

DATA STRATEGIE

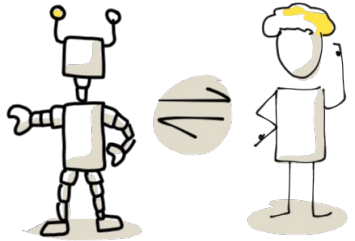


Jan Meskens
01 - 2024





Phd Human
Computer Interaction



Inspiring

SIEVAX



Consultant
/ Advisor



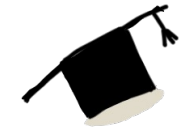
●●● Medium
@meskensjan



janmeskens



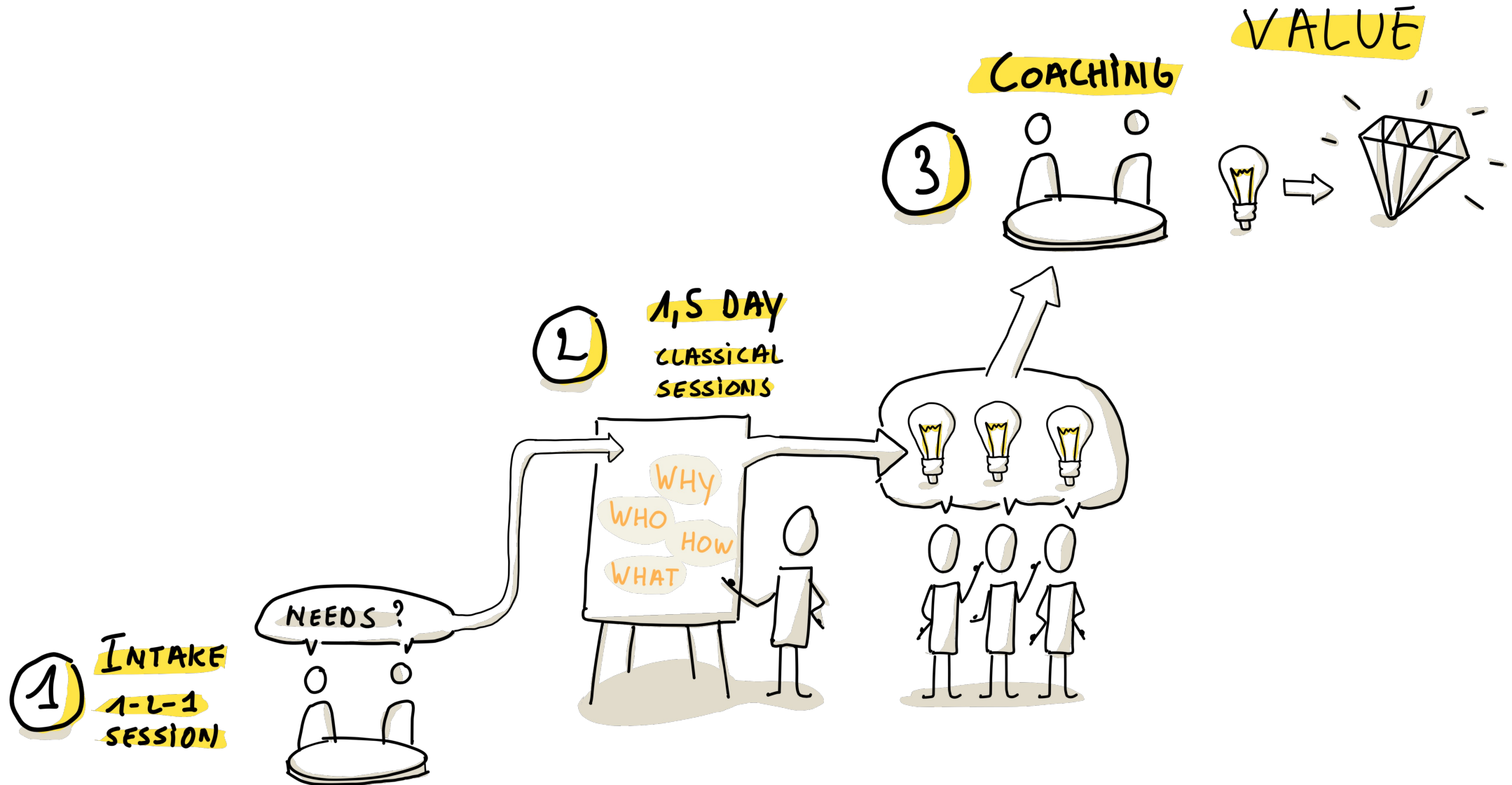
Jan@Sievax.be



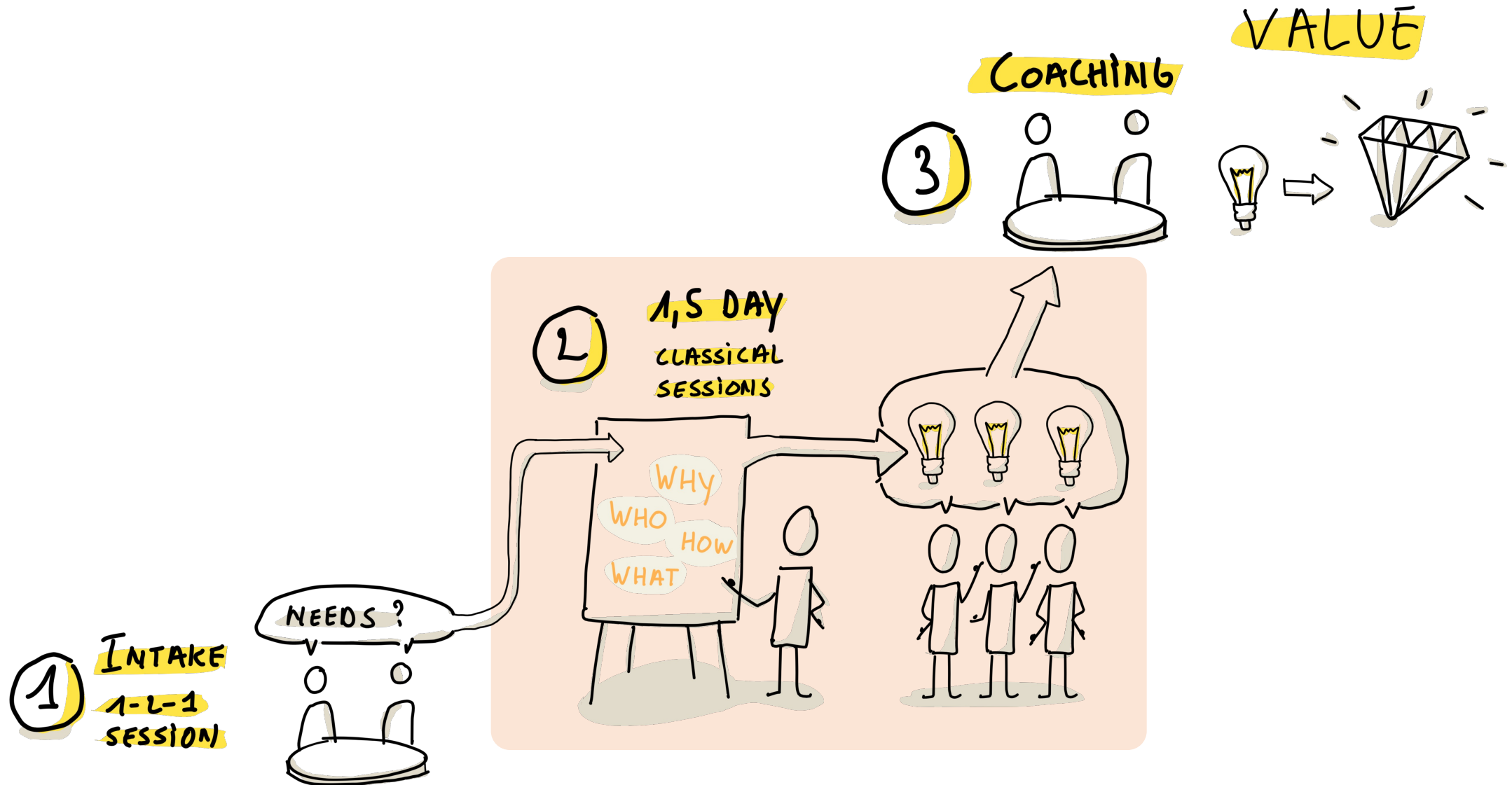
Coaching & Training



TRAINING OVERVIEW



TRAINING OVERVIEW

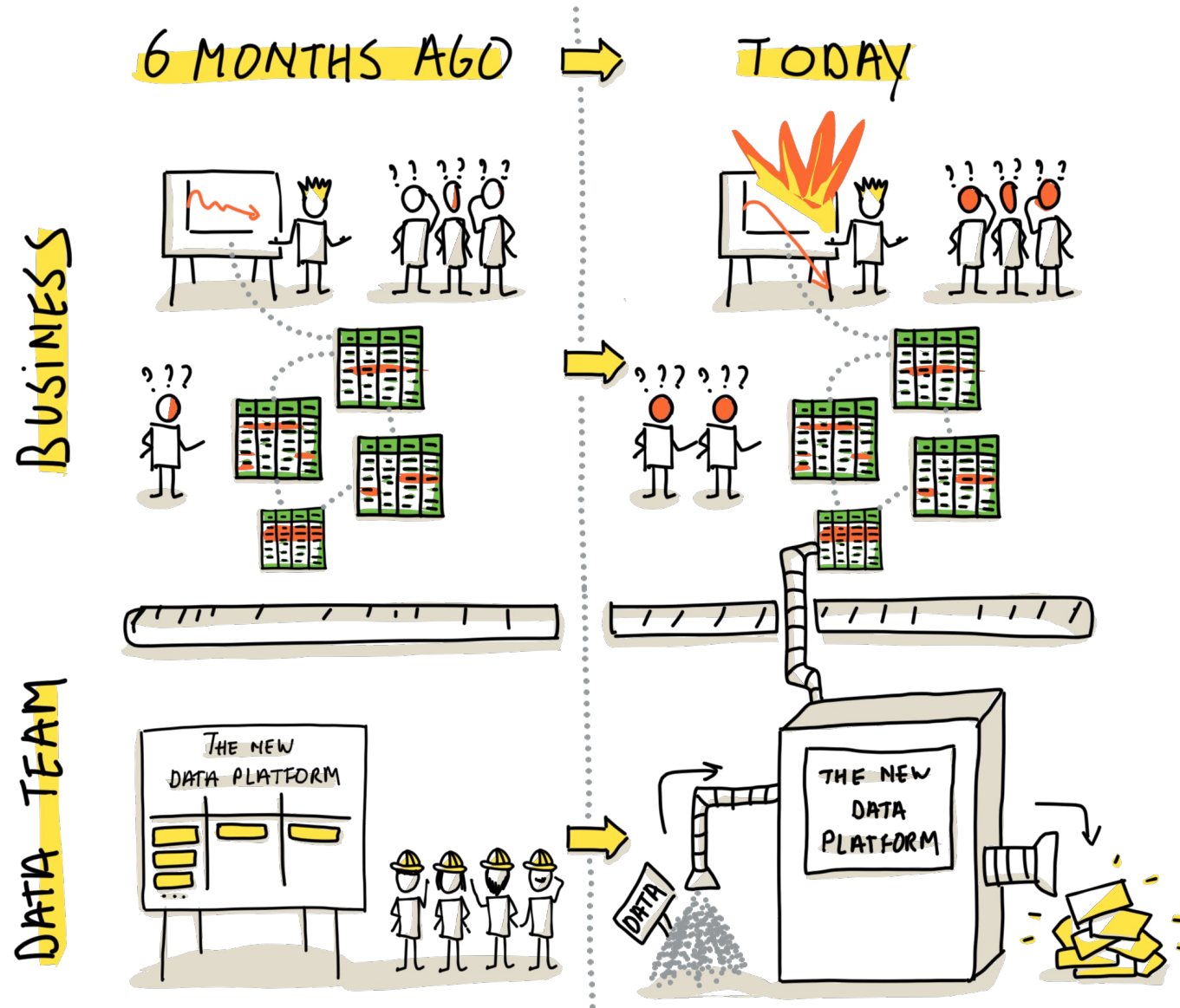


DATA STRATEGY

A scenic mountain landscape with a winding road and a small village. The foreground shows a paved road that curves through a valley. In the middle ground, there are rolling hills and a small cluster of buildings. The background features majestic mountains under a cloudy sky.

WHY THIS TRAINING?

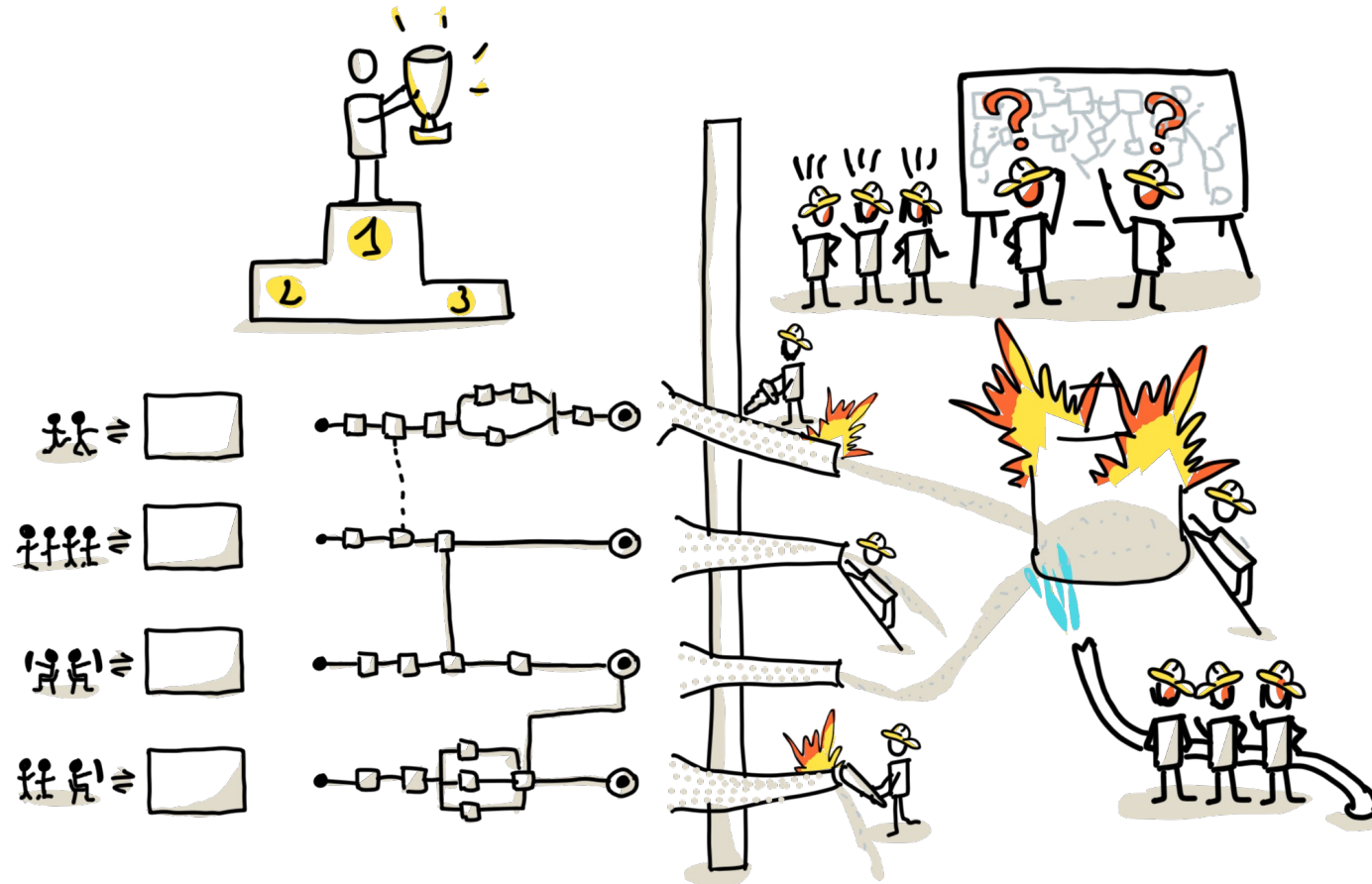
EXPERIENCE 1: BUILDING SOMETHING THAT ISN'T USED



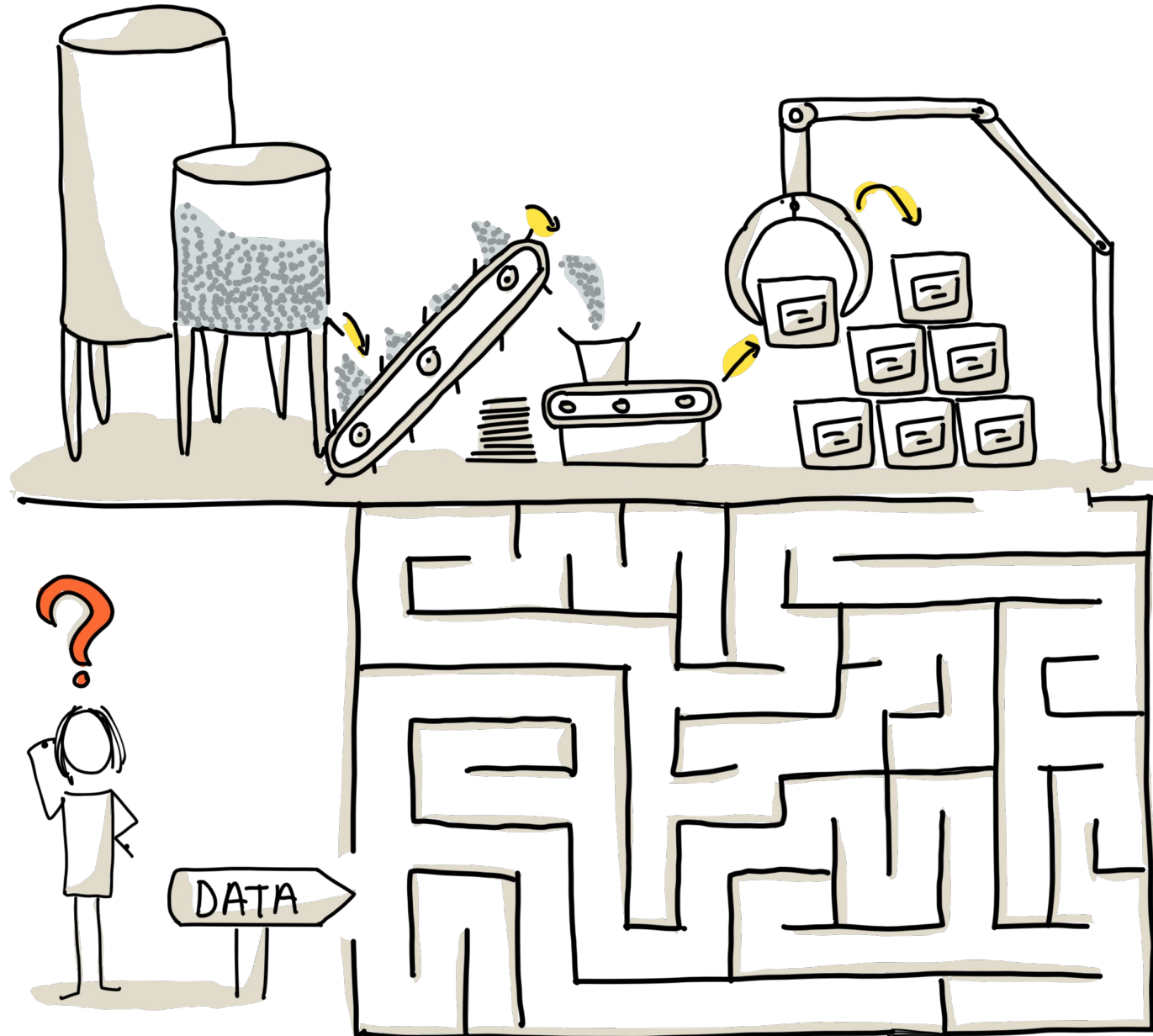
EXPERIENCE 2: DIGITAL TRANSFORMATIONS WITHOUT DATA

PARTY
IN THE FRONT

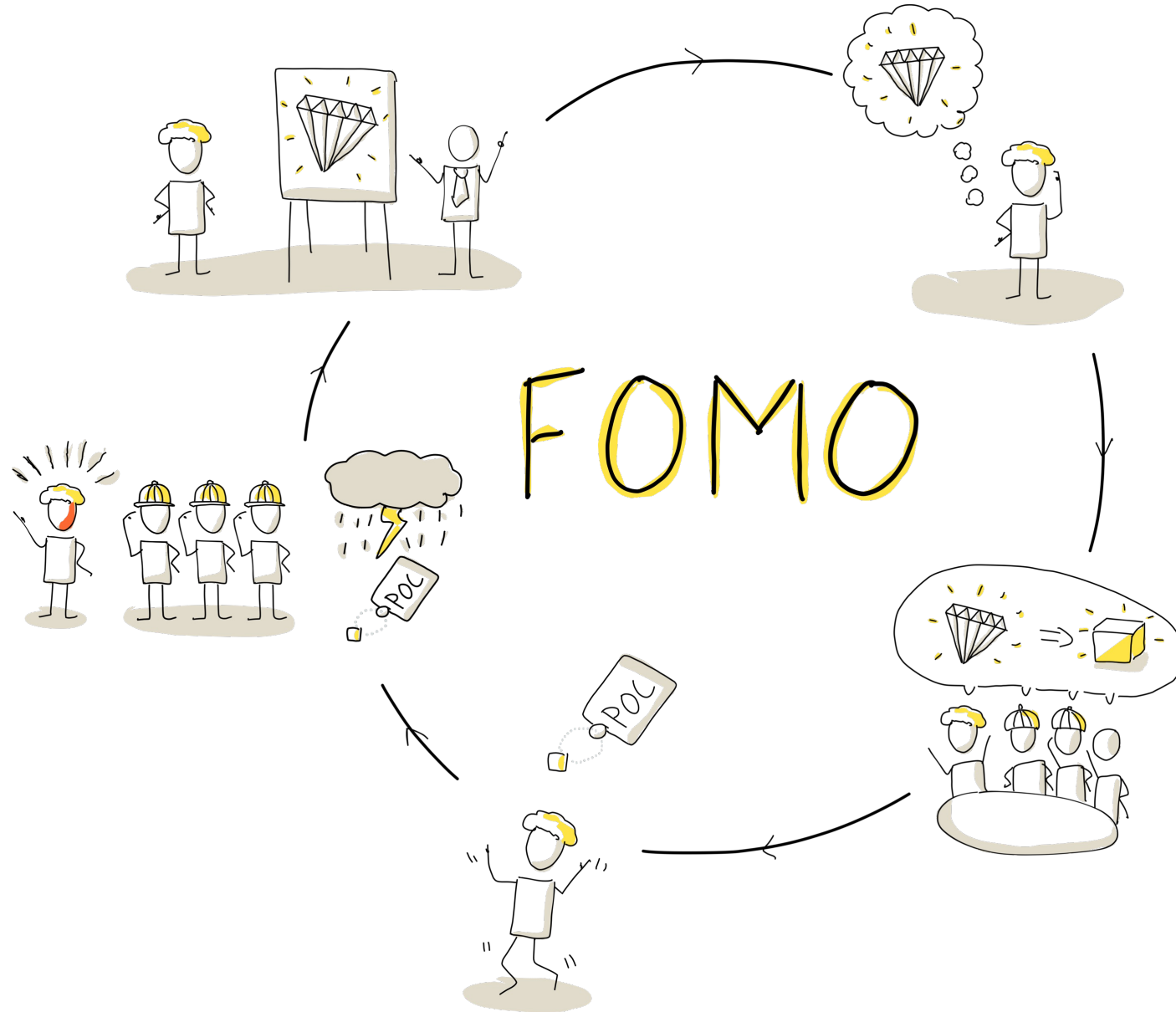
FIREFIGHTING
IN THE BACK



EXPERIENCE 3: DATA IS NEEDED BUT CAN'T BE FOUND



EXPERIENCE 4: FEAR OF MISSING OUT CYCLE



EXPERIENCE 4: FEAR OF MISSING OUT CYCLE



A scenic mountain landscape with a winding road and a small village. The foreground shows a paved road that curves through a valley. In the middle ground, there is a small cluster of buildings, possibly a village or a farmstead. The background features rolling hills and mountains under a hazy sky. The overall tone is serene and natural.

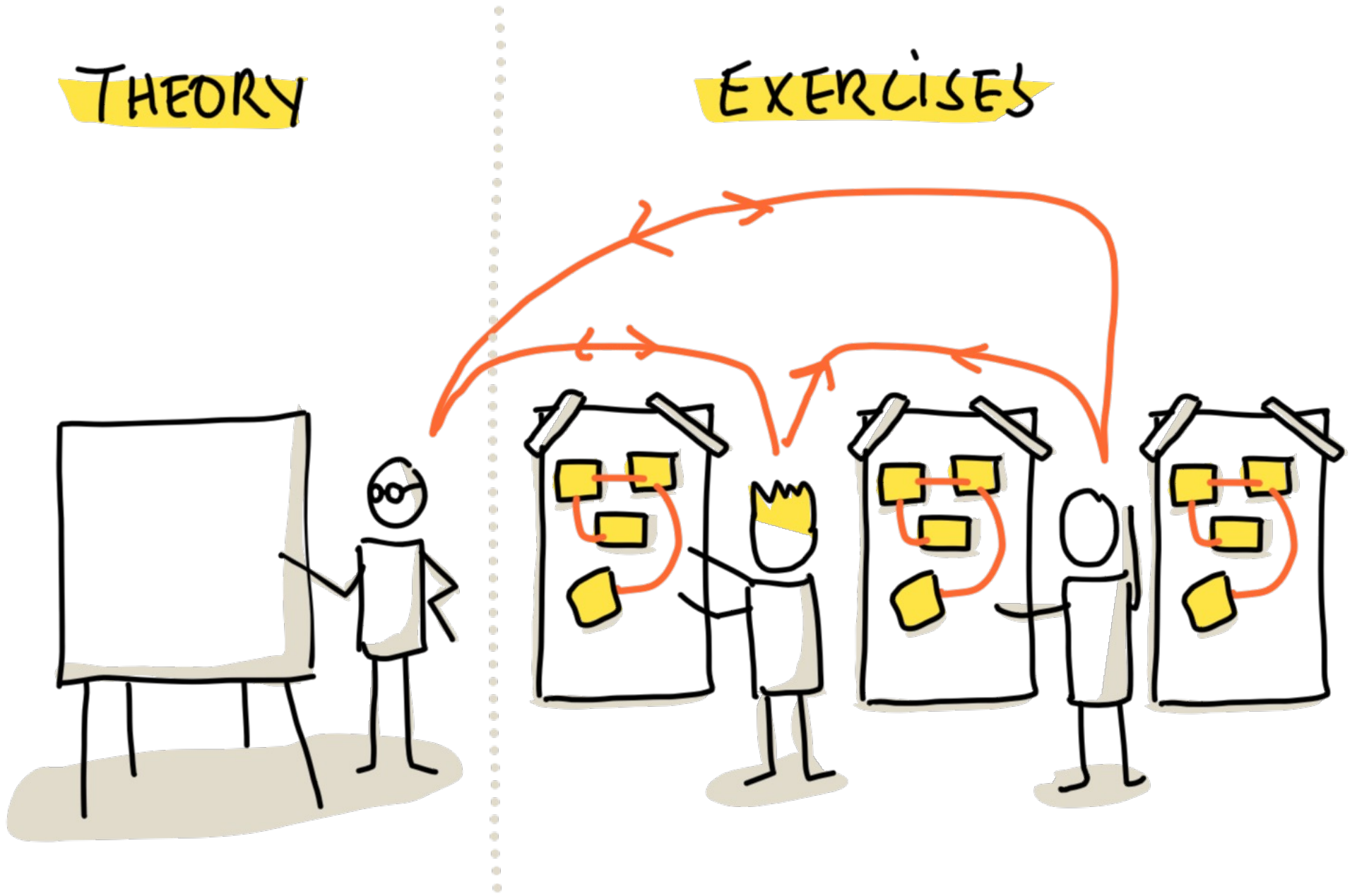
**WE SHOULD BE ABLE
TO DO BETTER!**

DATA STRATEGY TRAINING

APPROACH

THEORY

EXERCISES



Nov 7, 2023

Data Driven Management: The Why, Who, What and How?

While many organizations aspire to become data-driven, a significant portion of them tends to focus narrowly on the technical aspects of data, treating it primarily as a technical asset. Consequently, their...



Data Strategy 10 min read



Aug 23, 2023

Once You POC, You (Can't) Stop!

Every week, the realm of data technology introduces fresh products and services. Each tool arrives with an enticing proposition: "Embrace the new and fancy Product <X> – your ultimate problem solver!" And...



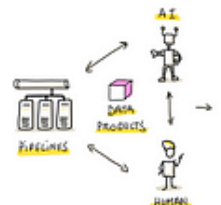
Data Strategy 6 min read



Aug 9, 2023

Data Usability: How to Build Better Data Products?

Looking ahead to the future, the utilization of data is poised to become a prevailing norm, empowering intelligent workflows and effortless human-machine interactions. McKinsey's projection posits that by...



Data Strategy 8 min read



Published in **The Modern Scientist** · Jan 17

Data Ingestion — Part 2: Tool Selection Strategy

This article is the second one in my series on data ingestion. For an introduction to the topic and to explore 'data ingestion patterns', you can refer to my first article in this series. Data ingestion tools play a...



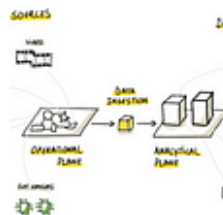
Data Engineering 15 min read



Published in **The Modern Scientist** · Nov 27, 2023

Data Ingestion — Part 1: Architectural Patterns

Over the course of two articles, I will thoroughly explore data ingestion, a fundamental process that bridges the operational and analytical worlds. Ingestion is critical for transporting data from a...

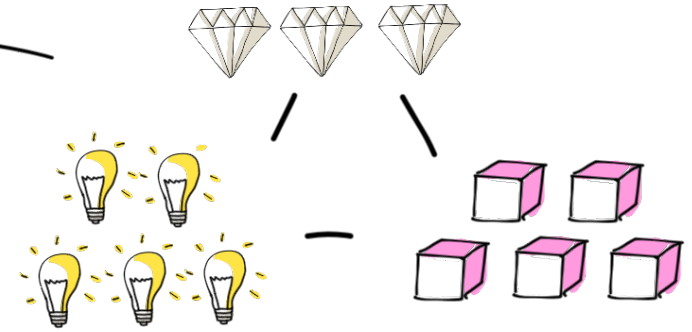
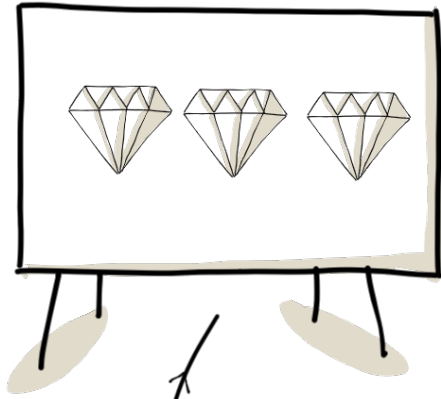


Data Engineering 11 min read



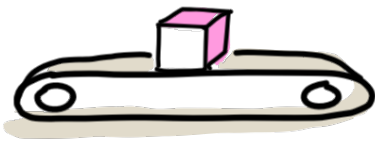
TABLE OF CONTENTS

BUSINESS STRATEGY



DATA PRODUCT

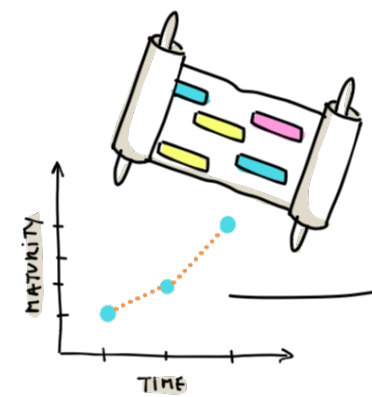
WHY-WHAT-WHO-HOW



DELIVERY

	1	2	3	4	5
CRIT 1		●			
CRIT 2			●		
CRIT 3	●				
CRIT 4				●	

DATA MATURITY

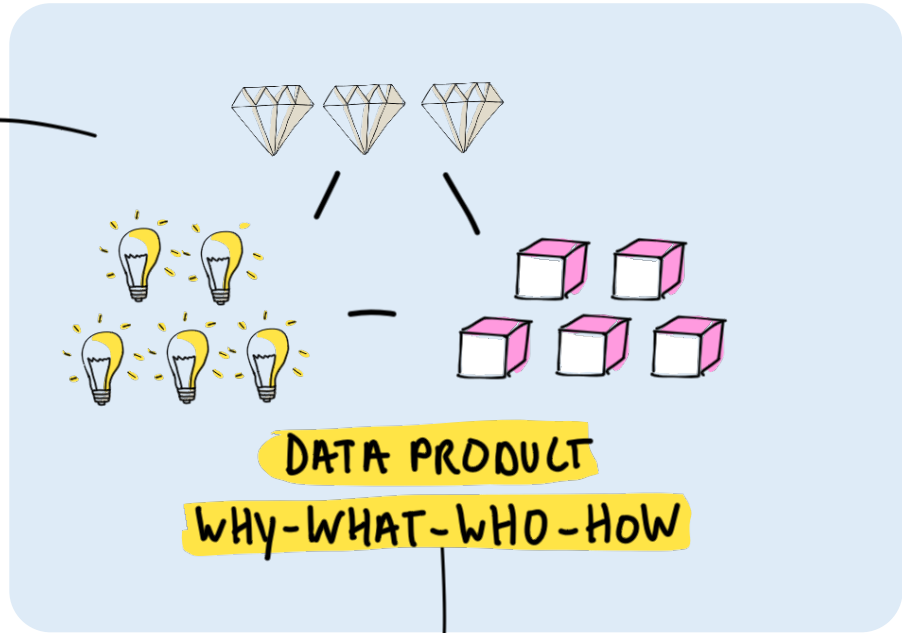
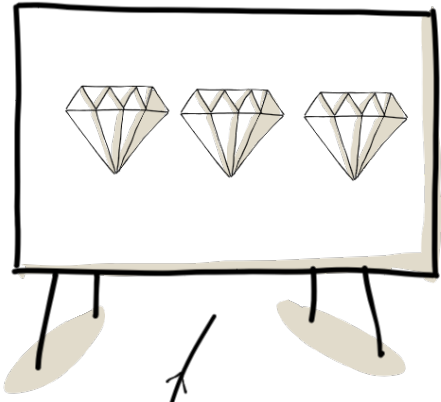


STRATEGY + ROADMAP



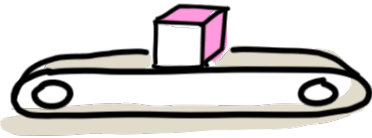
TABLE OF CONTENTS

BUSINESS STRATEGY



DATA PRODUCT

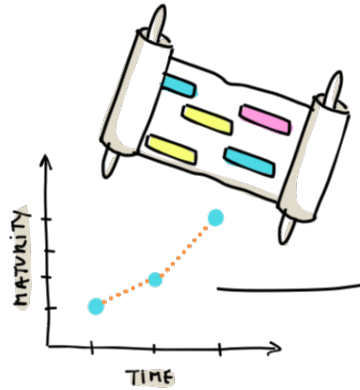
WHY-WHAT-WHO-HOW



DELIVERY

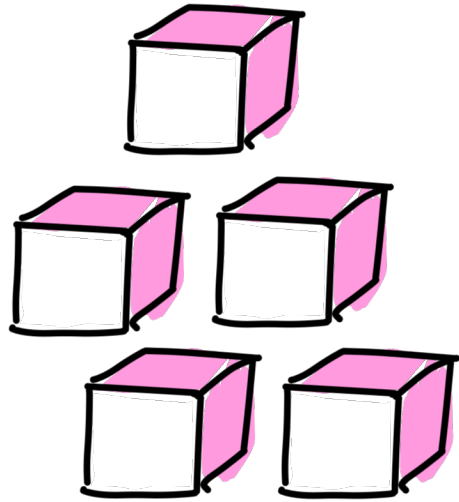
	1	2	3	4	5
CRIT 1		●			
CRIT 2			●		
CRIT 3	●				
CRIT 4				●	

DATA MATURITY



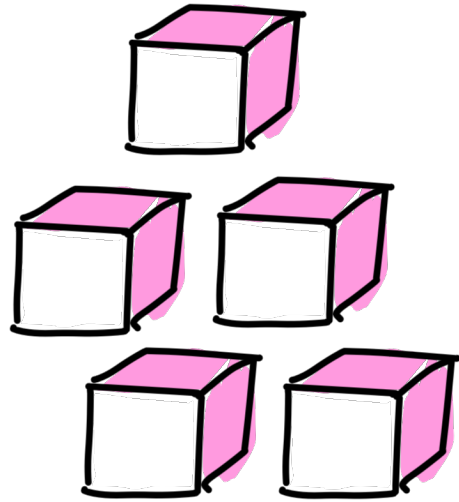
STRATEGY + ROADMAP





DATA PRODUCTS

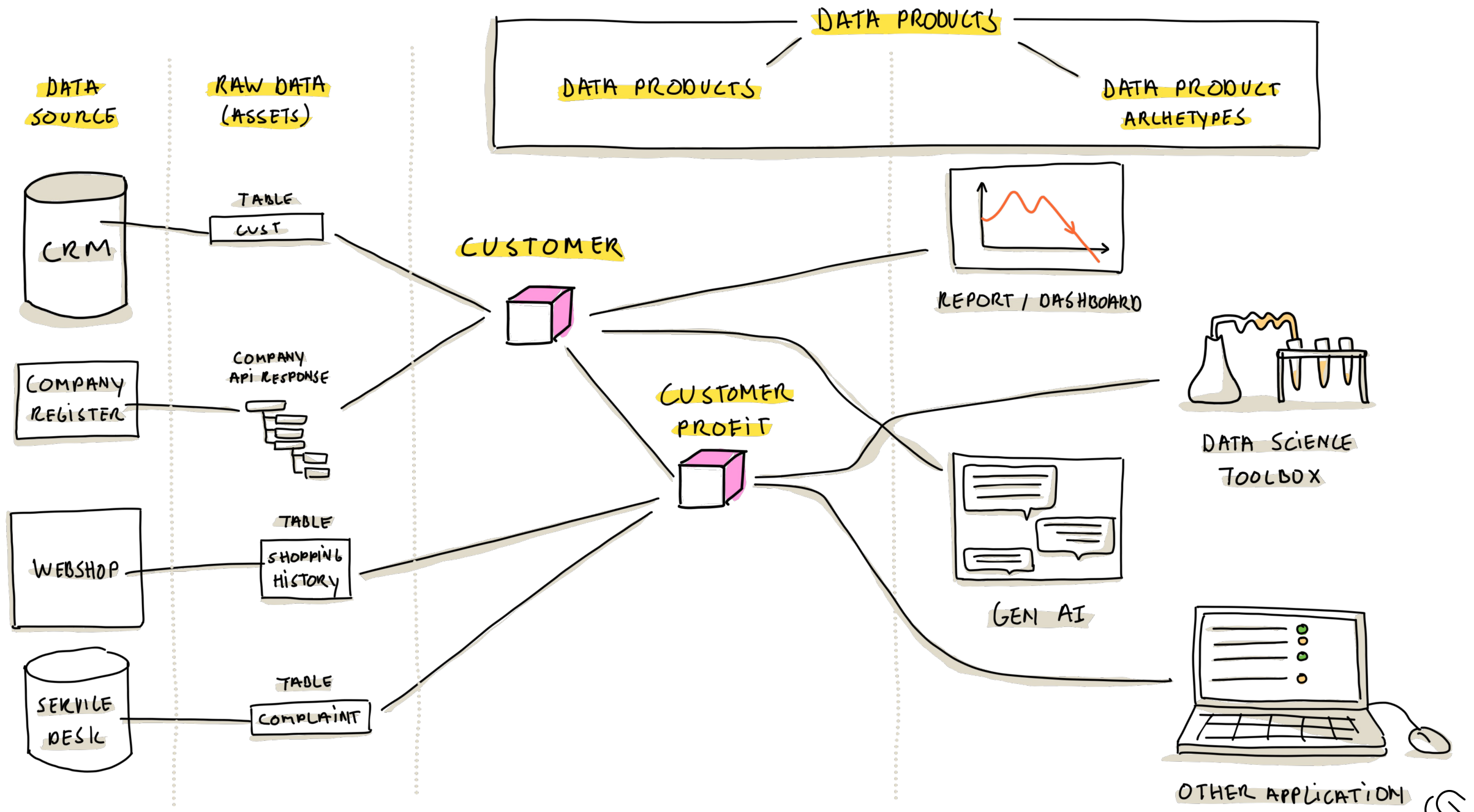




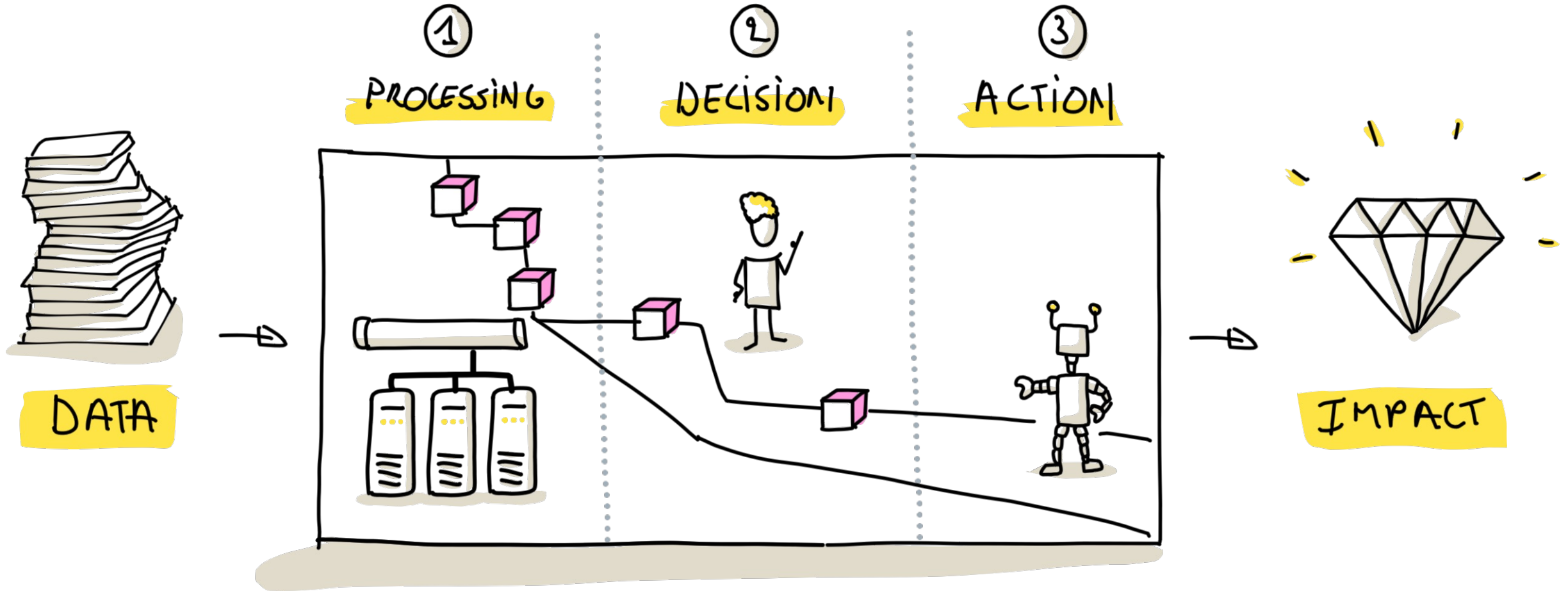
DATA PRODUCTS

= A **REUSABLE** DATA ASSET THAT BUNDLES **DATA** TOGETHER WITH **EVERYTHING NEEDED** TO MAKE IT INDEPENDENTLY **USABLE** BY AUTHORIZED CONSUMERS.

