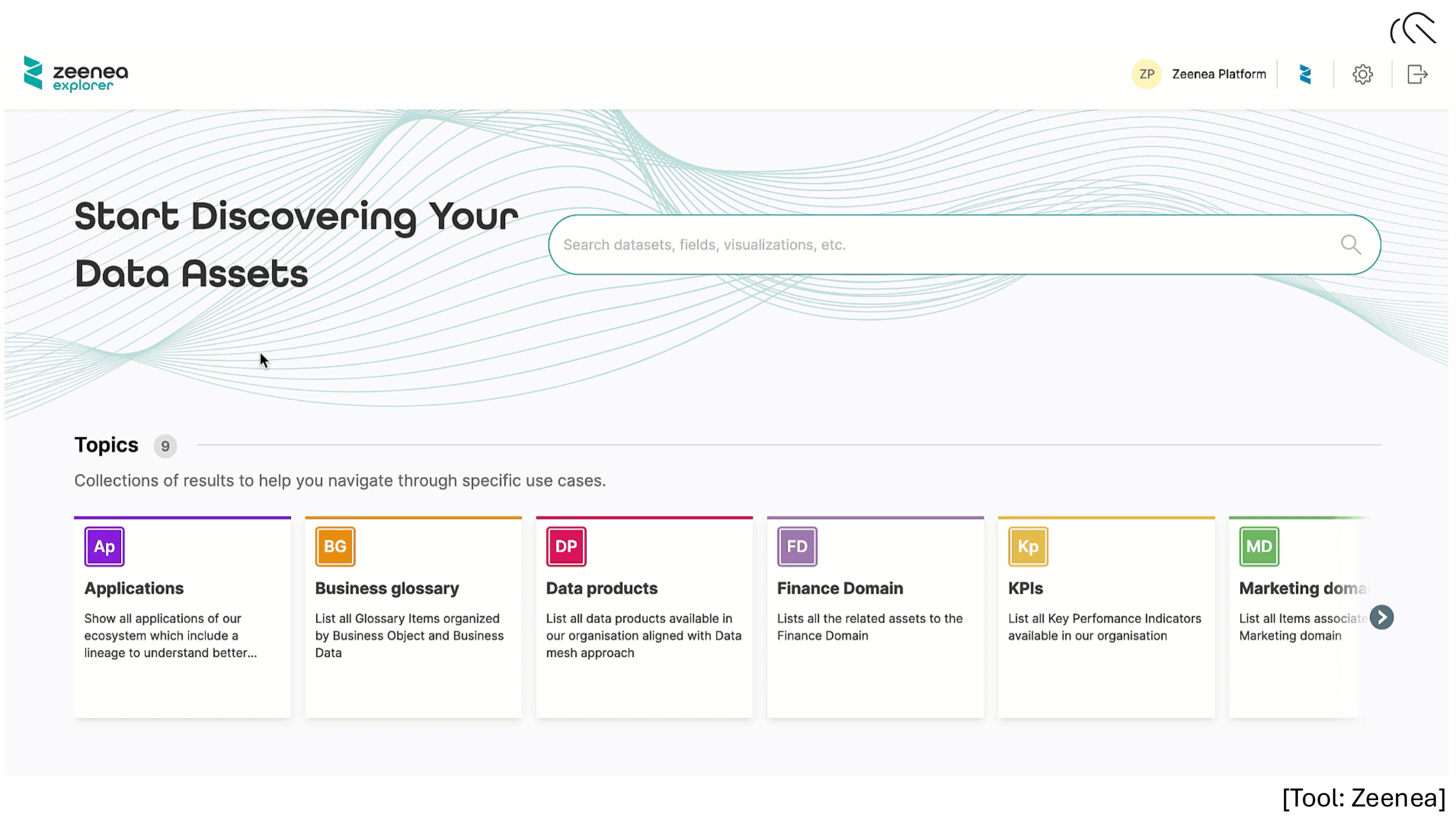
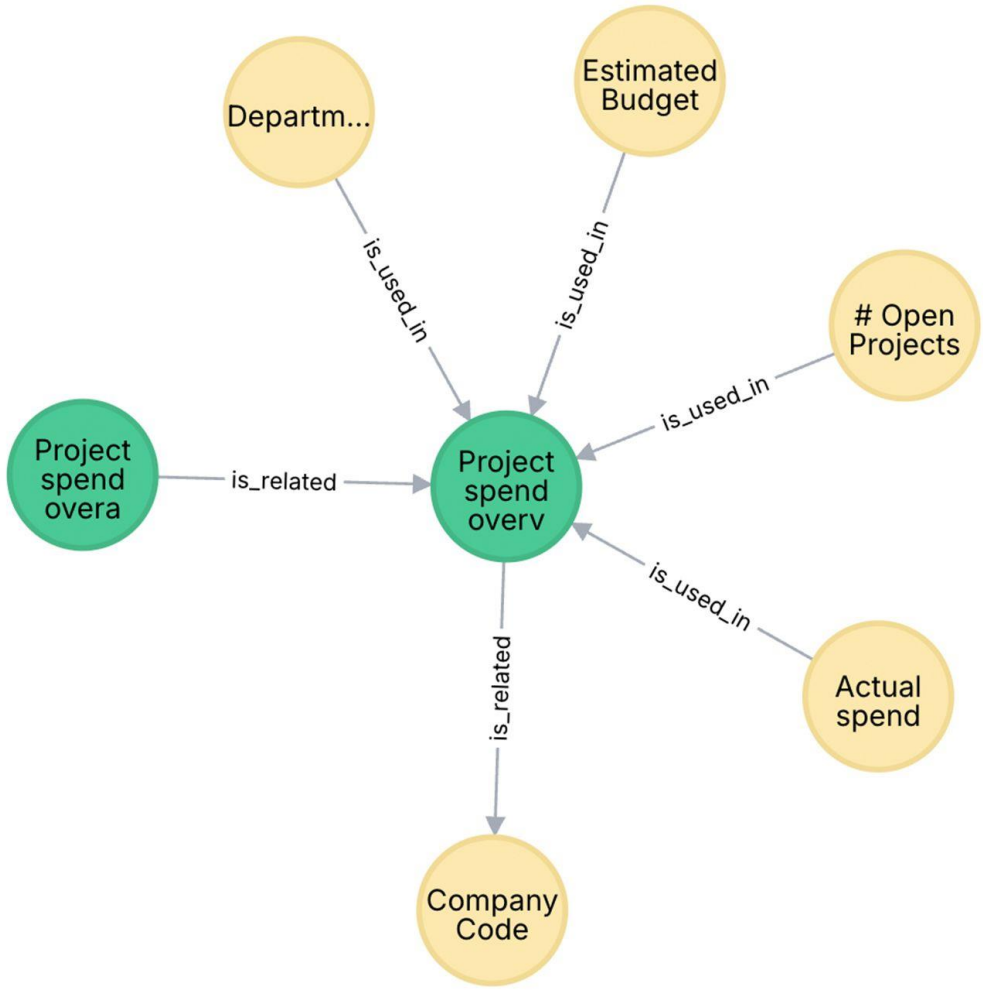


Figure 10-2. An overview showing all the major metadata management subject areas for the enterprise. These subject areas are nonexhaustive. Each organization uses the areas differently.




zeenea explorer
ZP
Zeenea Platform
Start Discovering Your Data Assets
Search datasets, fields, visualizations, etc.
Topics 9
Collections of results to help you navigate through specific use cases.
Ap
Applications
Show all applications of our ecosystem which include a lineage to understand better...
BG
Business glossary
List all Glossary Items organized by Business Object and Business Data
DP
Data products
List all data products available in our organisation aligned with Data mesh approach
FD
Finance Domain
Lists all the related assets to the Finance Domain
Kp
KPIs
List all Key Perfomance Indicators available in our organisation
MD
List all Items associate
Marketing domain

[Tool: Zeenea]



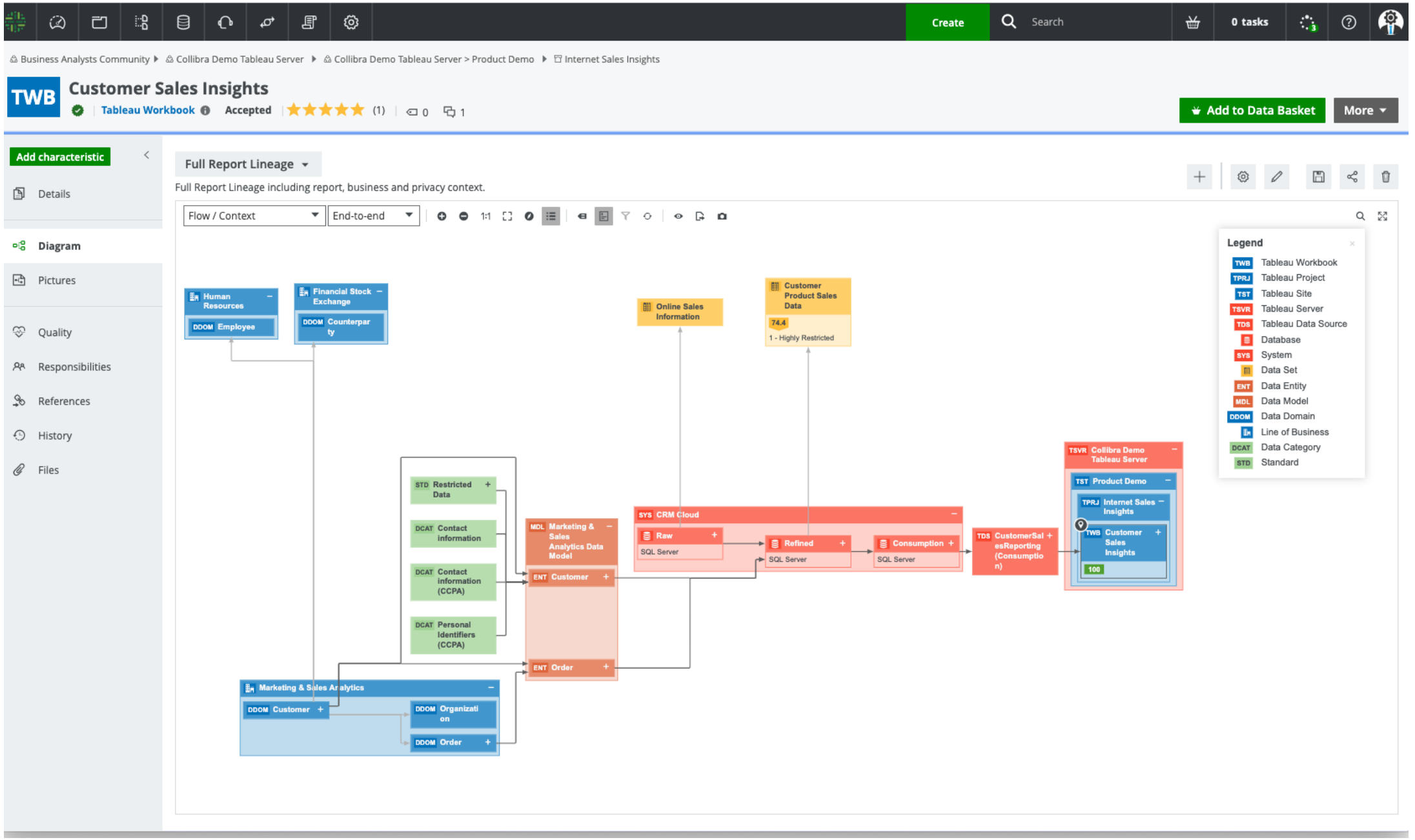
Asset Types

- [x] Report
- [x] Definition
- [x] Webpage

Relation Type

- [x] is used in / uses
- [x] is related to
- [x] is parent of / is child of
- [x] mentions / is mentioned in

[Tool: Dscribe]