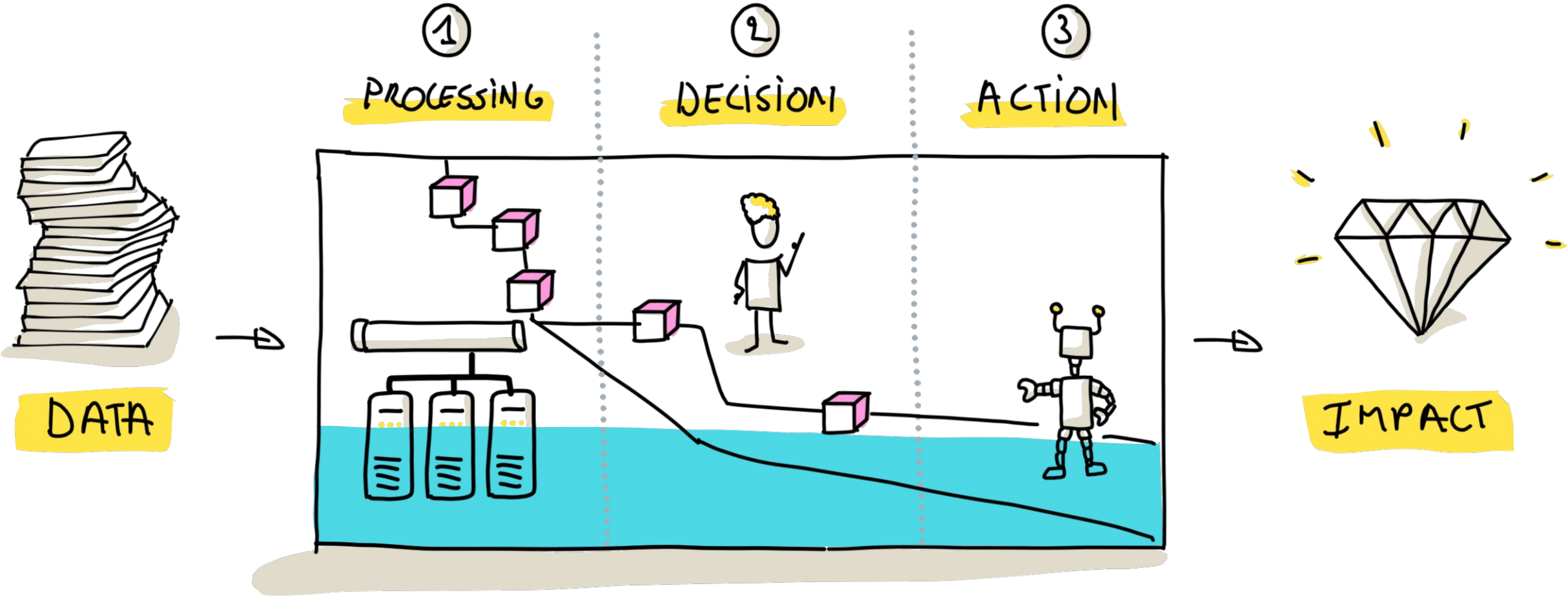




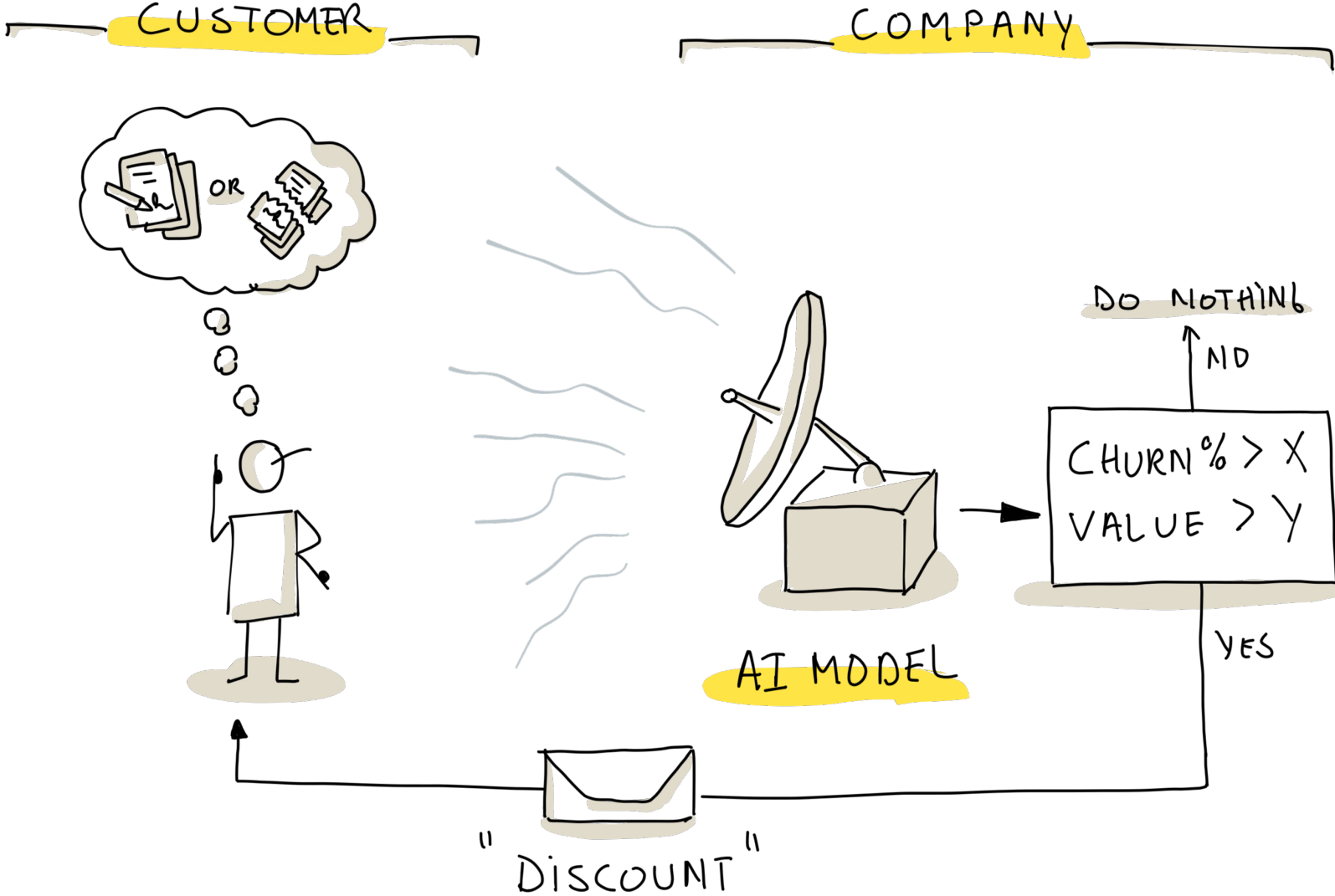
Data Product Types



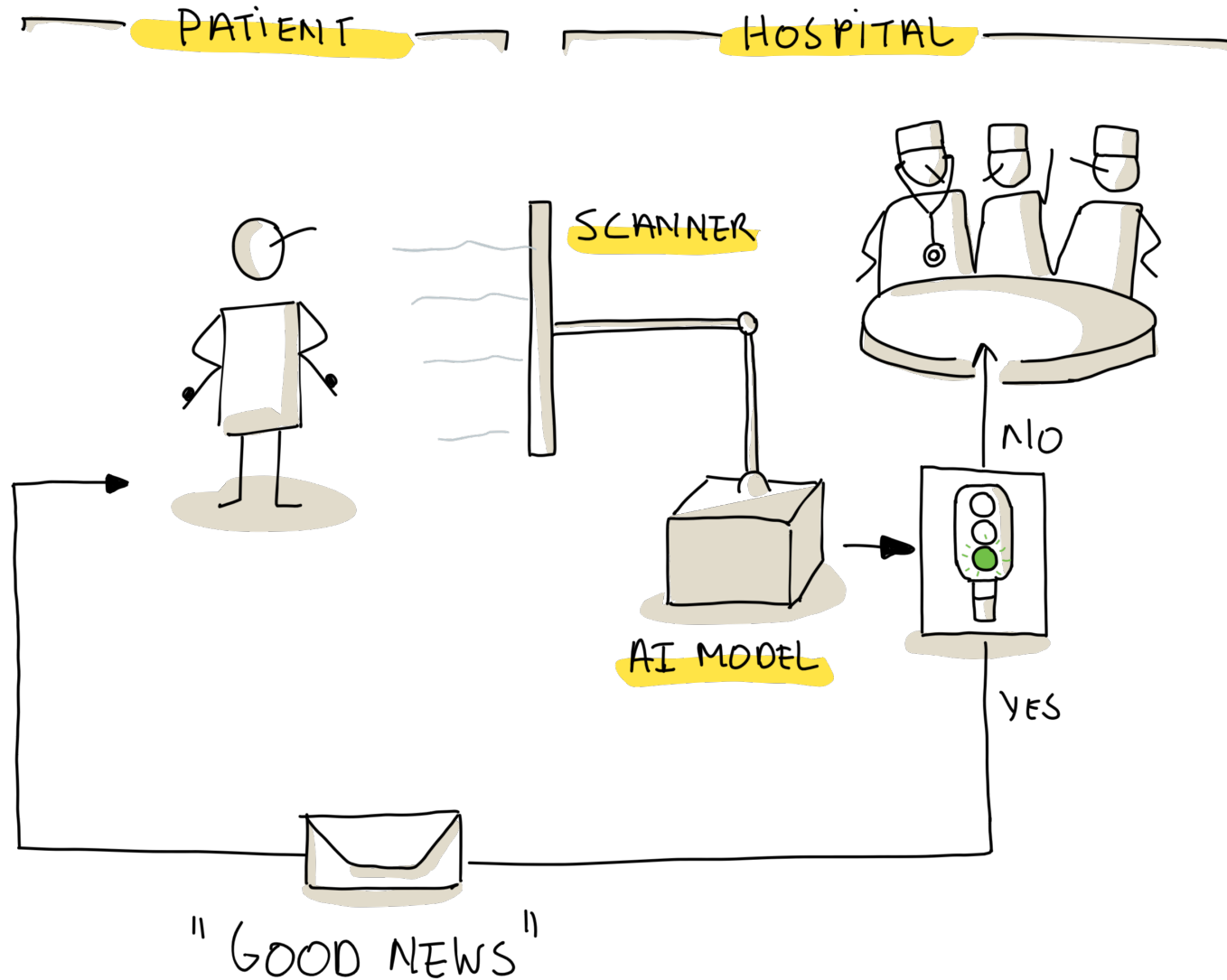
Autonomous Data Products



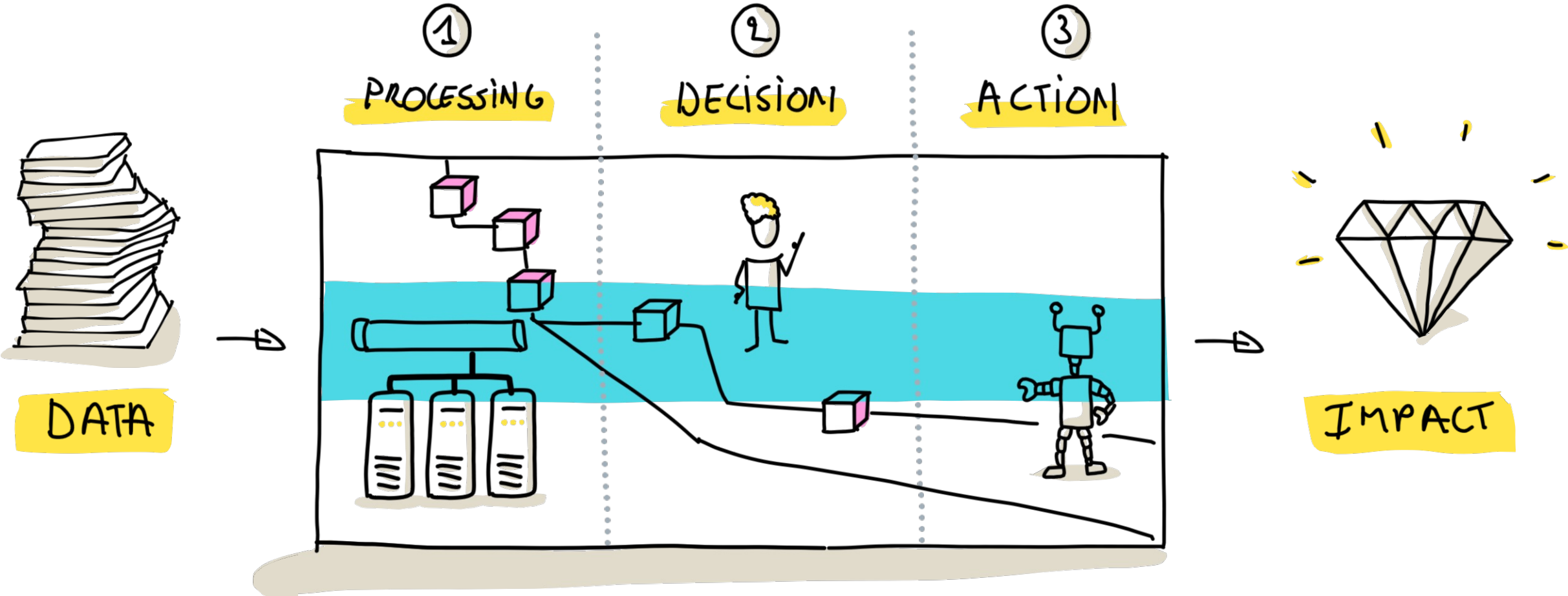
AI Model To Detect Churn



AI Model To Analyse Scans



Semi-Autonomous Data Products



Generate Social Media Posts With ChatGPT

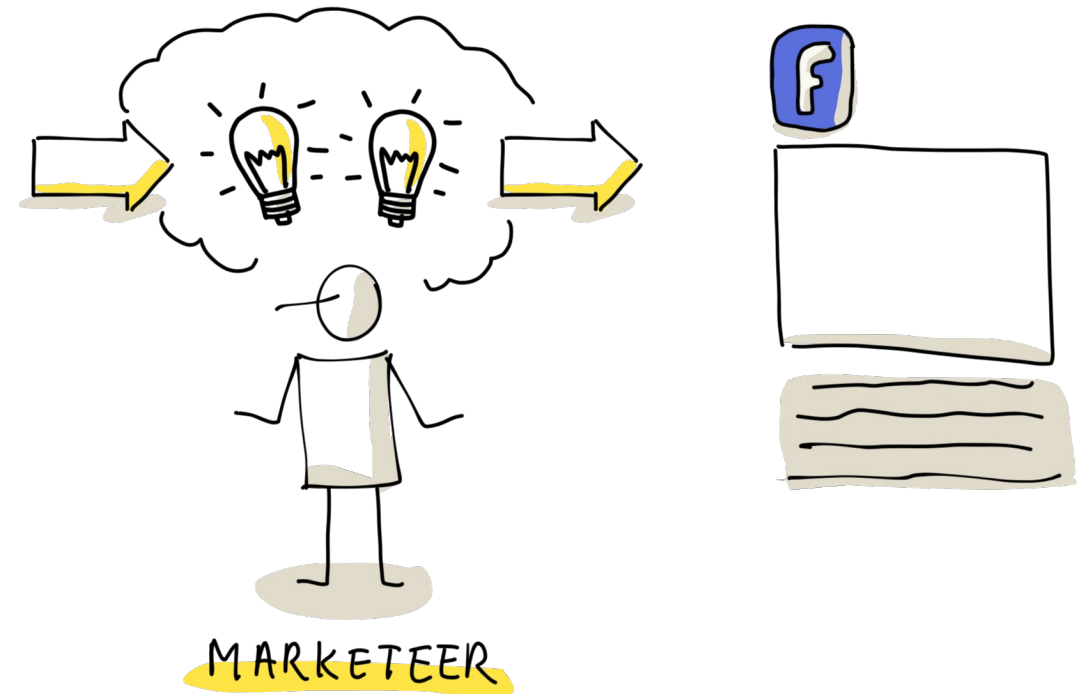
T I have a food shop. Because of the good weather, I would like to promote my BBQ assortment (meat and vegetarian) on facebook. Could you generate 5 ideas for a facebook post?



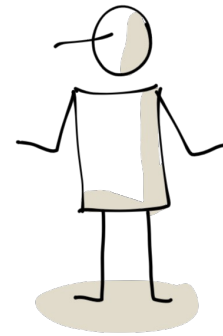
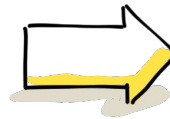
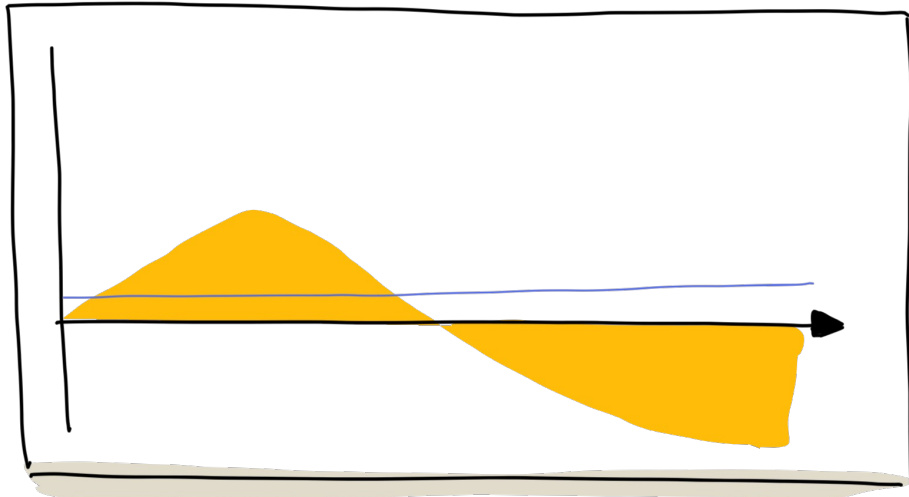
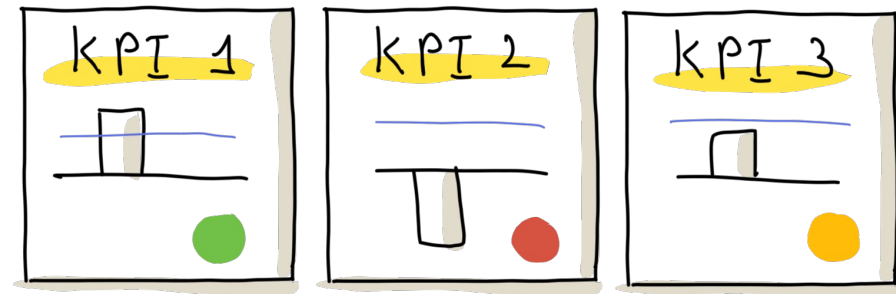
Sure! Here are five ideas for a Facebook post to promote your BBQ assortment:



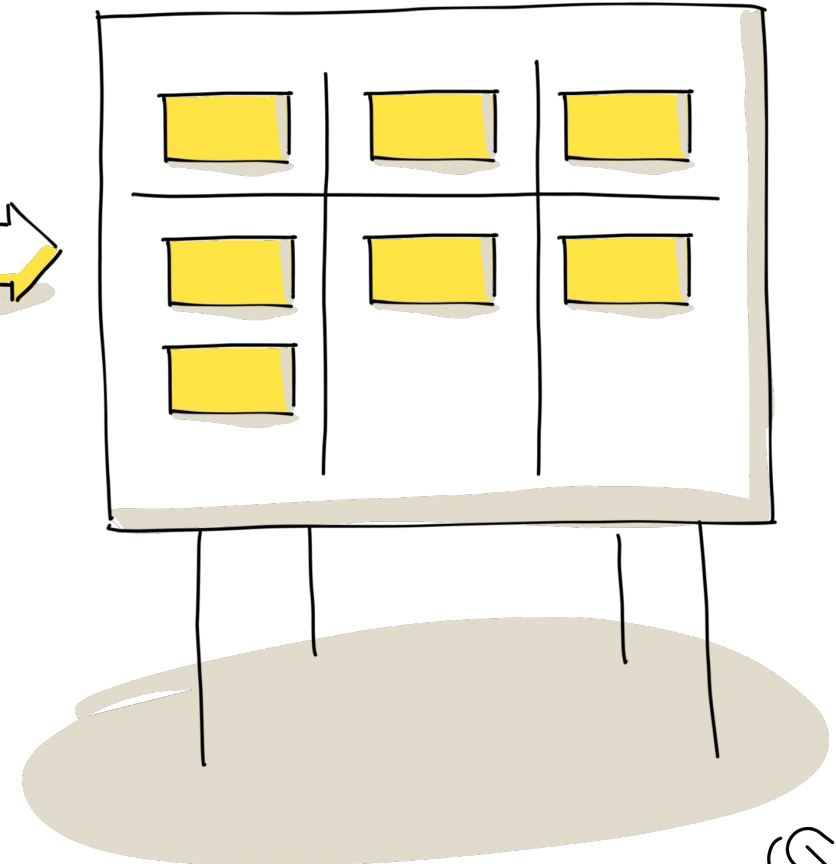
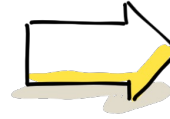
1. "🔥 Embrace the sizzling season with our mouthwatering BBQ assortment! 🍖🥗
Whether you're a meat lover or a vegetarian, we've got you covered. Indulge in the juiciest grilled meats and the freshest veggies, all infused with our signature flavors. Come on down and experience BBQ bliss at [Your Shop Name]!"
2. "🍷🍖 It's time to fire up the grill and treat yourself to a fantastic BBQ feast! From tender marinated steaks to flavorful veggie skewers, our BBQ assortment is a true delight for everyone. Grab your friends and family and join us for an unforgettable outdoor dining experience at [Your Shop Name]. Don't miss out on the ultimate summer indulgence!"
3. "🔥🥗 Ready for a BBQ adventure? Our diverse assortment has something for everyone! Sink your teeth into succulent BBQ chicken wings, savor the rich smokiness of our pulled pork sandwiches, or enjoy our delicious grilled veggie medley. Don't forget to top it off with our irresistible homemade BBQ sauce. Join us at [Your Shop Name] and let the grilling season begin!"
4. "🍔🥗 Craving a BBQ experience that caters to all tastes? Look no further! At [Your Shop Name], we believe in bringing people together through delicious food. Our BBQ assortment features a tempting array of juicy burgers, tender chicken kebabs, grilled portobello mushrooms, and so much more. Join us this weekend and let the flavors of summer ignite your taste buds!"
5. "🍷🔥 Calling all food lovers! Elevate your summer dining with our delectable BBQ assortment. Sink your teeth into our perfectly seasoned BBQ ribs, relish the smoky goodness of our BBQ jackfruit sliders, or enjoy the burst of flavors in our grilled vegetable platter. Don't settle for ordinary, indulge in extraordinary BBQ at [Your Shop Name]. Your taste buds will thank you!"



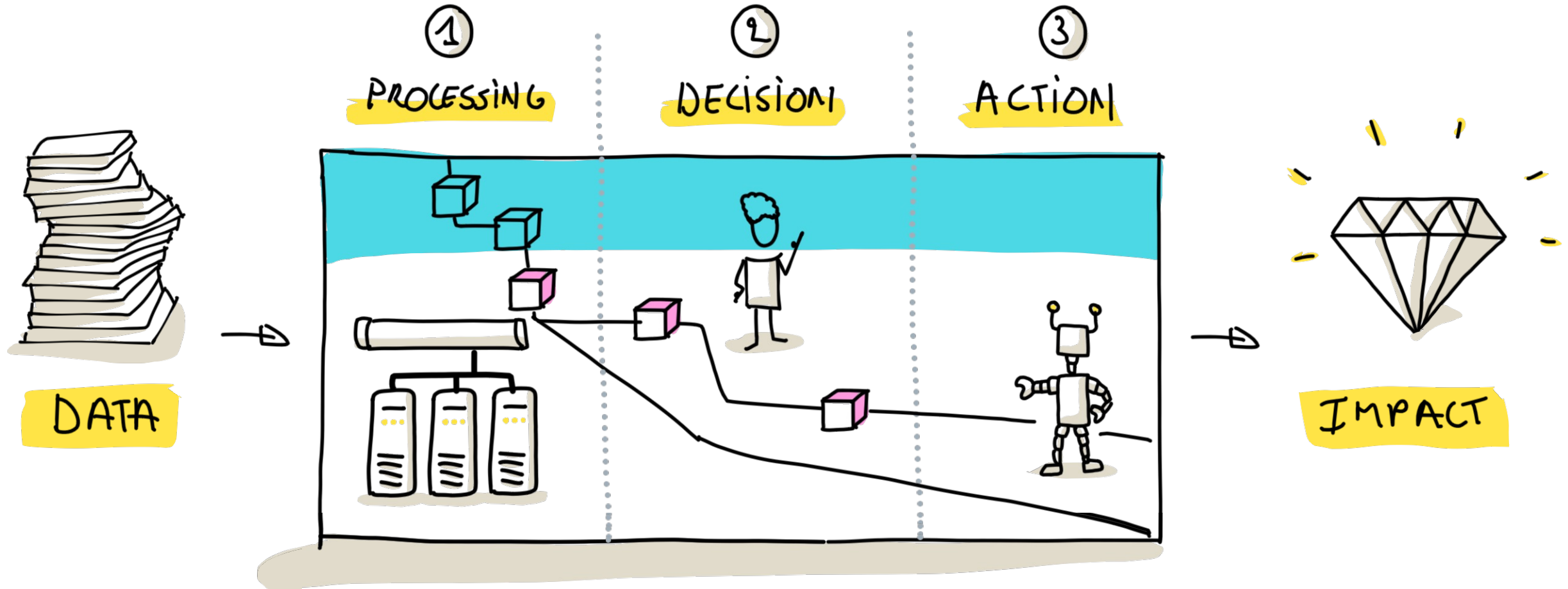
Process Monitoring And Continuous Improvement

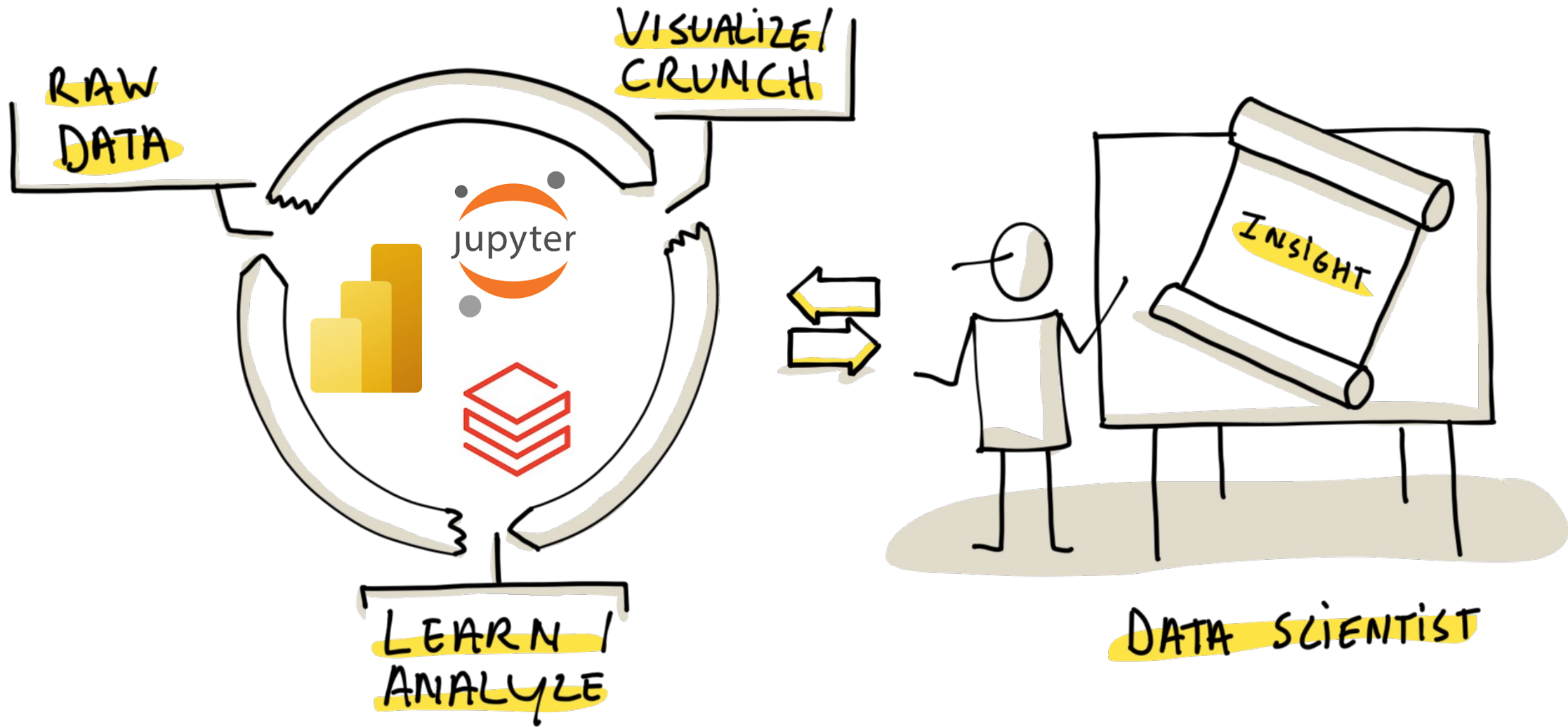


OPERATIONS
TEAM LEAD

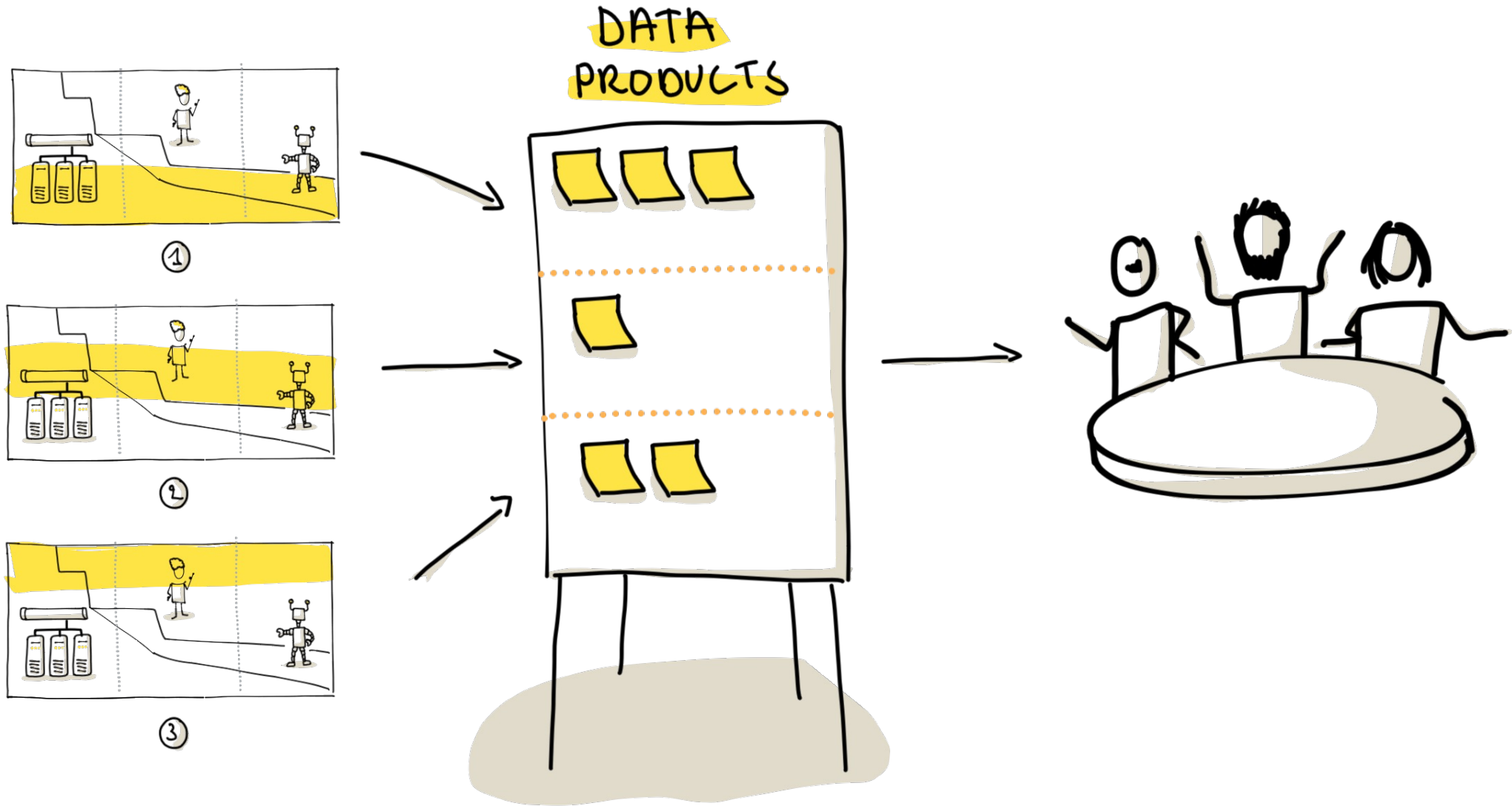


Raw Data Products

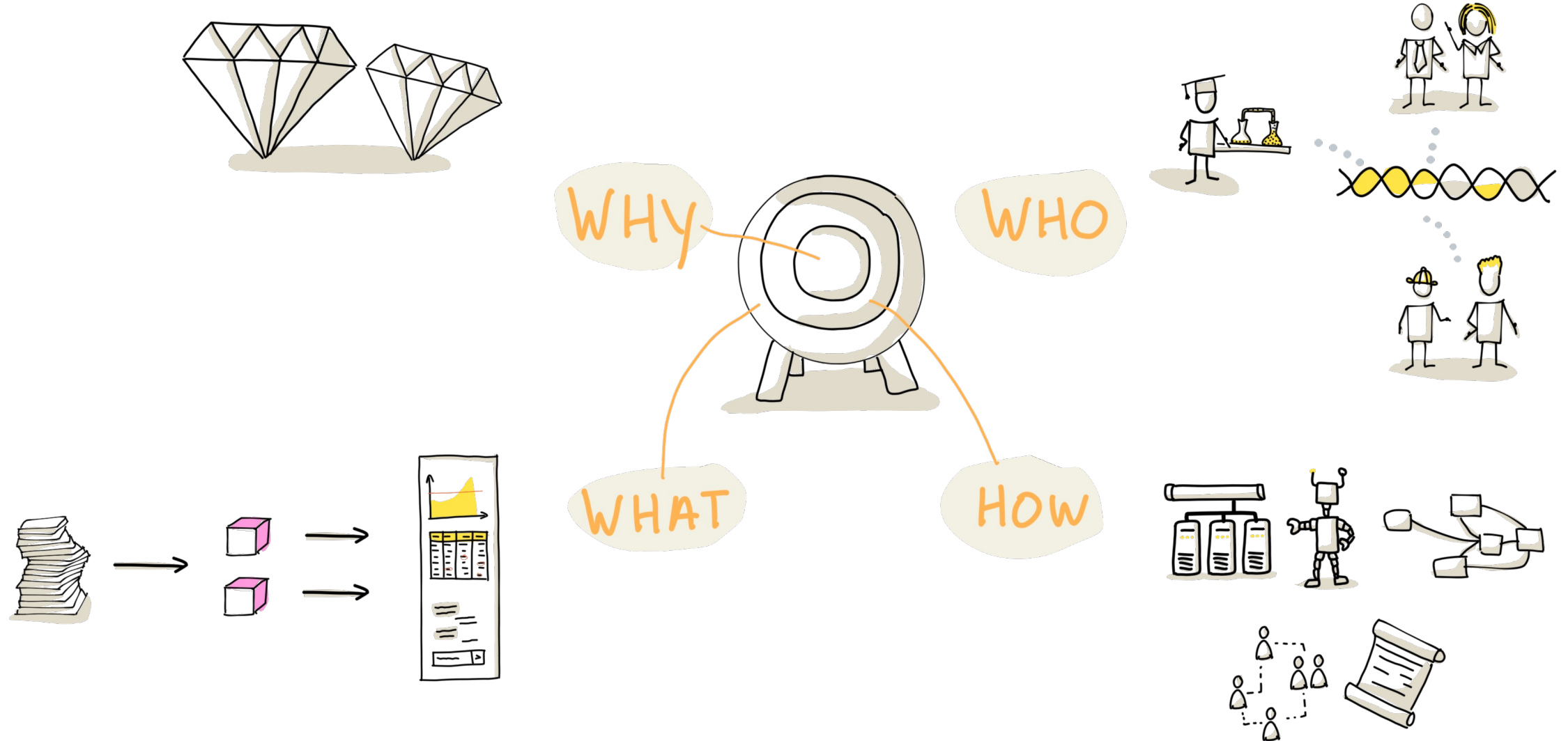




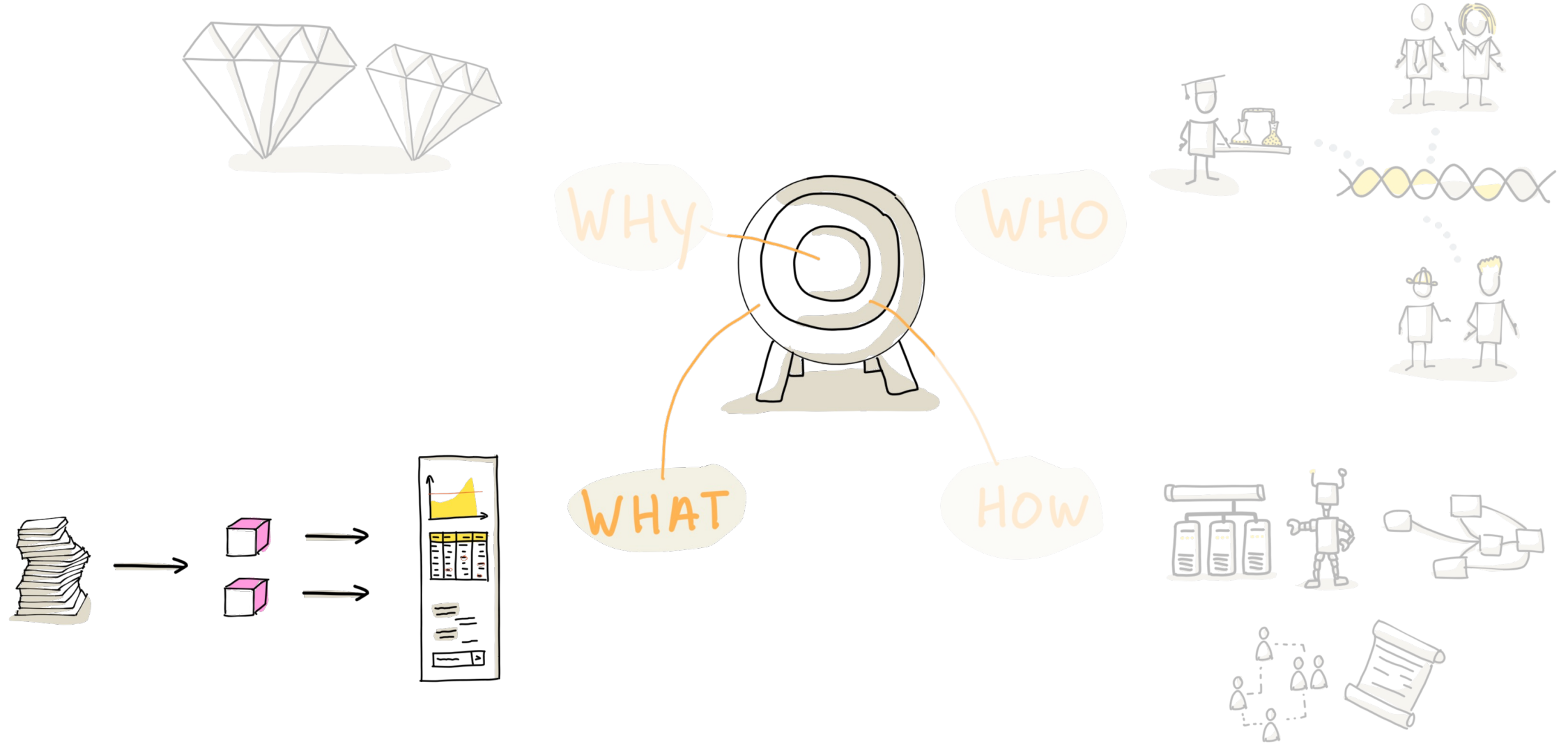
EXERCISE: LIST 1 OR MORE DATA PRODUCTS PER CATEGORY

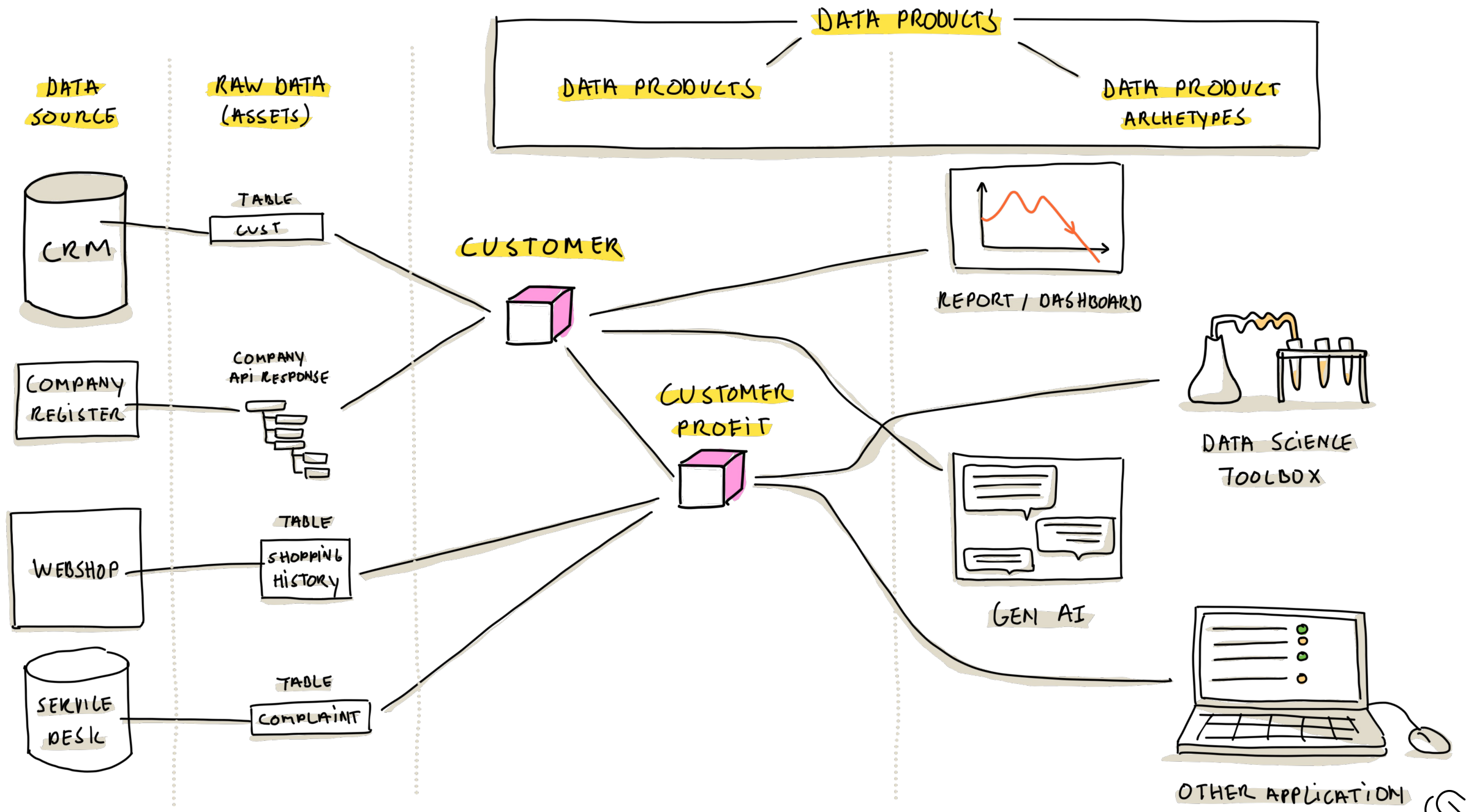


4 DATA QUESTIONS

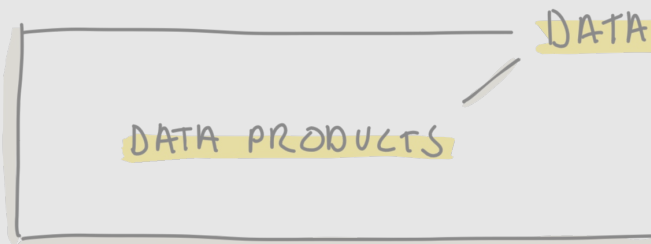
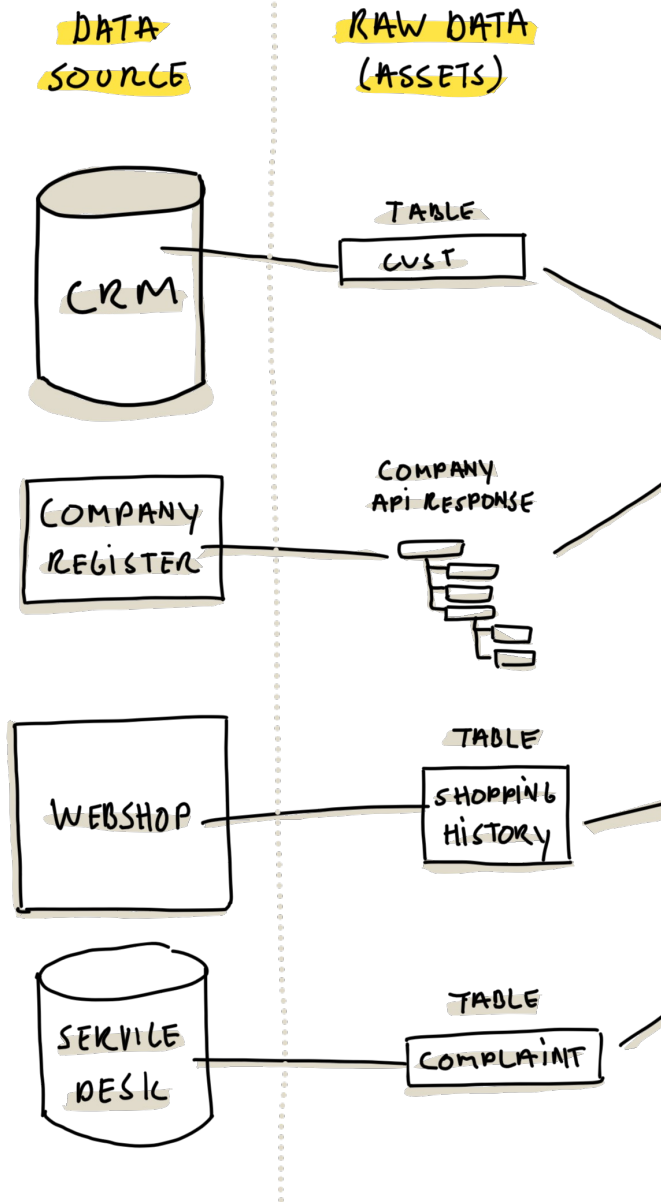


1 WHAT





1. DATA



CUSTOMER



CUSTOMER PROFIT



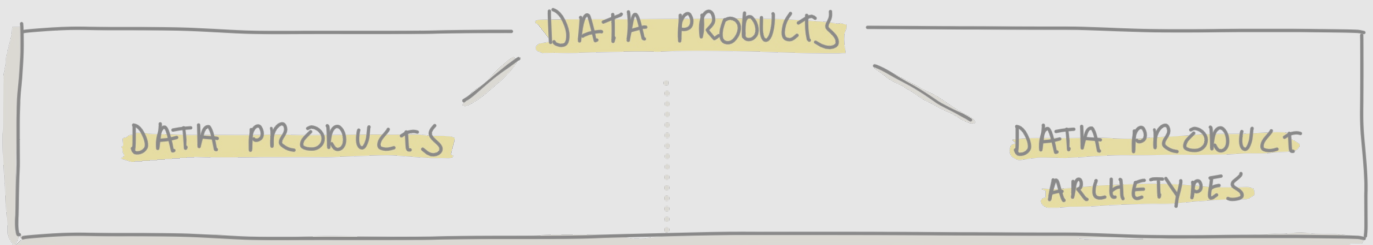
2. DATA PRODUCT ARCHETYPE



1. DATA

DATA SOURCE

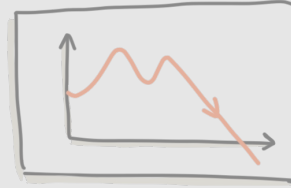
RAW DATA (ASSETS)