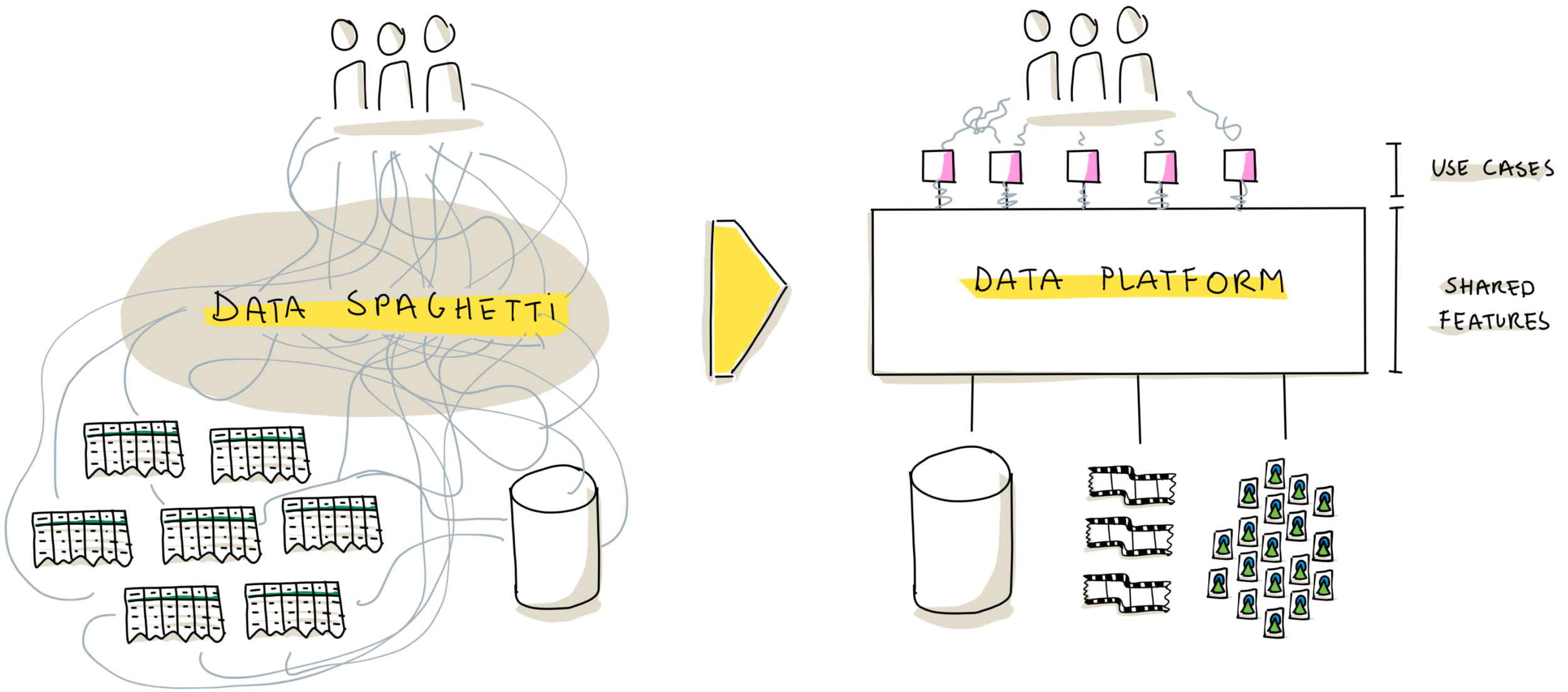


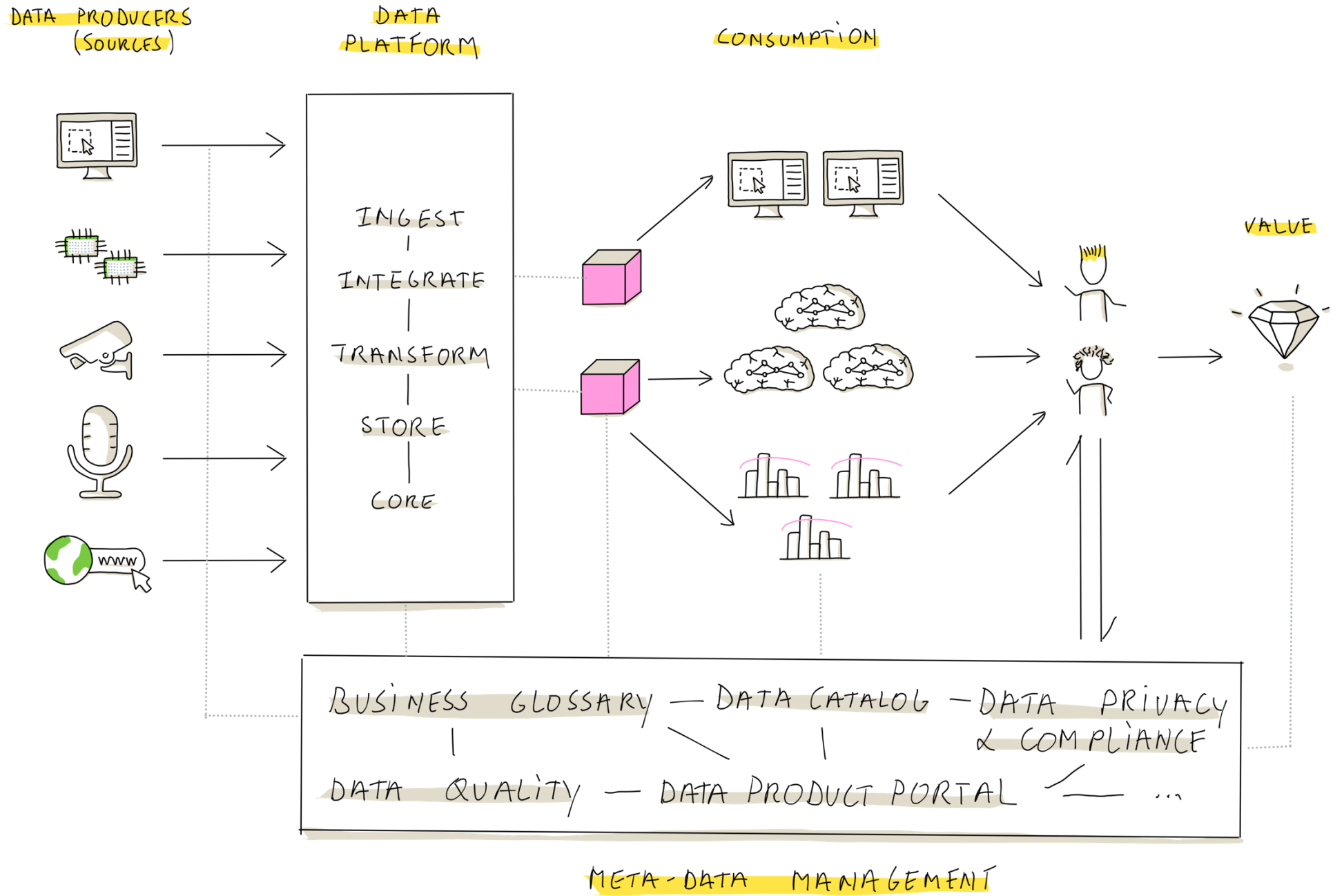
[Tool: Collibra]


Data Quality
Data Architecture
Meta Data
Data Modeling & Design
Data Warehousing & Business Intelligence
Data Governance
Data Storage & Operations
7
8
9
Reference & Master Data
Data Security
Documents & Content
Data Integration & Inter-operability

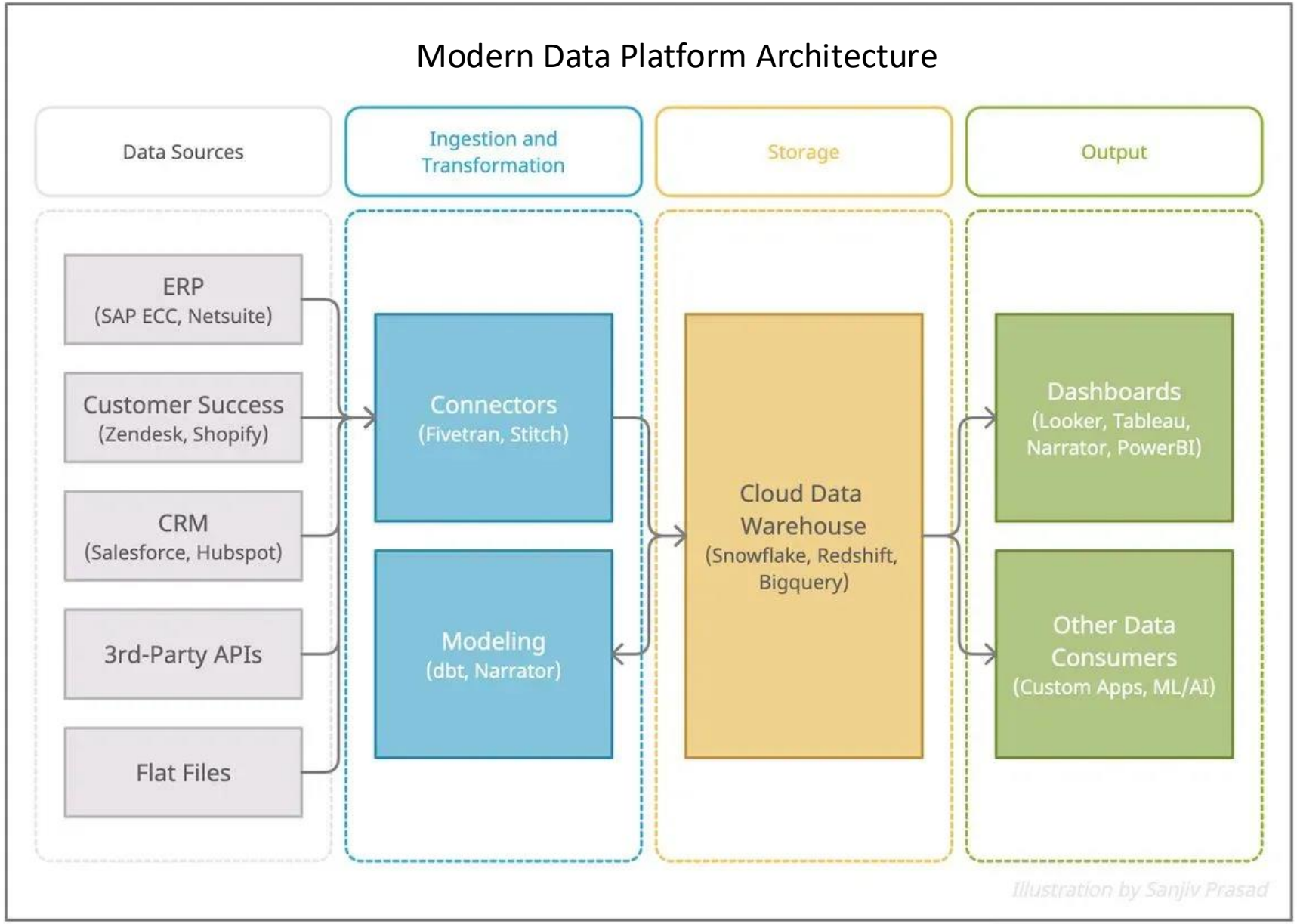
Data Management

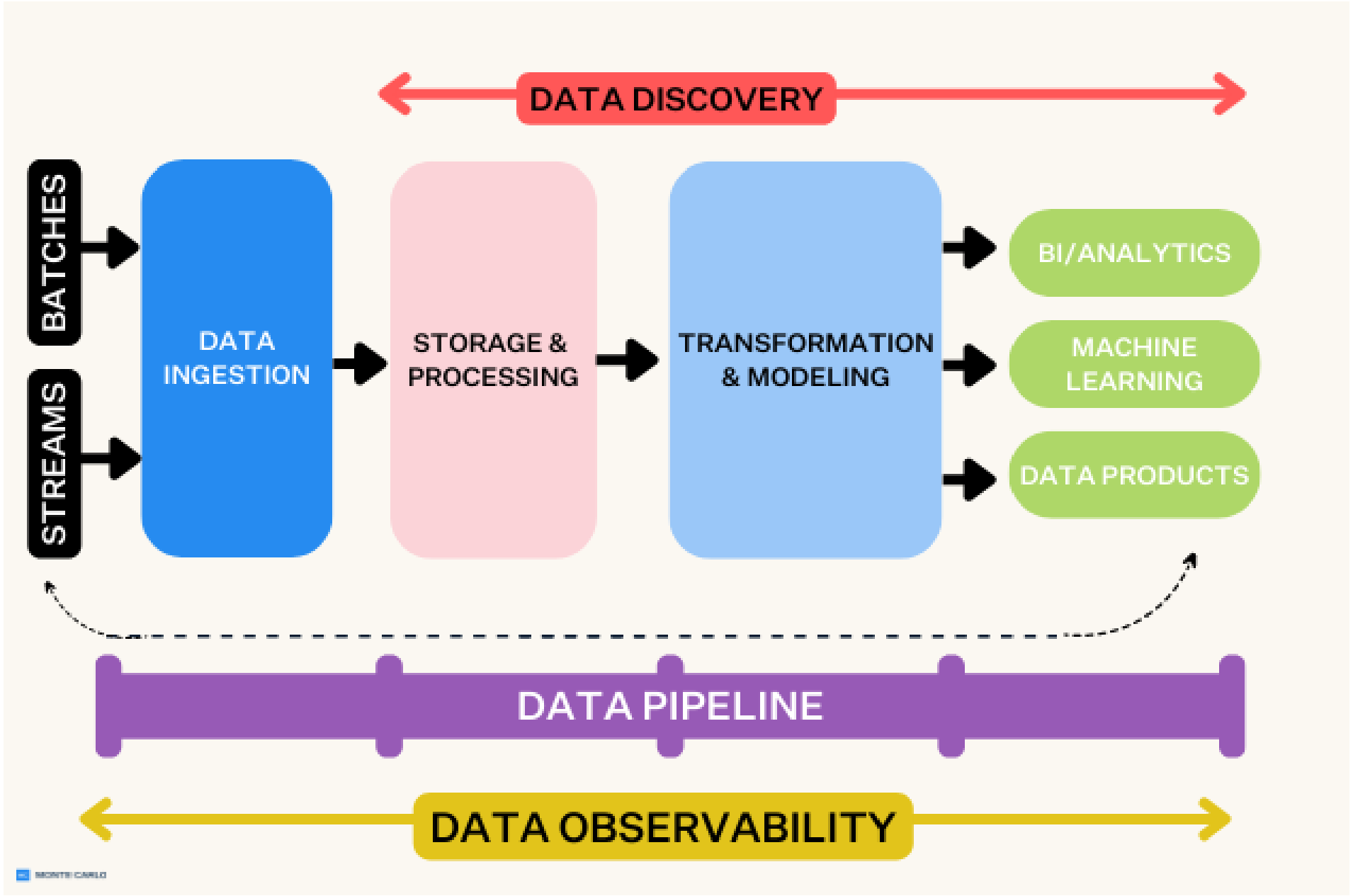
- Defining Data
- Data Producers
- Big Data Vs
- The Bigger Picture
- Data Management
- **Data Technology**