CUSTOMER



CUSTOMER PROFIT



REPORT / DASHBOARD



DATA SCIENCE TOOLBOX



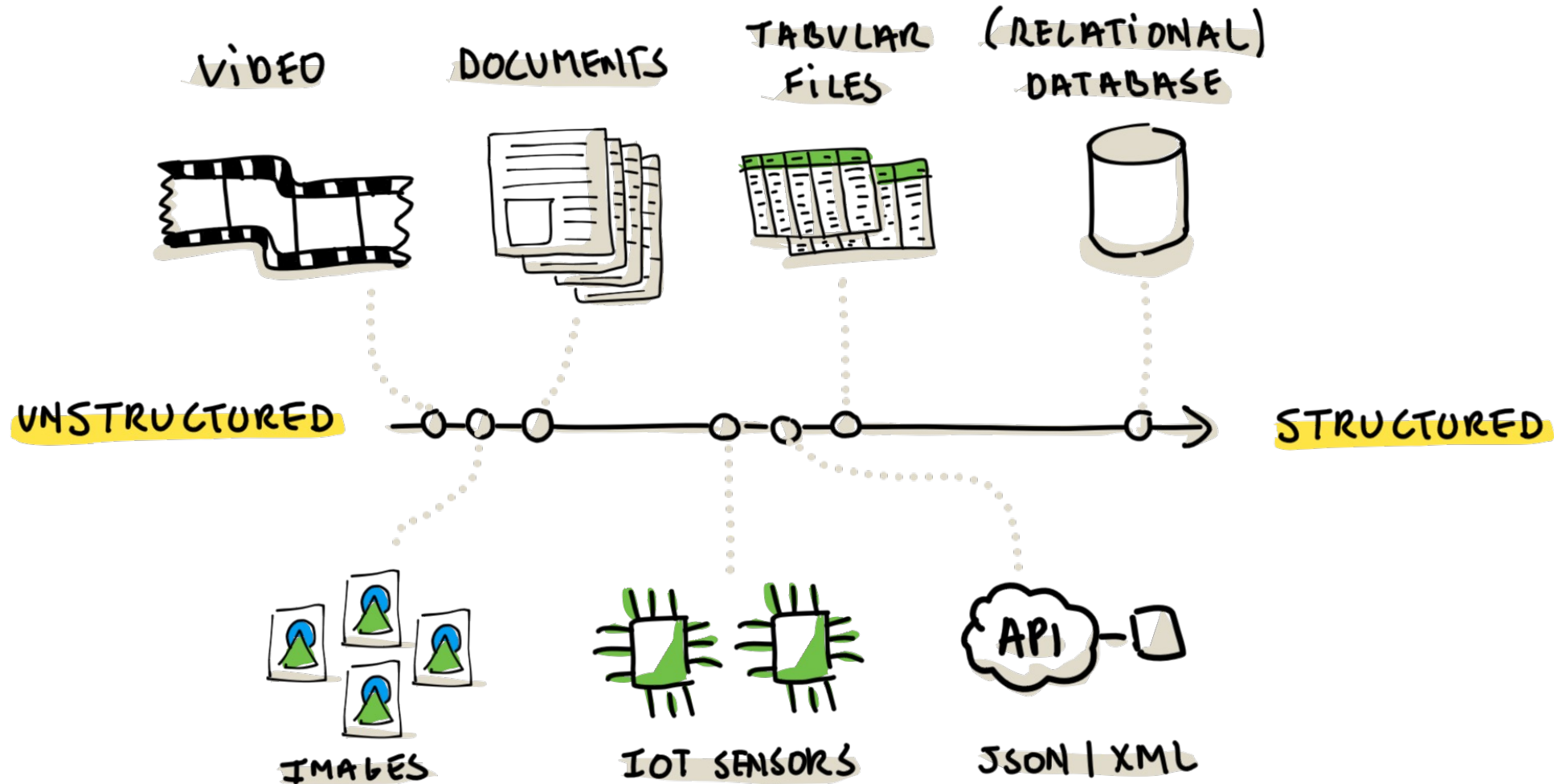
GEN AI



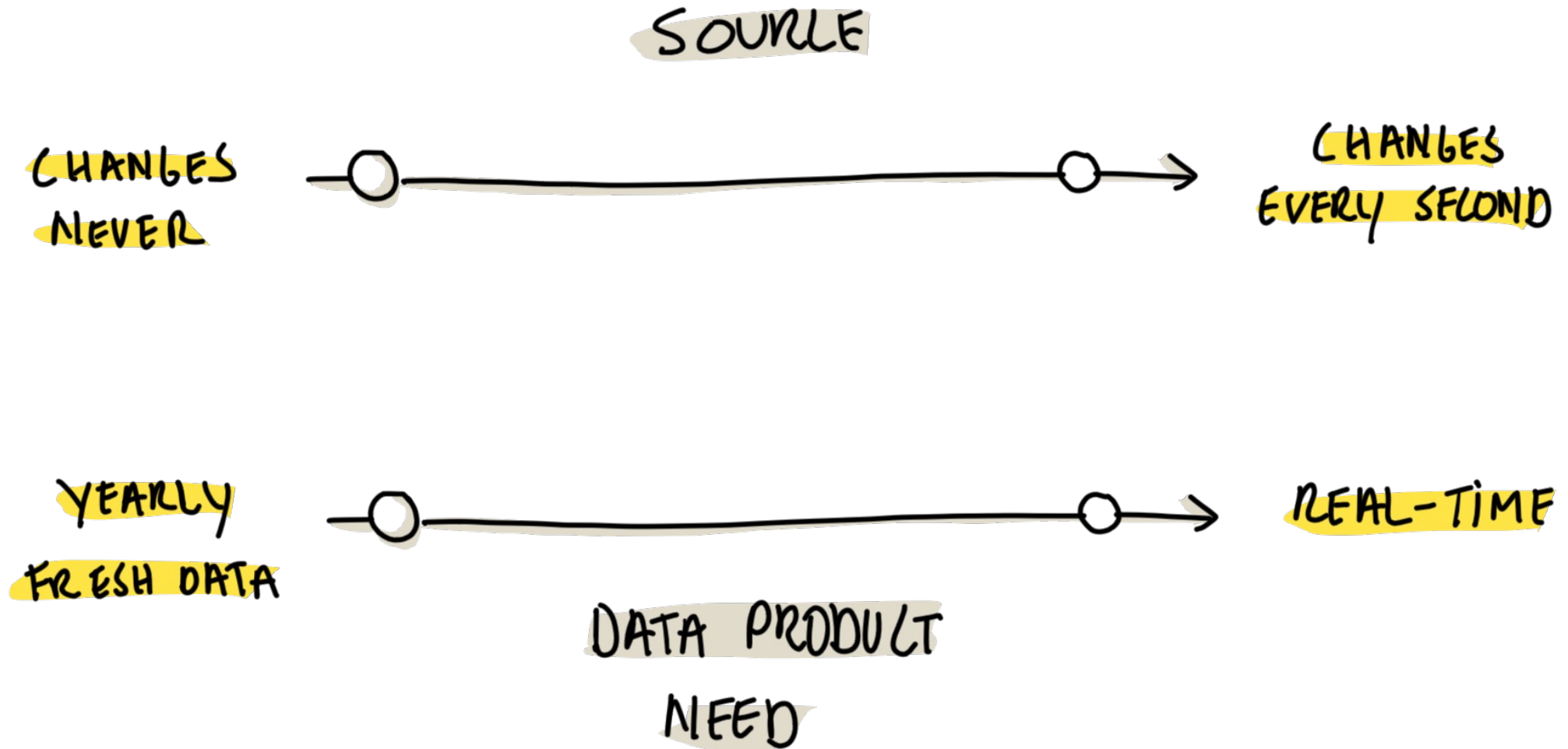
OTHER APPLICATION



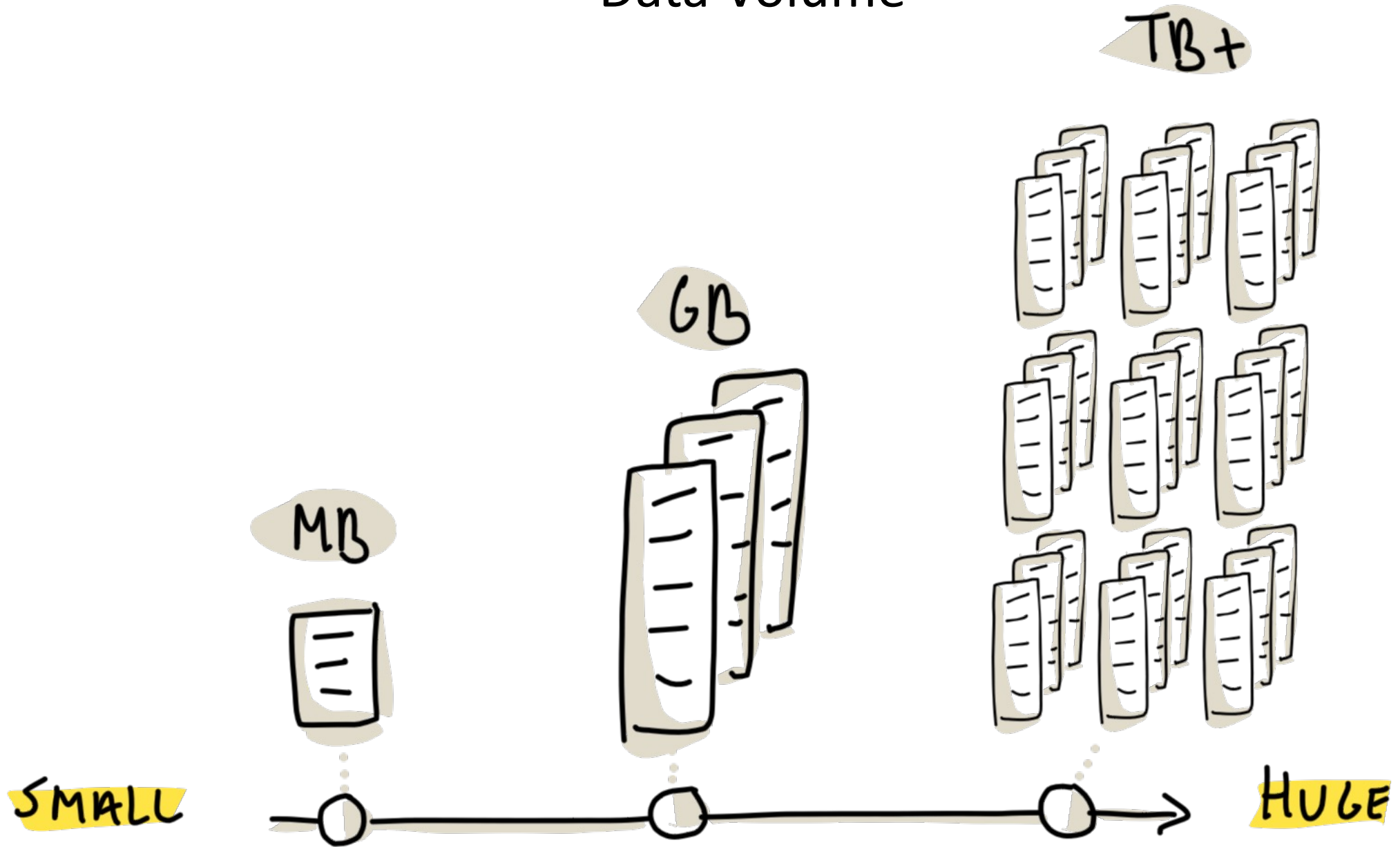
Type of Data



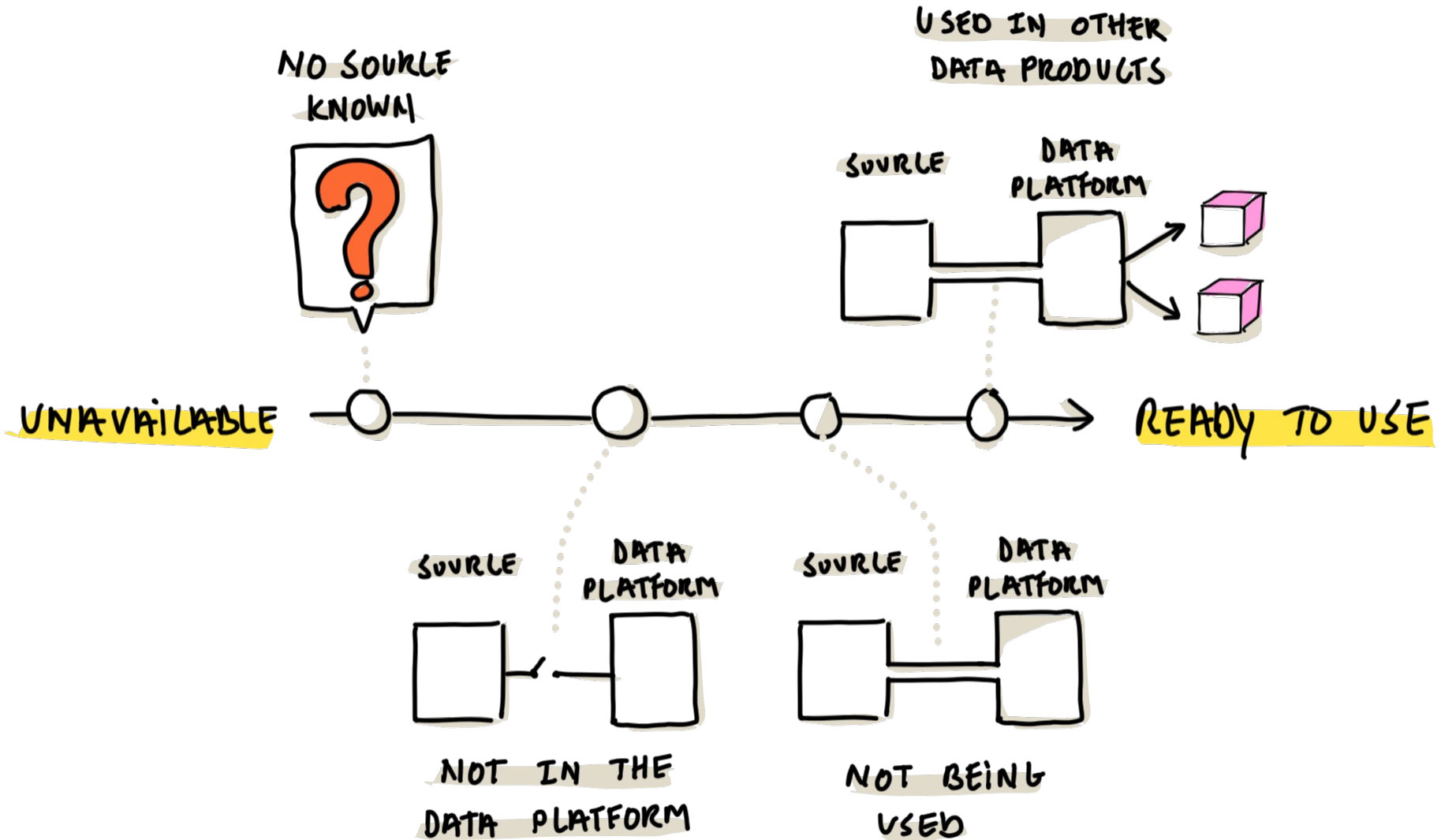
Data Velocity



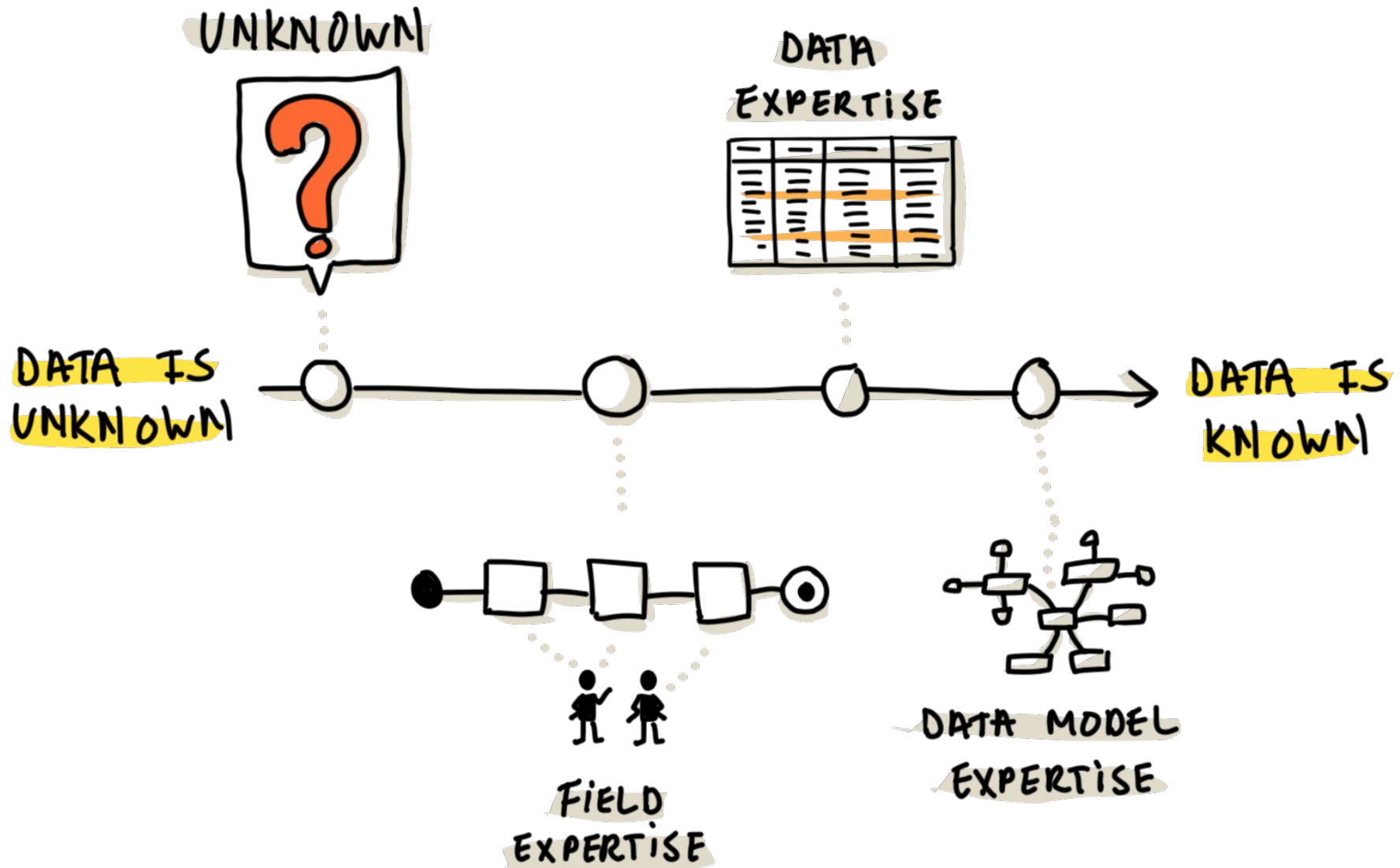
Data Volume



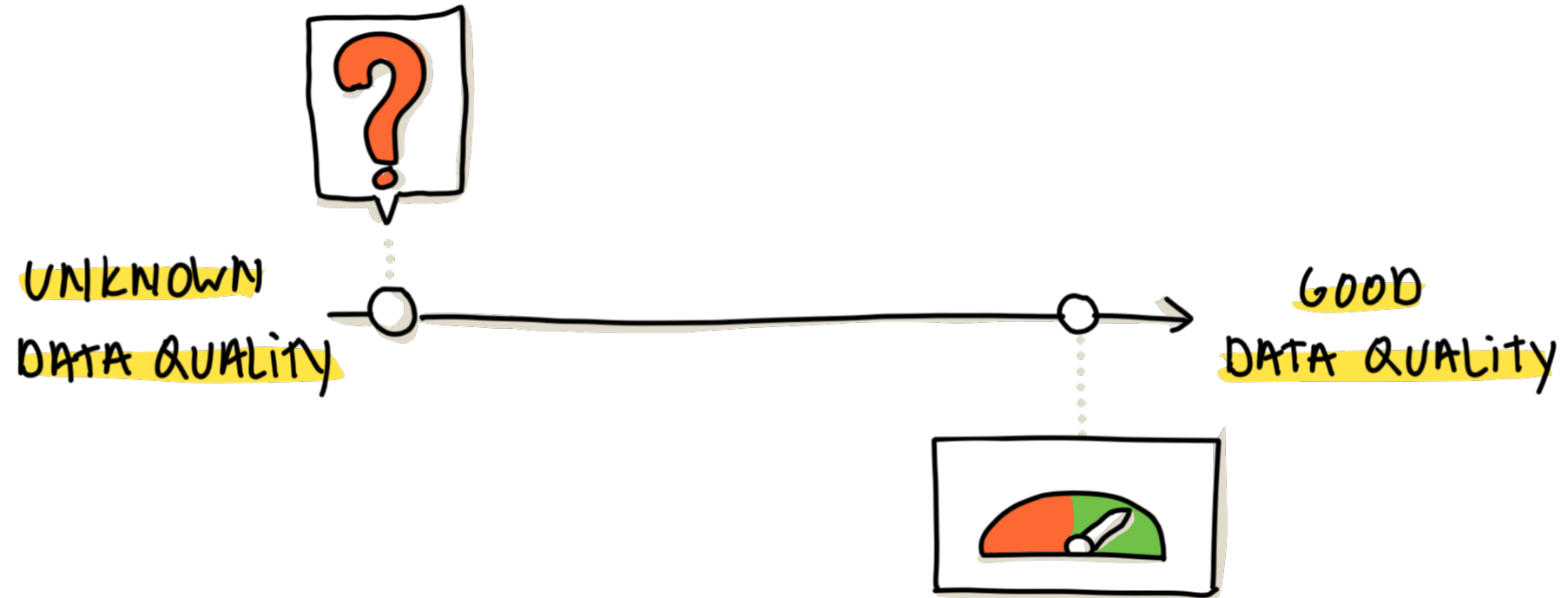
Data Availability



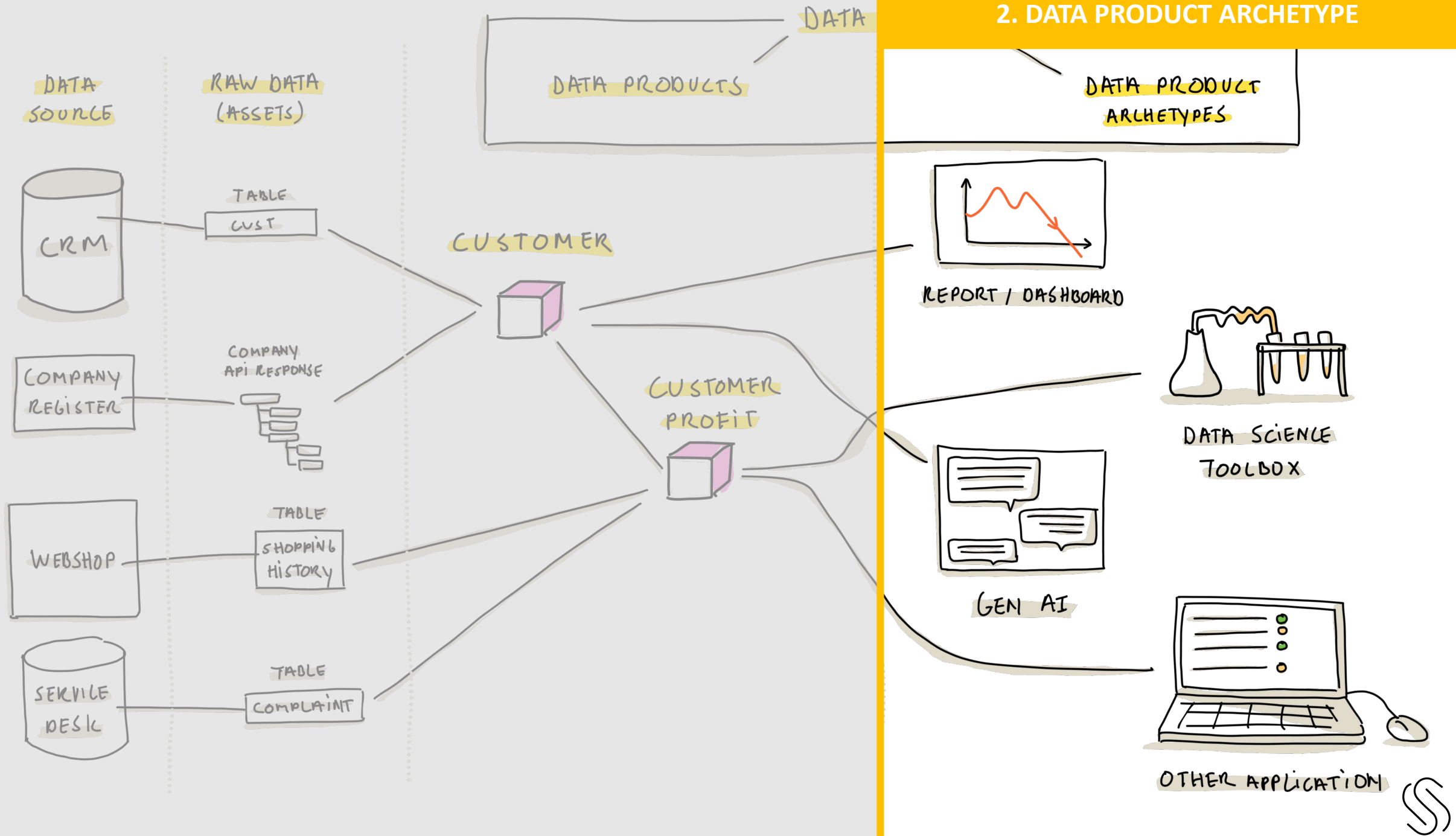
Data Knowledge



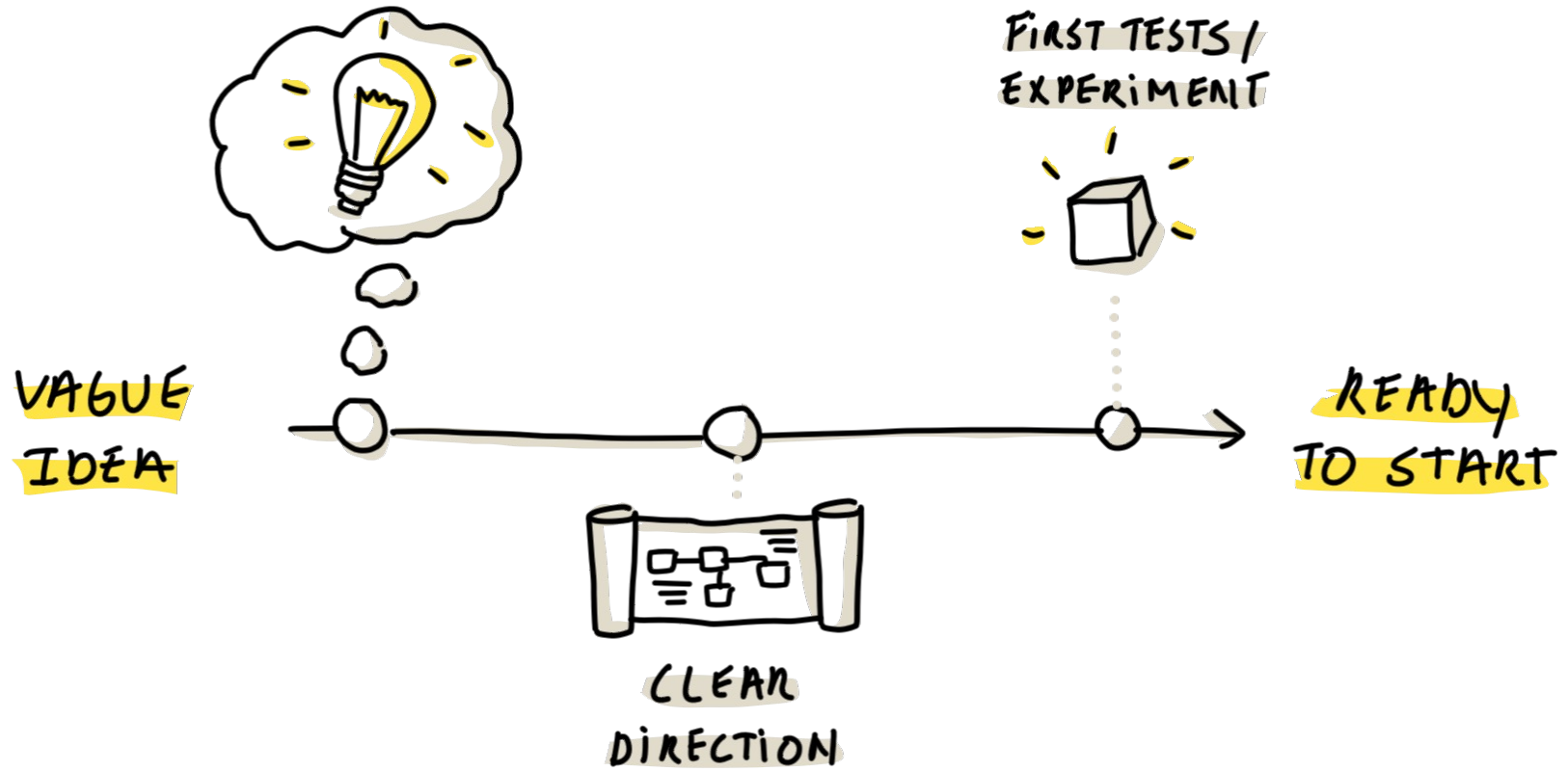
Data Quality



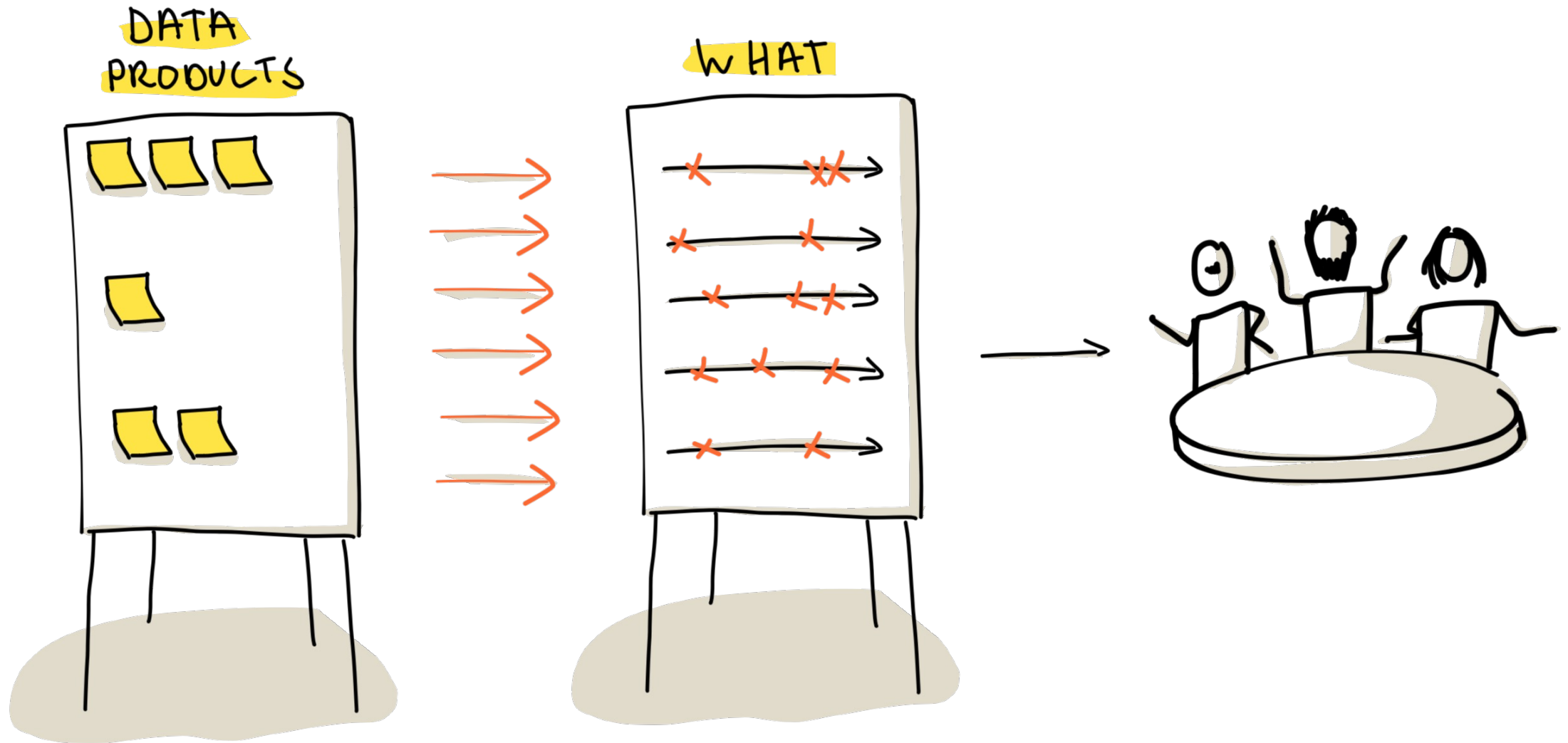
2. DATA PRODUCT ARCHETYPE



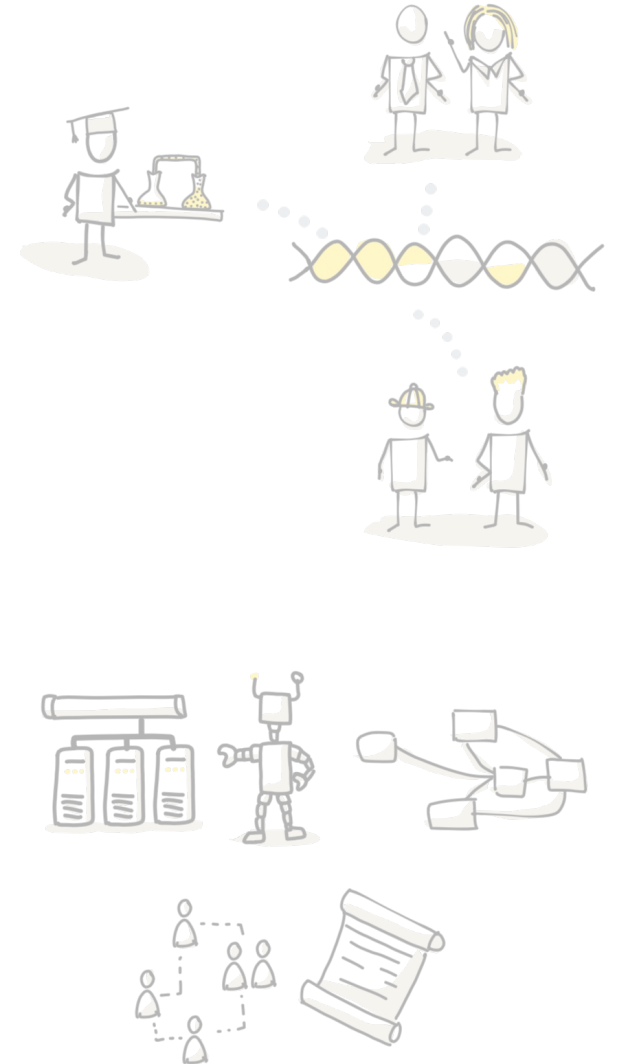
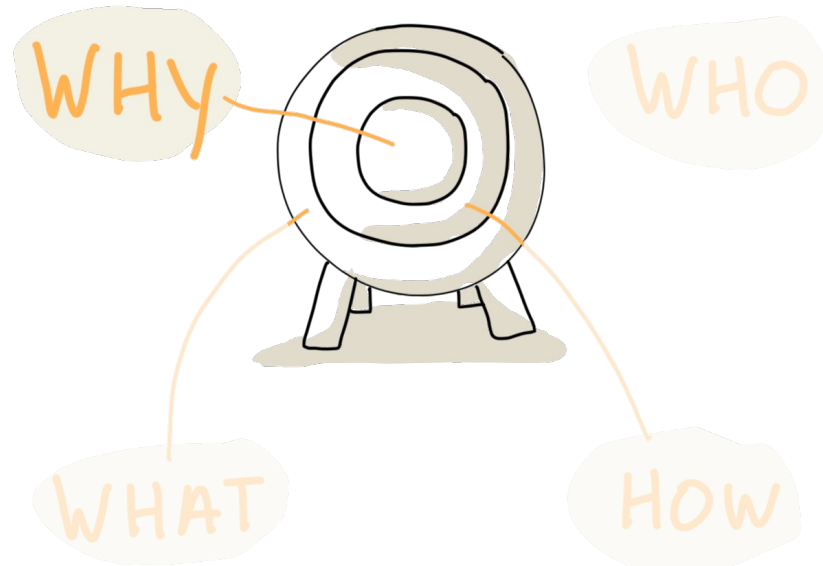
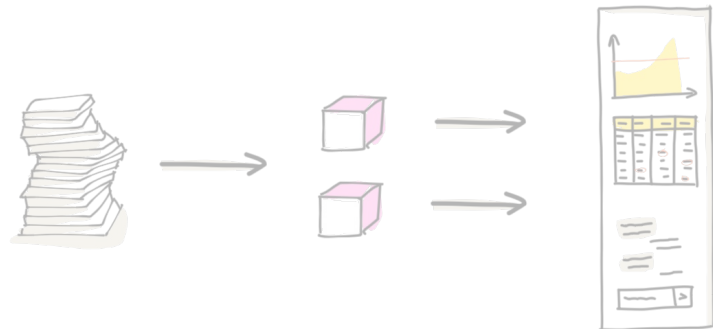
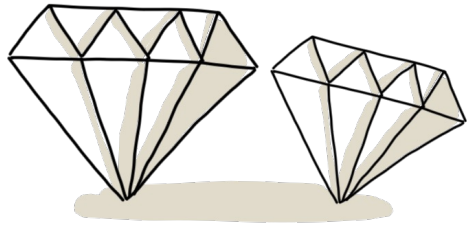
Readiness

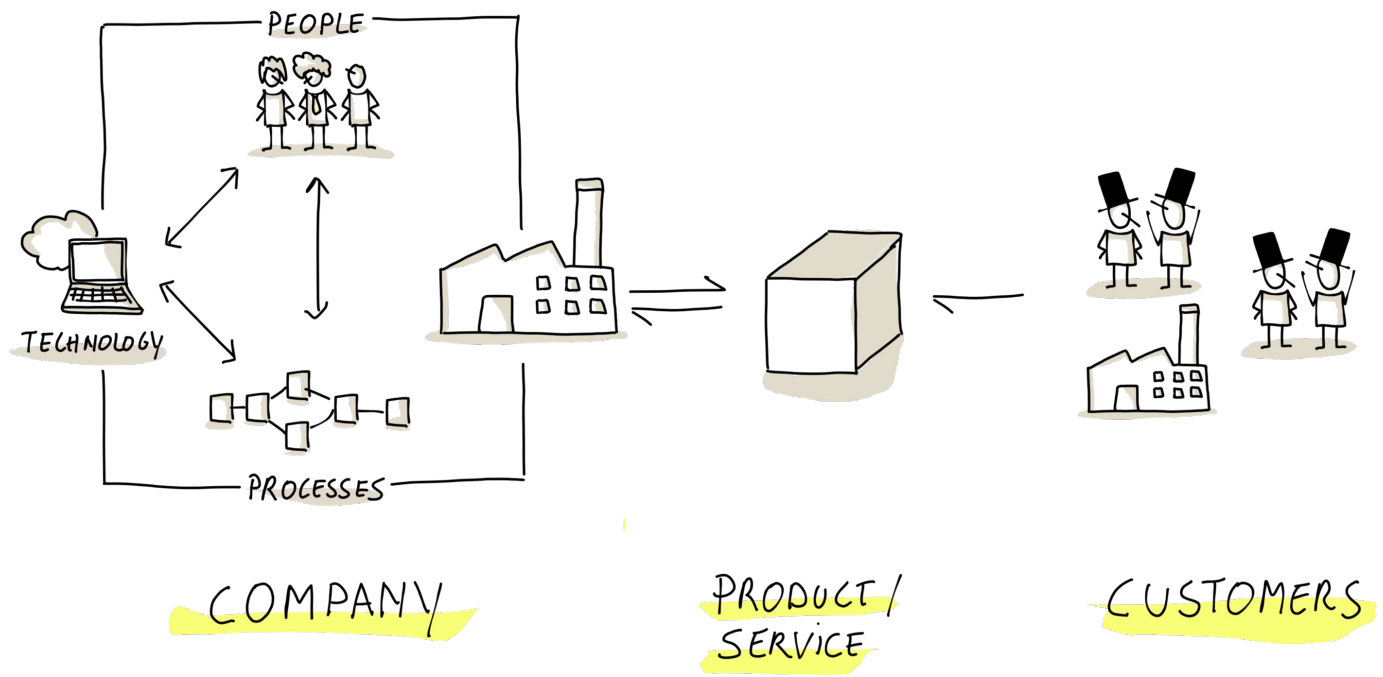


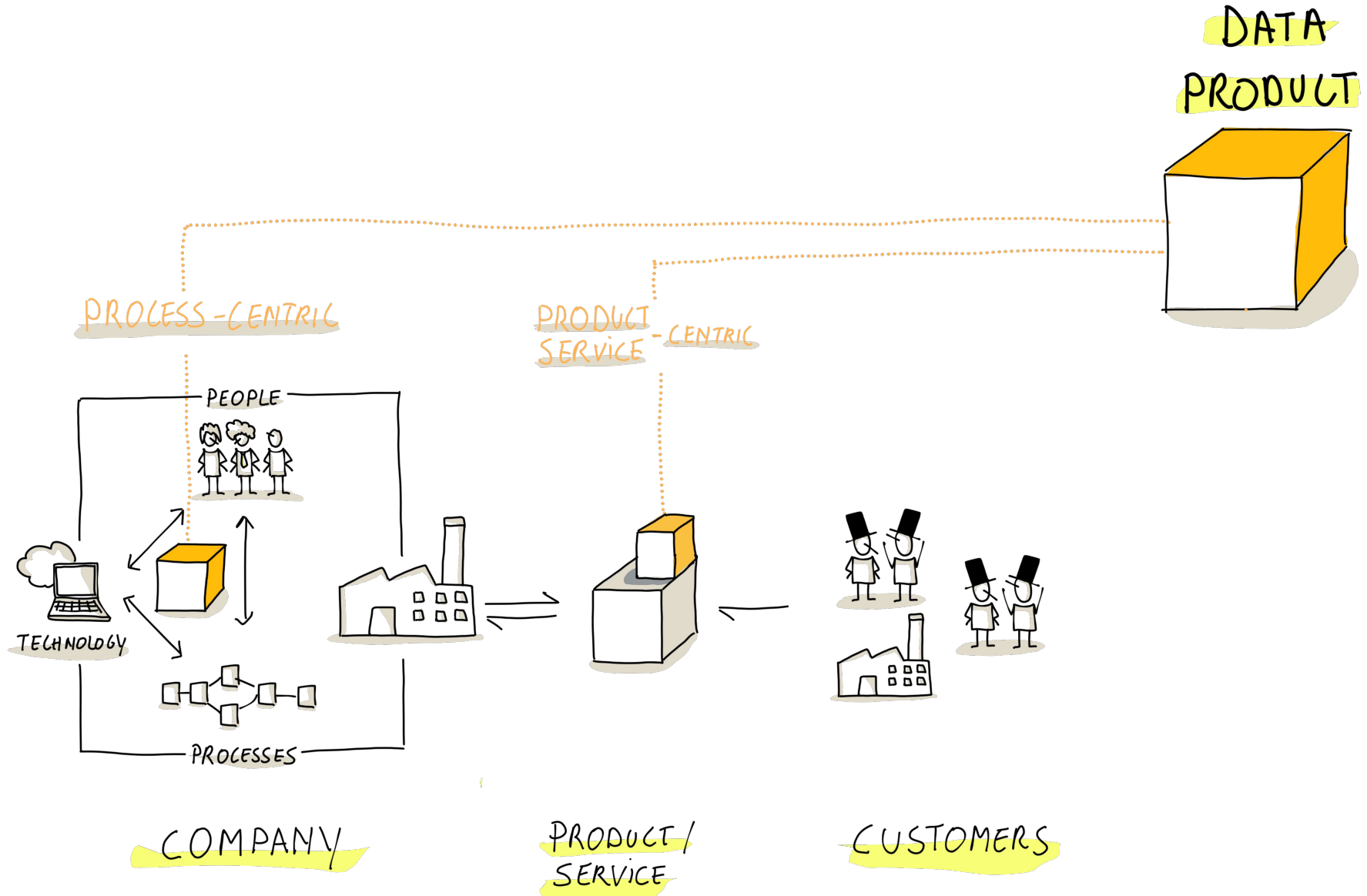
EXERCISE: SCORE THE 'WHAT' FOR EVERY DATA PRODUCT



2 WHY







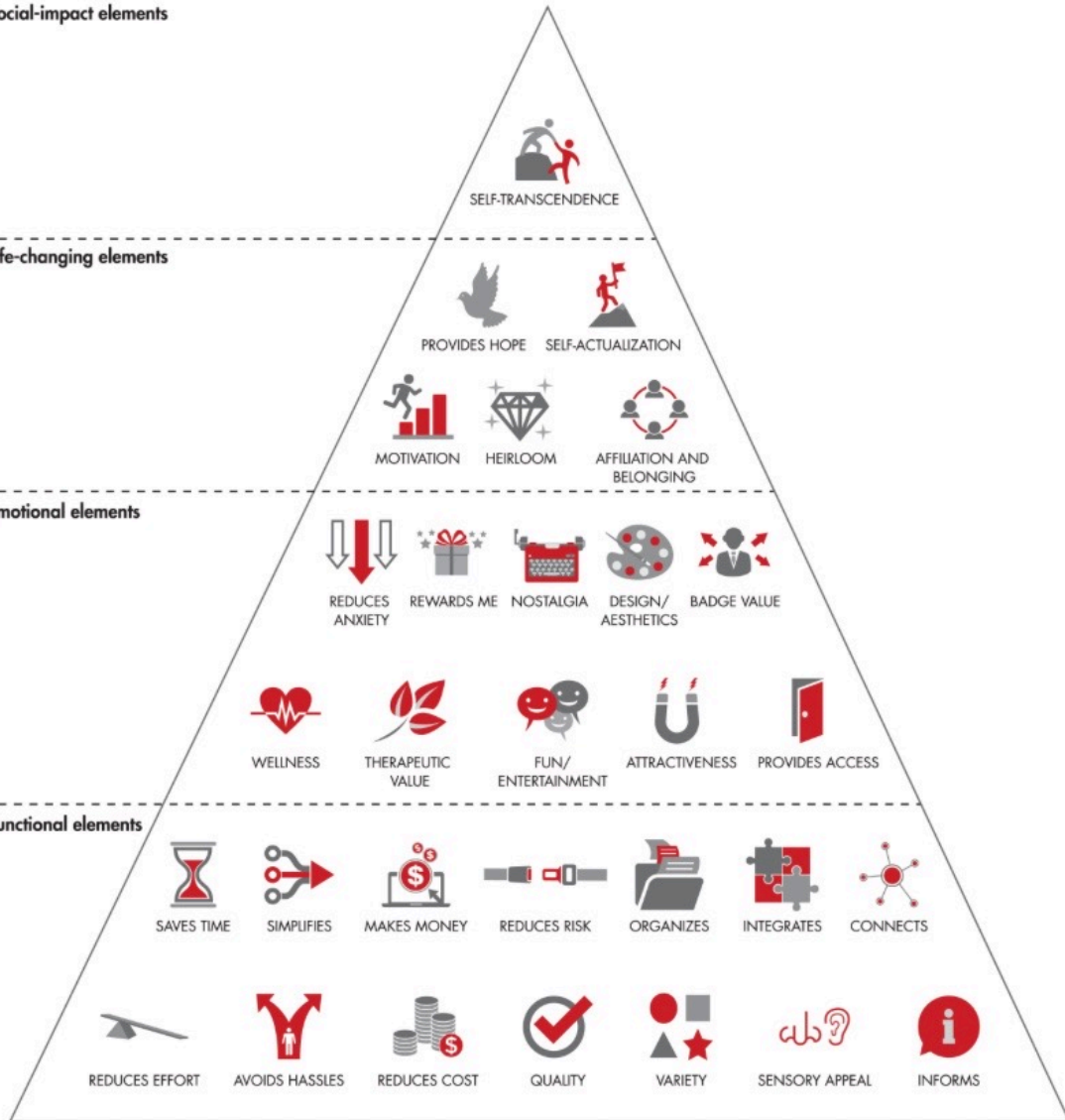
B2C

Social-impact elements

Life-changing elements

Emotional elements

Functional elements



B2B

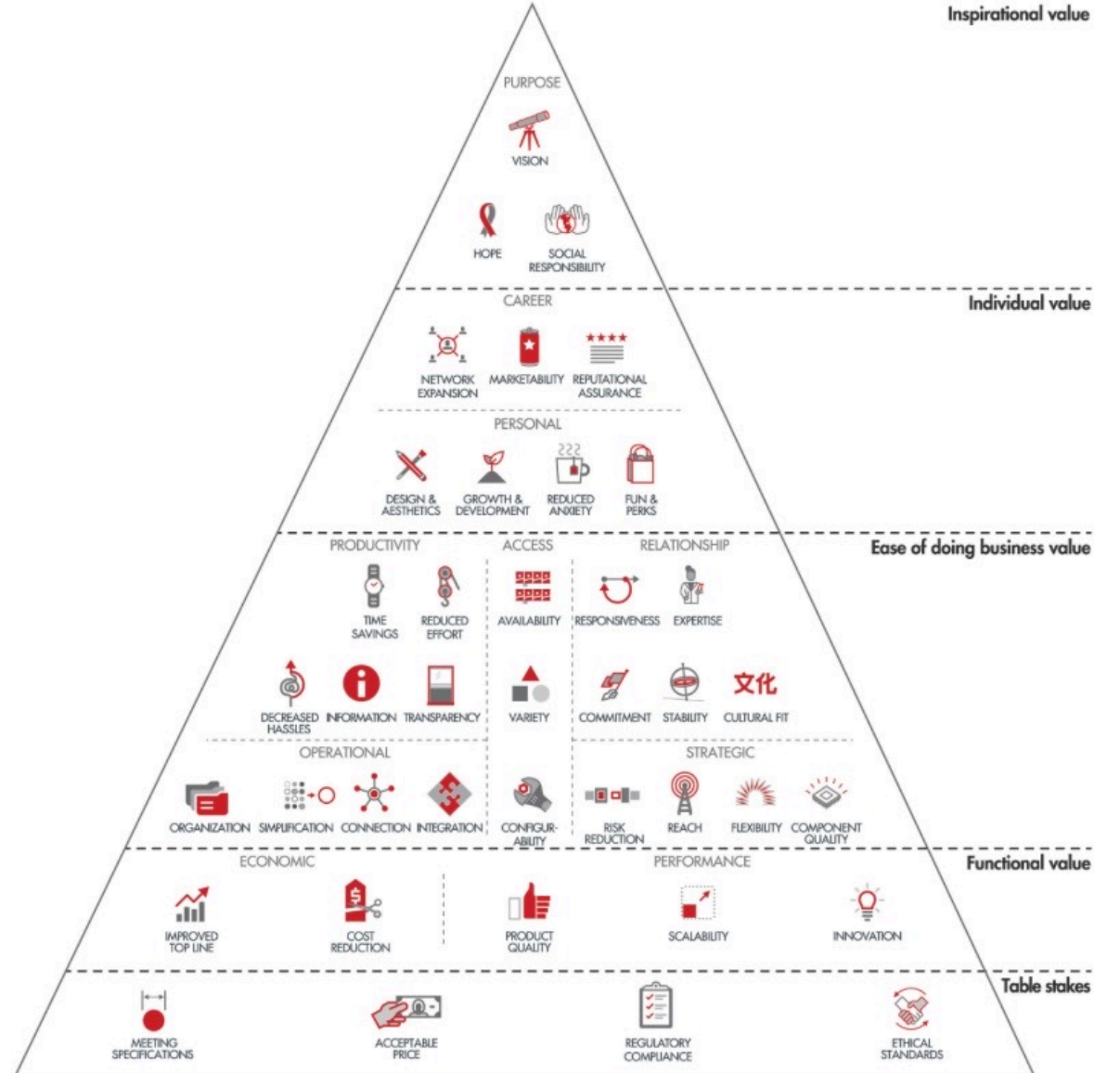
Inspirational value

Individual value

Ease of doing business value

Functional value

Table stakes



B2C

B2B

SAVES TIME

Social-impact elements

Life-changing elements

Emotional elements

Functional elements

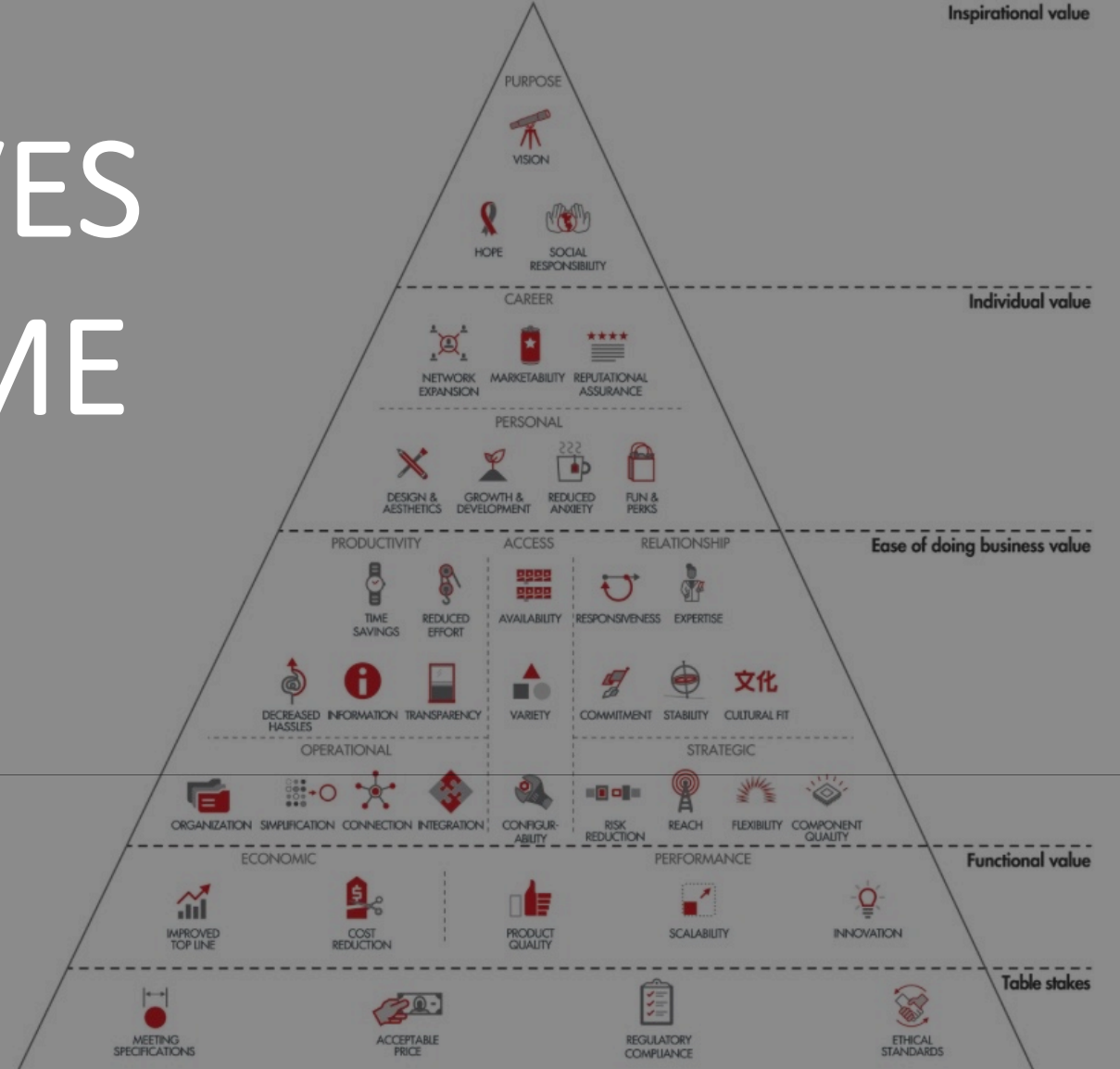
Inspirational value

Individual value

Ease of doing business value

Functional value

Table stakes



trivago

trivago



Park Hotel

★★★★ Hotel

Rome, Italy

8.7 Excellent (1402 reviews)



Hotel site

\$220

per night

View deal >

Hotels.com

\$250

per night

Our lowest price

\$220

per night



B2C

B2B

MAKES MONEY

Social-impact elements

Life-changing elements

Emotional elements

Functional elements

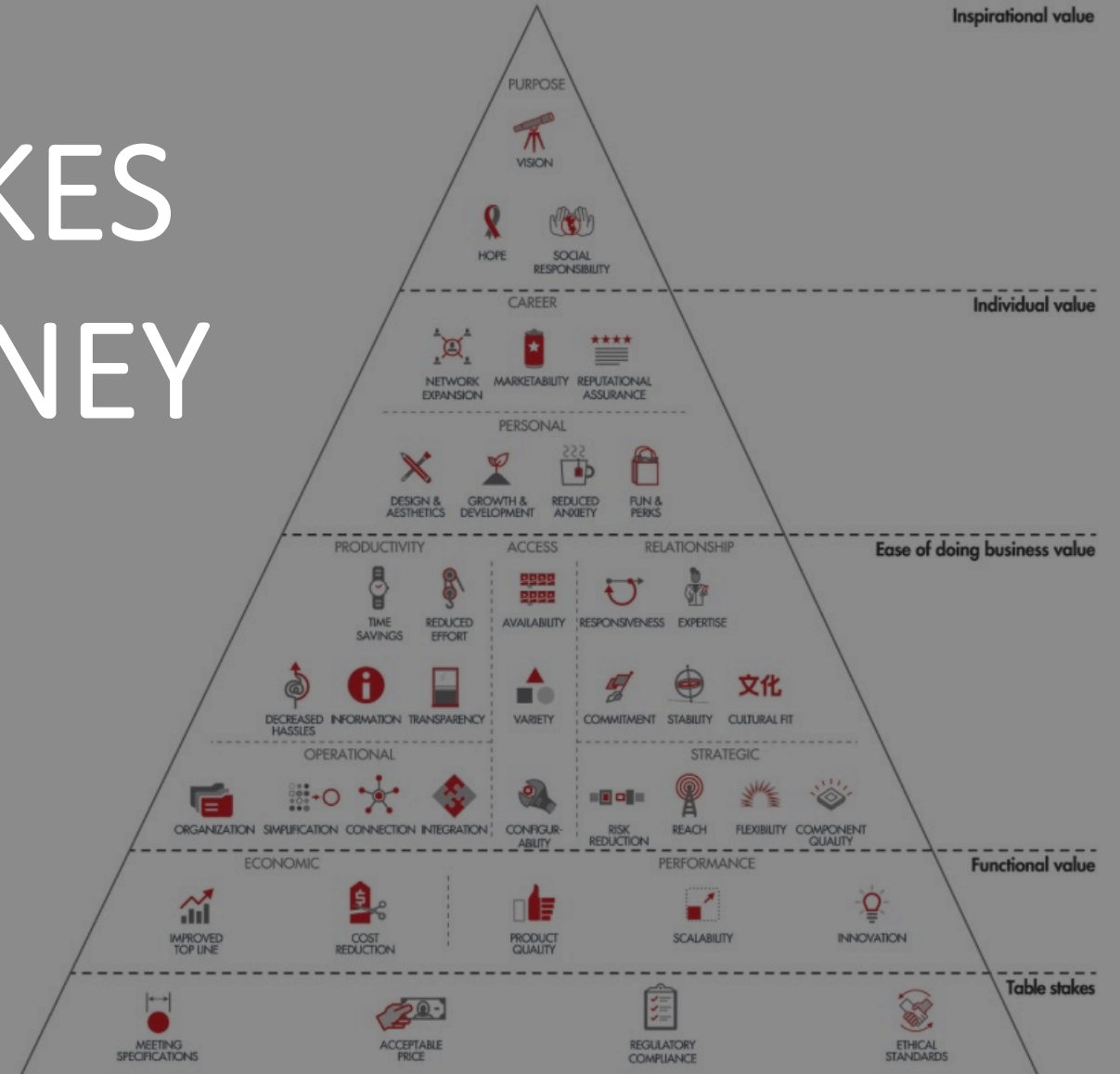
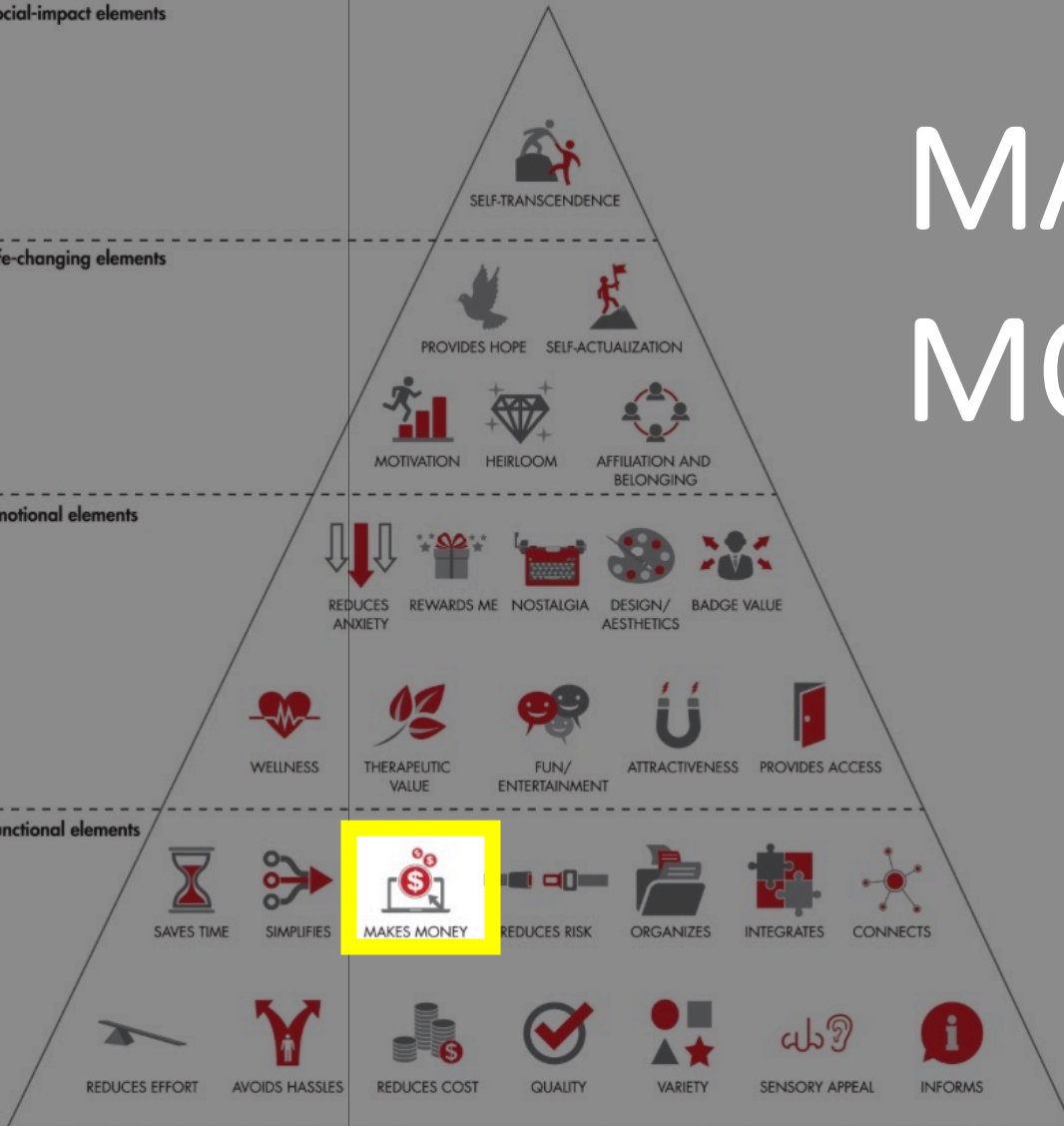
Inspirational value

Individual value

Ease of doing business value

Functional value

Table stakes





B2C

B2B

NOSTALGIA

Social-impact elements

Life-changing elements

Emotional elements

Functional elements

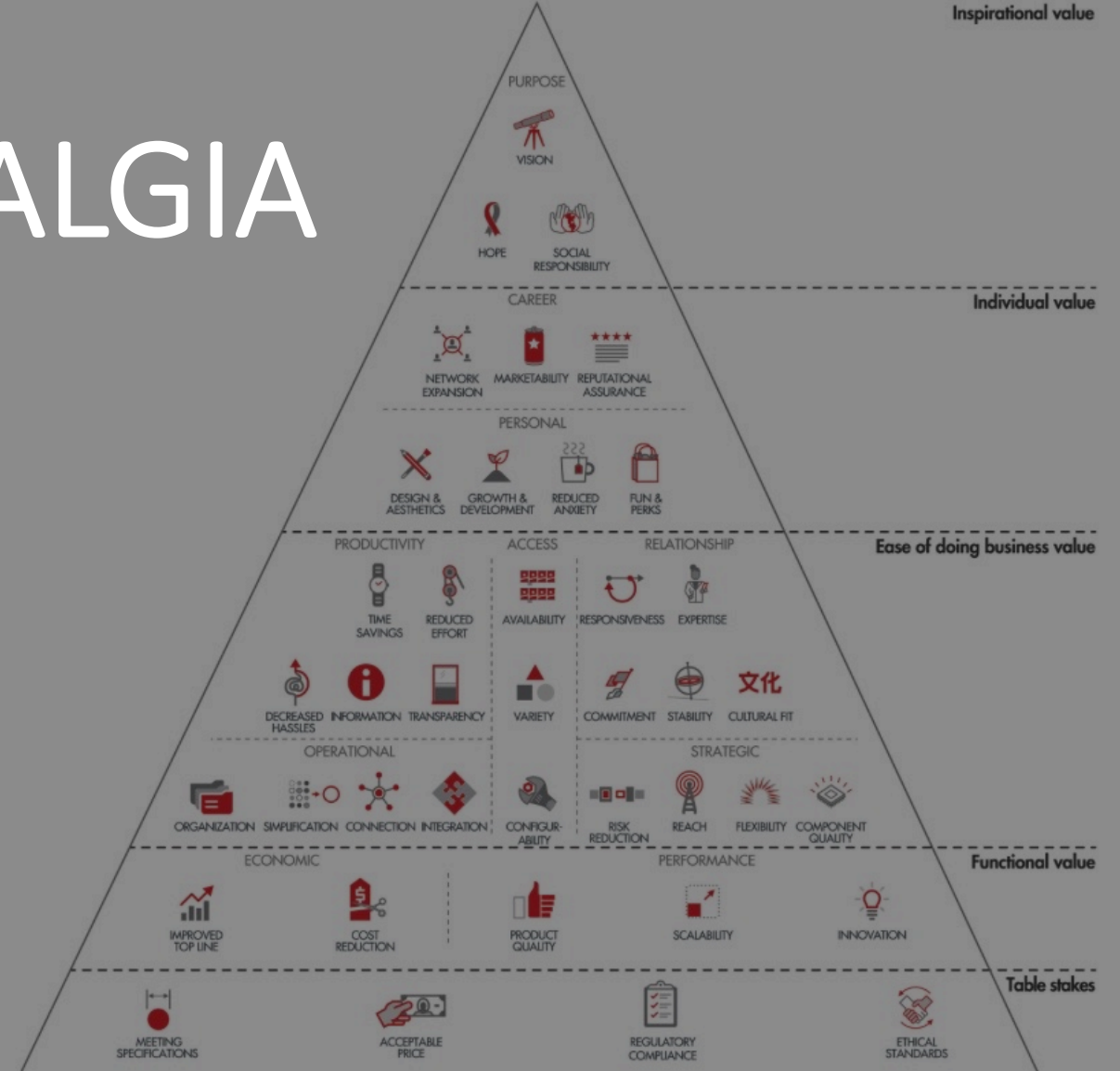
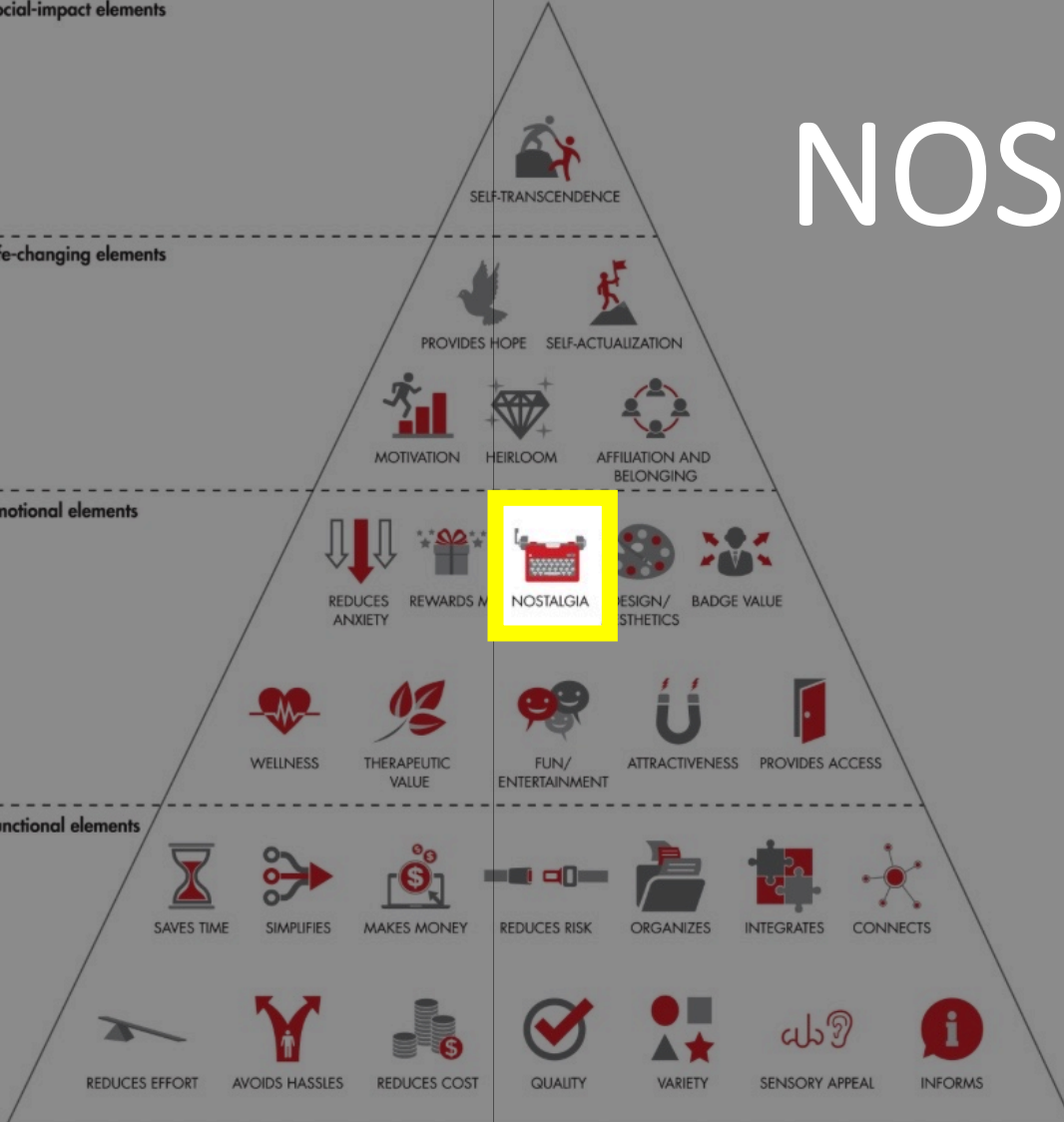
Inspirational value

Individual value

Ease of doing business value

Functional value

Table stakes



AI YearBook



B2C

B2B

MOTIVATION

Social-impact elements

Life-changing elements

Emotional elements

Functional elements

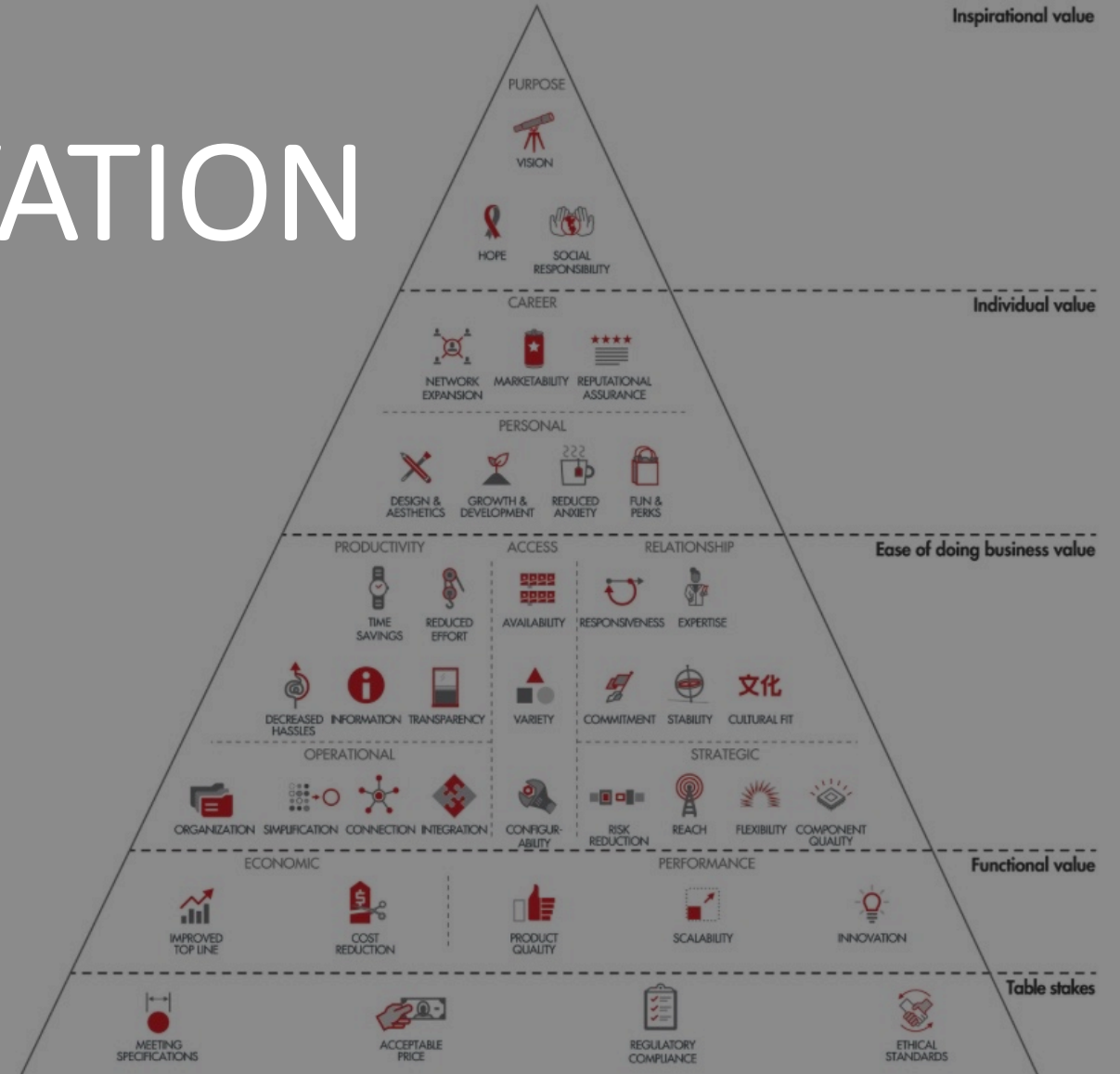
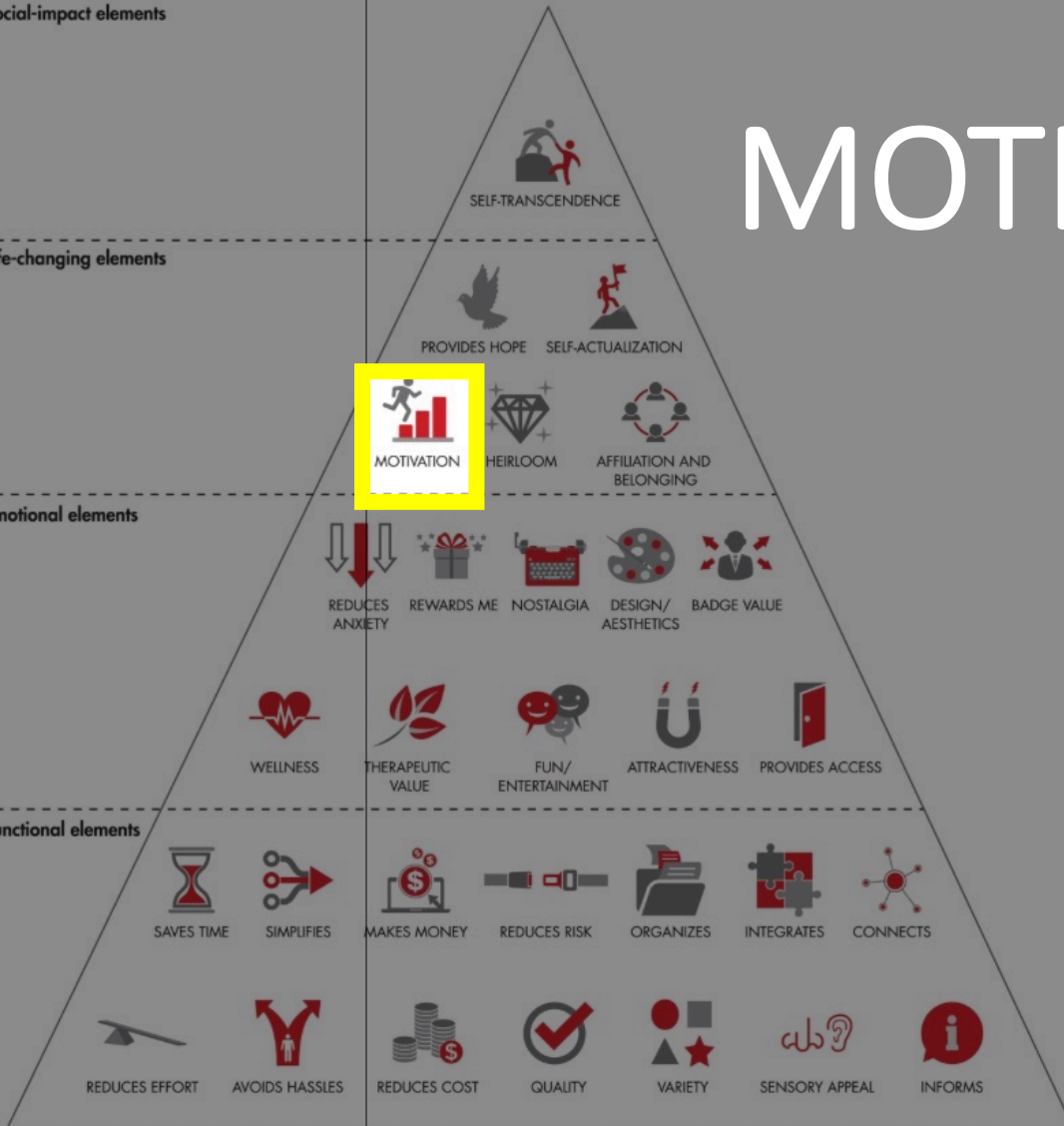
Inspirational value

Individual value

Ease of doing business value

Functional value

Table stakes



AI optimized training

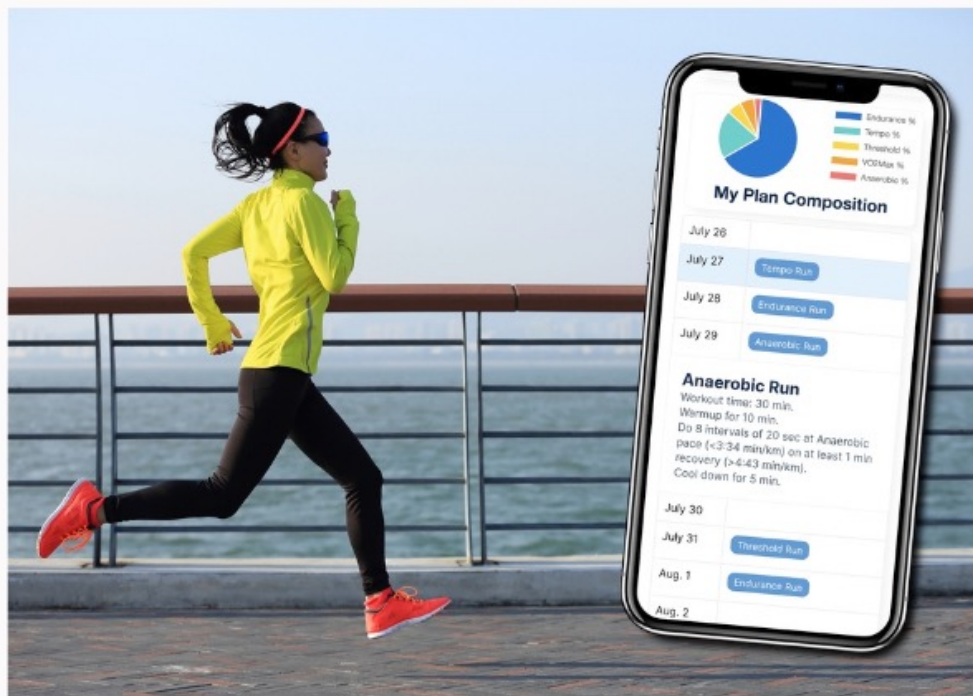
A scientific, data-driven training plan that actually improves your performance - we have the data to back it up. Artificial intelligence is your edge.

Whether you're a runner, cyclist, or triathlete, AI Endurance helps you get the best results from the time you invest in training. Our app creates personalized training plans using AI, to prepare you for any race or to simply keep you in shape.

Try 14 days FREE with no upfront payment information required!

[Get Started](#)

[Learn More](#)



B2C

B2B

ETHICAL STANDARDS

Social-impact elements

Life-changing elements

Emotional elements

Functional elements

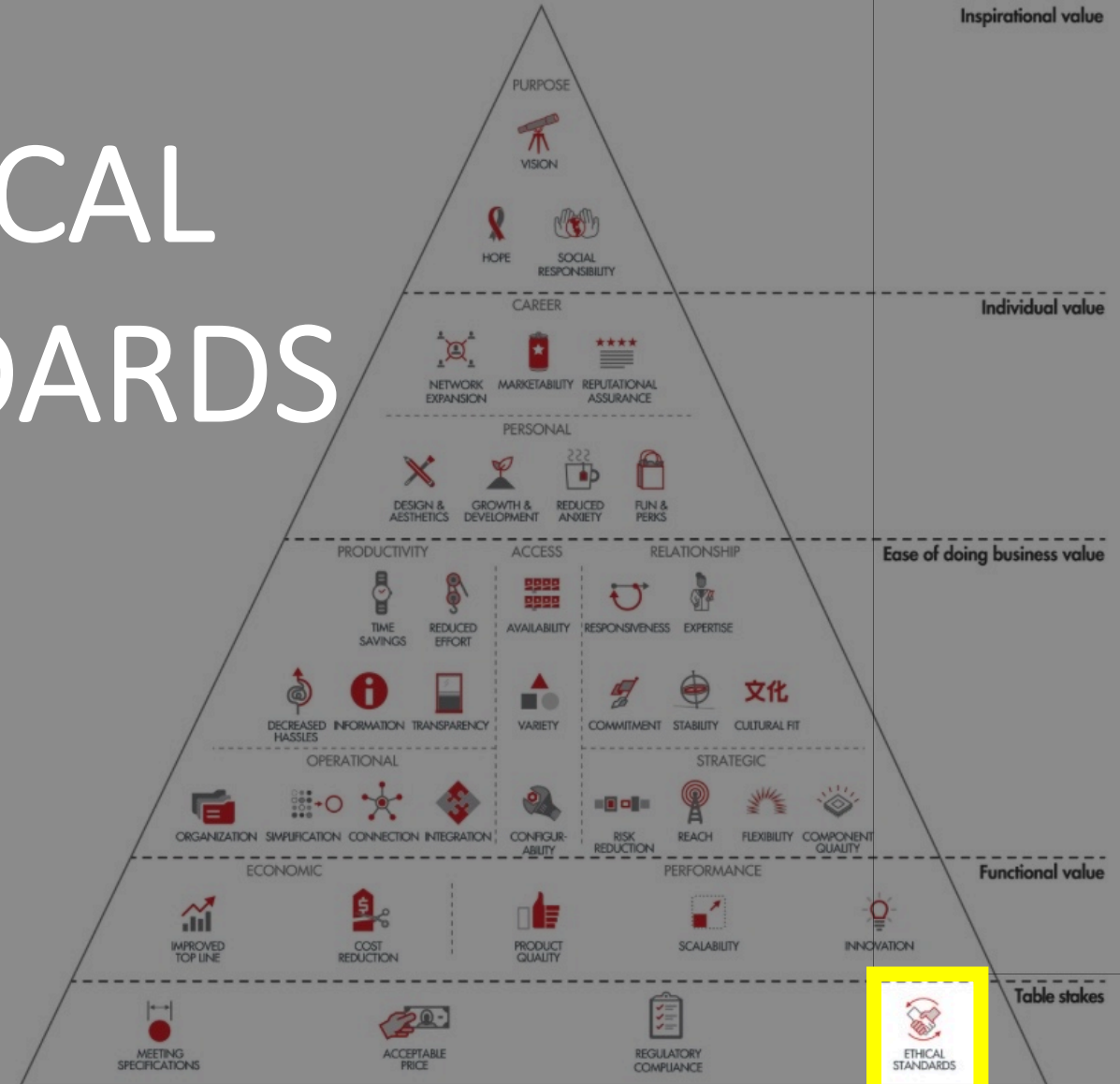
Inspirational value

Individual value

Ease of doing business value

Functional value

Table stakes

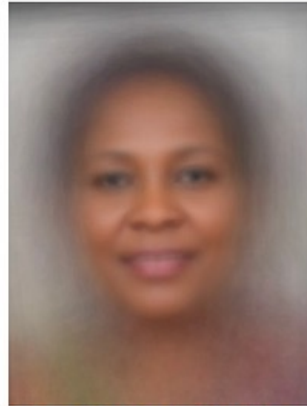


98.7%



**DARKER
MALES**

68.6%



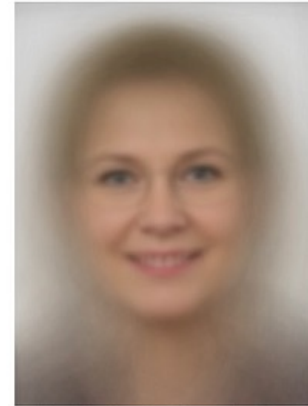
**DARKER
FEMALES**

100%



**LIGHTER
MALES**

92.9%



**LIGHTER
FEMALES**

B2C

B2B

TRANSPARENCY

Social-impact elements

Life-changing elements

Emotional elements

Functional elements

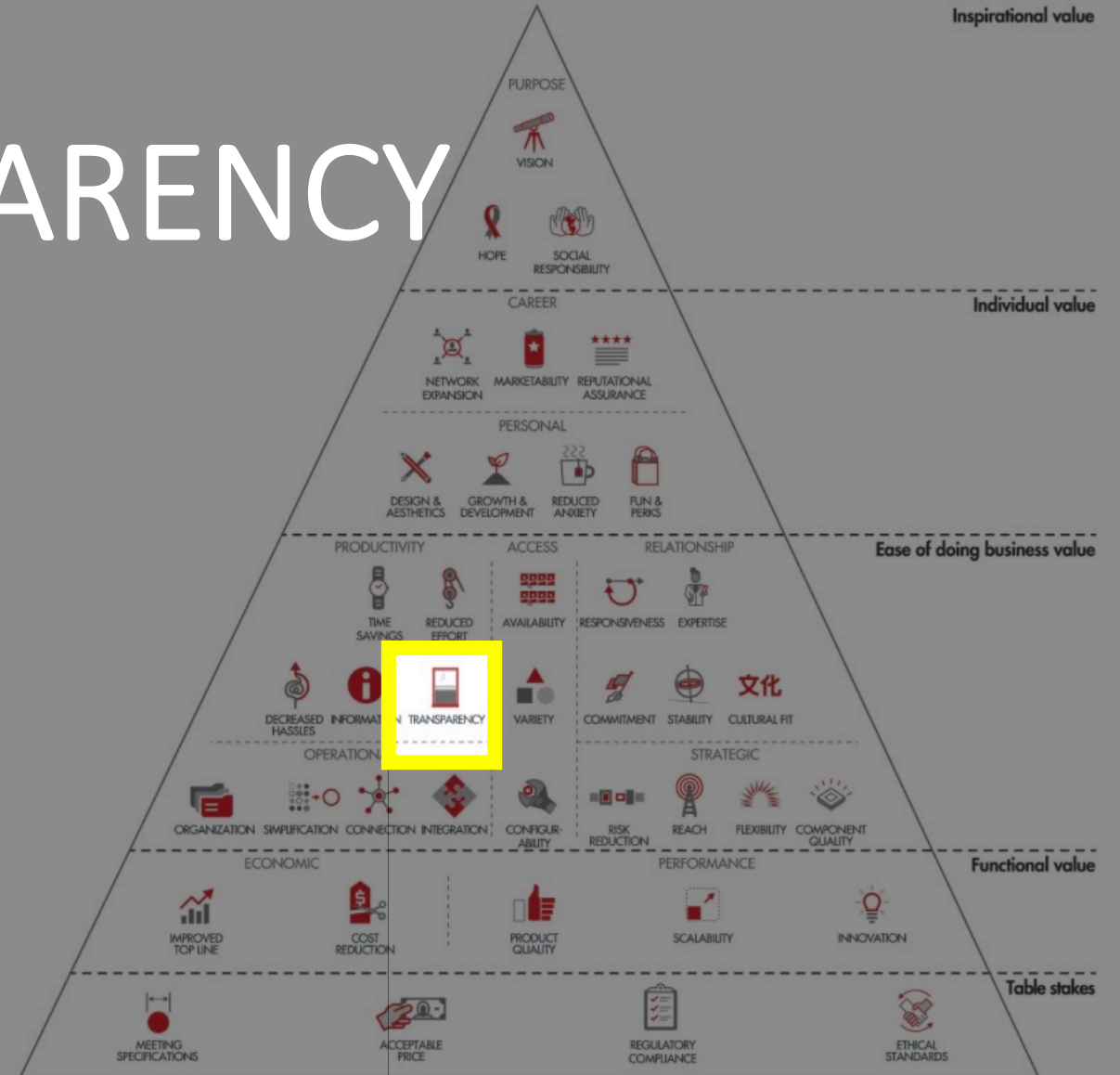
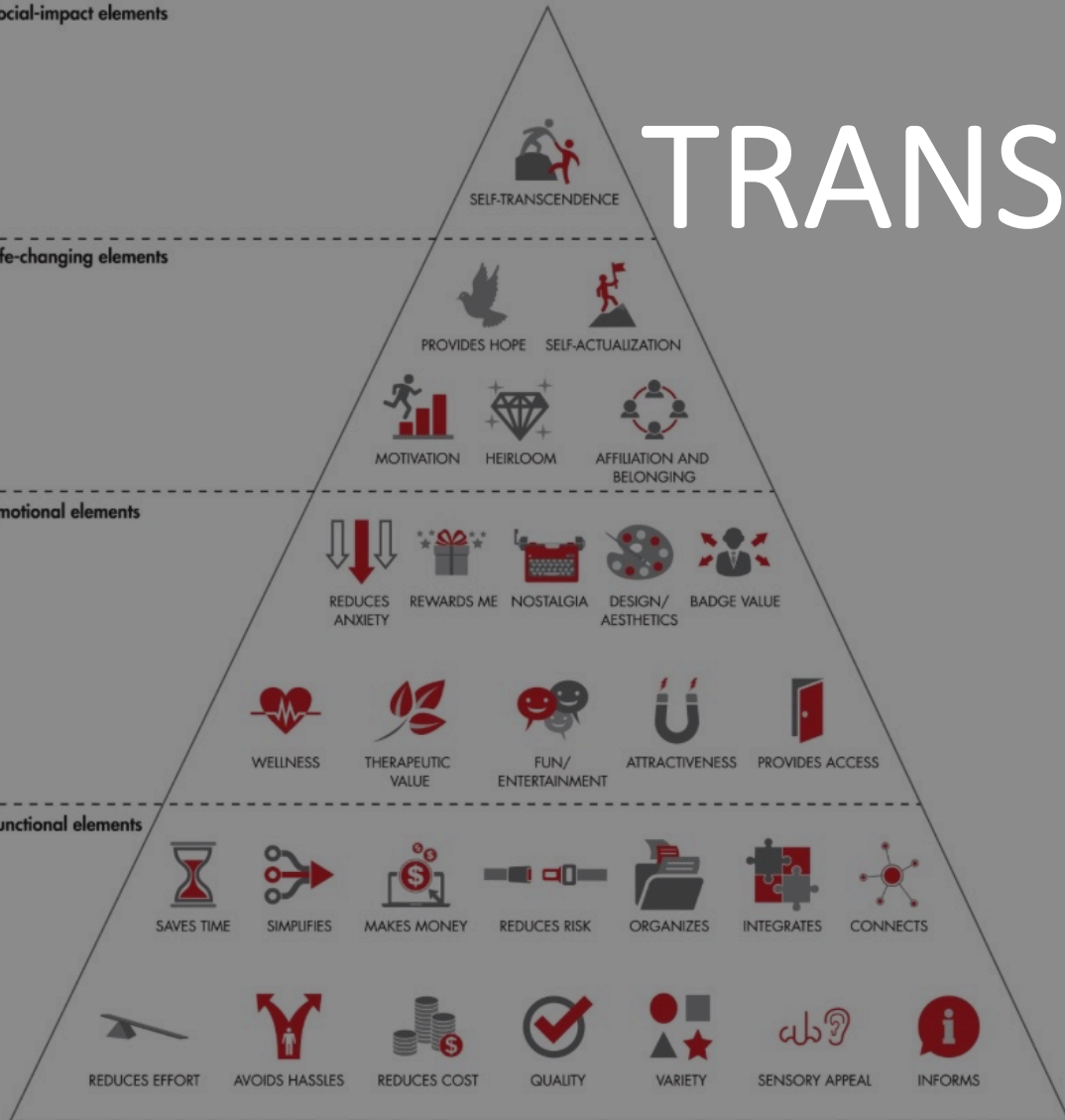
Inspirational value

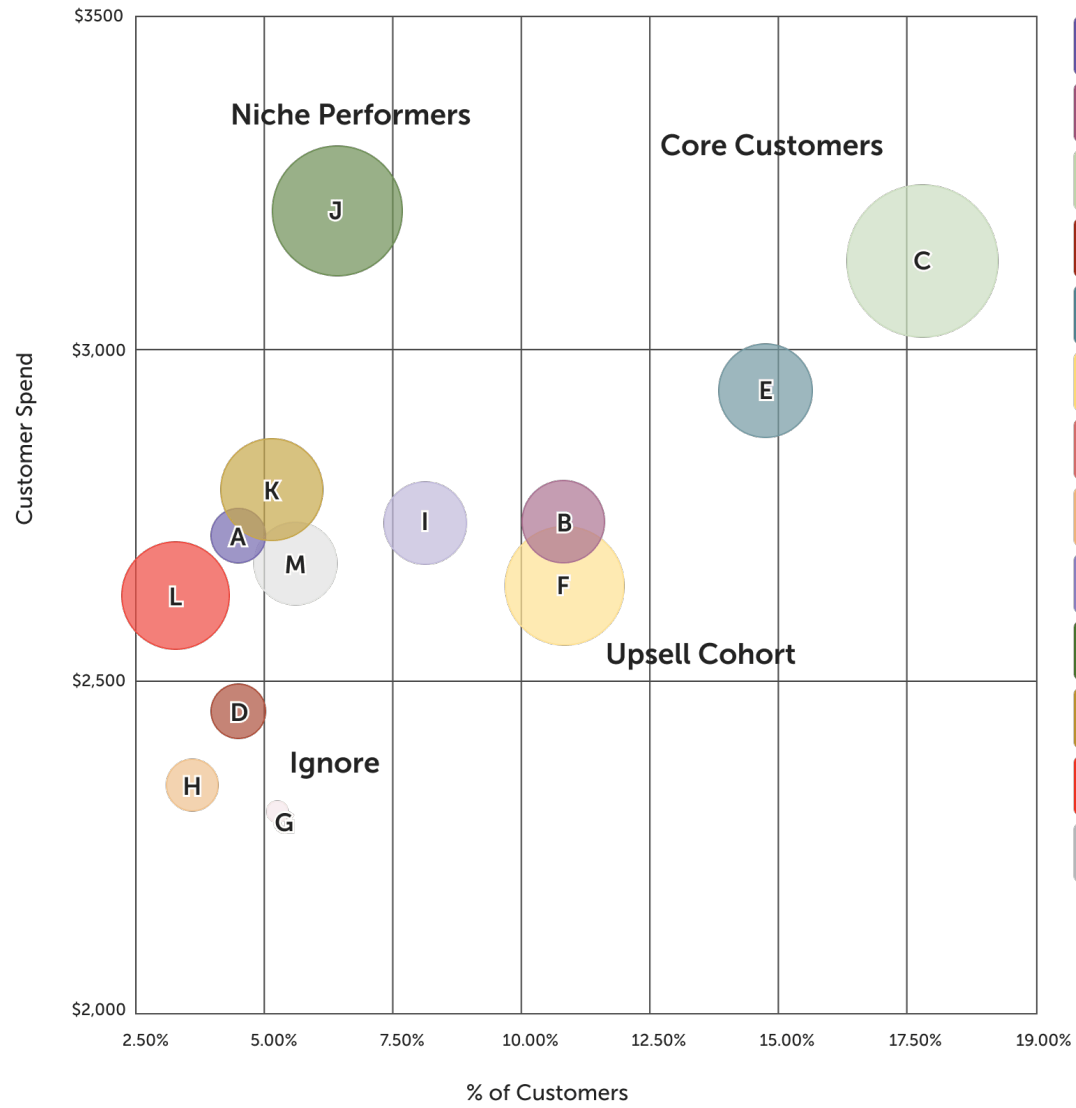
Individual value

Ease of doing business value

Functional value

Table stakes





A	Ultra Wealthy
B	Elite Families
C	Upper-Middle Families
D	Educated Urbanites
E	Middle Age Suburban
F	Middle Class Suburbs
G	Retirement Years
H	Young Professionals
I	Young Starters
J	Country
K	Rural Low Income
L	Hispanic Families
M	Urban Low Income

Large Market
>10% of national pop.

Medium Market
>5% of national pop.

Niche Market
<5% of national pop.



B2C

B2B

AVAILABILITY

Social-impact elements

Life-changing elements

Emotional elements

Functional elements

Inspirational value

Individual value

Ease of doing business value

Functional value

Table stakes



Jimmy 5:41 PM

How can I change my password?



Pete the Chatbot 5:42 PM

Please give me your email address.
I'll send you a special link.

Jimmy 5:41 PM

jimmy@sample.com



Pete the Chatbot 5:42 PM

Thanks! Please check your
email inbox! 📧

× support

× password

× newsletter



B2C

B2B

PRODUCTIVITY

Social-impact elements

Life-changing elements

Emotional elements

Functional elements

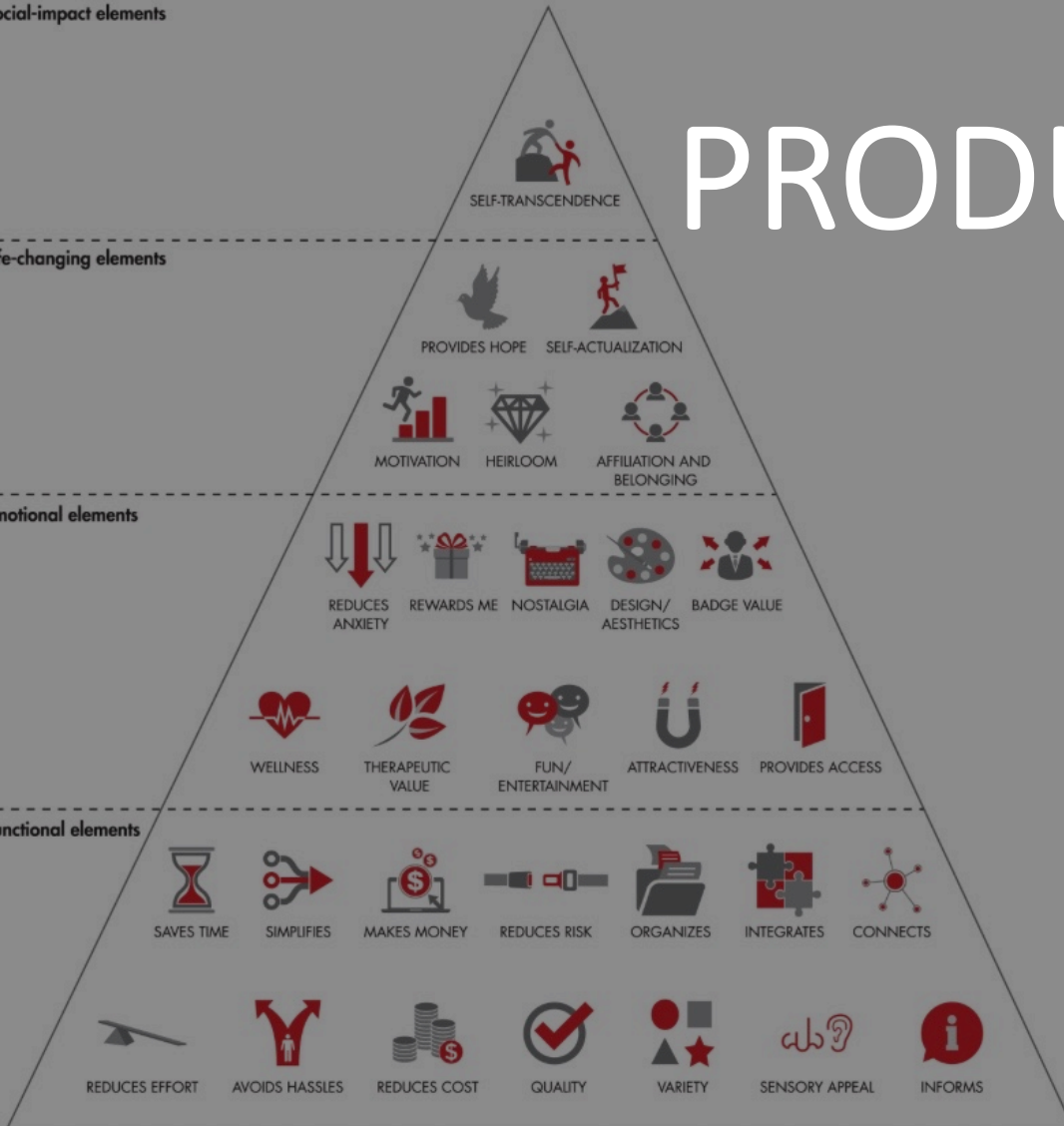
Inspirational value

Individual value

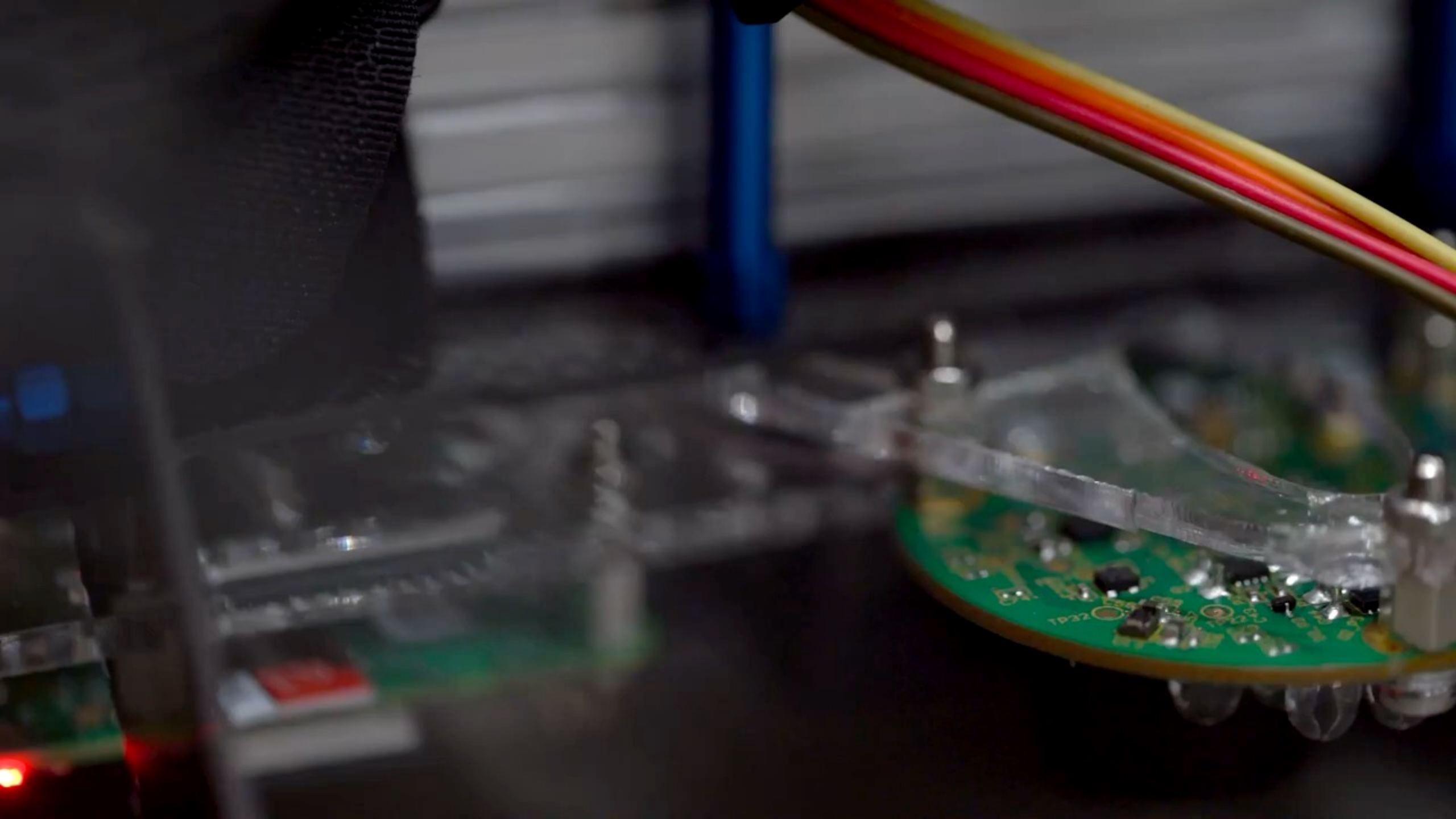
Ease of doing business value

Functional value

Table stakes







B2C

B2B

Social-impact elements

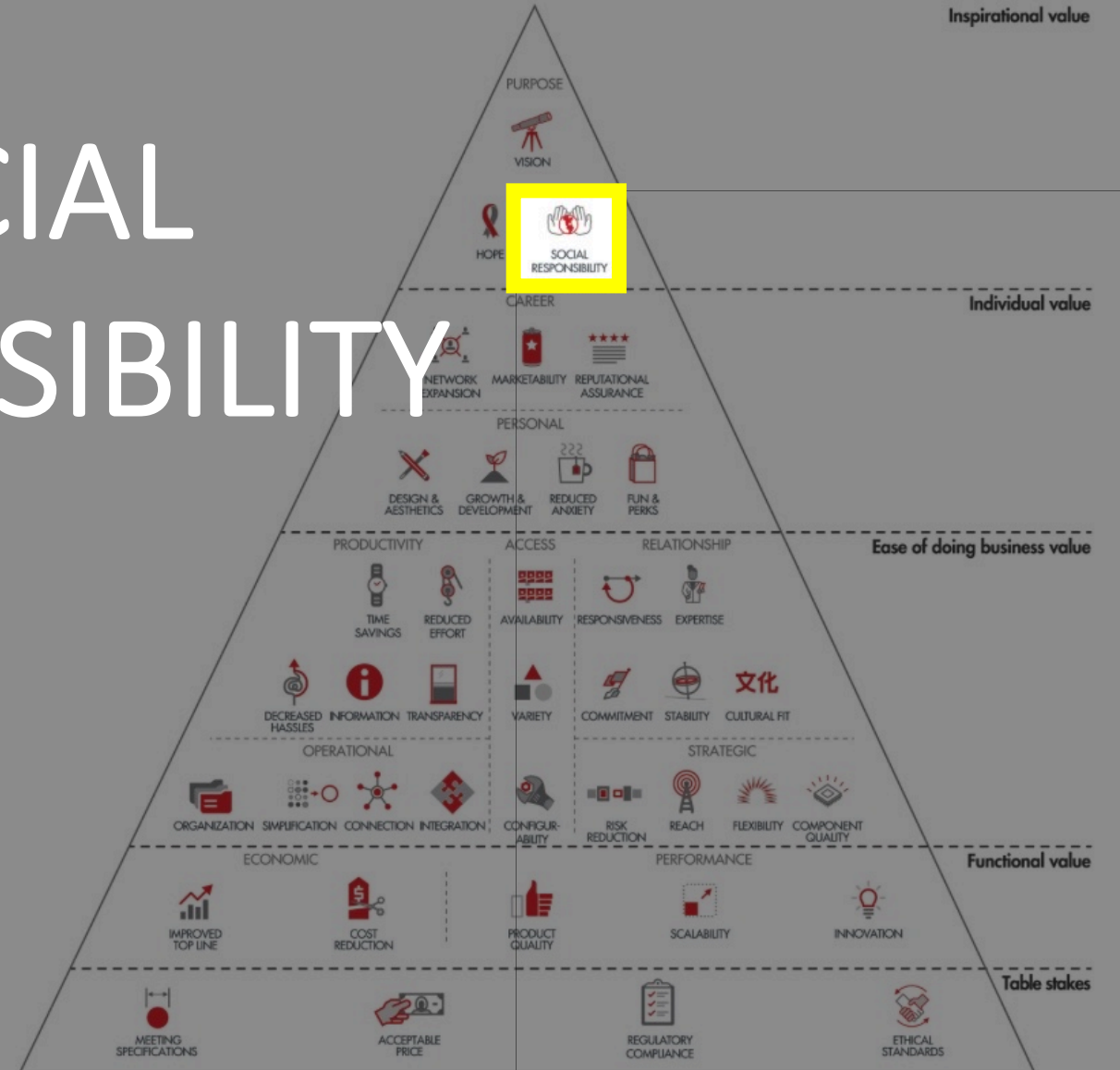
Inspirational value

Life-changing elements

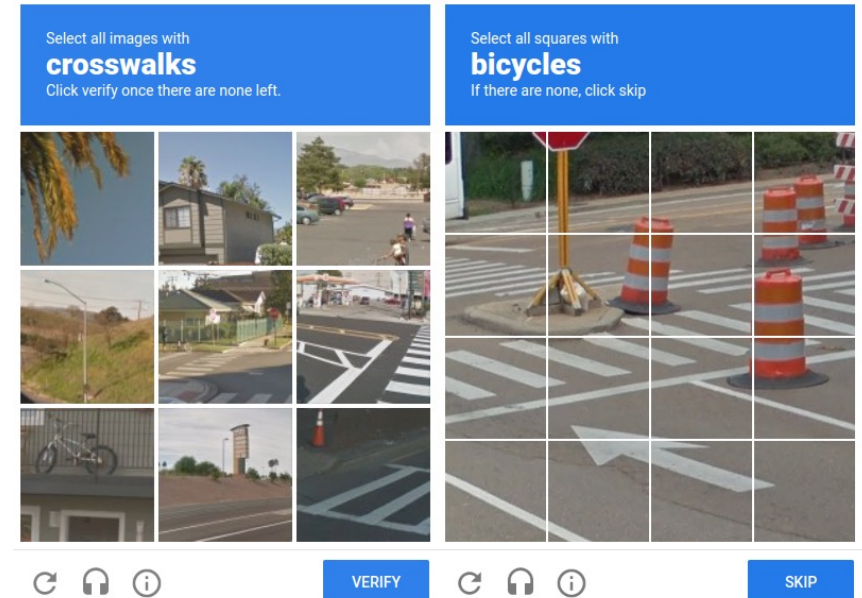
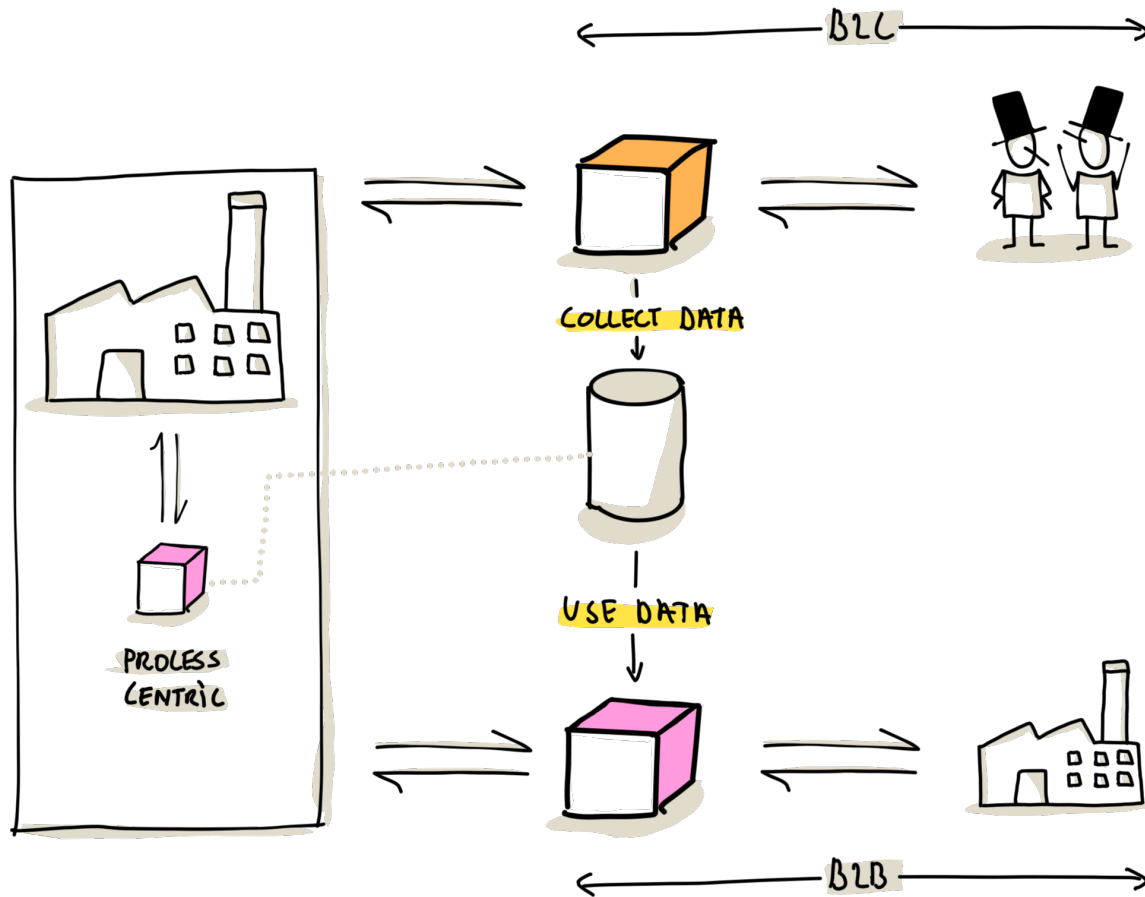
Emotional elements