DATA SPAGHETTI
DATA PLATFORM
USE CASES
SHARED FEATURES

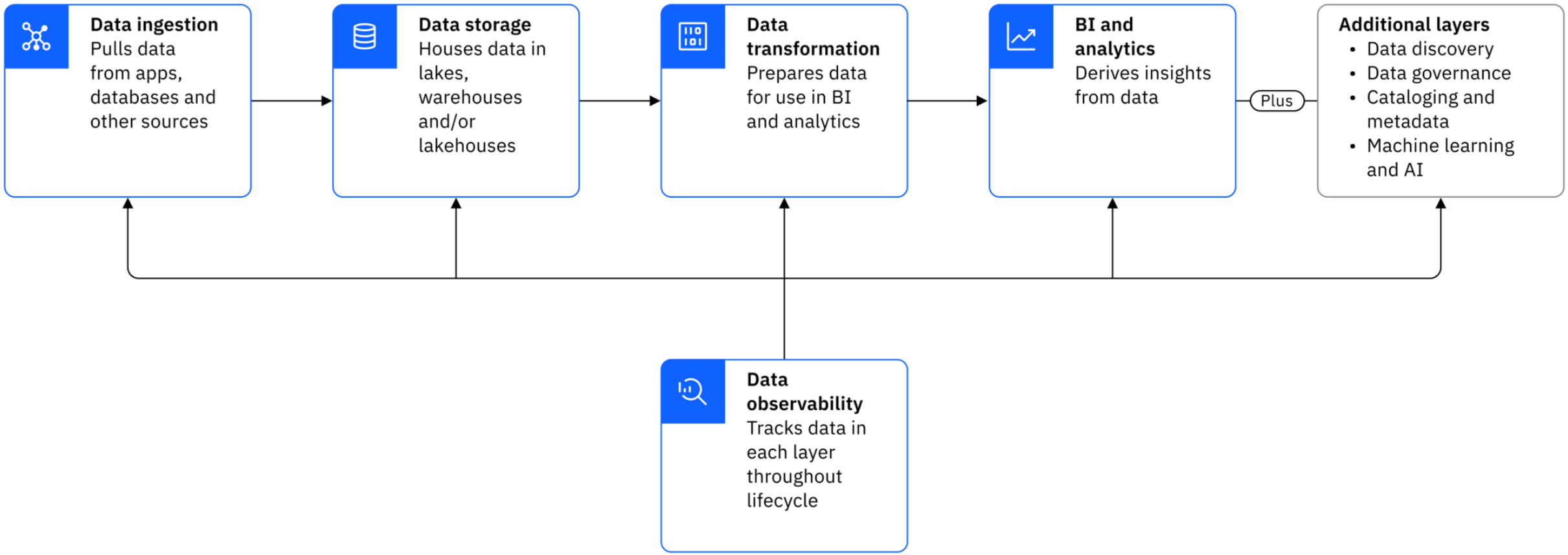

DATA PRODUCERS (SOURCES)
DATA PLATFORM
CONSUMPTION
VALUE
INGEST
INTEGRATE
TRANSFORM
STORE
CORE
www
BUSINESS GLOSSARY
DATA CATALOG
DATA PRIVACY & COMPLIANCE
DATA QUALITY
DATA PRODUCT PORTAL
...
META-DATA MANAGEMENT