Functional elements

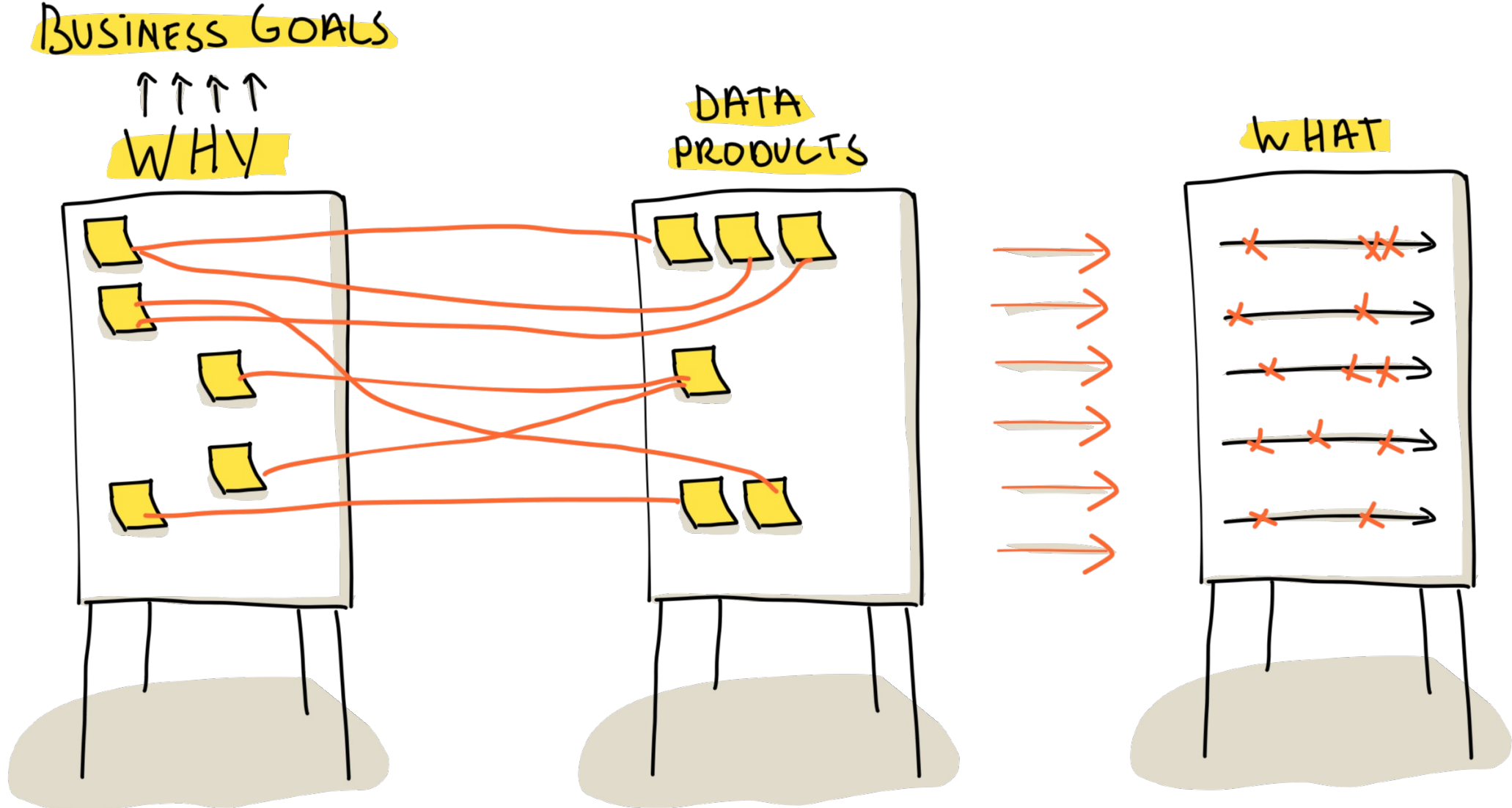
SOCIAL RESPONSIBILITY



Everything Connects...



EXERCISE: FIND THE VALUE FOR YOUR DATA PRODUCTS



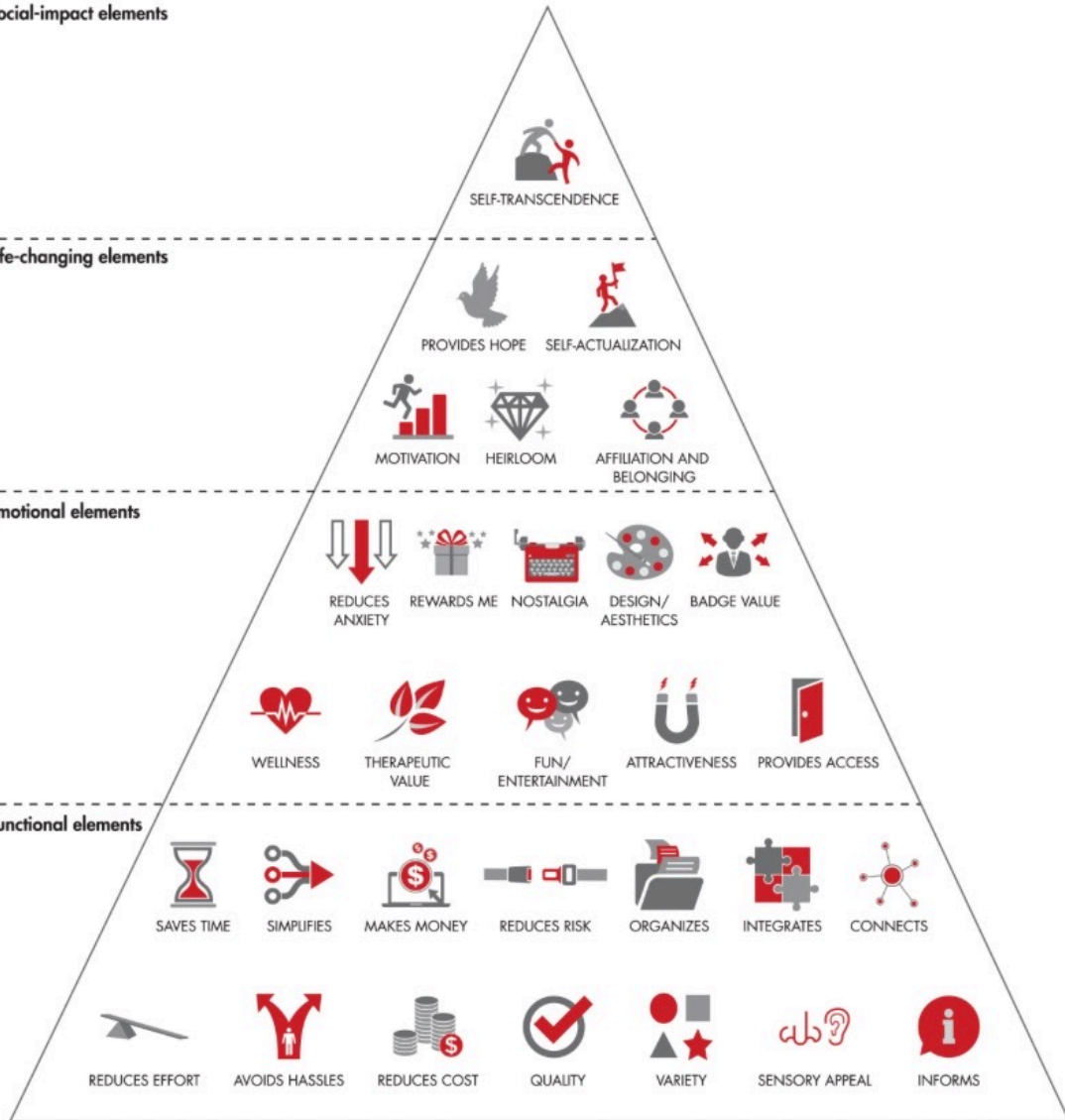
B2C

Social-impact elements

Life-changing elements

Emotional elements

Functional elements



B2B

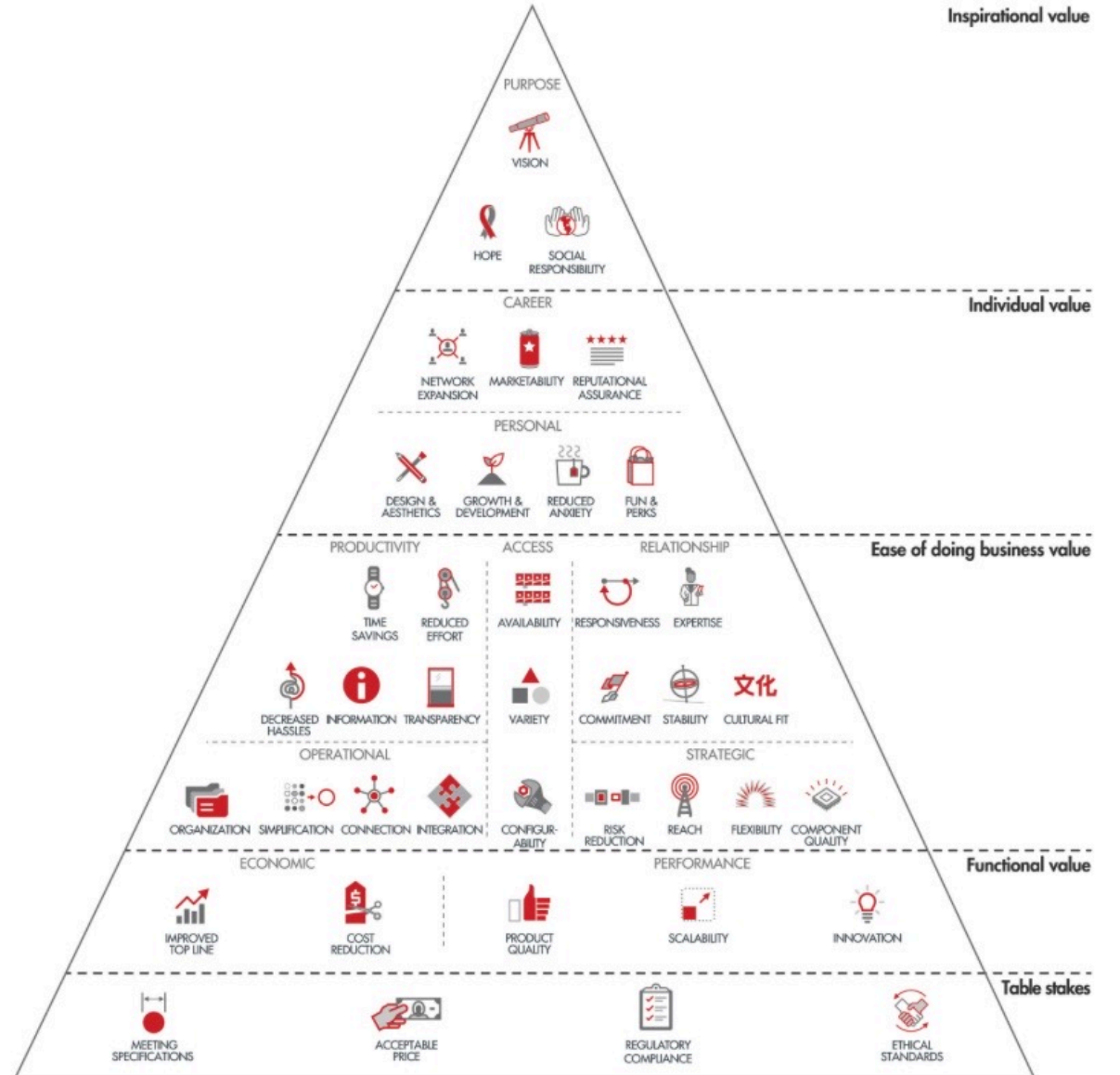
Inspirational value

Individual value

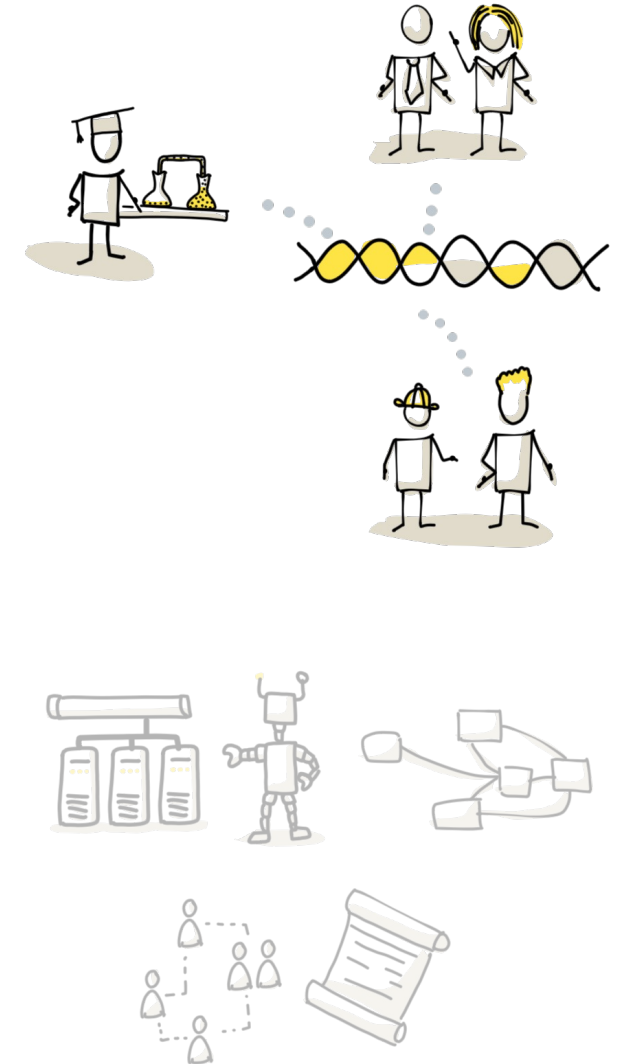
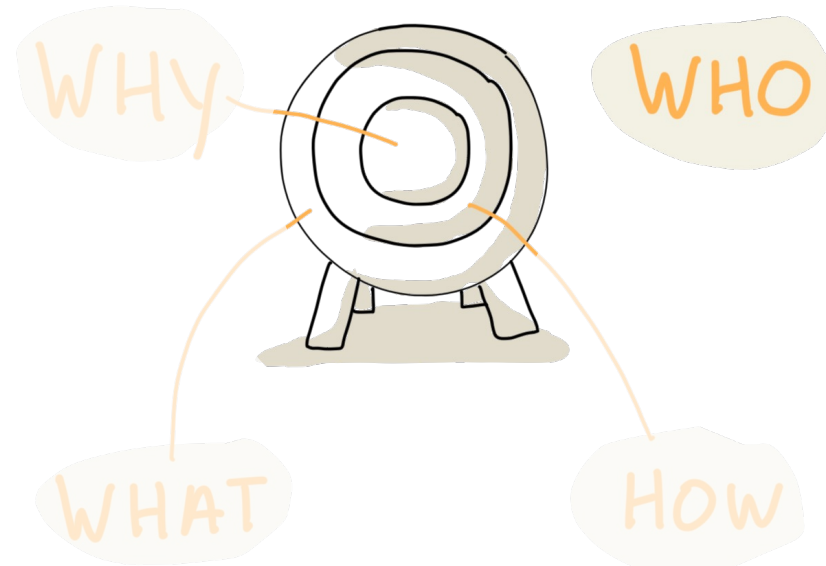
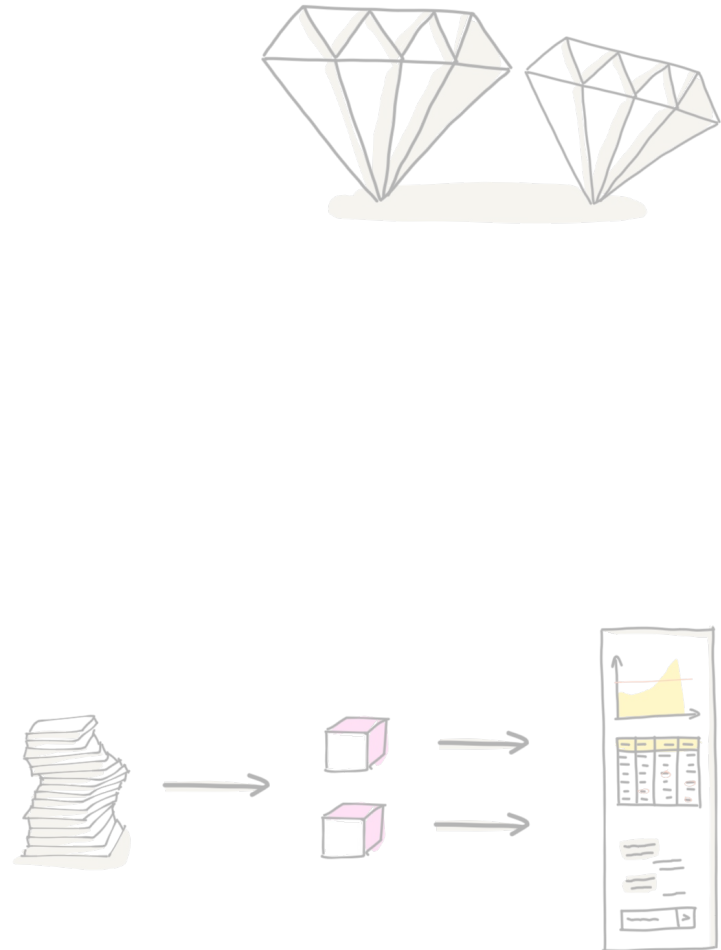
Ease of doing business value

Functional value

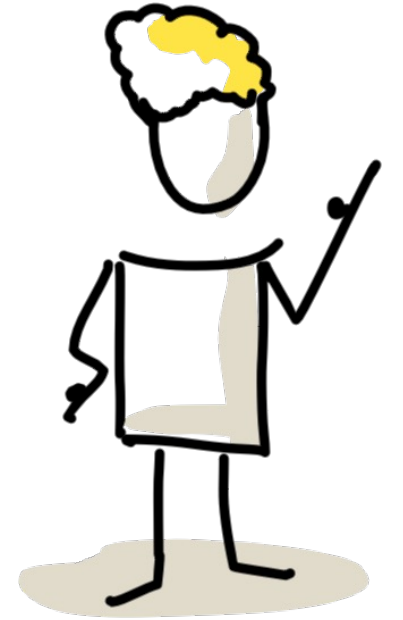
Table stakes



3 WHO



DATA USABILITY

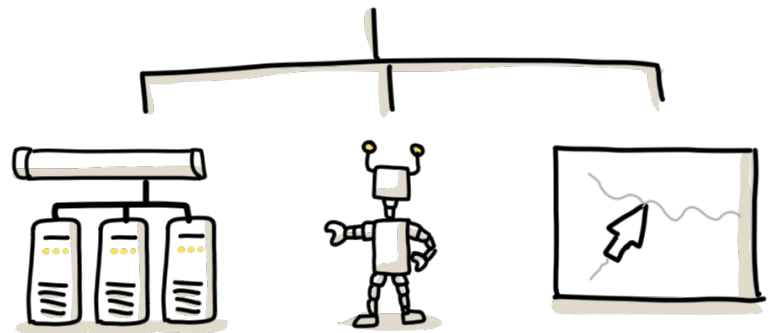
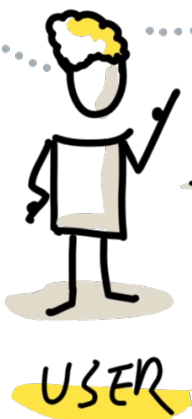
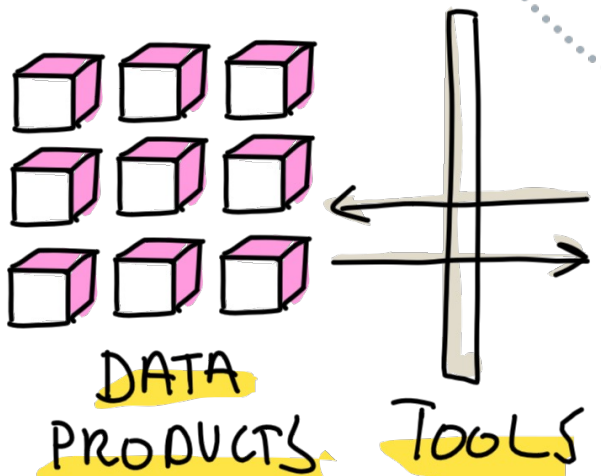
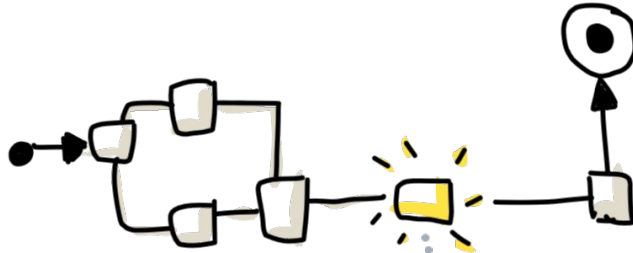


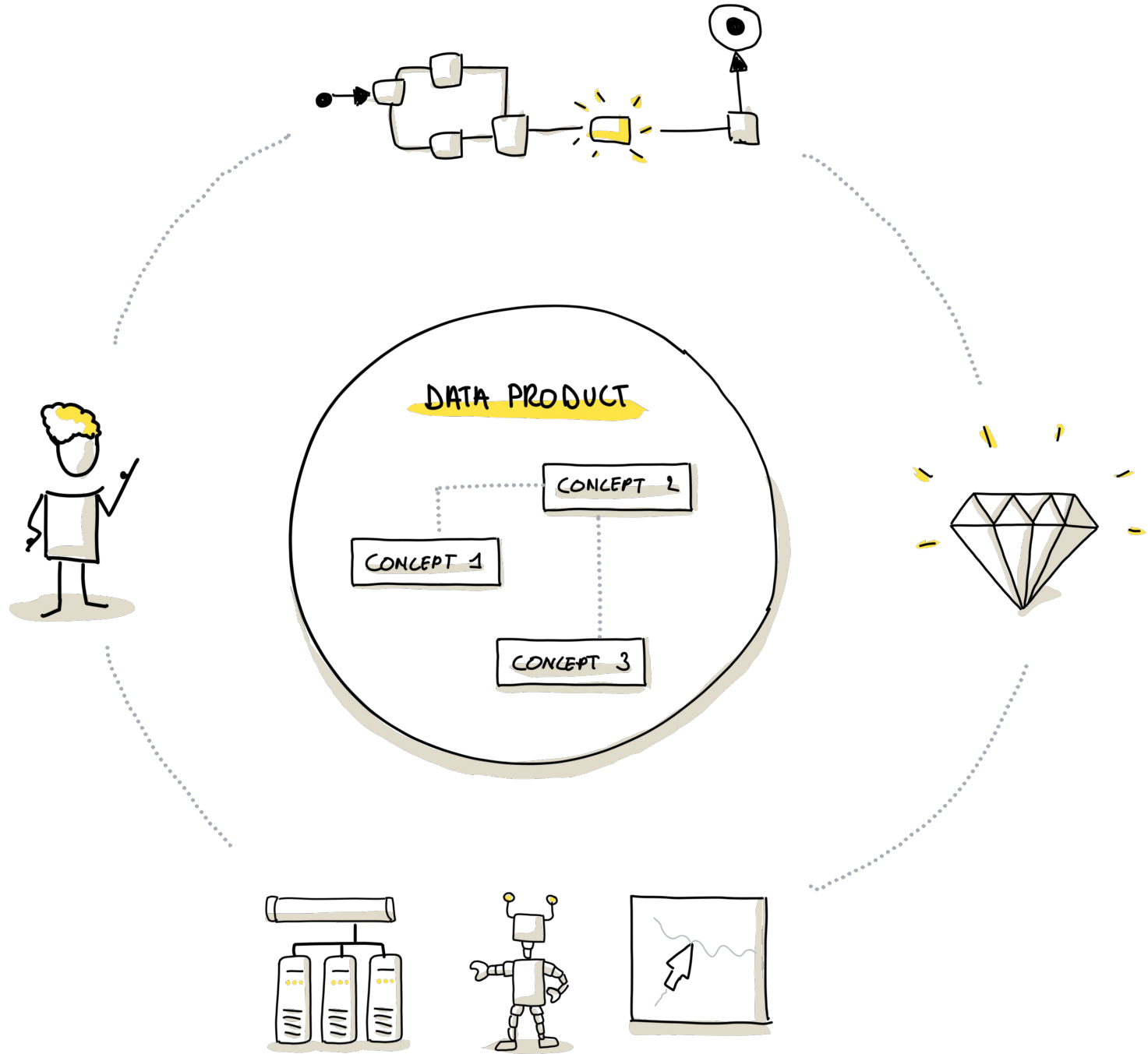
“A measure of how well **a user** in a specific context can **use data** to achieve a defined goal **effectively, efficiently, and satisfactorily.**”

PERSONA

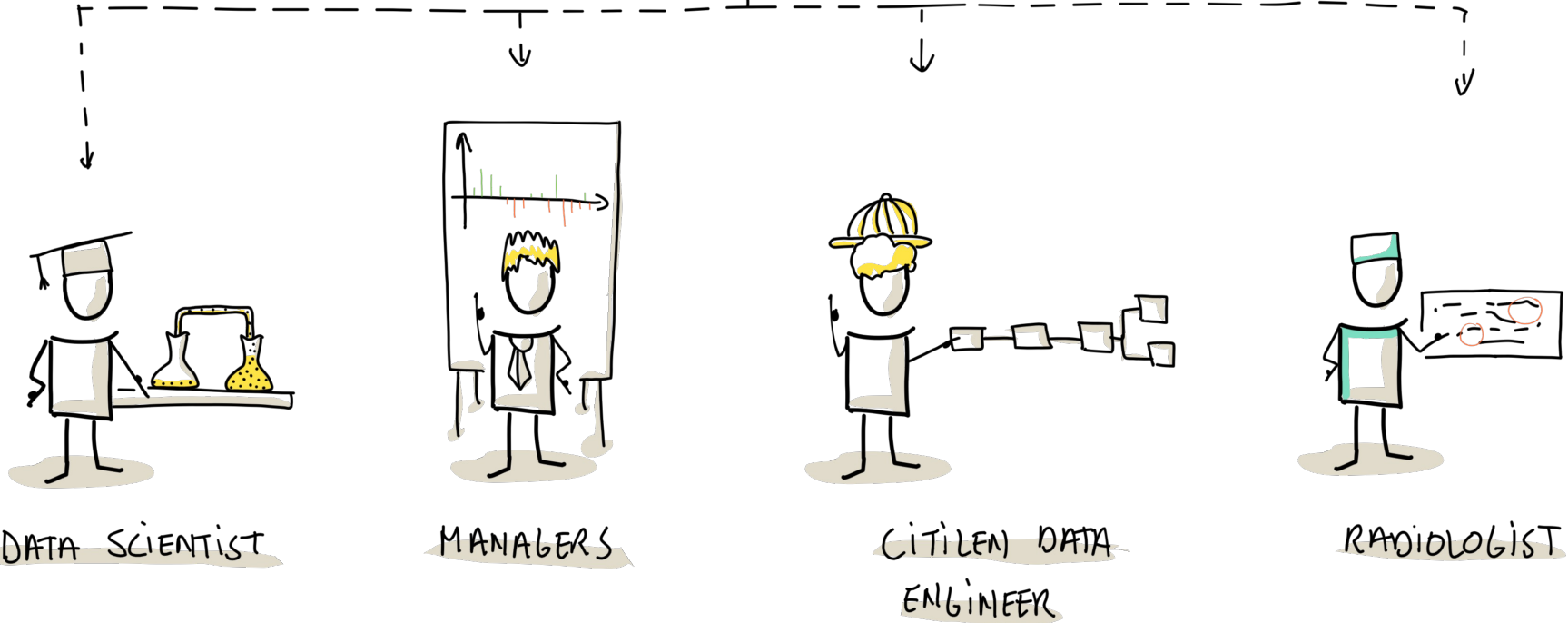
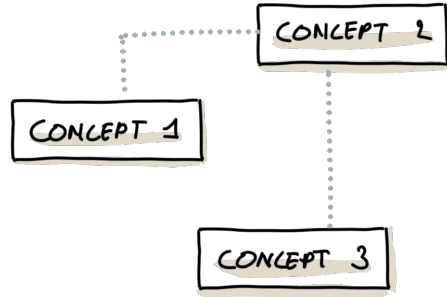


PROCESS & CONTEXT





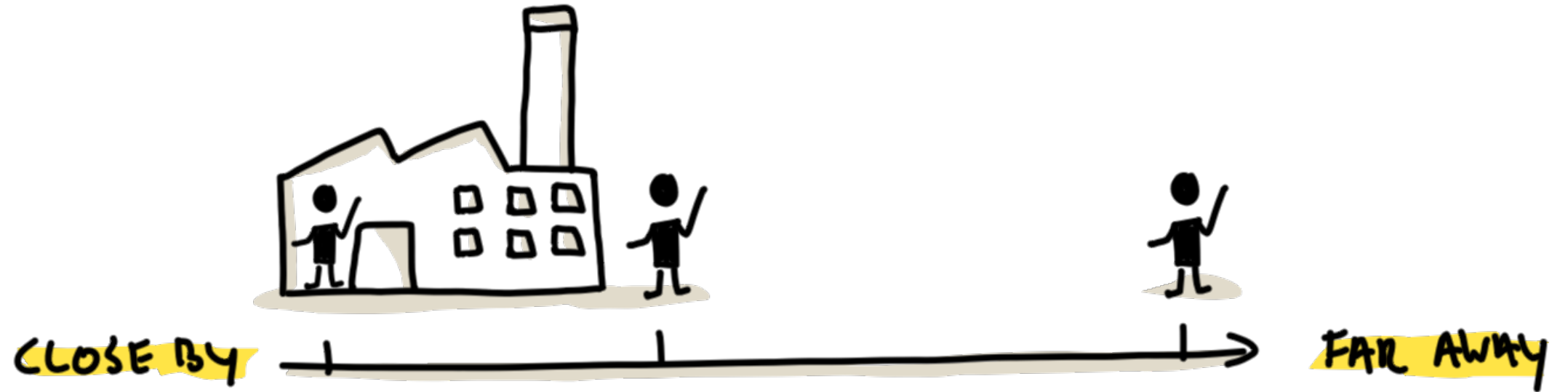
DATA PRODUCT



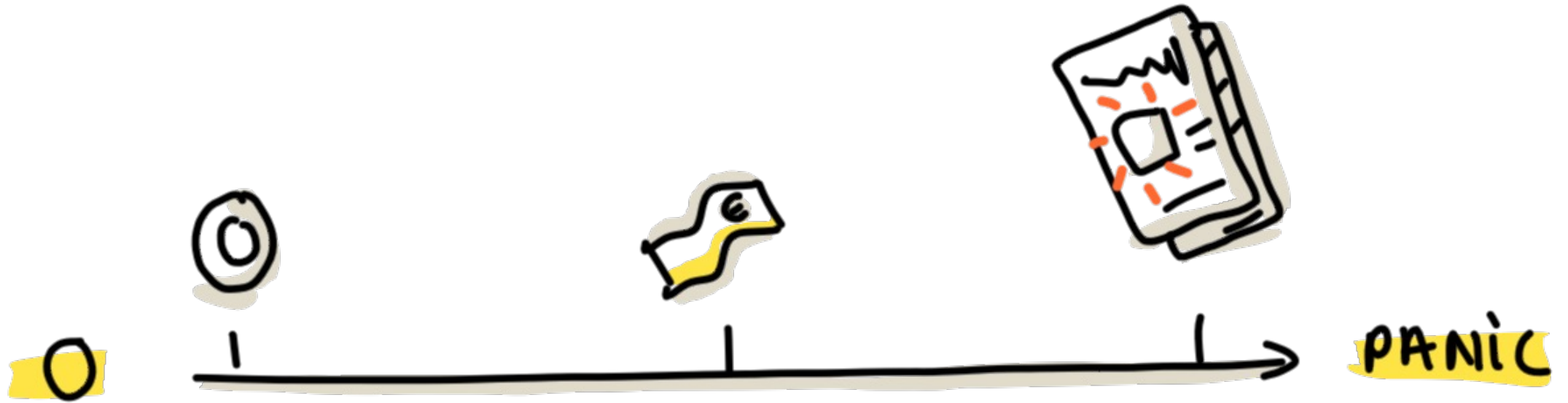
DATA PRODUCT CONSUMPTION



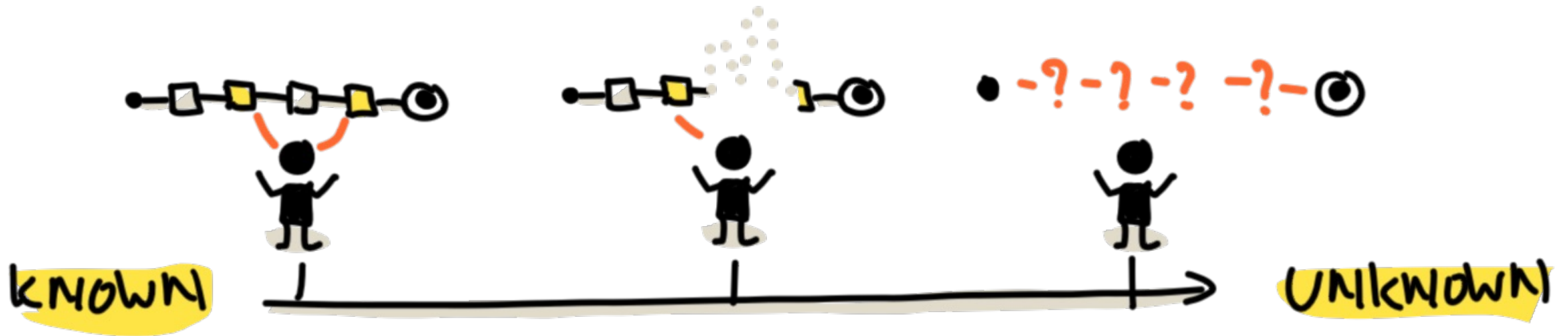
Reachability



Cost of Error

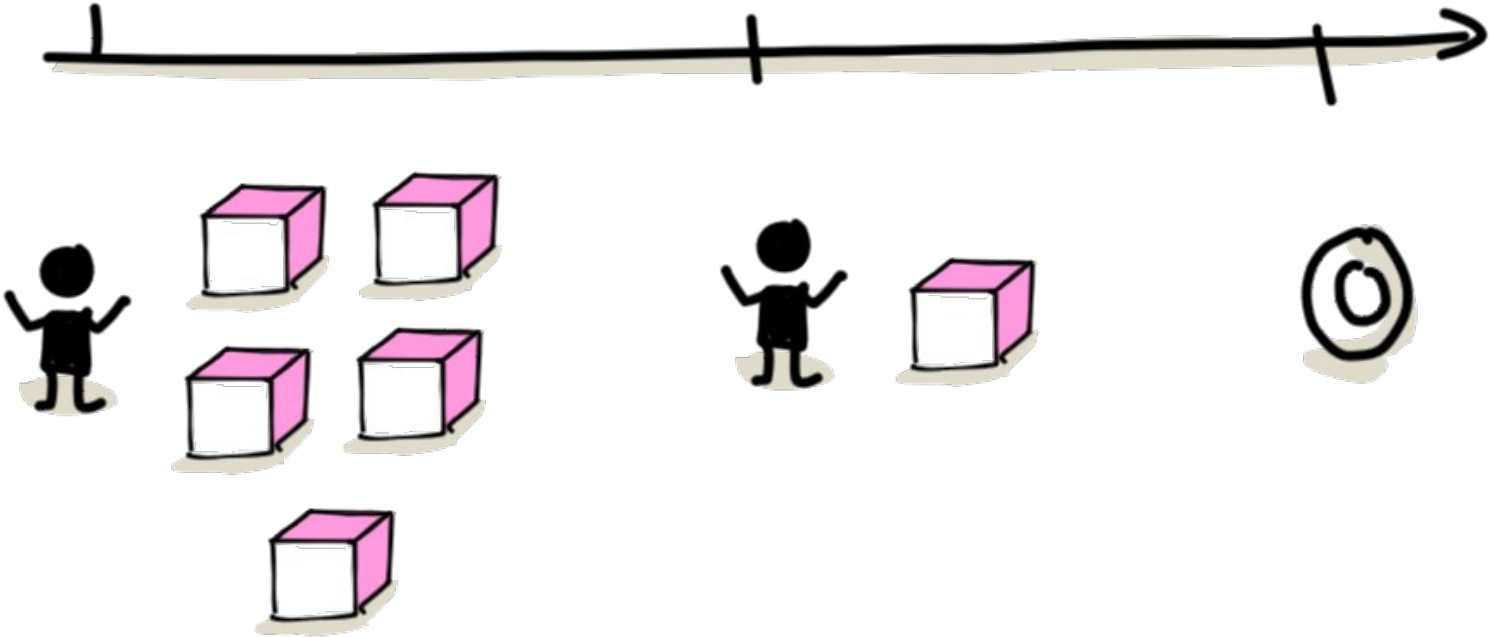


Knowledge of Context of Data Use



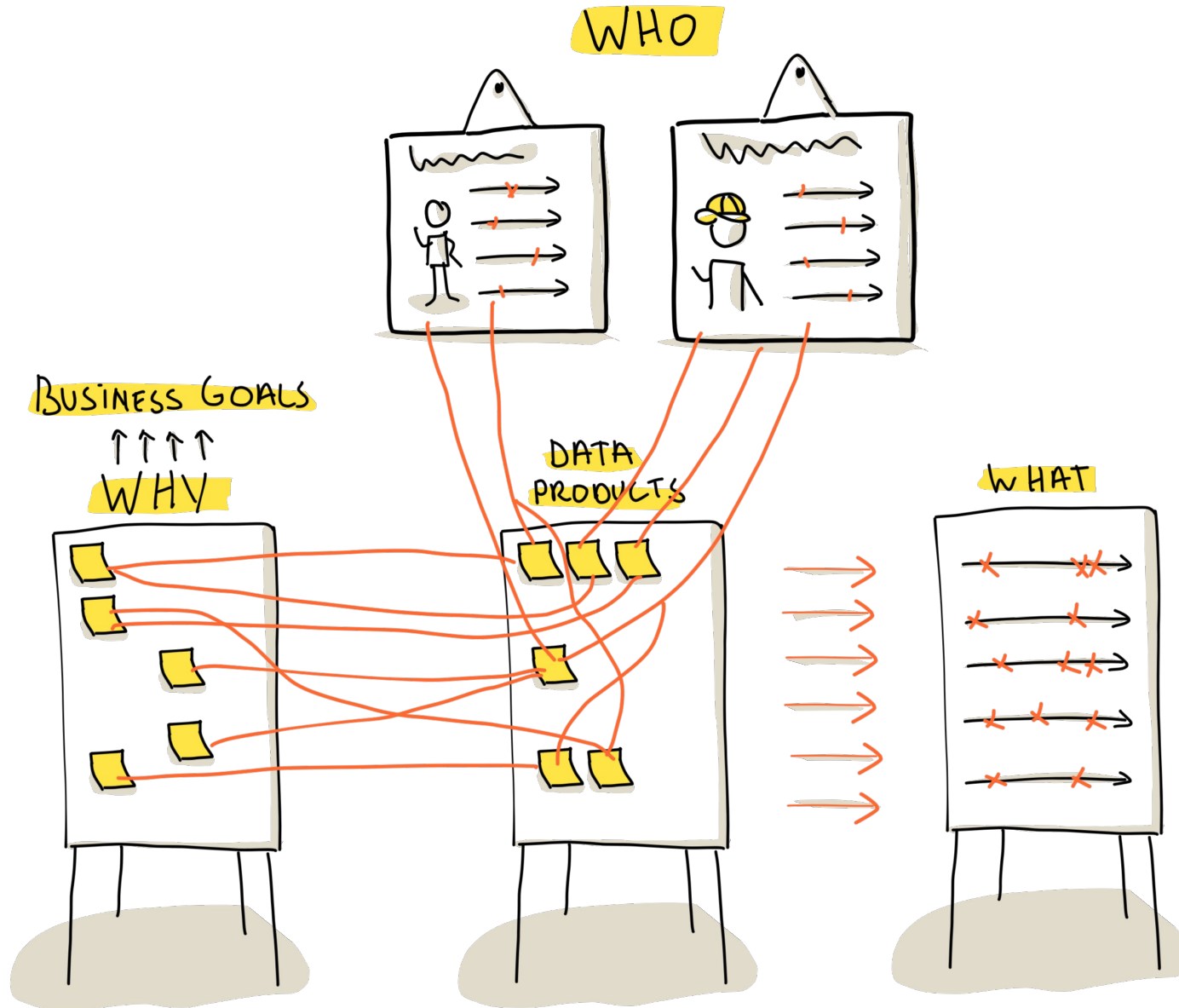
Collaboration

VERY OFTEN

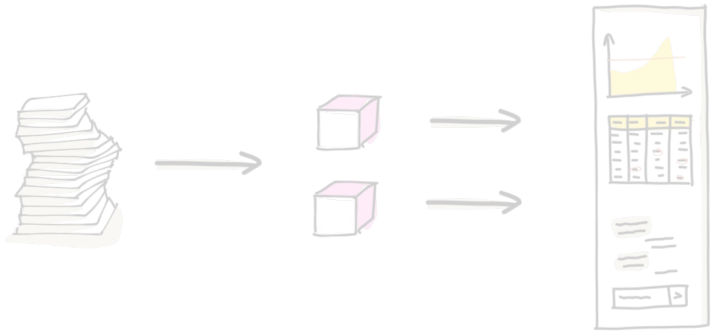
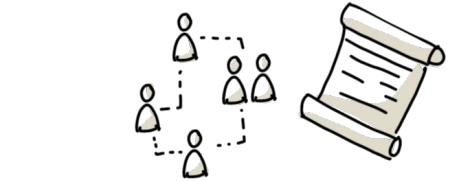
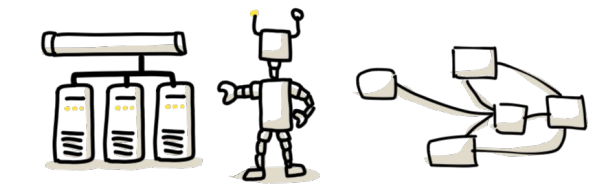
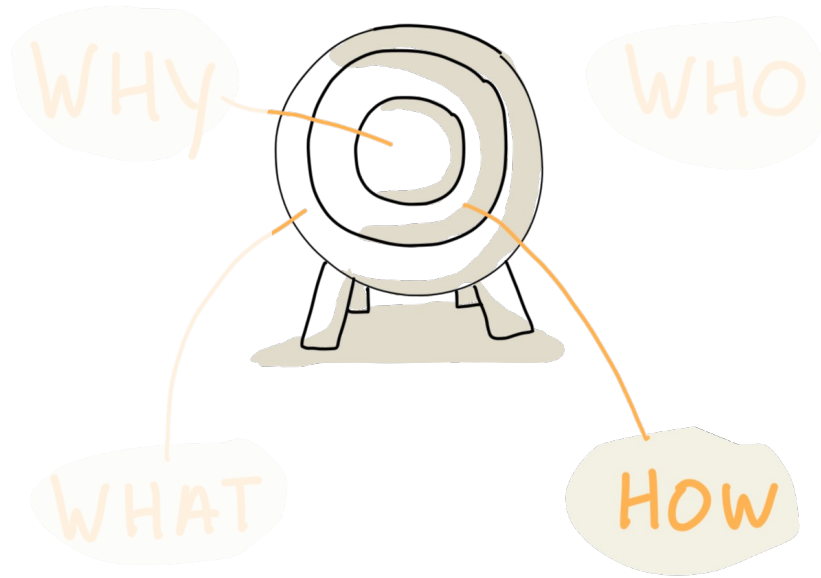
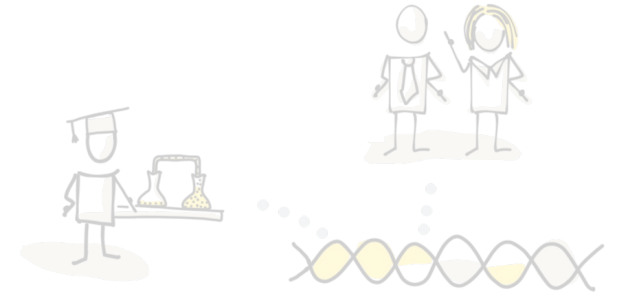
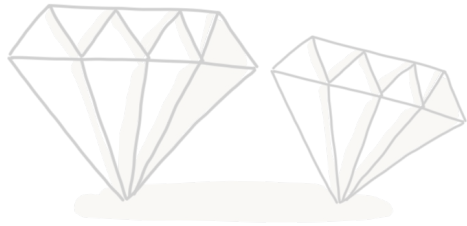


NEVER

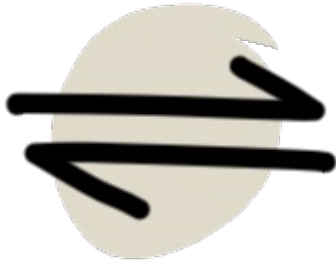
EXERCISE: WHO WILL USE YOUR PRODUCTS?



4 HOW

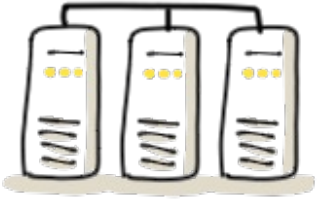


DATA PLATFORM

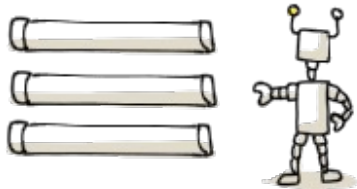


DATA GOVERNANCE

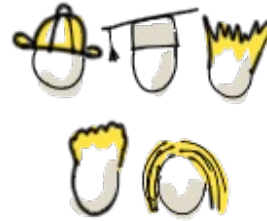
INFRASTRUCTURE



TECHNOLOGY



PEOPLE



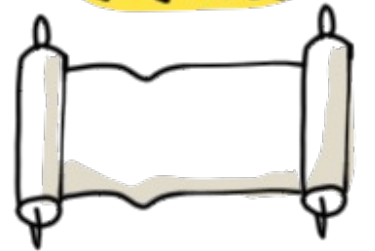
ORGANIZATION



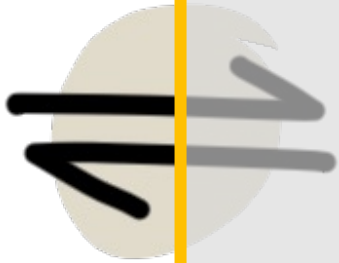
PROCESSES



POLICIES, STANDARDS & RULES

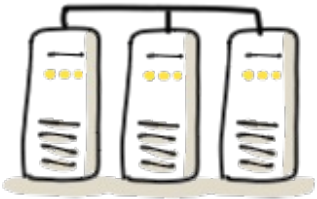


DATA PLATFORM

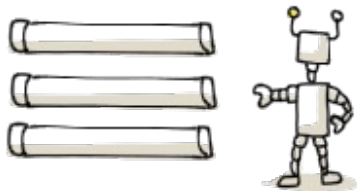


DATA GOVERNANCE

INFRASTRUCTURE



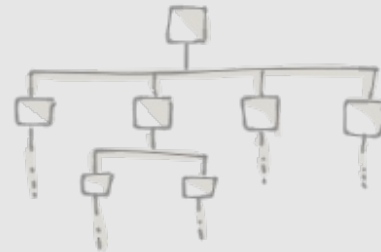
TECHNOLOGY



PEOPLE



ORGANIZATION



PROCESSES



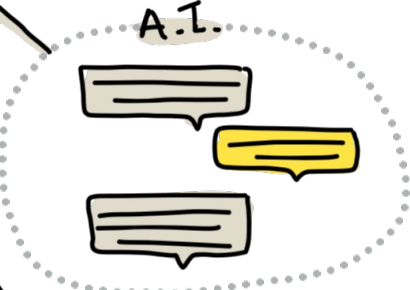
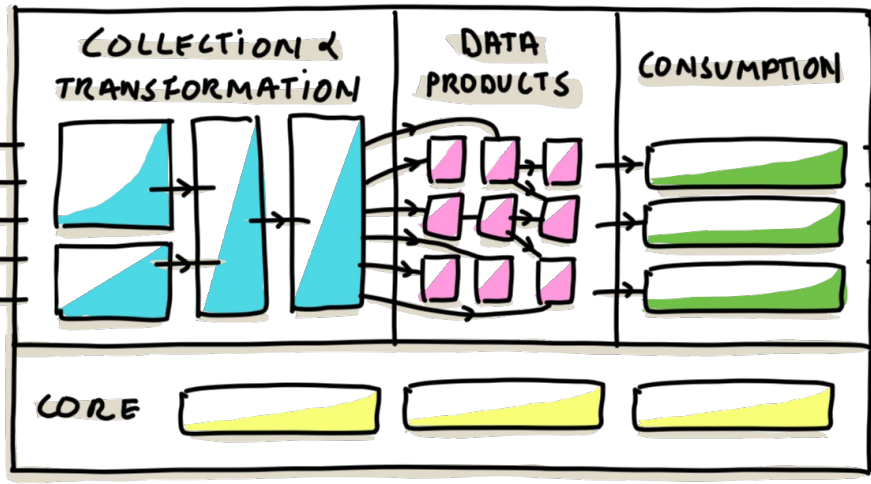
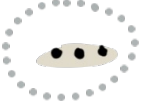
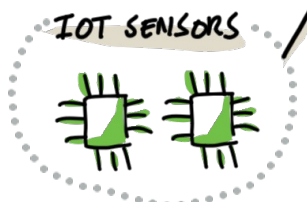
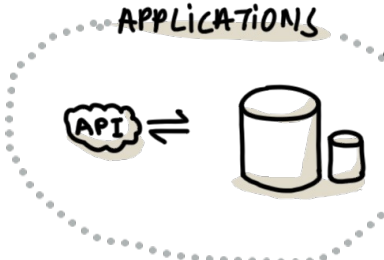
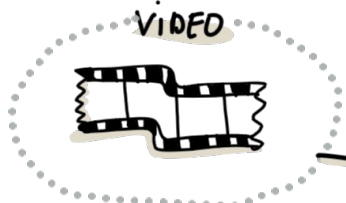
POLICIES, STANDARDS & RULES



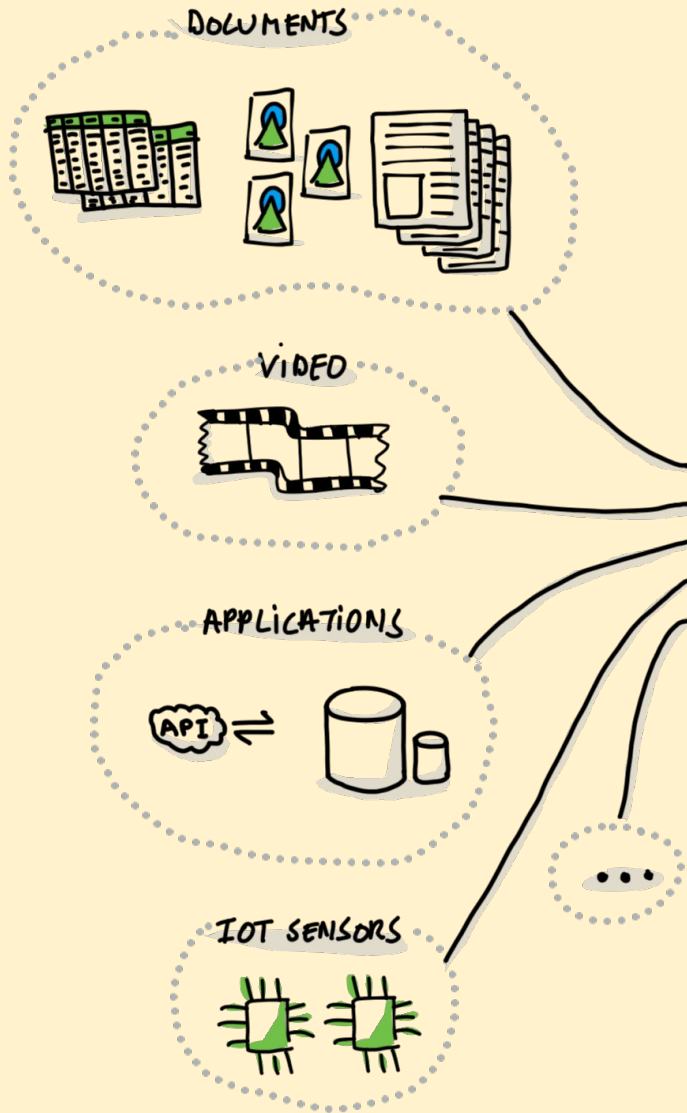
DATA SOURCES

DATA PLATFORM

CONSUMERS

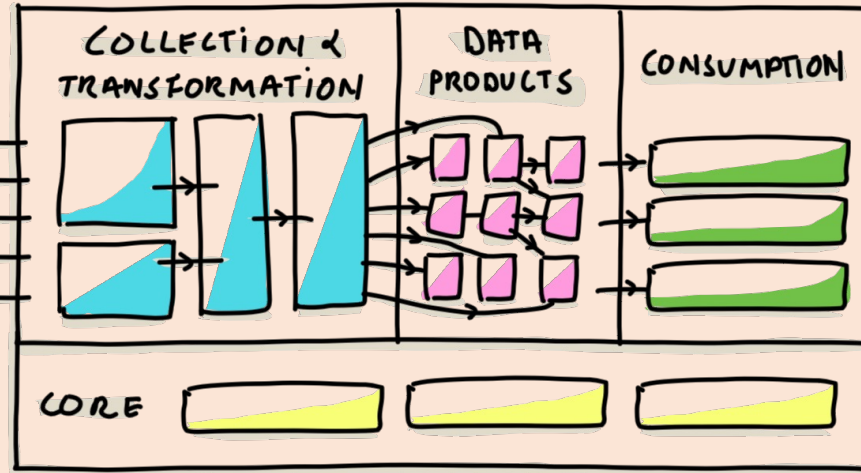


DATA SOURCES



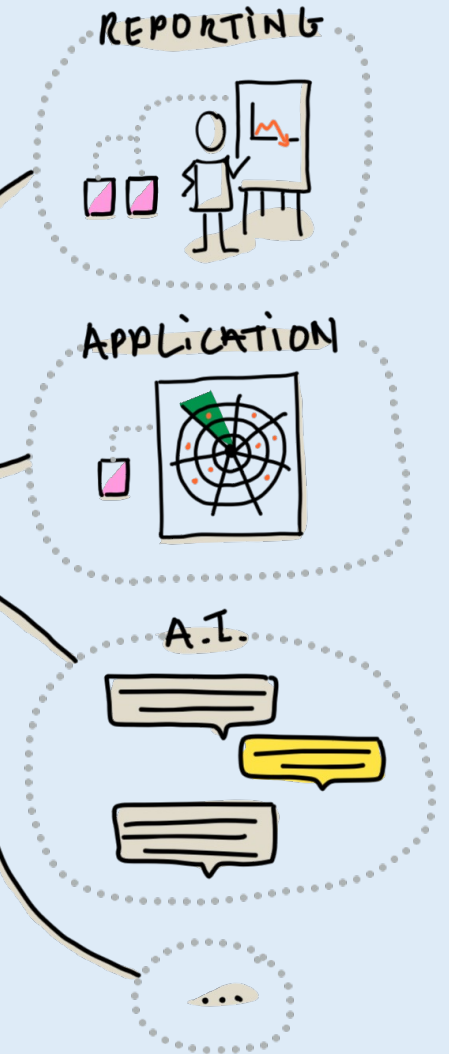
“Operational Plane”

DATA PLATFORM



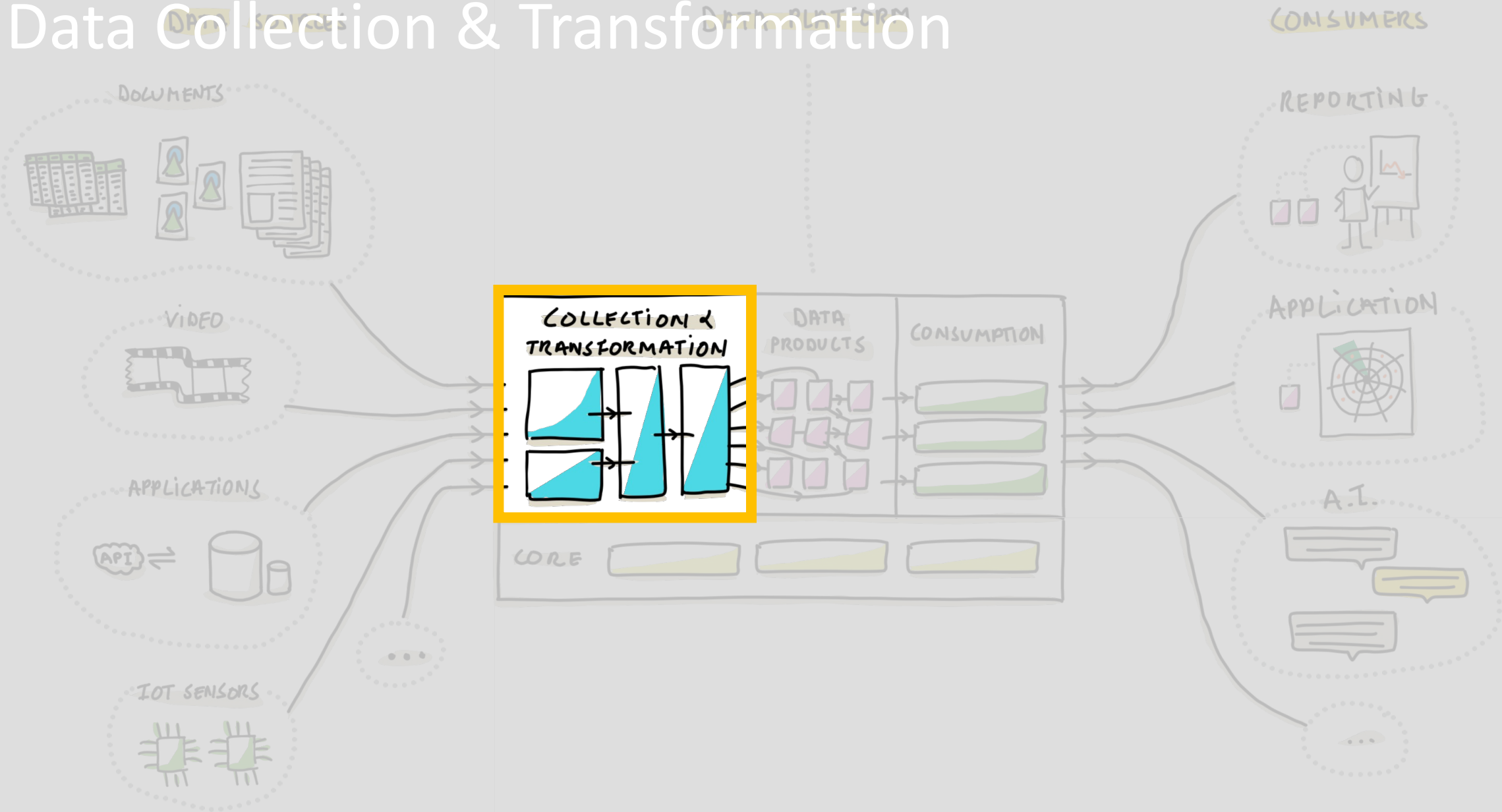
“Analytical Plane”

CONSUMERS



“Operational / Analytical Plane”

1. Data Collection & Transformation

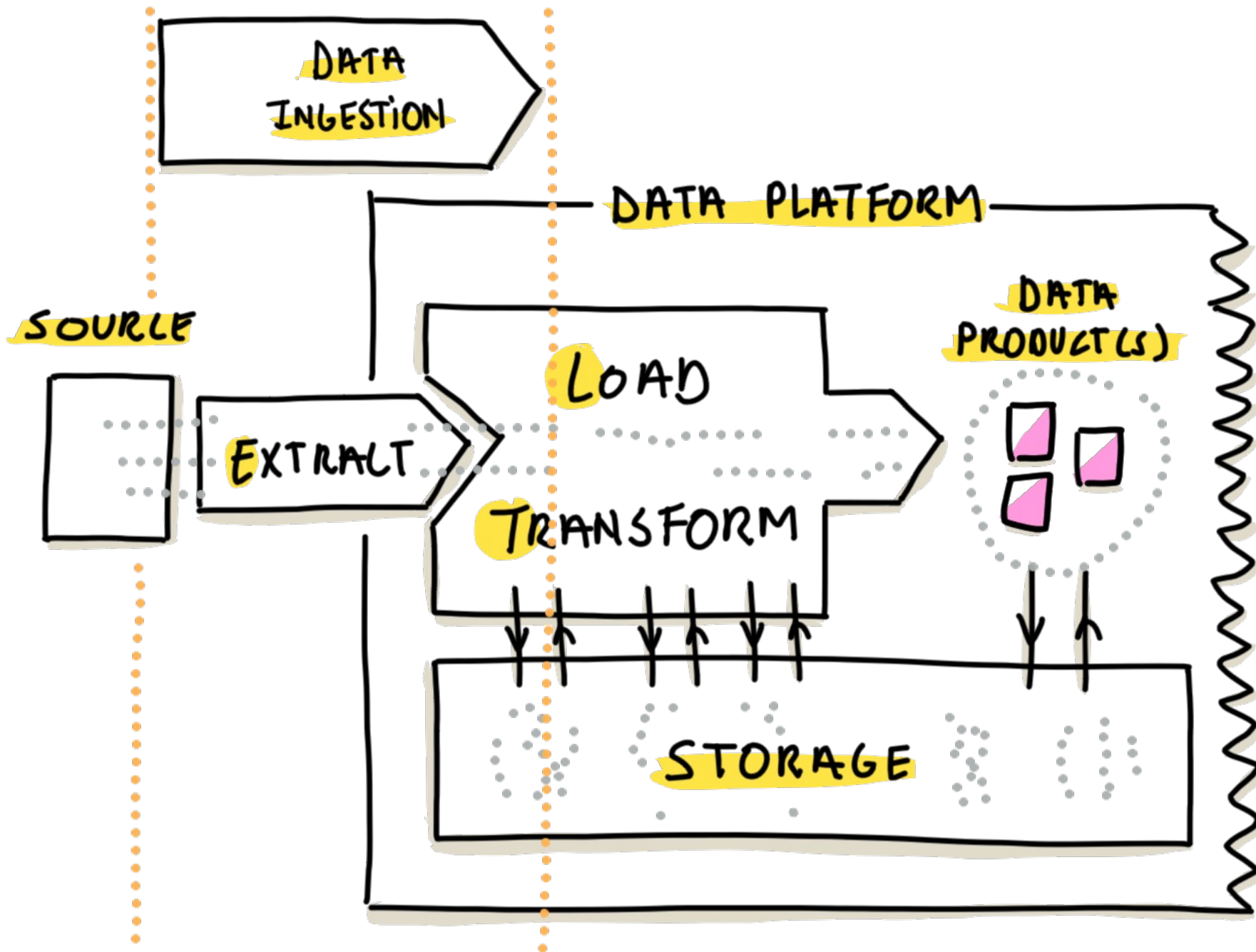


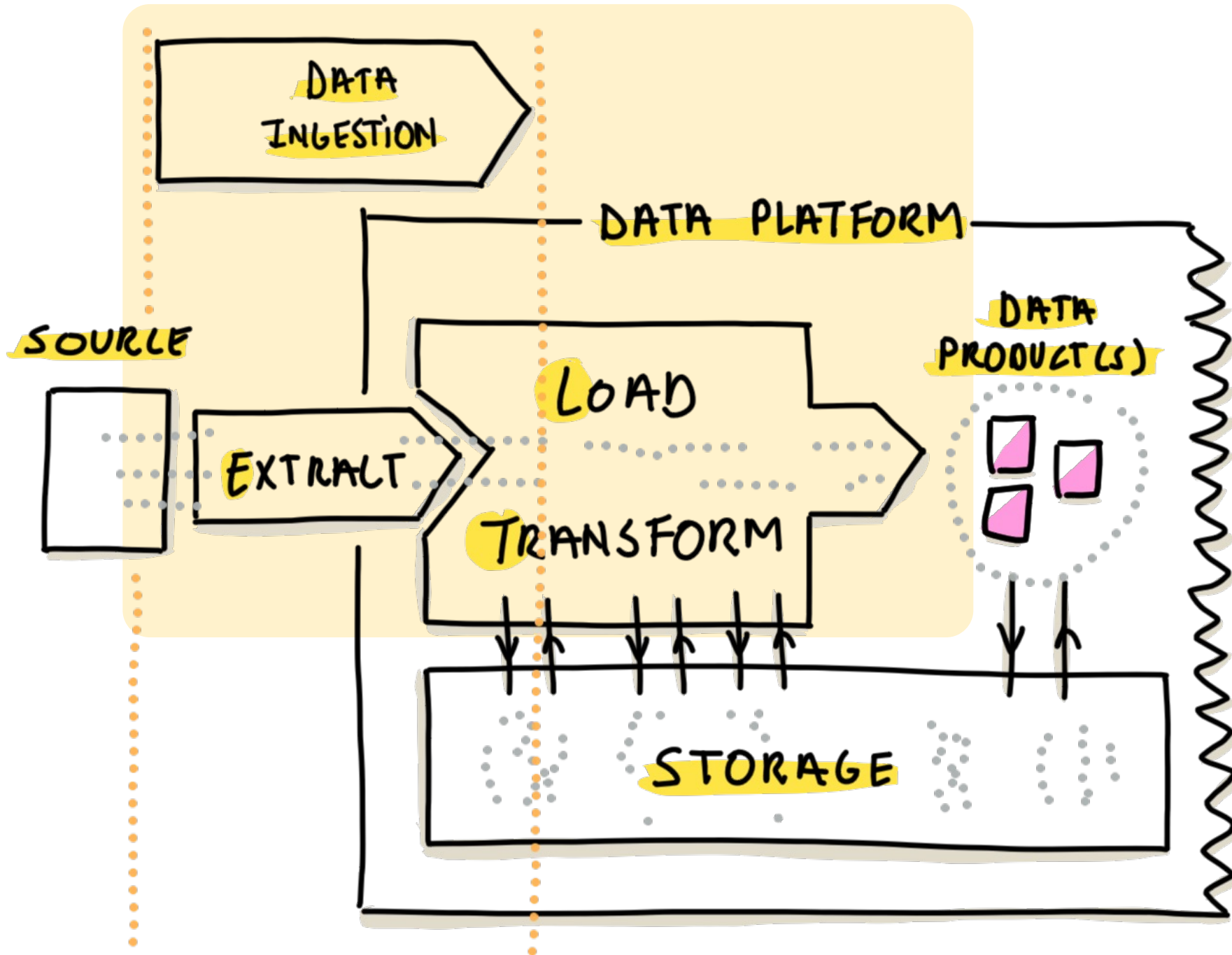
“Operational Plane”

“Analytical Plane”

“Operational / Analytical Plane”

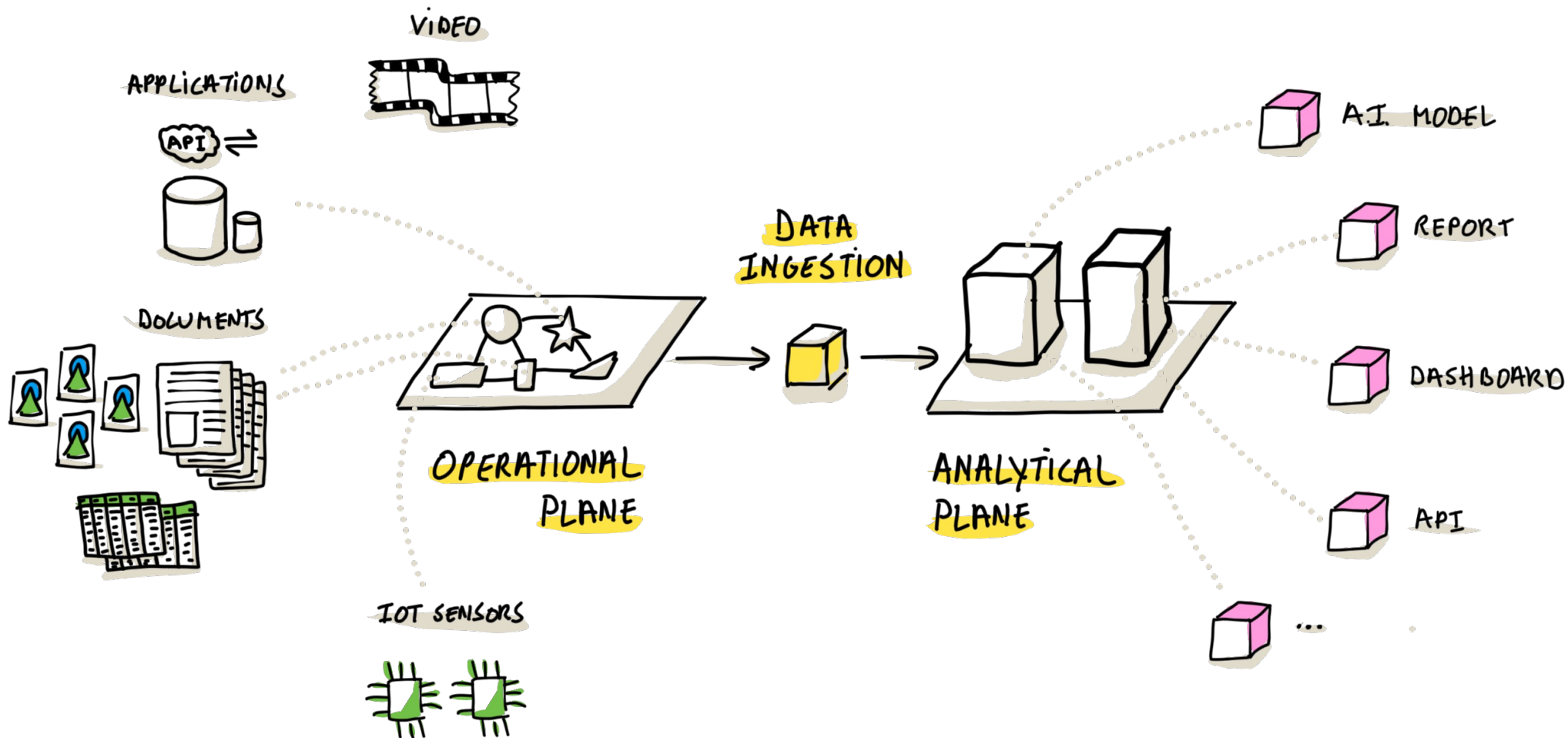






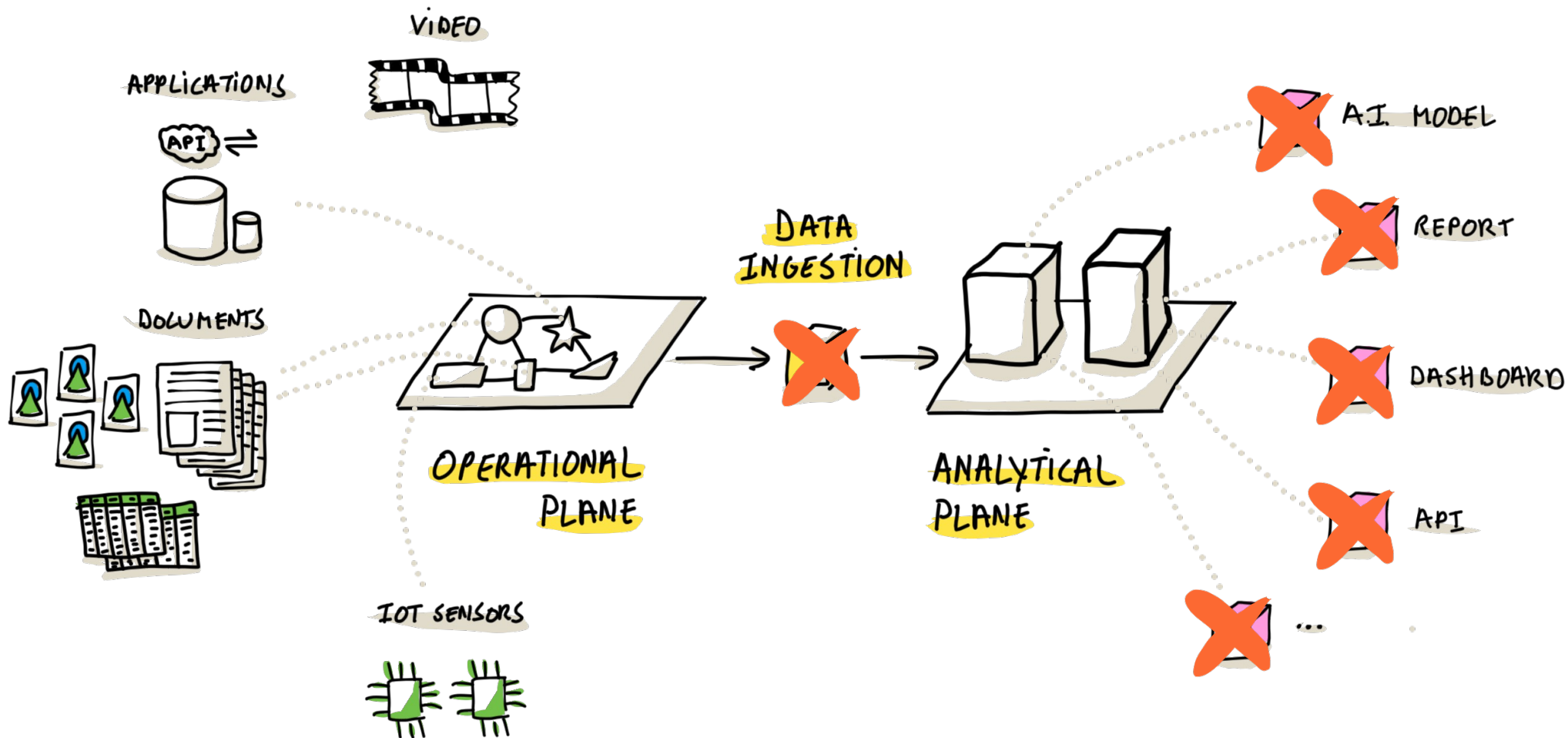
DATA SOURCES

DATA PRODUCTS

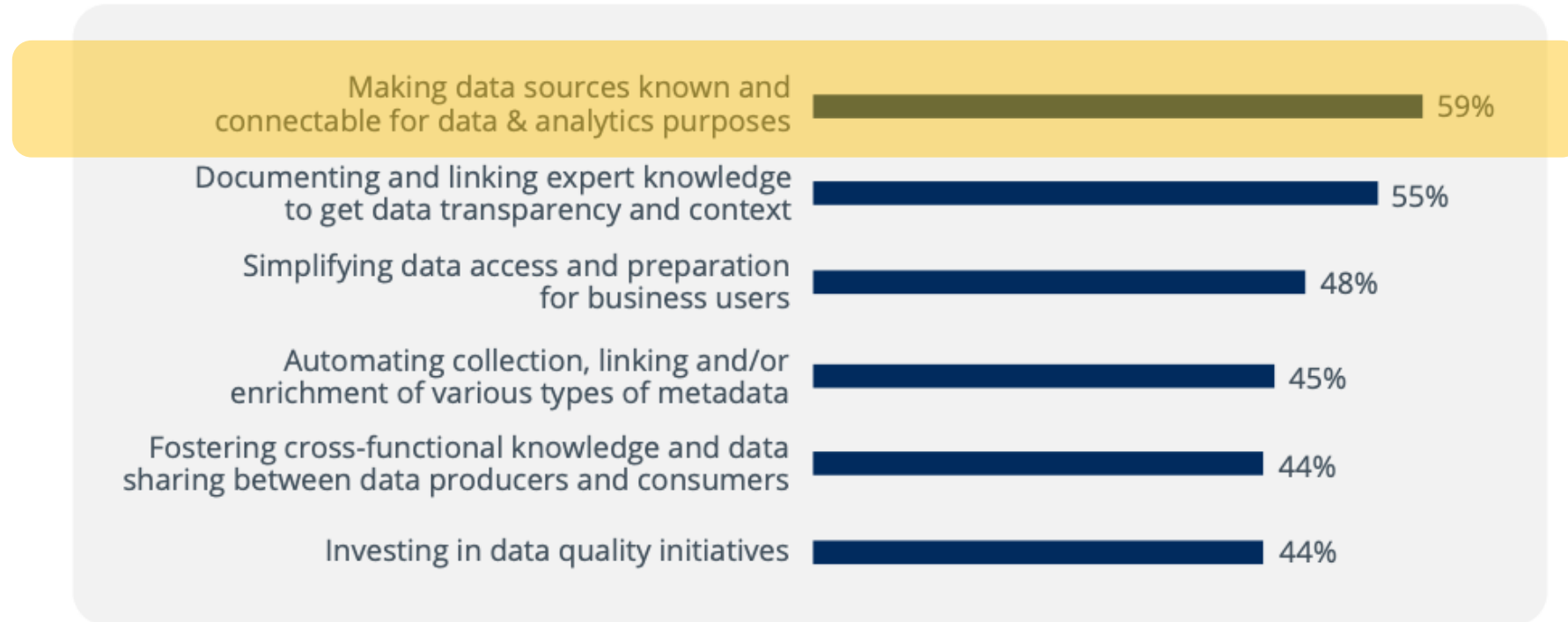


DATA SOURCES

DATA PRODUCTS



Data Ingestion = One of the Biggest Data Access Challenges



What do you see as the biggest challenges in data access? Top 6 (n=369)

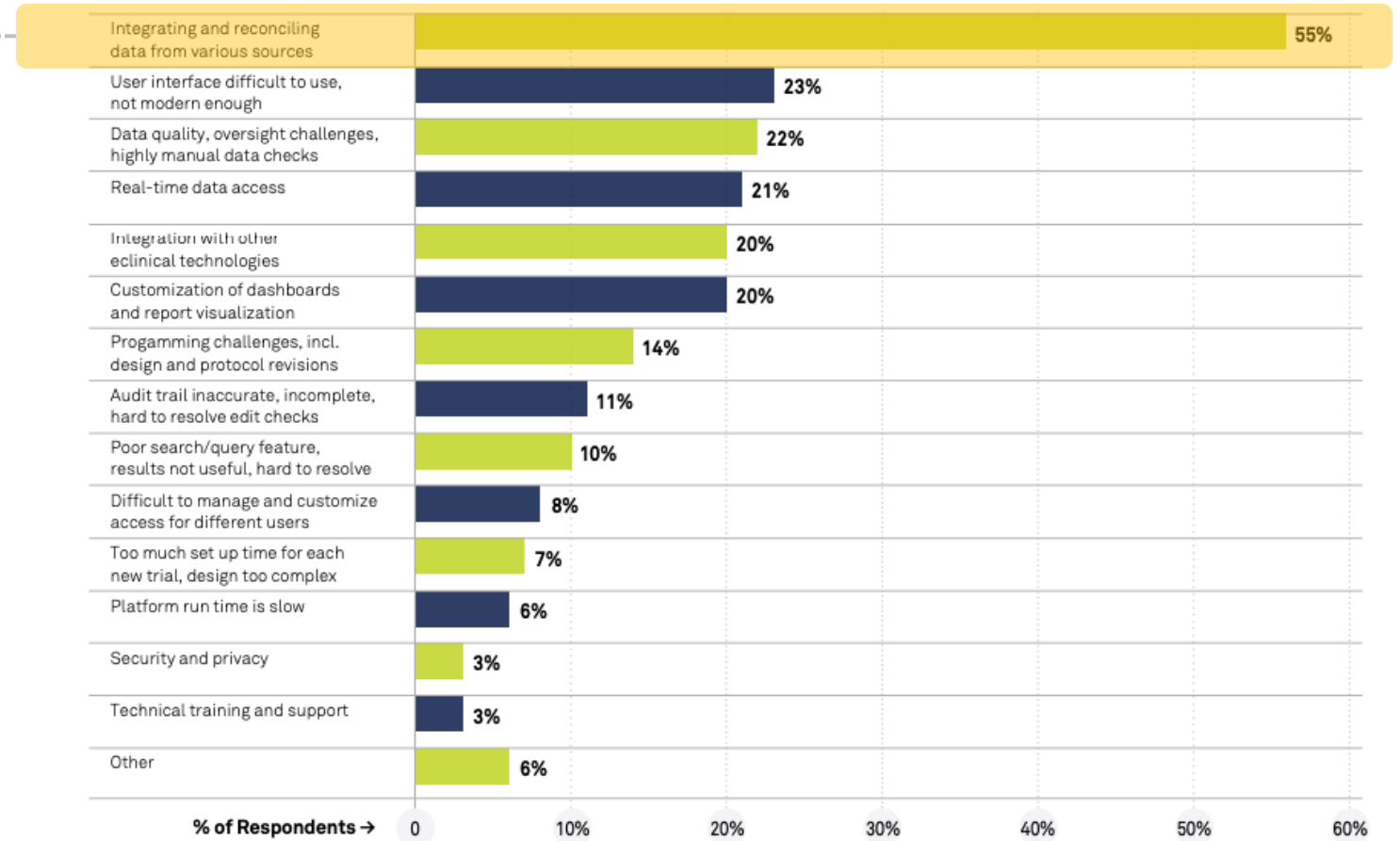


Data Ingestion = Pain Point #1 in CDMS

“Data Integration and Reconciliation Pain Points Persist Respondents described several pain points related to how they manage clinical data.

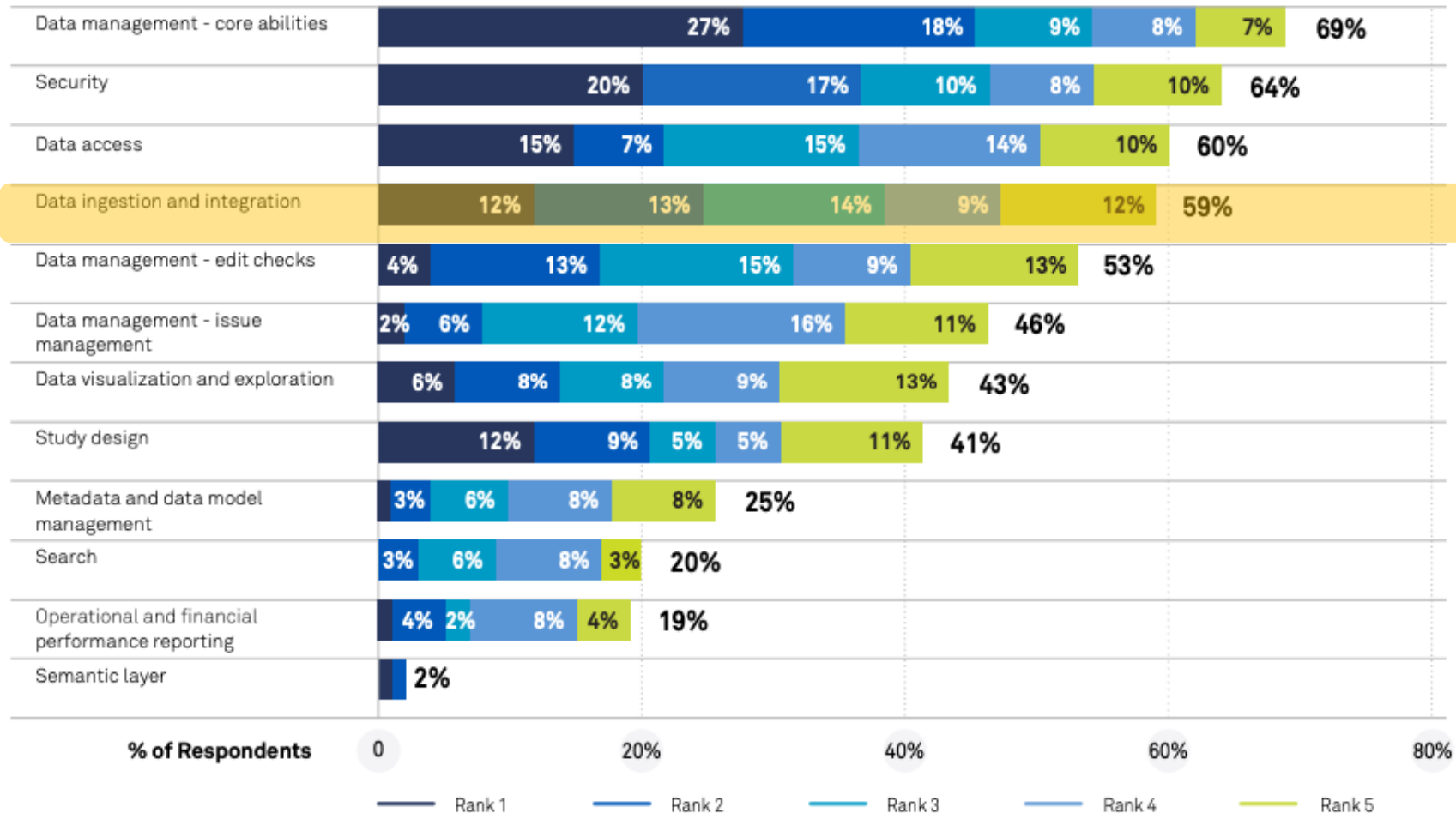
The most mentioned challenges included ingesting, integrating, and aggregating data from various sources.”

Figure 1: Pain Points in Clinical Data Management



Data Ingestion = Top Feature Request in CDMS

Figure 2: Ranking Importance of CDMS Features.

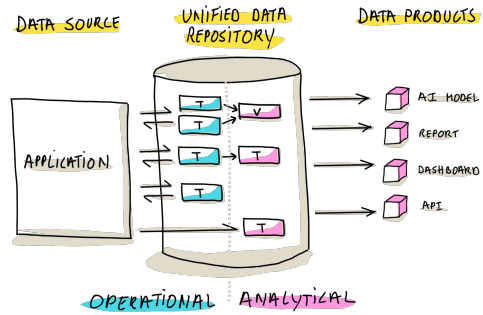


Definition

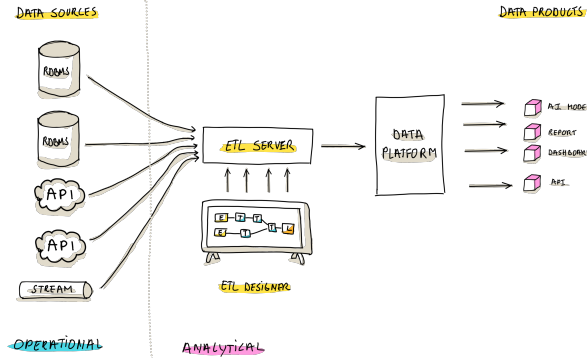
Data ingestion refers to the tools & processes used to **collect data from various sources and move it to a target site**, either in **batches or in real-time**. The data ingestion layer is **critical to your downstream** data science, BI, and analytics systems which depend on **timely, complete, and accurate data**.

Data Ingestion Patterns

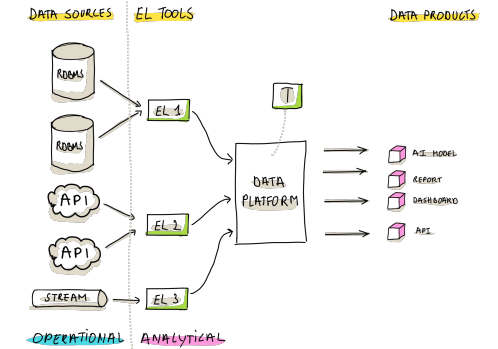
Pattern #1
UNIFIED DATA REPOSITORY



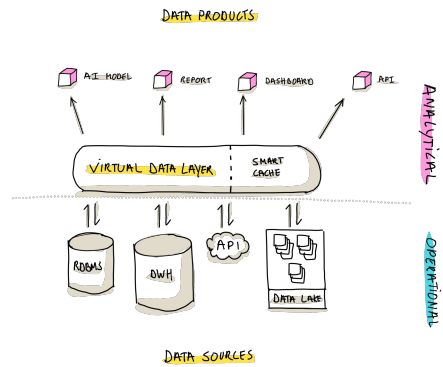
Pattern #2
ETL – Extract Transform Load



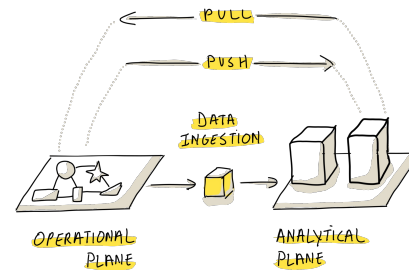
Pattern #3
ELT – Extract Transform Load



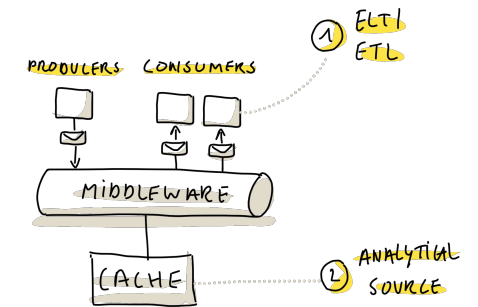
Pattern #4
Data Virtualization



Pattern #5
Push vs Pull

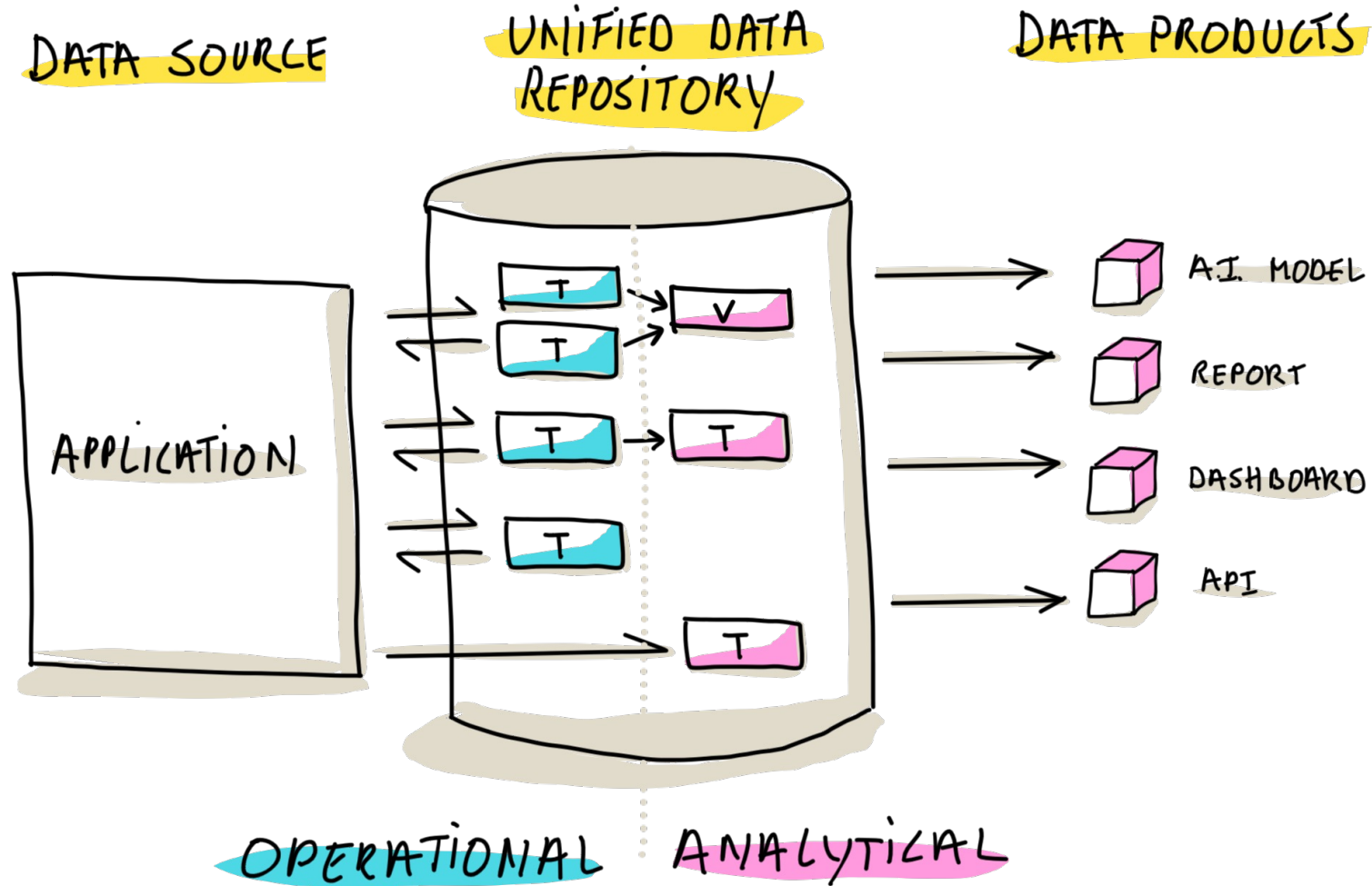


Pattern #6
Streaming Data



Pattern #1

UNIFIED DATA REPOSITORY



Unified Data Repository = A single storage system caters to both the operational application needs and analytical processing

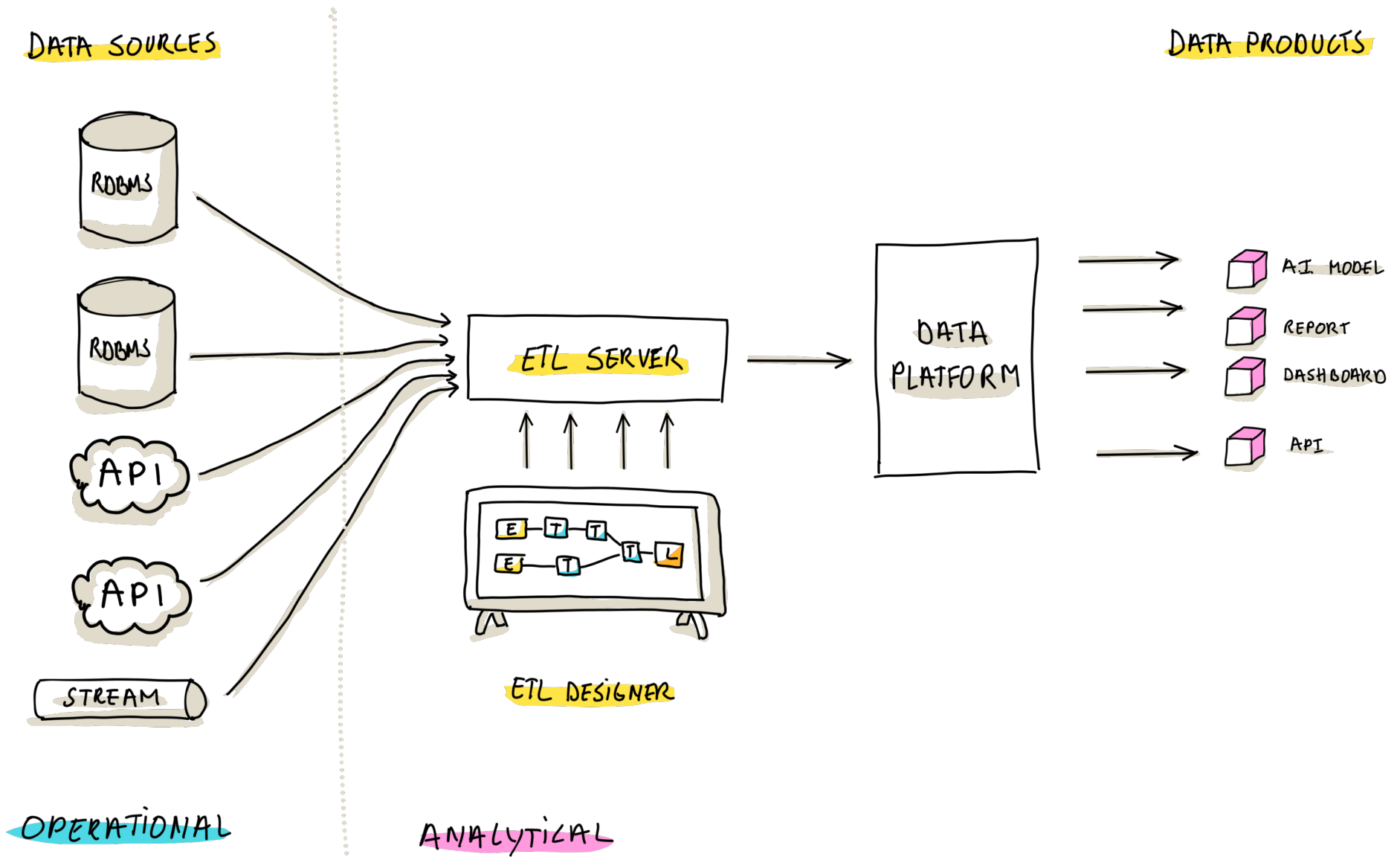
Specificities:

- Typically a Relational Database Management System (RDBMS).
- The same database is utilized for both everyday operations and data analysis
- Two prevalent sub-patterns:
 1. **Virtualization**
 2. **Duplication and Transformation**



Pattern #2

ETL – Extract Transform Load



ETL (Extract, Transform, Load) = *a well-established paradigm in data processing*

3 steps:

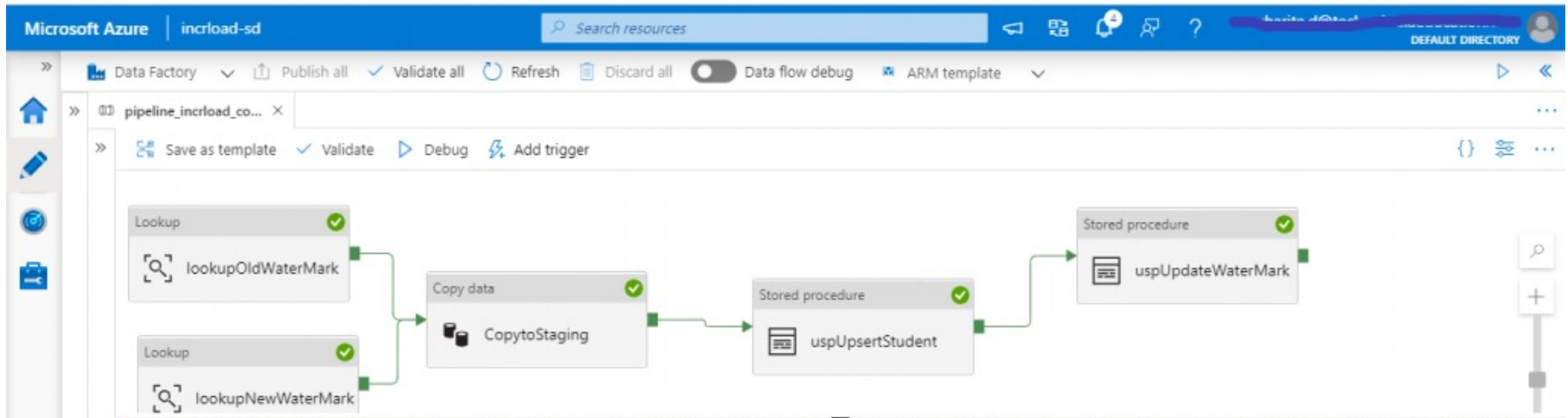
1. **Extract** : Data is harvested from its source
2. **Transform** : Refined on an ETL server
3. **Load** : The polished output is deposited into an analytics-focused database.