Modern Data Platform Architecture

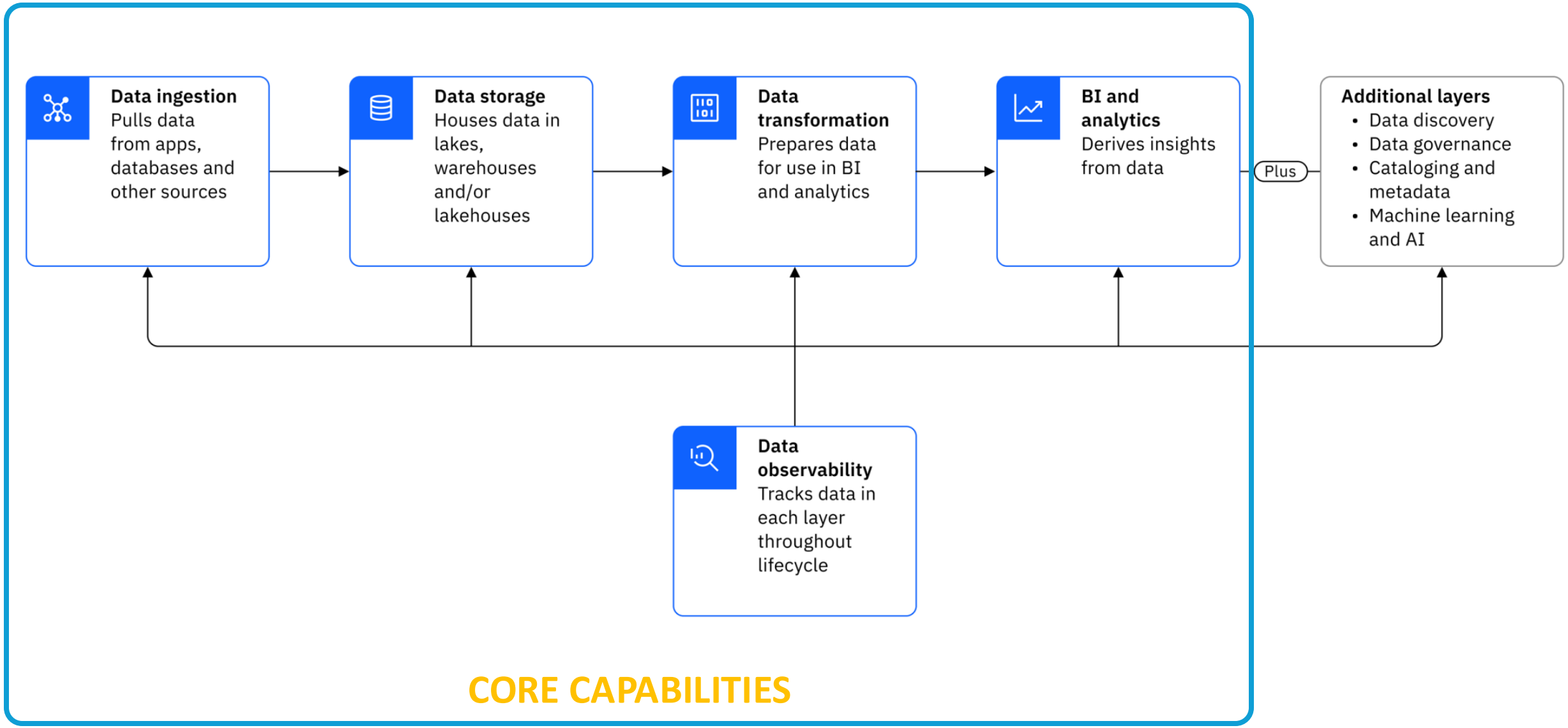




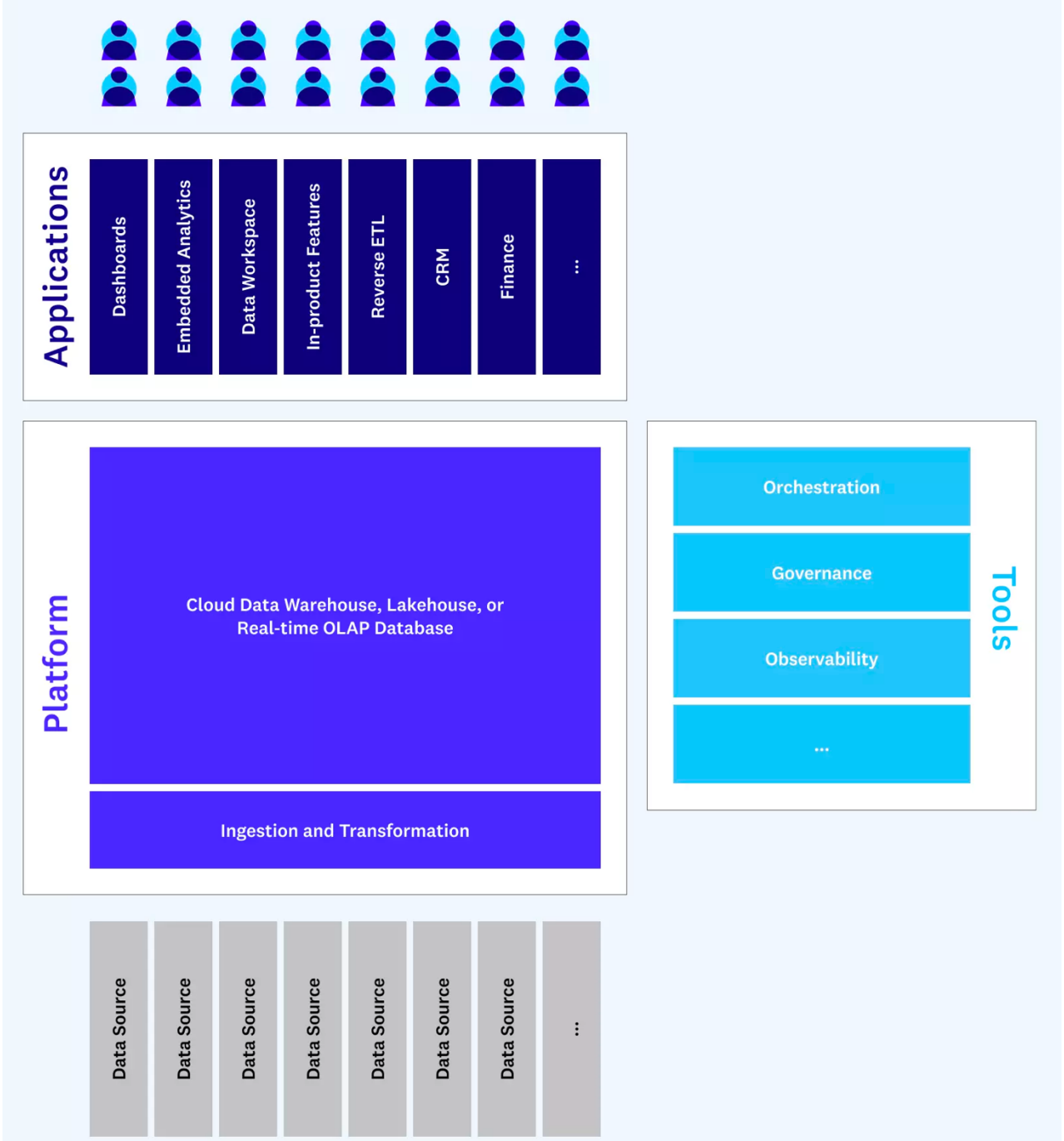
[Montecarlo – What is a Data Platform and how to Build One?]



[IBM: Data Platforms]



[IBM: Data Platforms]



[Atlan: What is a Data Platform]

[Atlan: What is a Data Platform]

Data Management

- Defining Data
- Data Producers
- Big Data Vs
- The Bigger Picture
- Data Management
- Data Technology