ETL tools have a graphical interface where users can interlink Extract, Transform, and Load operations within an intuitive visual workflow. These processes are often further customizable through scripting or direct SQL queries.



Pattern #2

ETL – Extract Transform Load



Extract Transform

Load

Pattern #2

ETL – Extract Transform Load

Microsoft Azure | incload-sd

Data Factory | Publish all | Validate

pipeline_incload_co...

Save as template | Validate

Lookup
lookupOldWaterMark

Lookup
lookupNewWaterMark

Copy data
CopytoStaging

Stored procedure
uspUpsertStudent

Stored procedure
uspUpdateWaterMark

Extract Transform

Load

General Settings User properties

Source dataset * SqlServerTable1 Open + New Preview data

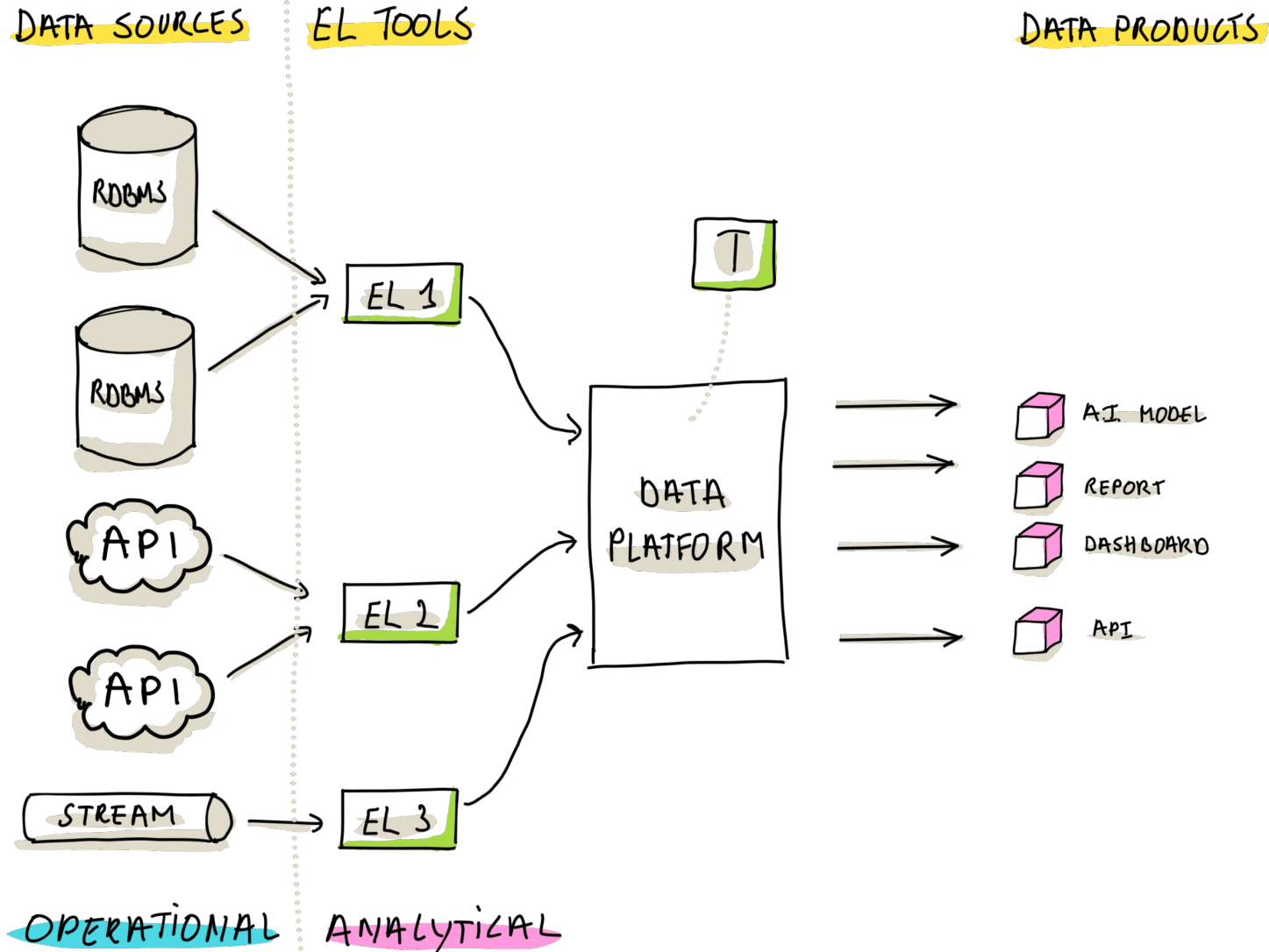
Use query Table Query Stored procedure

Query
SELECT
MAX(@{pipeline().parameters.waterMarkCol}) AS NewwaterMarkVal FROM
@{pipeline().parameters.srcTableName}

Query timeout (minutes) 120

Pattern #3

ELT – Extract Transform Load



ELT, sharing the basic steps of ETL, diverges by restructuring and redefining these processes.

In ELT:

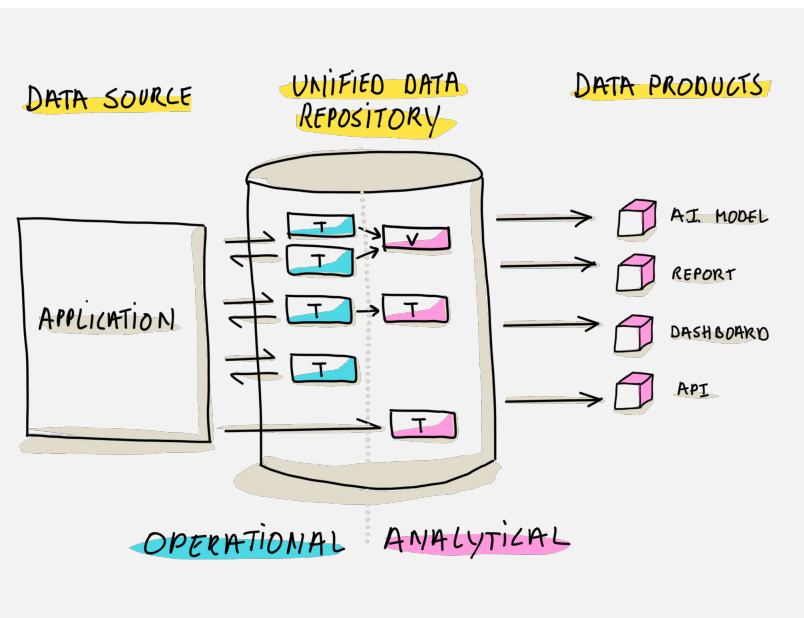
- 1. EL — Extract and Load operations** are carried out first, transferring raw data directly to the data platform without immediate transformation.
- 2. T — Transformation** occurs subsequently, converting raw data into actionable insights. Crucially, transformation tasks can operate independently and on different schedules from the extraction and loading.



EXERCISE 5: STRENGTHS & WEAKNESSES

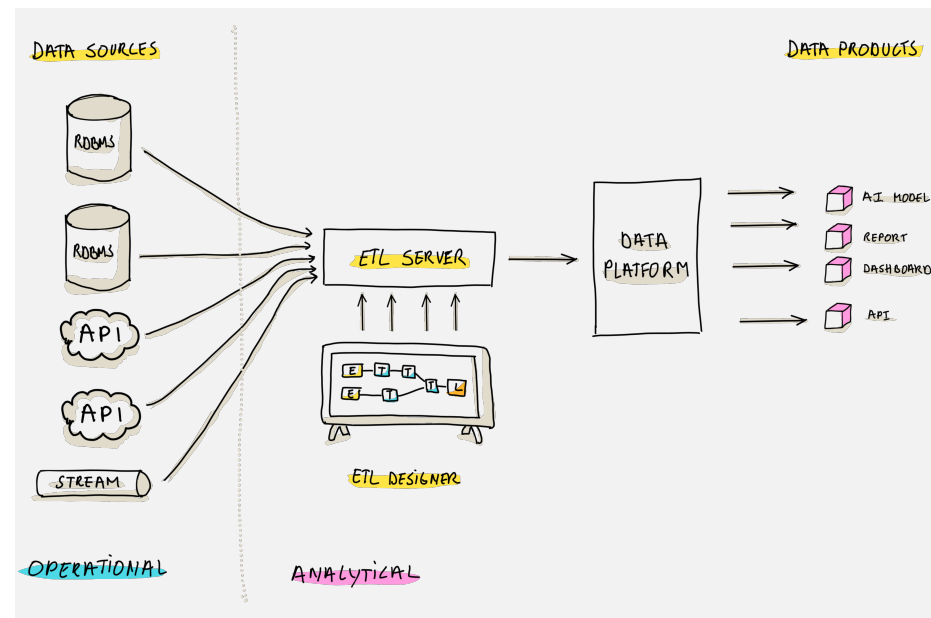
#1

Unified Data Repository



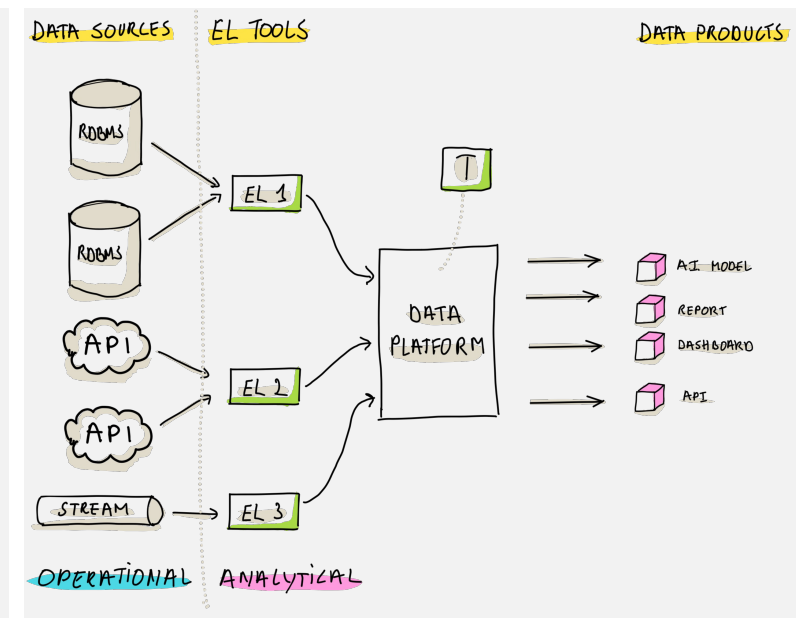
#2

ETL



#3

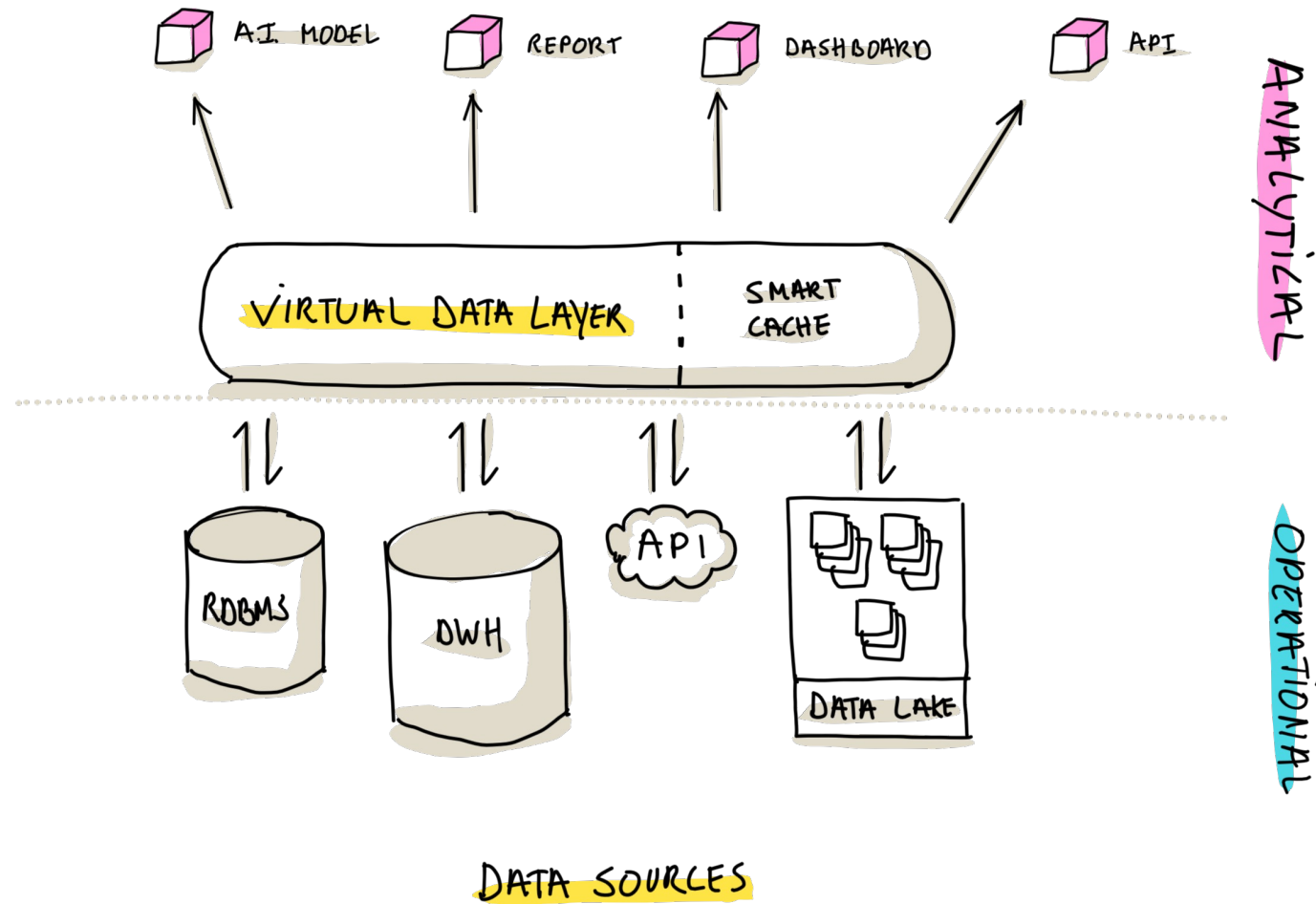
ELT



Pattern #4

Data Virtualization

DATA PRODUCTS



Data Virtualization = specialized software to establish a virtualized data layer over multiple underlying data sources. This intermediary layer allows for the execution of queries that are partially processed by the original data sources, integrating the results into a cohesive dataset for analysis.

- Inspired by the Unified Data Repository (Pattern #1)
- **Pro / Cons?**



BI integration

Looker Power BI Tableau ThoughtSpot Metabase

Client support

dbt Airflow python jupyter R Apache Spark JDBC ODBC

Starburst Enterprise



Analytics engine

MPP query engine Data products Fault-tolerant execution
Query optimizer Elastic auto scaling Smart indexing & caching Metrics & logging



Global security

Fine-grained access control End-to-end encryption Data masking Event logging Query auditing



Data lakes



Relational DBs



NoSQL stores



Real-time analytics



Applications

Any data source anywhere

Cross-cloud and region analytics

On Premise

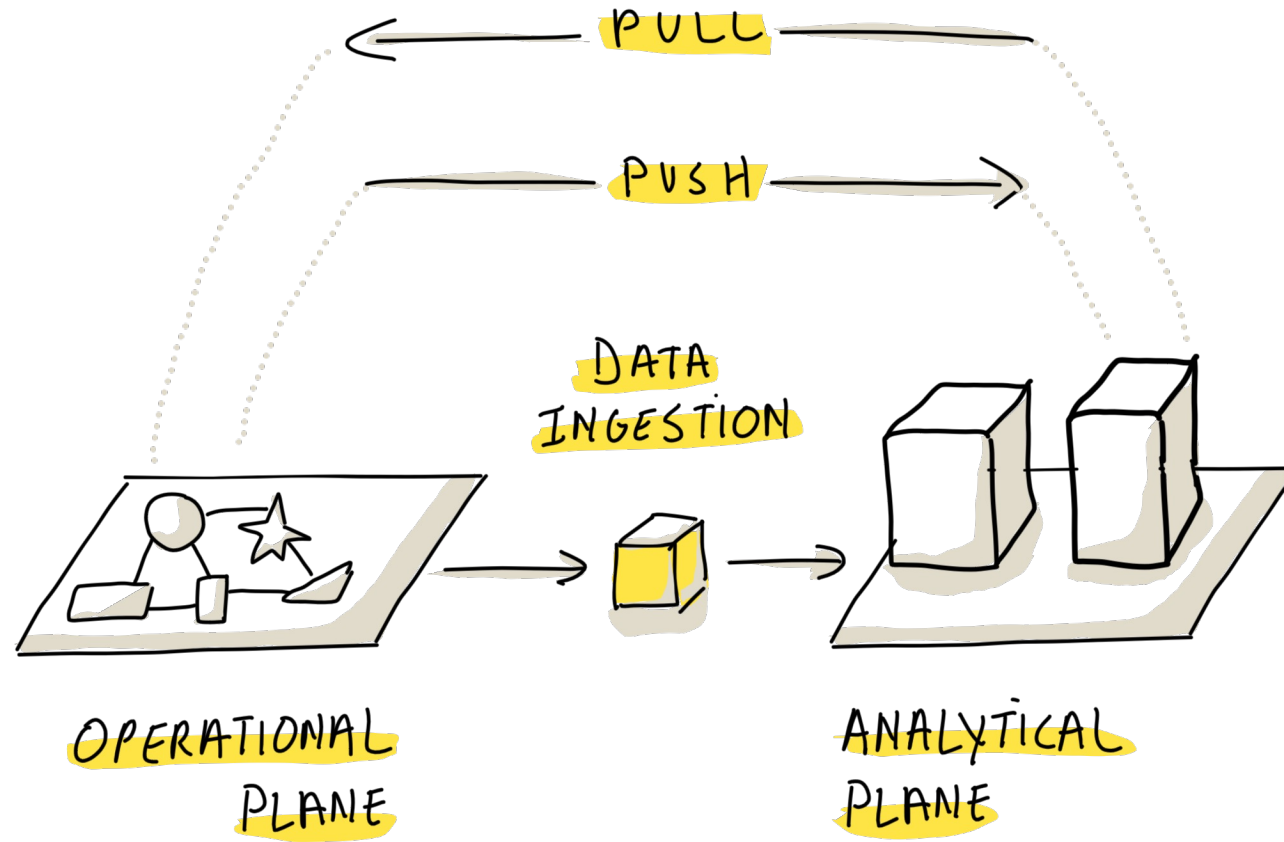
Hybrid

Cross-cloud



Pattern #5

Push vs Pull



Push = The operational plane initiates data transfer to an endpoint designated by the analytical plane.

- Often found within streaming architectures (discussed next) but is not confined to them.
- Software development teams are mostly responsible to implement the push mechanism
- **Pro/Cons ?**



Pattern #6

Streaming Data

**Streams
record history**



“The sequence of moves”

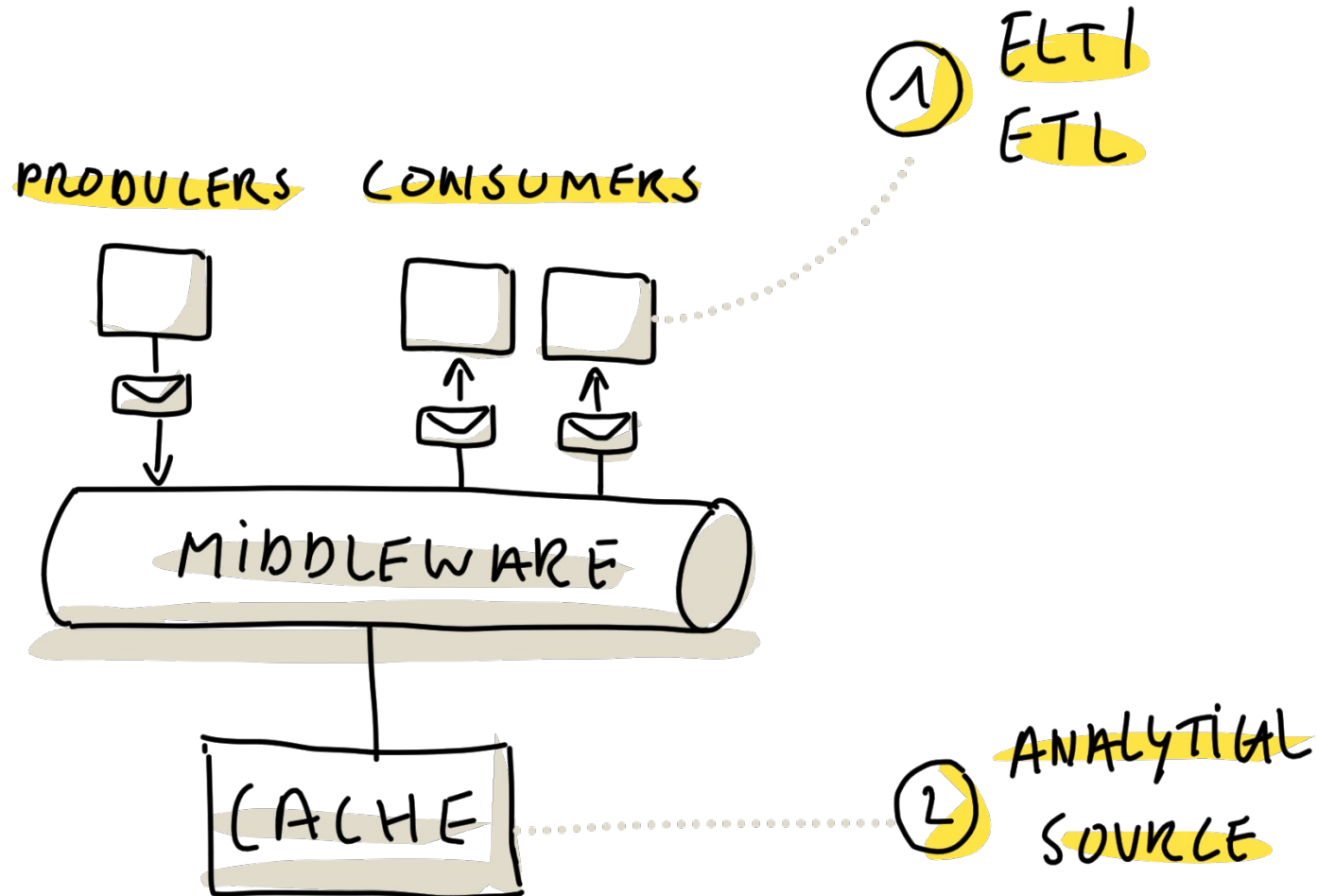
**Tables
represent state**



“The state of the board”



Pattern #6 Streaming Data



Stream processing = the continuous flow of data as it's generated, enabling real-time processing and analysis for immediate insights.

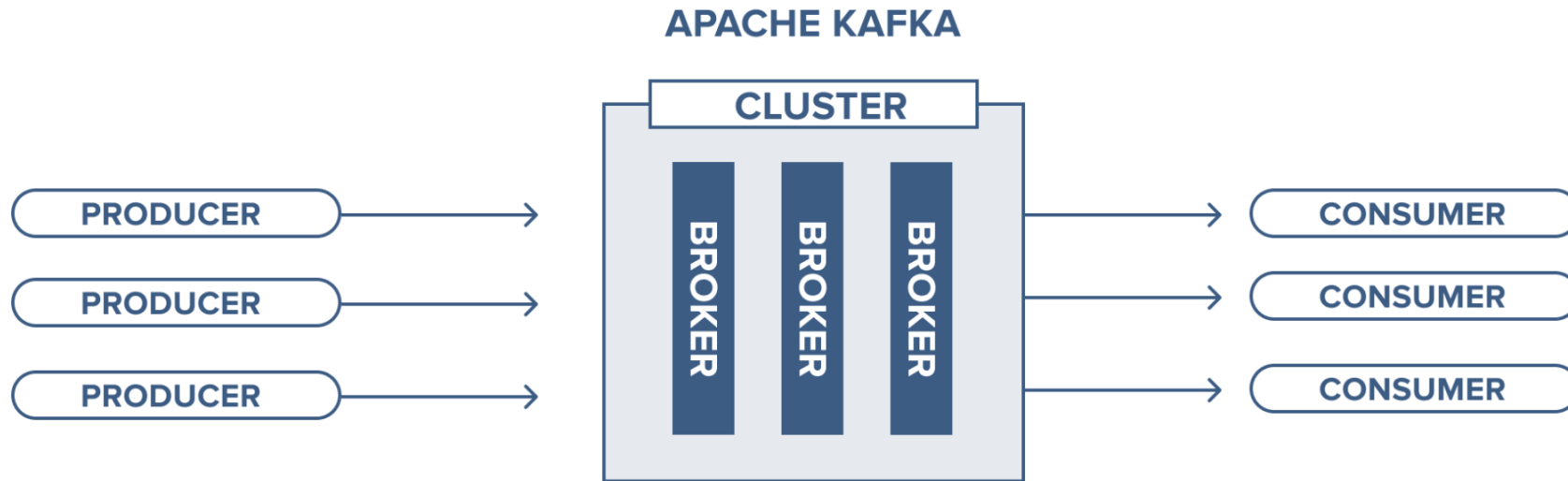
These systems are crucial for instant decision-making tasks and support high-volume, low-latency processing for activities like financial trades, real-time analytics, and IoT monitoring.

Two common approaches:

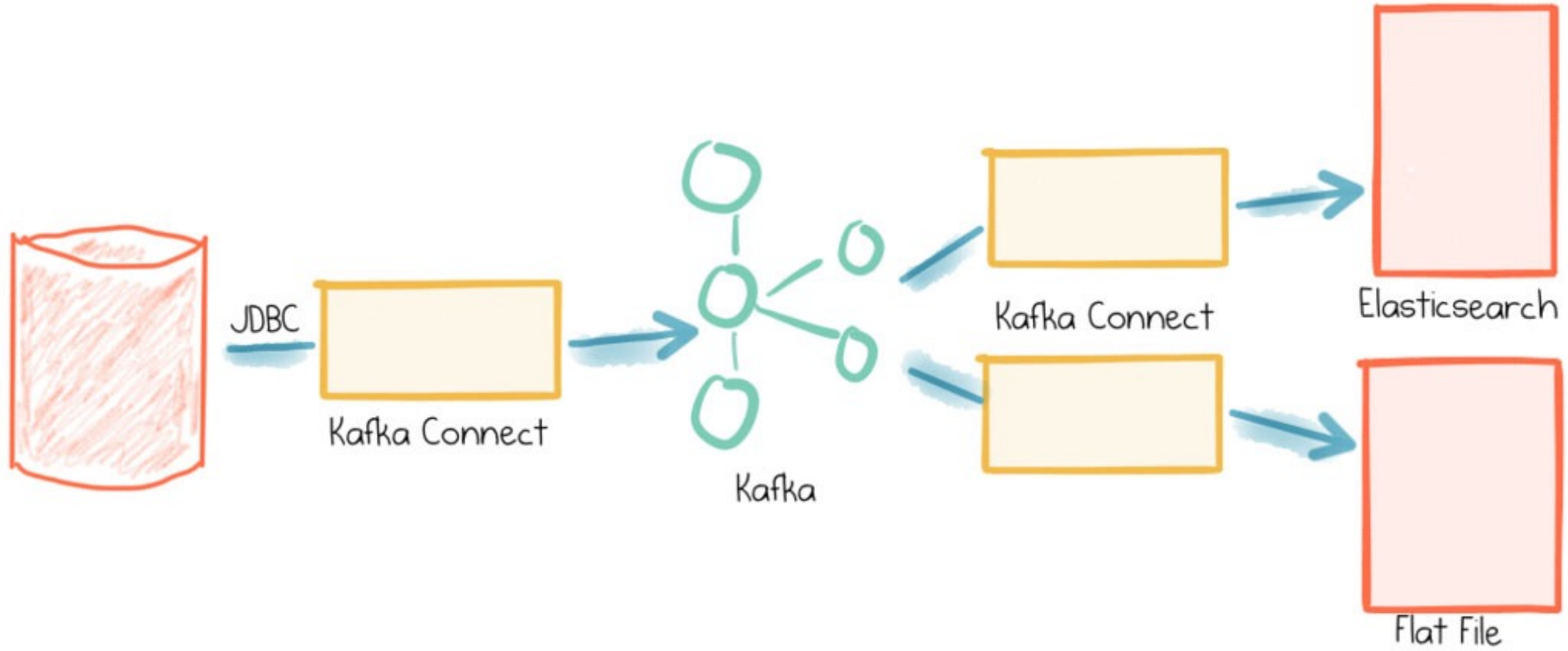
- ELT (or ETL) for streaming
- Leveraging streaming caches



Example: Kafka

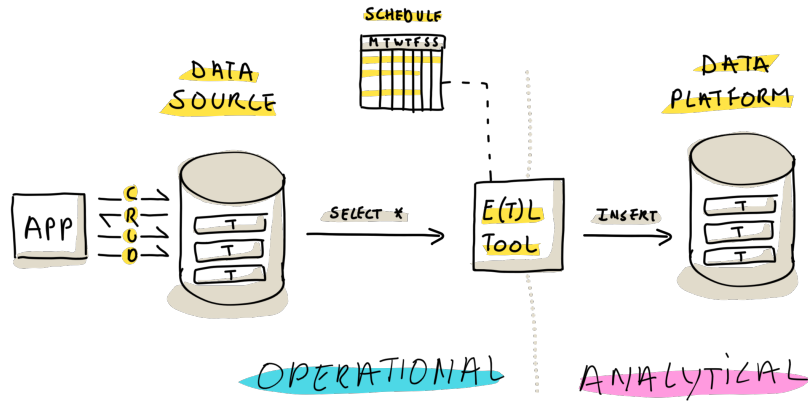


Example: Kafka Connect

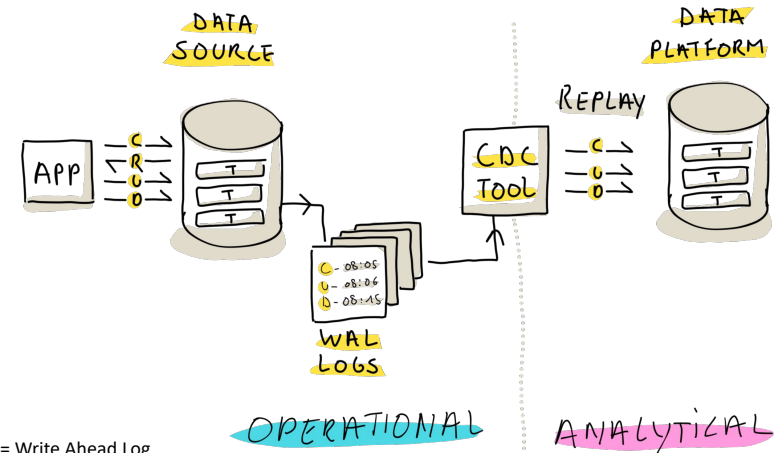


Data Ingestion Tool Flavors

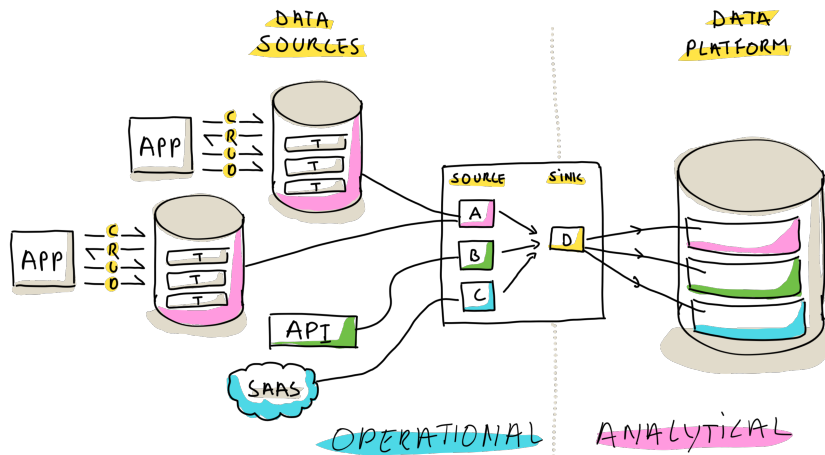
Flavor #1
Batch Loading



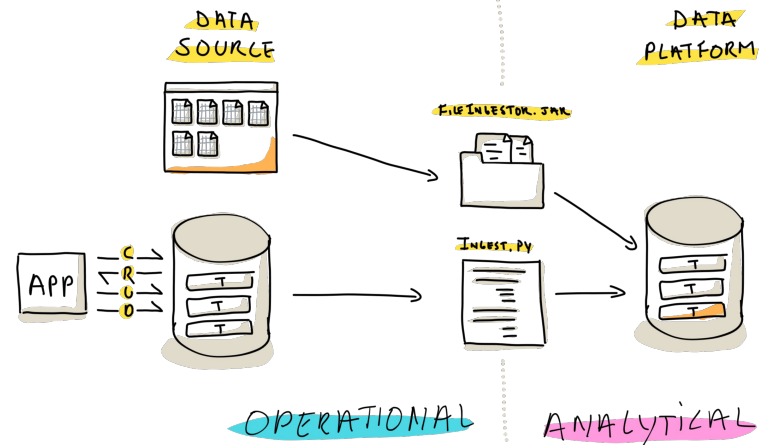
Flavor #2
CDC – Change Data Capture



Flavor #3
Connector Based

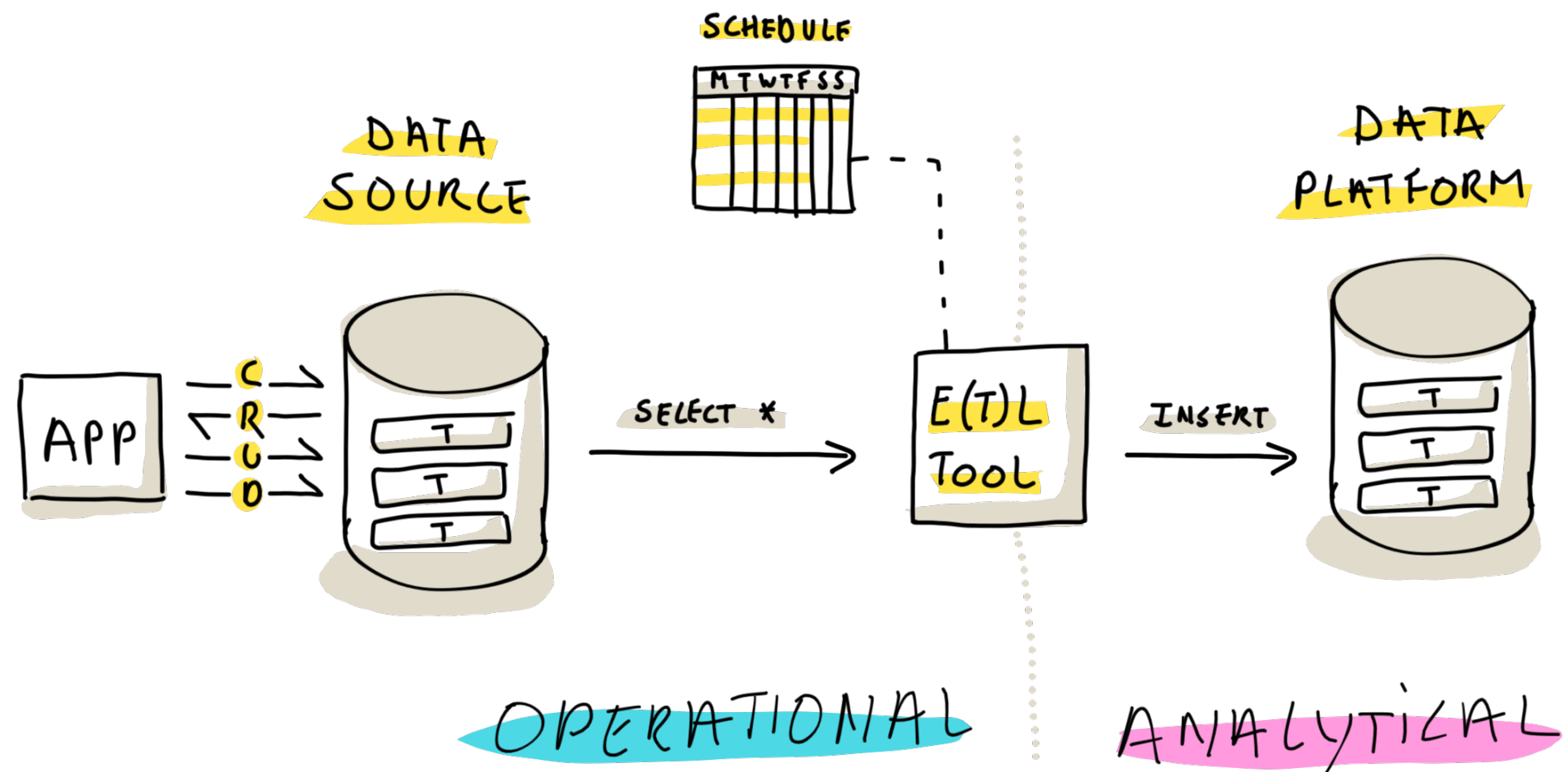


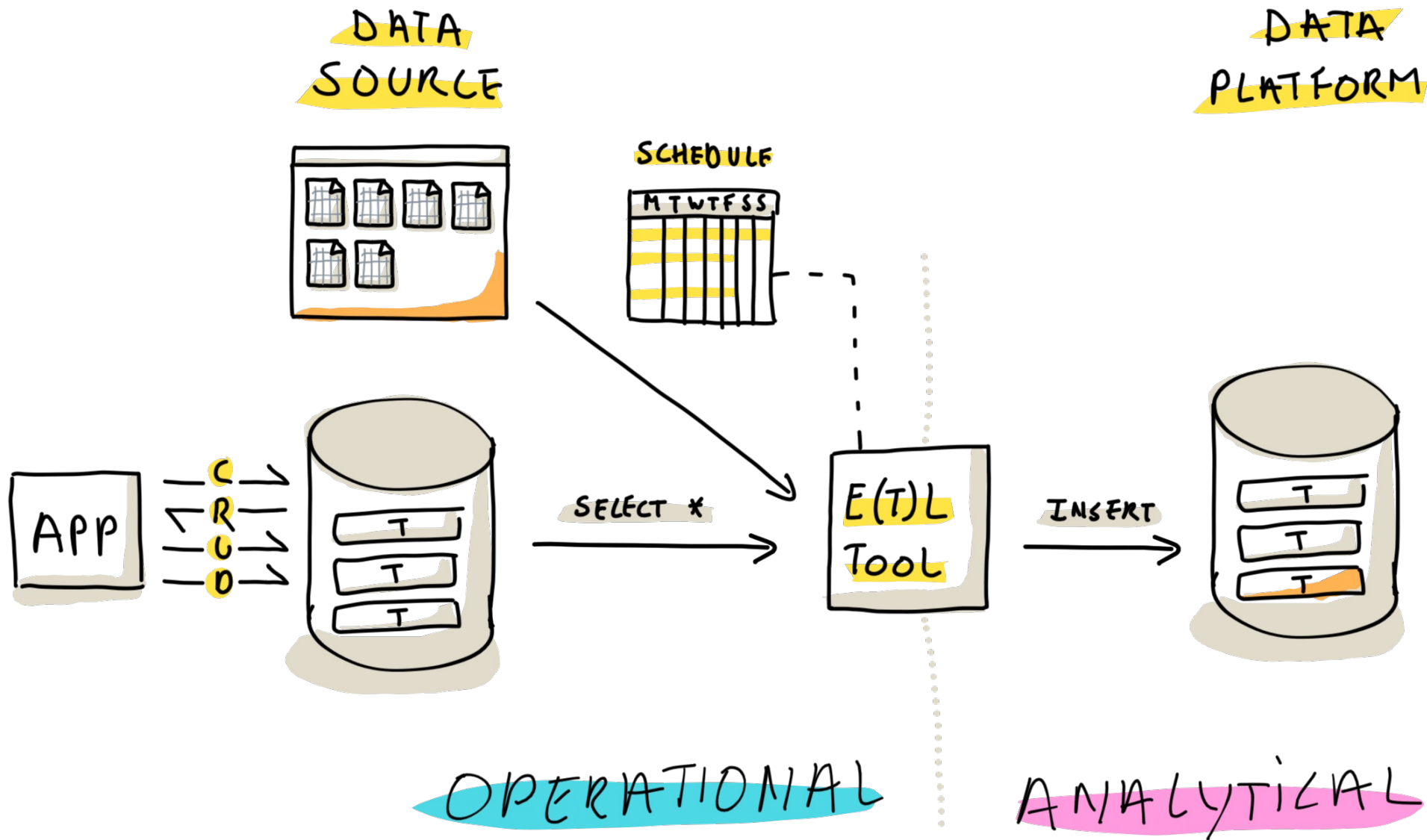
Flavor #4
Custom Builds



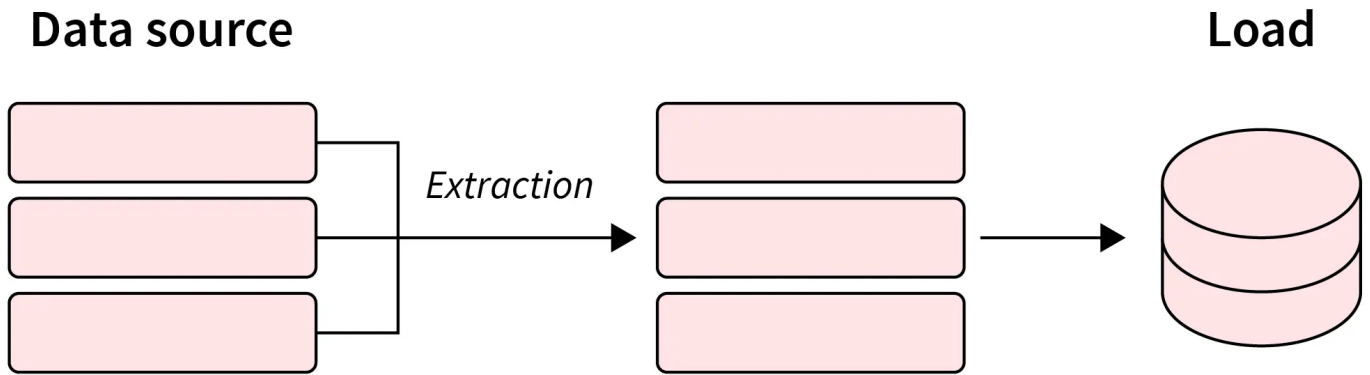
Flavor #1

Batch Loading

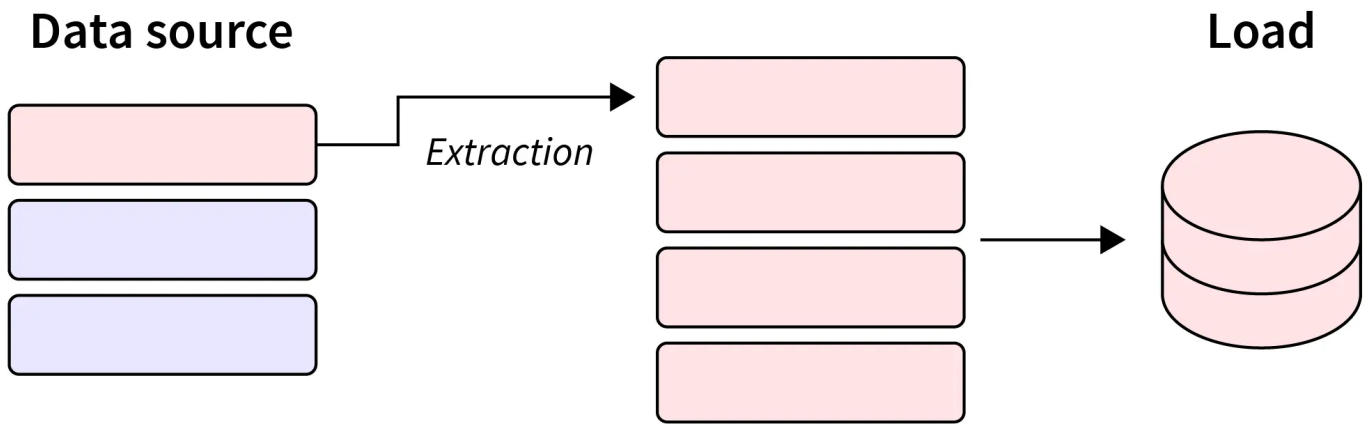




Full load
All available data is extracted at the same time























Incremental load
The occurring changes in the source data is incrementally extracted and loaded



EXERCISE 6: DESCRIBE INCREMENTAL LOADING

☒ m_transaction

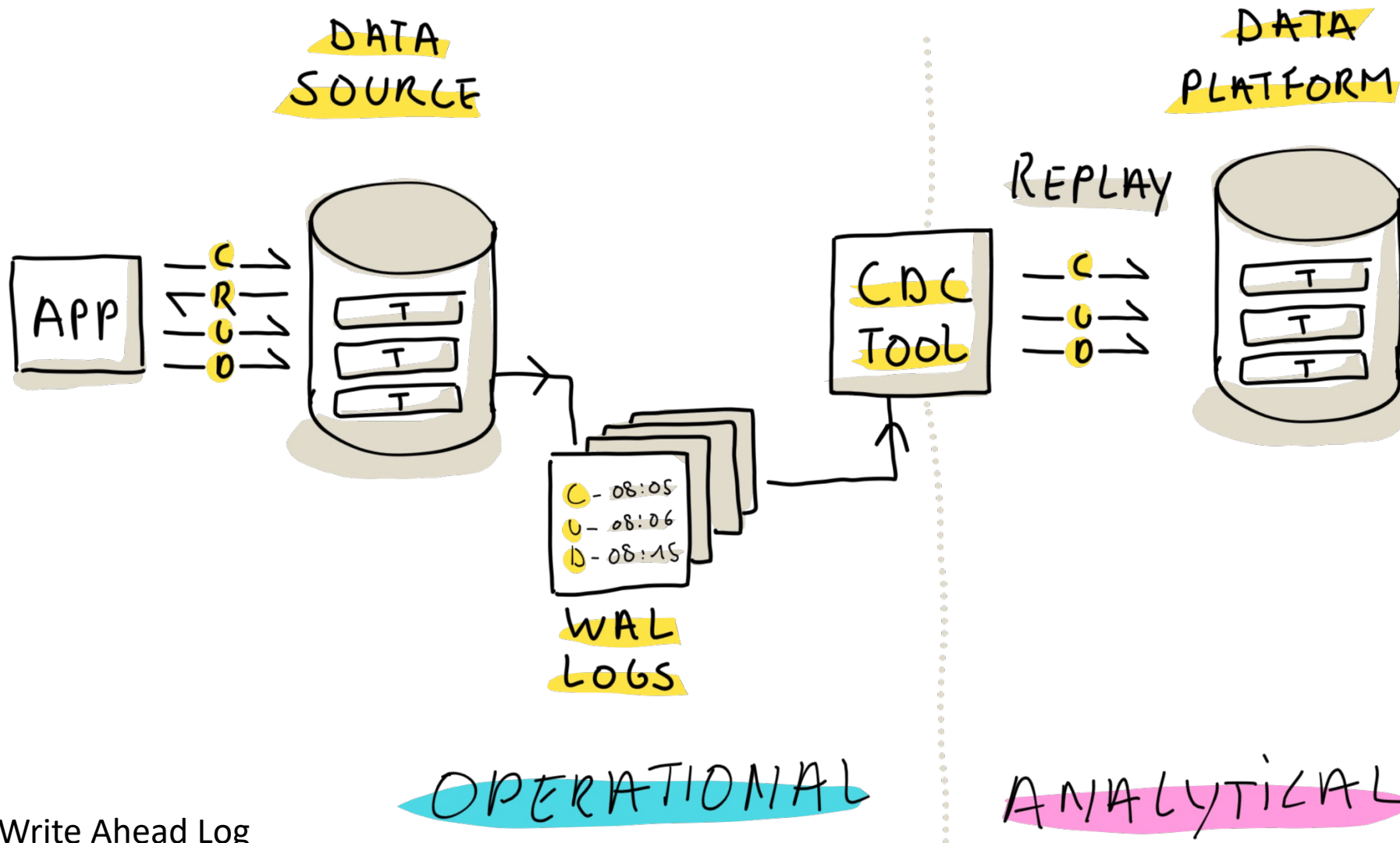
OFF ▾                    

	m_transaction_id	ad_client_id	ad_org_id	isactive	created	createdby	updated
1	1002856	1000010	1000061	Y	2020-03-18 14:40:01.253	100	2020-03-18 14:40:01.253
2	1002857	1000010	1000061	Y	2020-03-18 14:53:40.809	100	2020-03-18 14:53:40.809
3	1002858	1000010	1000061	Y	2020-03-18 15:20:00.275	100	2020-03-18 15:20:00.275
4	1002859	1000010	1000061	Y	2020-03-18 17:42:27.395	1000405	2020-03-18 17:42:27.395
5	1002860	1000010	1000129	Y	2020-03-18 19:50:49.07	100	2020-03-18 19:50:49.07
6	1002861	1000010	1000129	Y	2020-03-18 19:59:42.211	100	2020-03-18 19:59:42.211
7	1002862	1000010	1000129	Y	2020-03-18 20:00:39.243	100	2020-03-18 20:00:39.243
8	1002863	1000010	1000129	Y	2020-03-18 20:02:20.357	100	2020-03-18 20:02:20.357
9	1002864	1000010	1000129	Y	2020-03-18 20:10:12.598	100	2020-03-18 20:10:12.598
10	1002865	1000010	1000061	Y	2020-03-18 20:25:26.384	100	2020-03-18 20:25:26.384



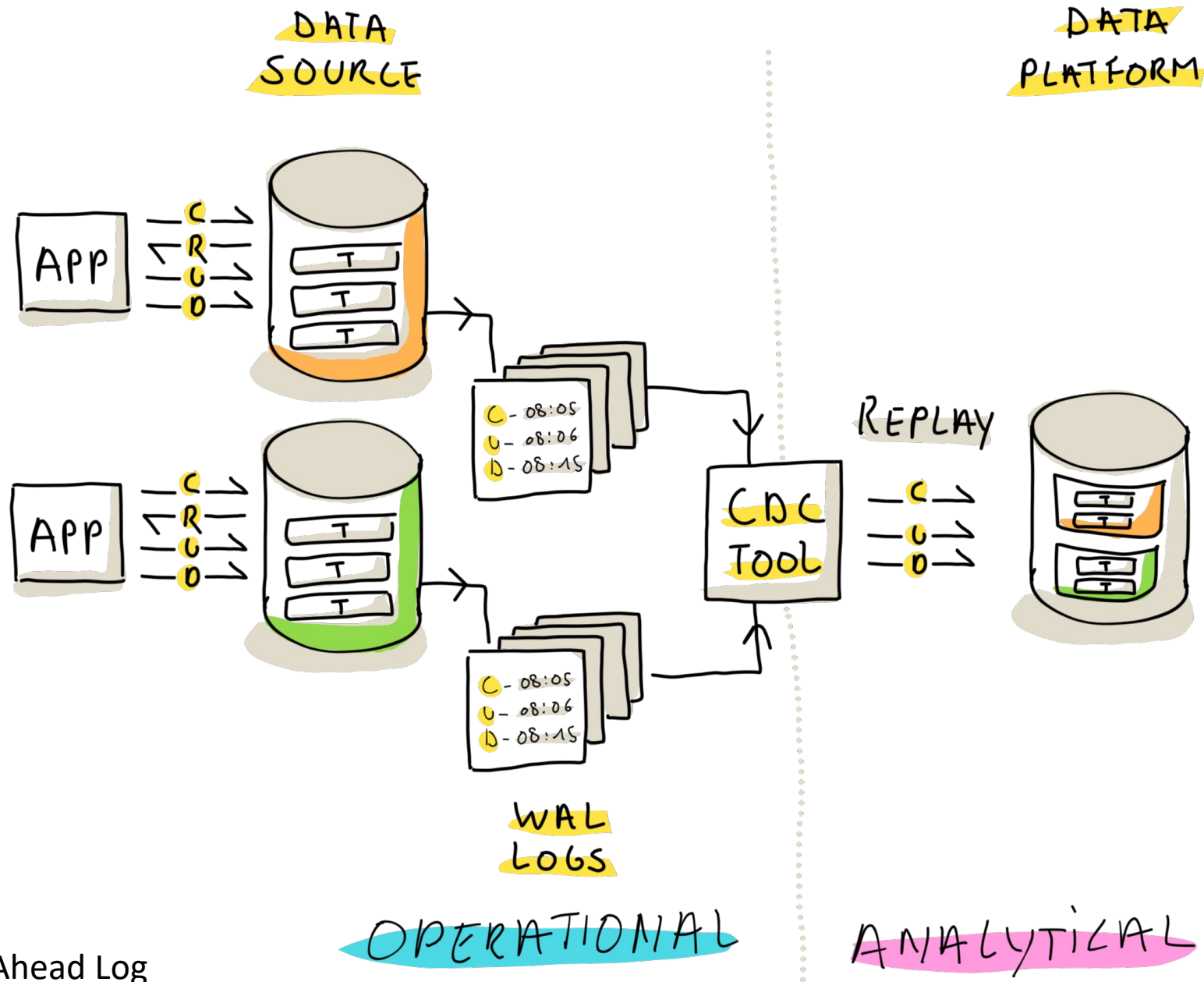
Flavor #2

CDC – Change Data Capture



WAL = Write Ahead Log

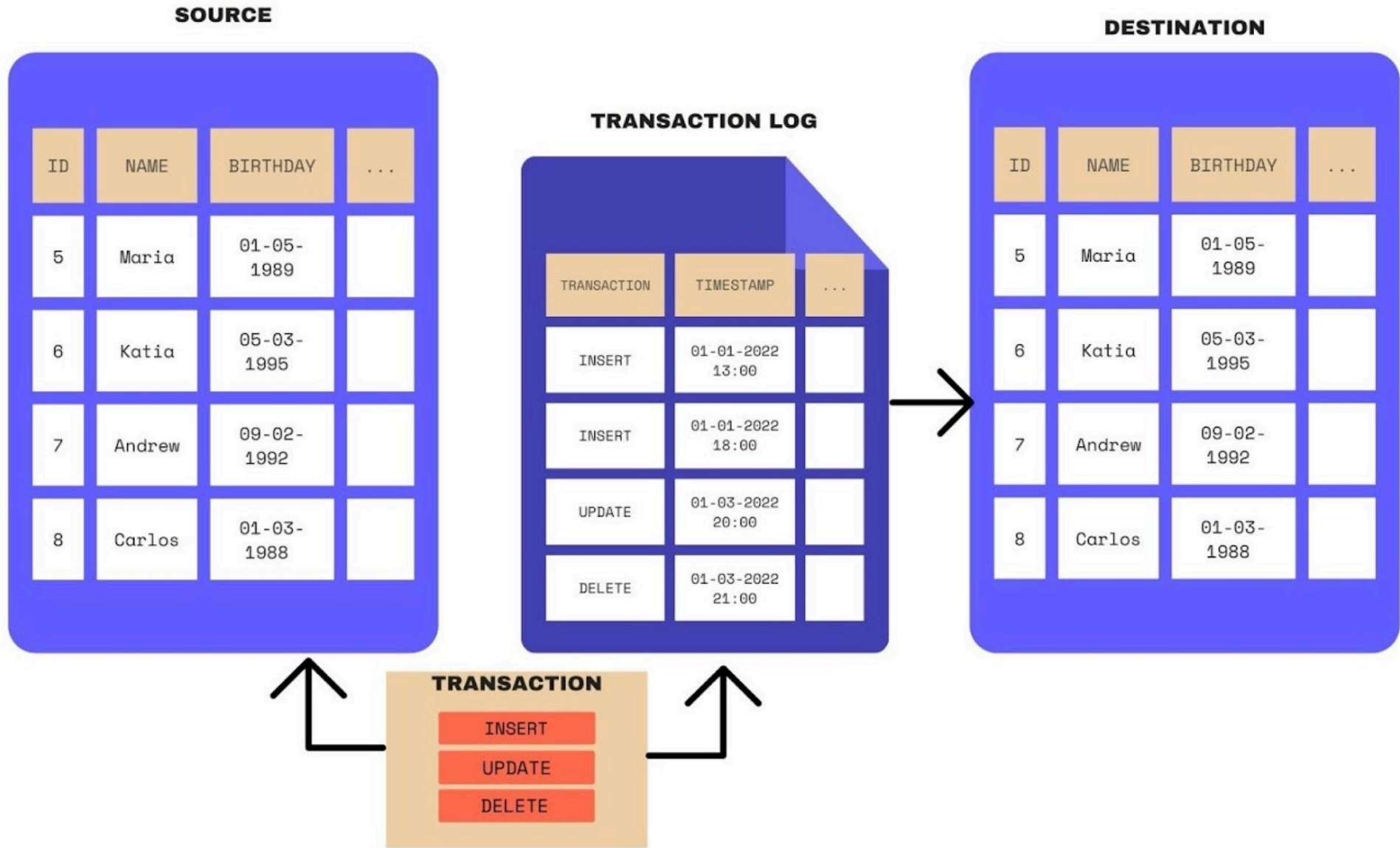


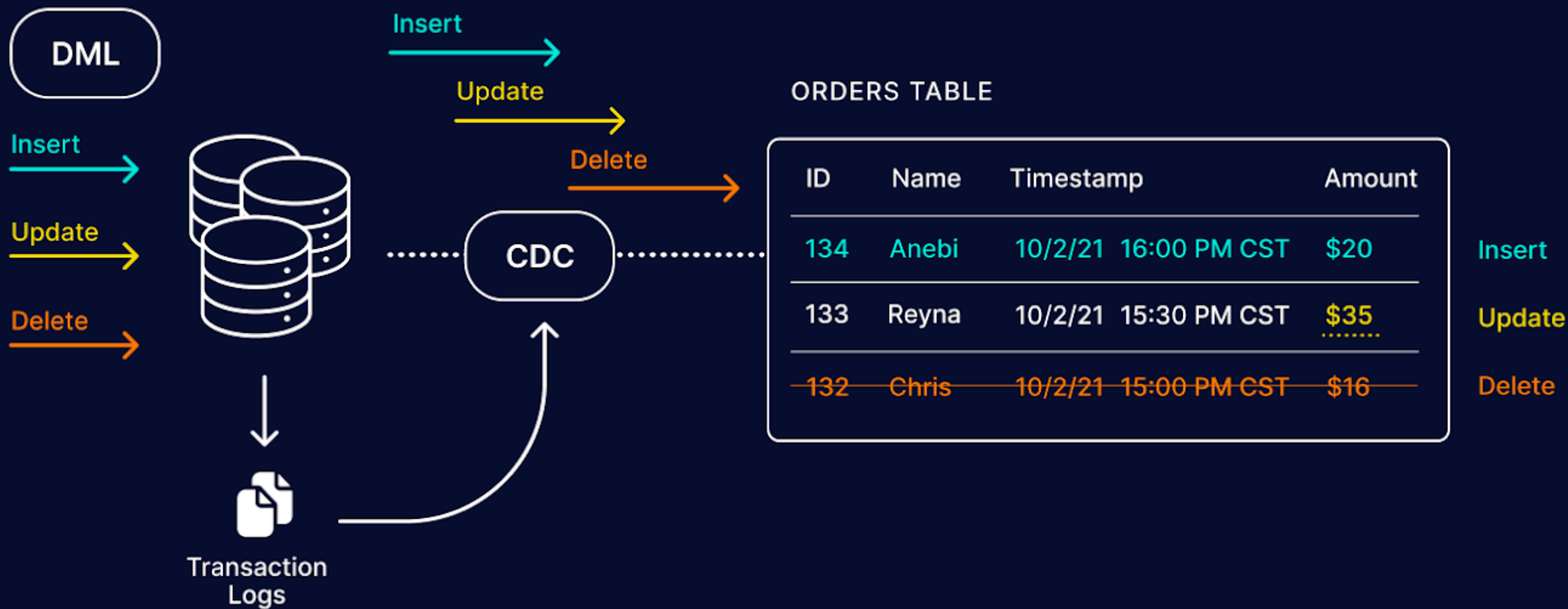


WAL = Write Ahead Log



Log-based CDC technique





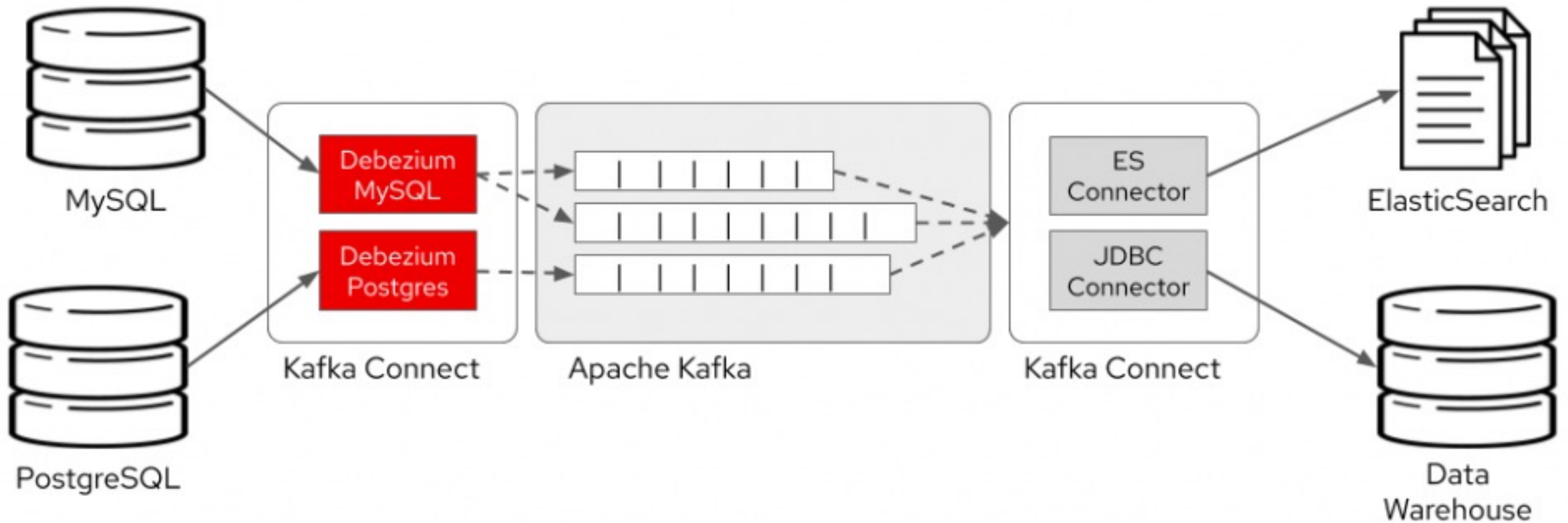
Change Data Capture (CDC) = a software that allows detecting and capturing changes made to data in a database and sending these changes, sometimes in real-time, to a downstream process or system. More specifically, CDC entails recording INSERT, UPDATE, and DELETE transactions applied to a table.

Various techniques exist: meta-data based, trigger based, **log based**

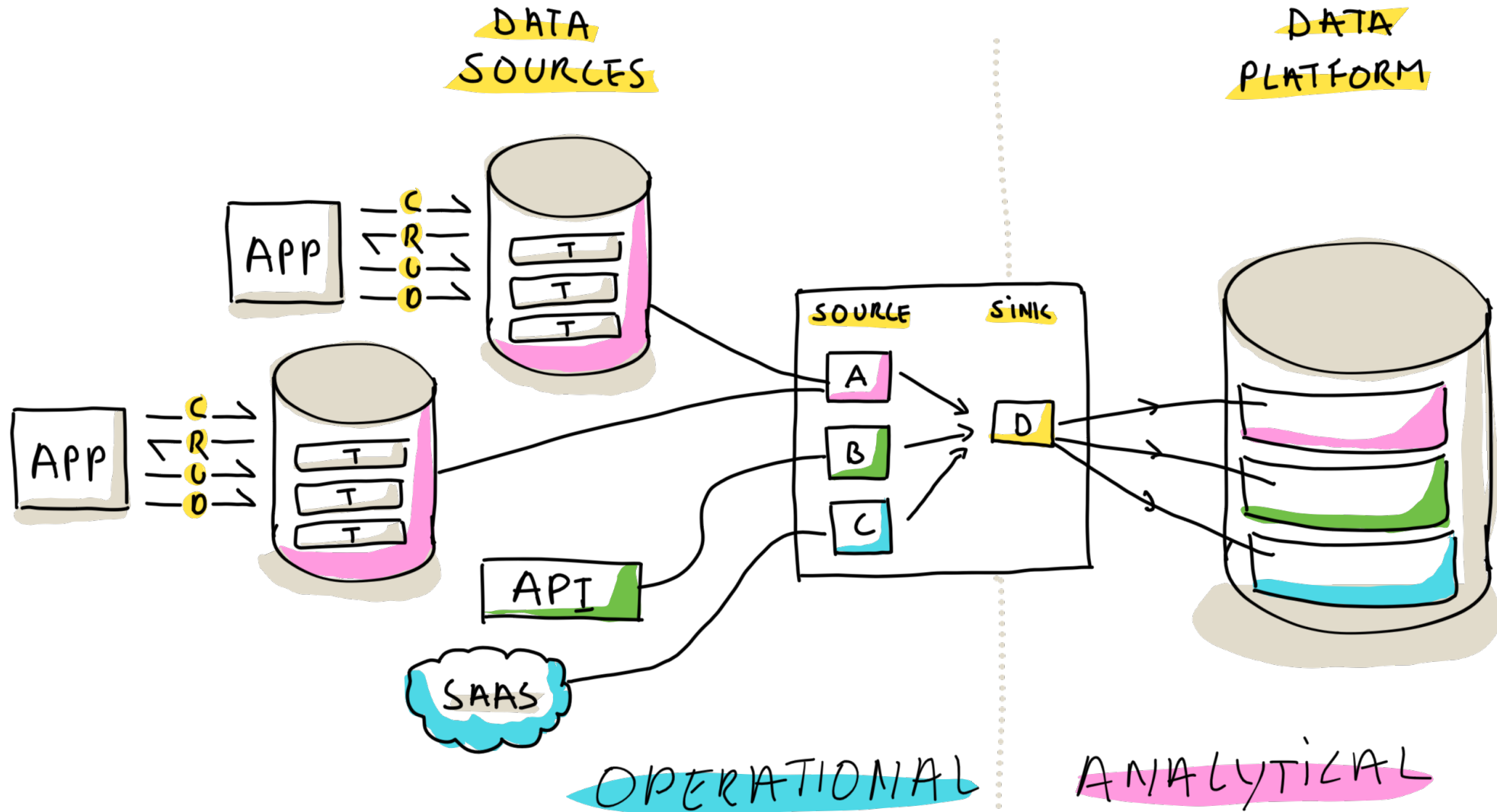
Log based CDC systems read data directly from the database Change Data Capture logs to identify changes in a database (not from the actual database)



CDC is based on transaction logs. These transactions can be incorporated in a **streaming architecture**.



Flavor #3 Connector Based

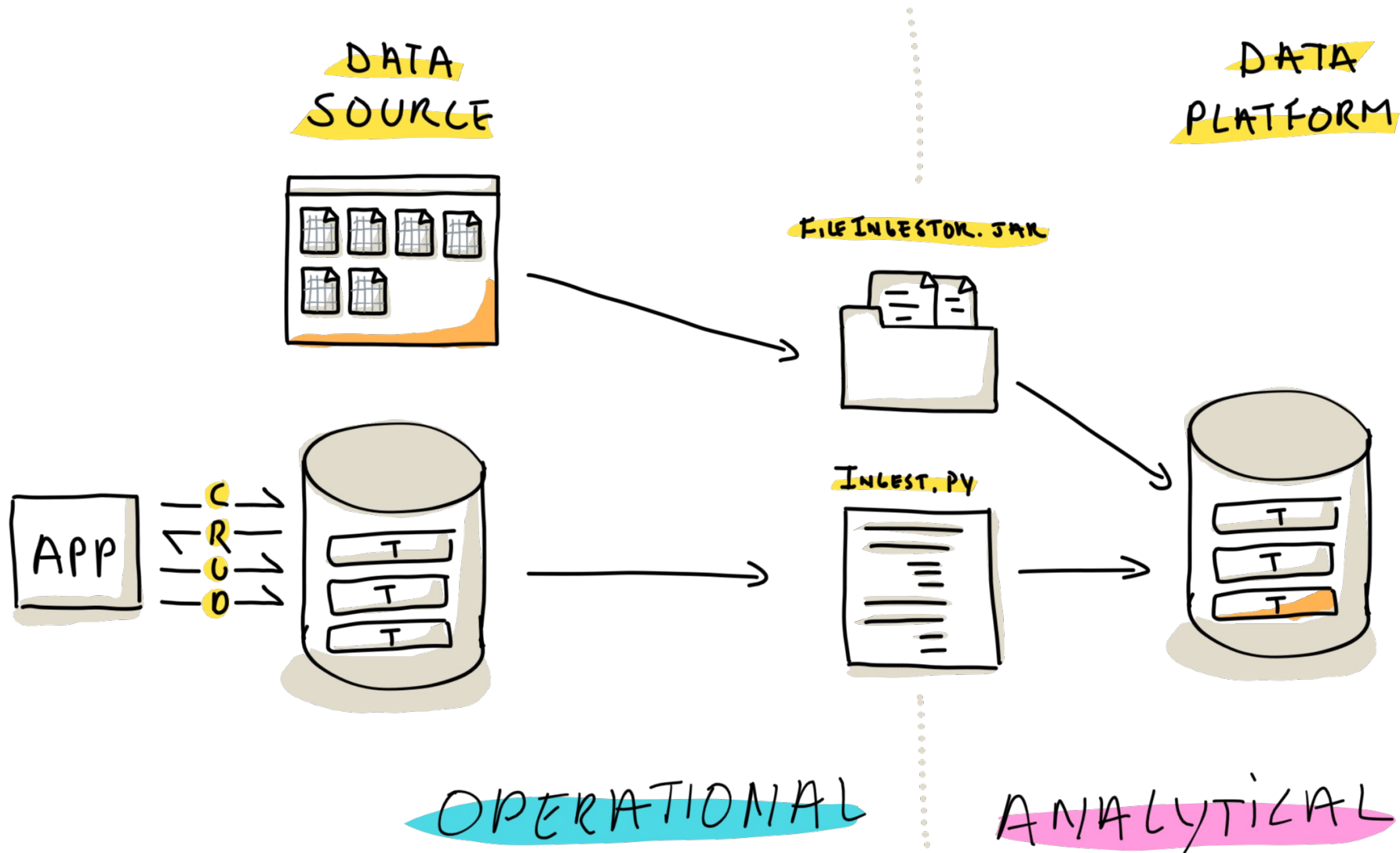


Pre-built source connectors are coupled with pre-built sink connectors, often using a graphical user interface.

- **Source connectors:** can be very diverse databases, APIs, SAAS-applications, applications, files, ...
- **Sink connectors:** mostly limited to (analytical) databases or data lakes.
- Highly **flexible** but **No control** over the individual connectors



Flavor #4 Custom Builds



Disadvantages:

- Building ingestion pipelines is usually more expensive than expected
- Specific programming knowledge & team needed
- High maintenance cost

Advantages:

- Full control
- Allows to ingest very specific / unique / exotic data sources



Building

VS

Buying



Customization and scale



Greater control

No license fees



Competitive edge



Easy to modify



Lower upfront cost



Rapid deployment

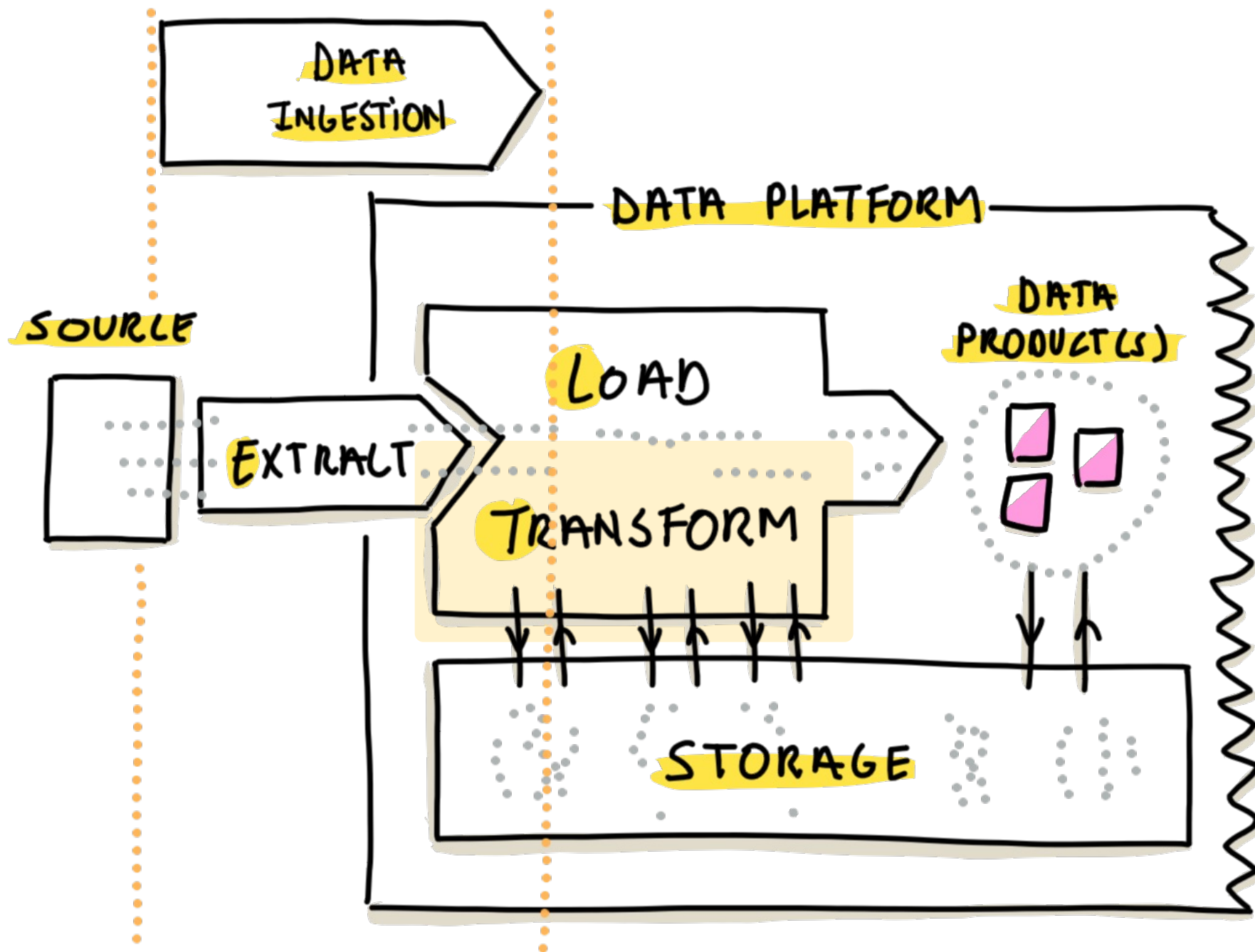


Updates and maintenance

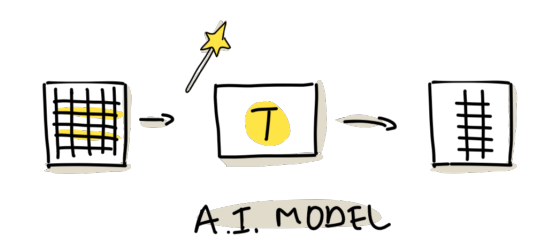
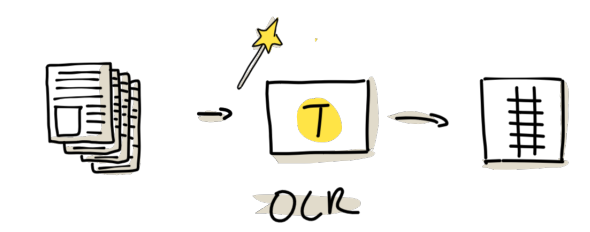
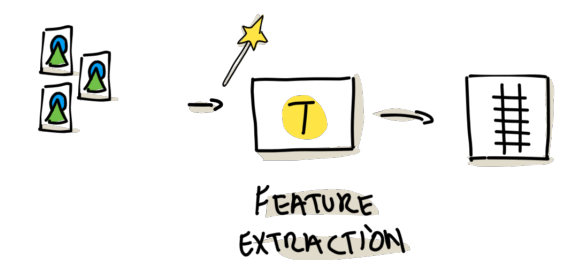
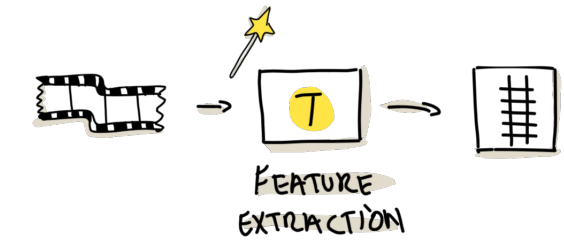
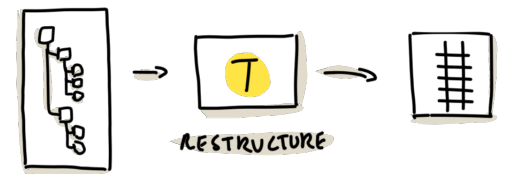
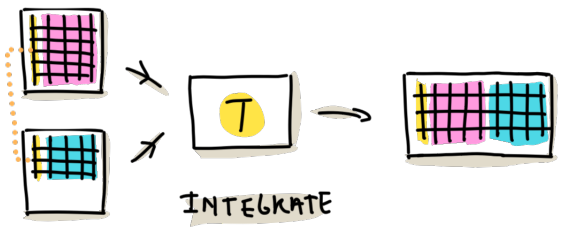
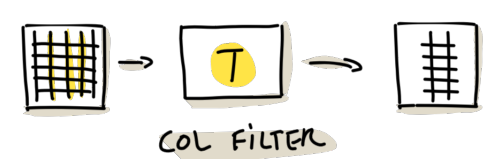
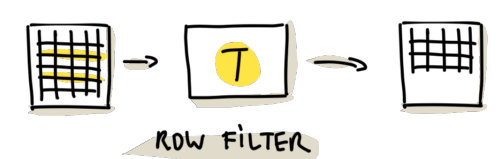
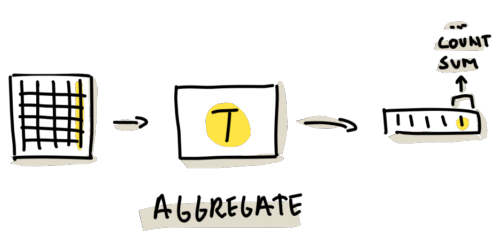
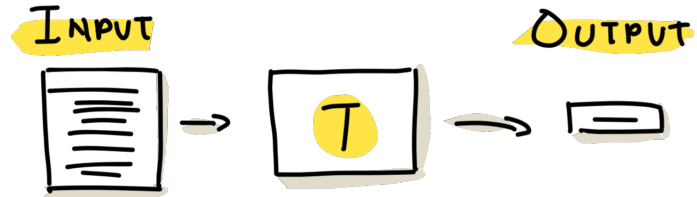


Has active userbase

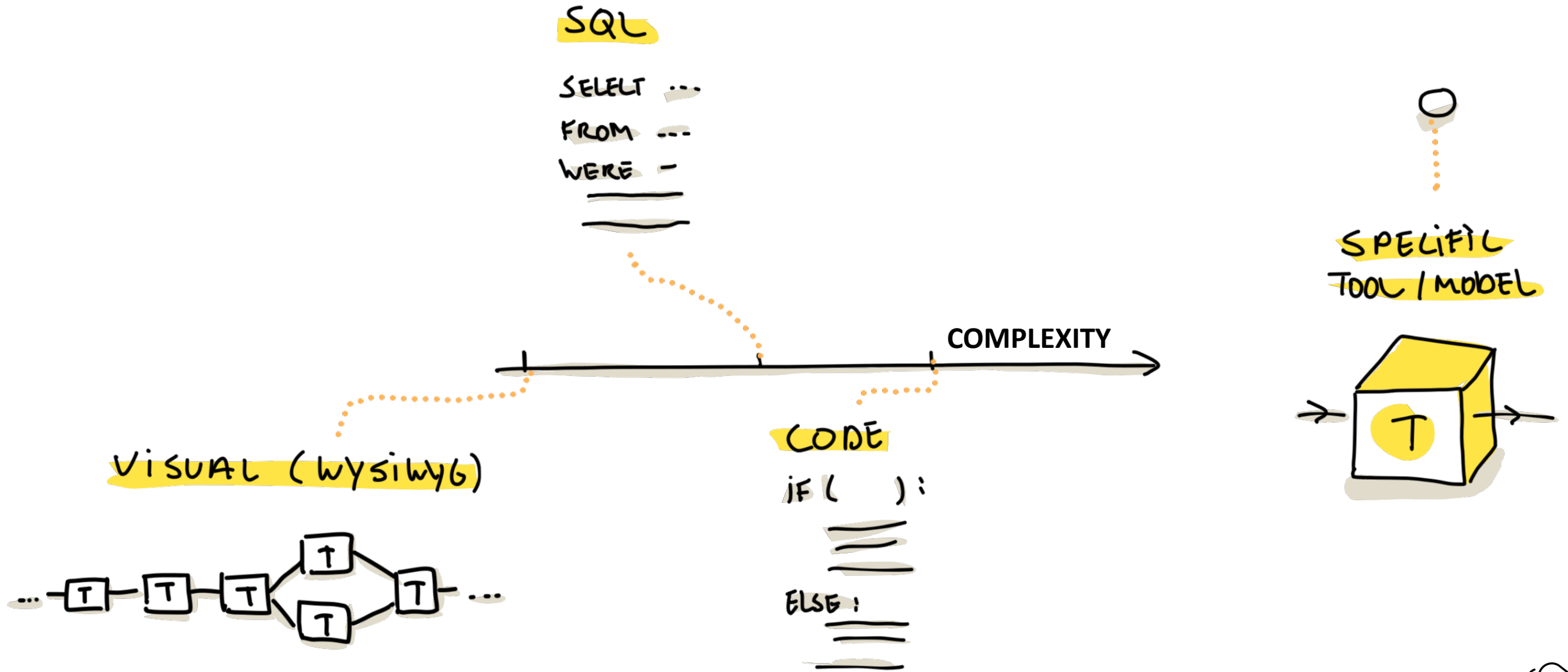




Data Transformations



Data Transformation Tool Flavors



Data Transformation Tool Flavors

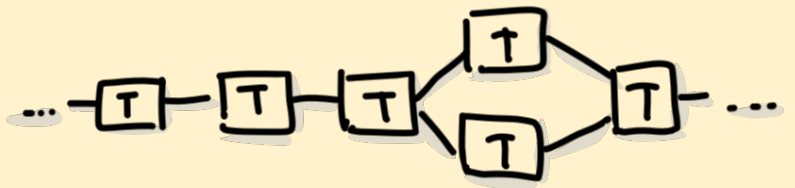
SQL

```
SELECT ...  
FROM ...  
WHERE -  
=====
```

○
.....
SPECIFIC
TOOL / MODEL

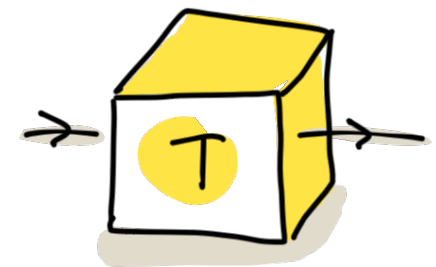
COMPLEXITY →

VISUAL (WYSIWYG)



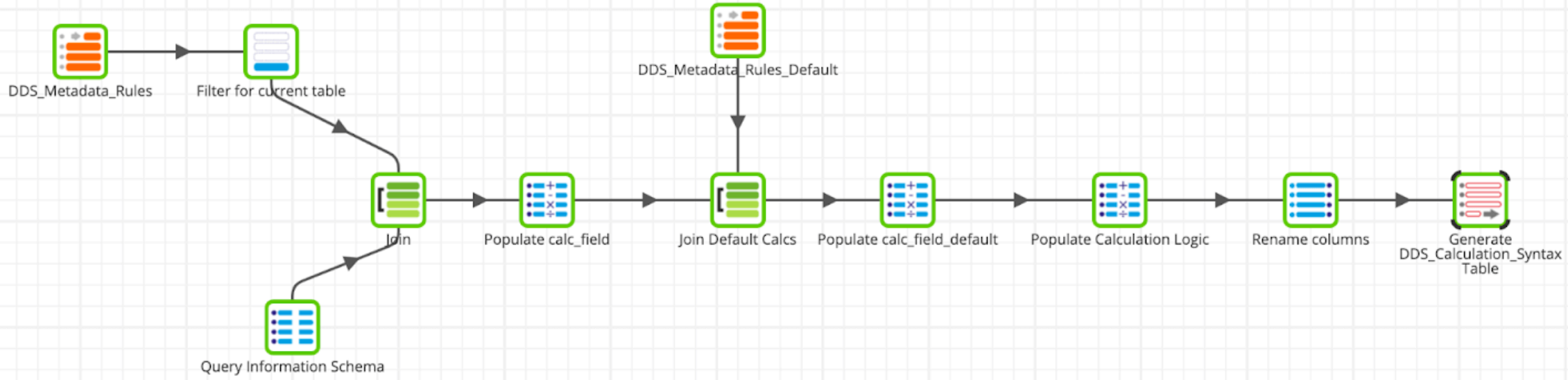
CODE

```
IF ( ) :  
=====  
ELSE :  
=====
```



Purpose:

The purpose of this transformation step is to formulate the appropriate syntax that should be applied onto the columns during the transformation step of the process. All of the syntax will be dynamically populated into a syntax metadata table called 'DDS_Calculation_Syntax'.



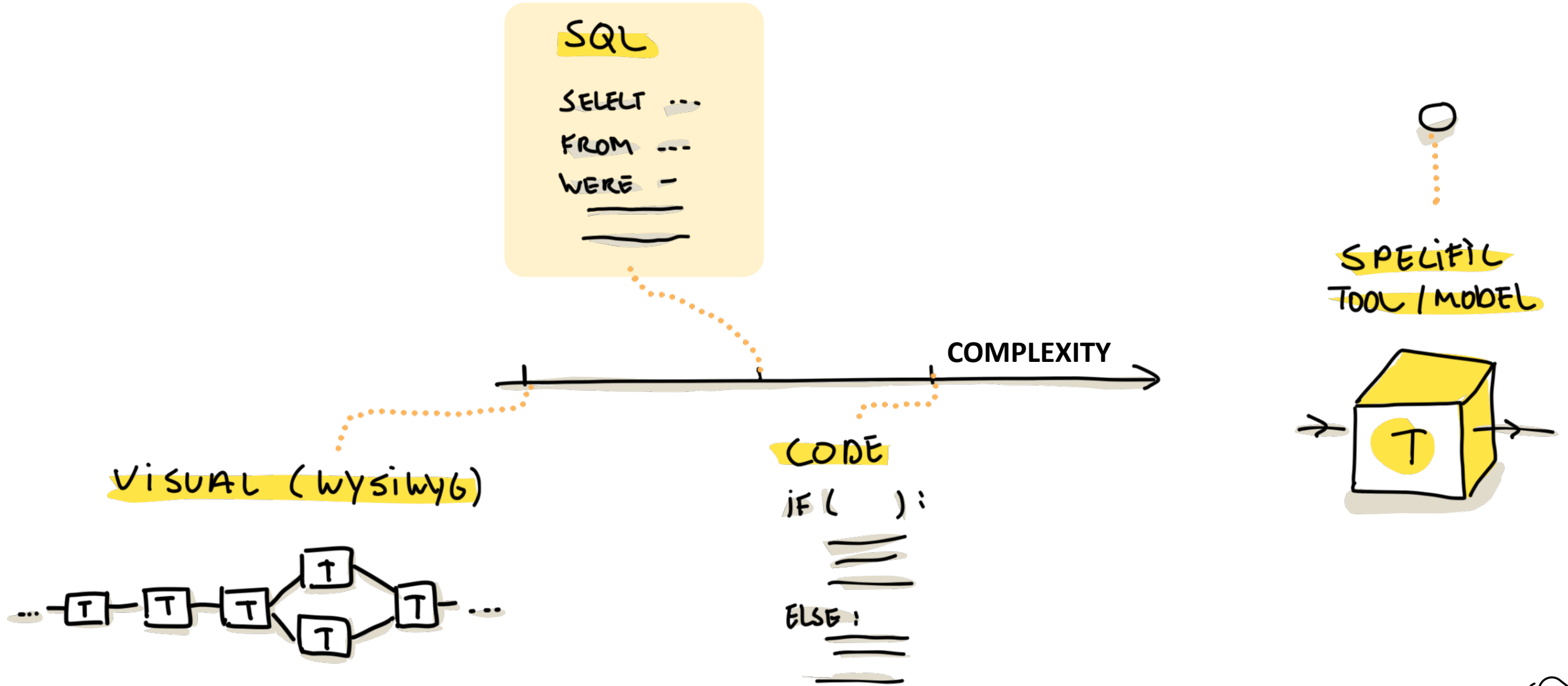
Properties | Export | Sample | Metadata | SQL | Plan | Help

Data | Row Count | Filter Not Set | Export

SCHEMA	TABLE_NAME	COLUMN_NAME	calculation
KB_AZURE_SNOW	DDS_STATE	STATEID	STATEID
KB_AZURE_SNOW	DDS_STATE	COUNTRYNAME	INITCAP(NVL(TRIM("COUNTRYNAME"),' '))
KB_AZURE_SNOW	DDS_STATE	STATECODE	UPPER(NVL(TRIM("STATECODE"),' '))
KB_AZURE_SNOW	DDS_STATE	FLAG	TO_BOOLEAN("FLAG")
KB_AZURE_SNOW	DDS_STATE	DATE	TO_TIMESTAMP_NTZ(TO_VARCHAR("DATE"), 'yyyymmdd')
KB_AZURE_SNOW	DDS_STATE	STATENAME	INITCAP(NVL(TRIM("STATENAME"),' '))



Data Transformation Tool Flavors



Project

view docs ?

Scratchpad 1

fact_employee_detail.sql

open pull request...

branch: jbarcheski_dev_demo

dbt_generic_demo

analysis

data

dbt_modules

logs

macros

models

sources

staging

warehouse

human_resources

dim_department.sql

dim_department.yml

dim_employee_department.sql

dim_employee_department.yml

fact_employee_detail.sql

fact_employee_detail.yml

purchasing

snapshots

target

tests

.gitignore

dbt_project.yml

packages.yml

```
31
32 final as (
33
34     select
35         to_binary(hex_encode('businessentityid'), 'HEX') as employee_sk,
36         e.businessentityid,
37         e.nationalidnumber as national_id,
38         e.loginid as login_id,
39         e.organizationnode as organization_node,
40         e.organizationlevel as organization_level,
41         e.jobtitle as job_title,
42         e.birthdate as birth_date,
43         e.maritalstatus as martial_status,
44         e.gender as gender,
45         e.hiredate as hire_date,
46         e.salariedflag as salaried_flag,
47         e.vacationhours as vacation_hours,
48         e.sickleavehours as sick_leave_hours,
49         e.currentflag as employee_current_flag,
50         e.rowguid as row_guid,
51         e.modifieddate as employee_modified_date,
52         datediff(year, hiredate, current_date()) as years_since_hire,
53         current_department_id,
54         current_shift_id,
55         current_department_start_date
56     from employees e
57     left join current_department d
58         on e.businessentityid=d.businessentityid
59 )
60
61 select * from final
```

Preview

Compile

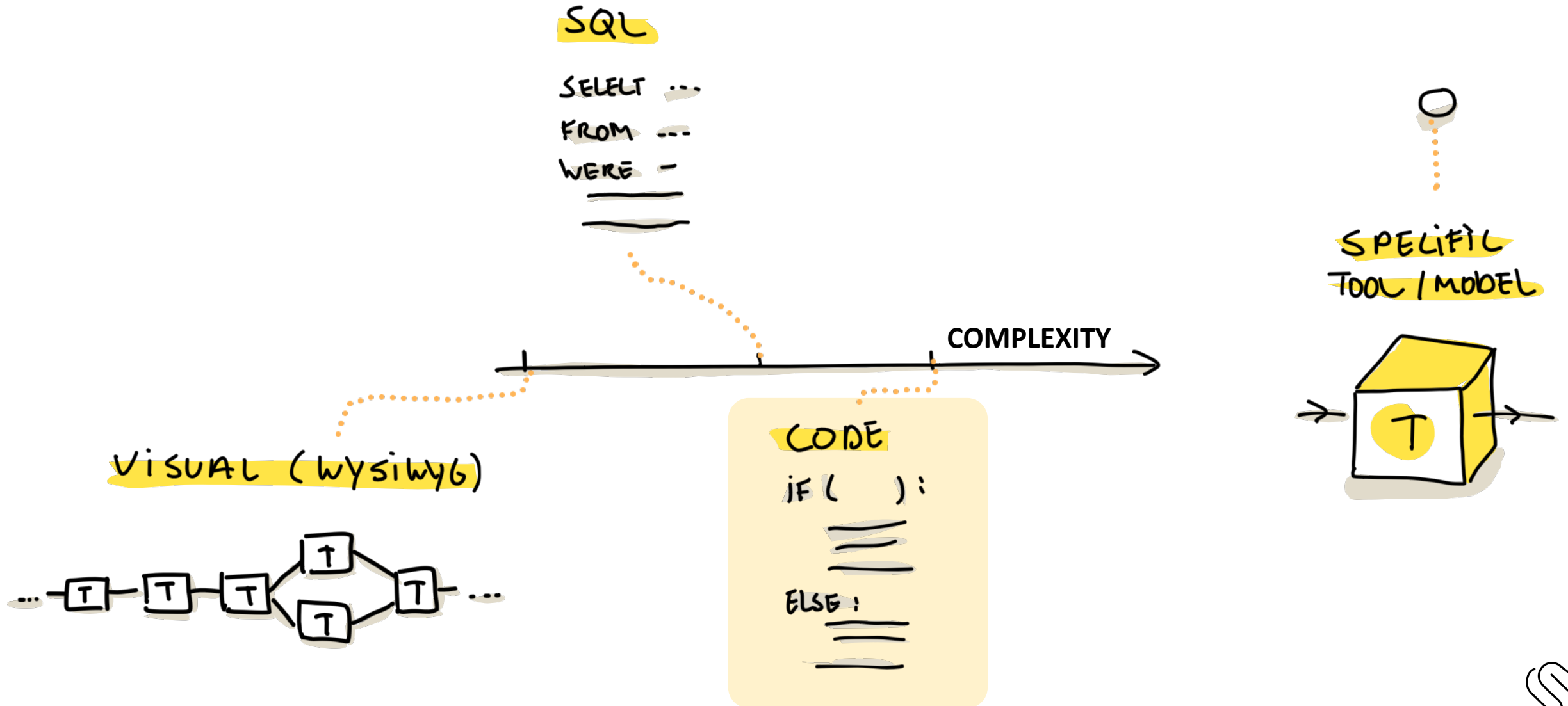
Query Results

Compiled SQL

Lineage



Data Transformation Tool Flavors





Background-color change



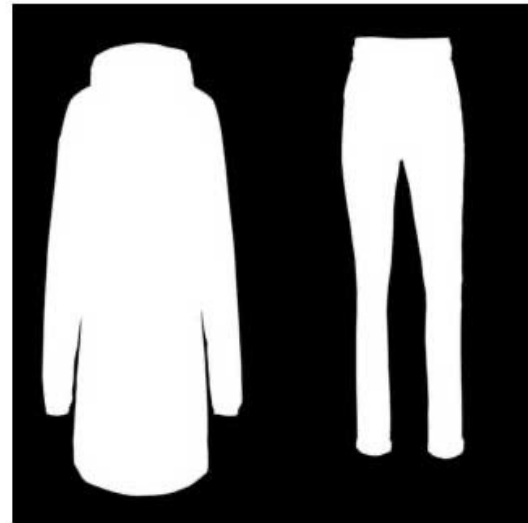
Background-color change

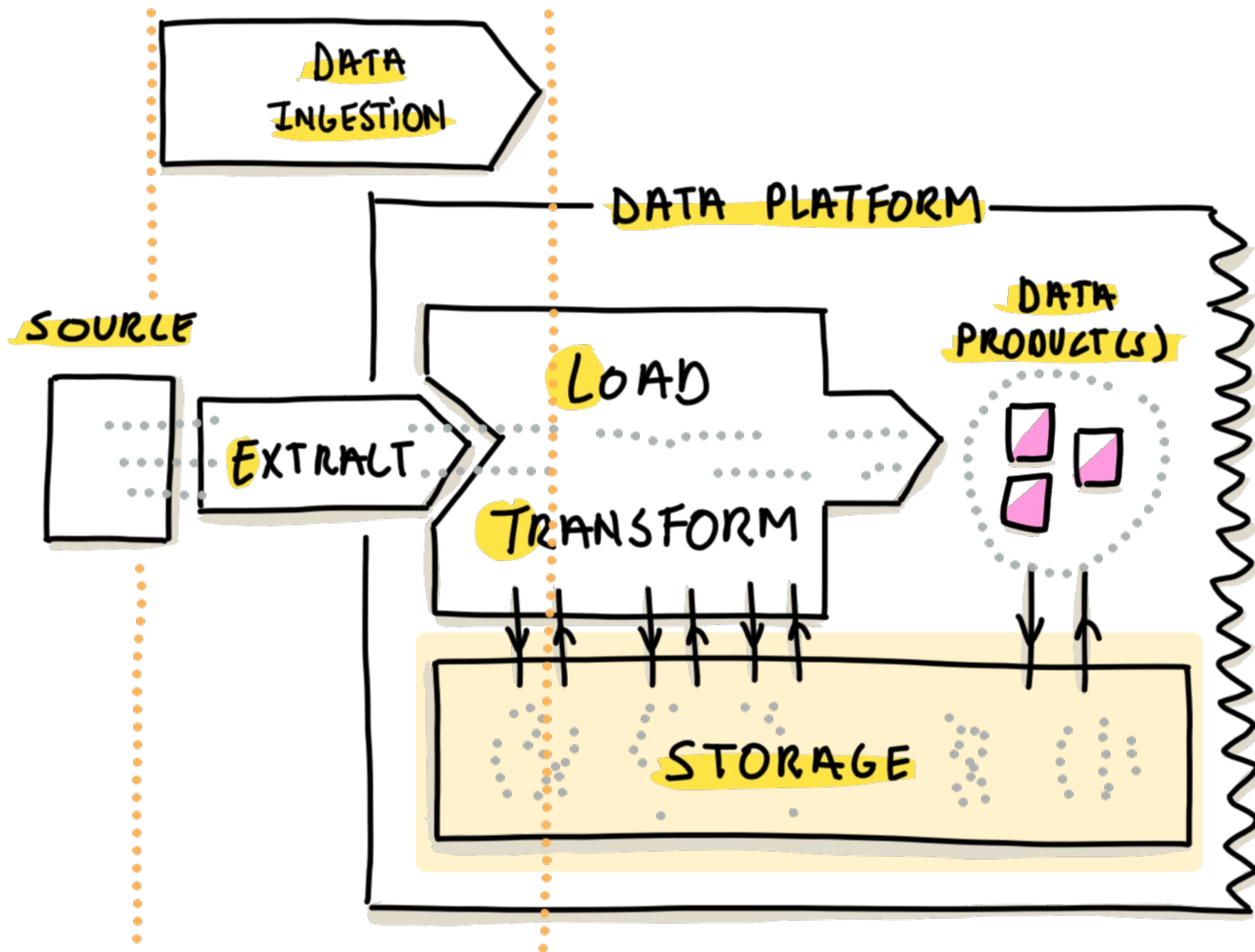


Cropping

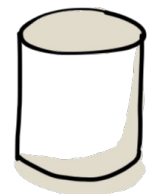


Combined images

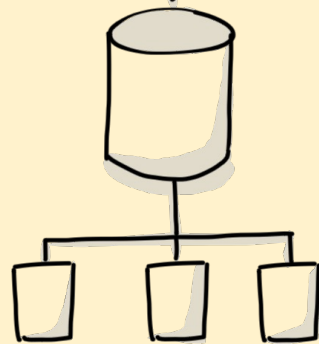




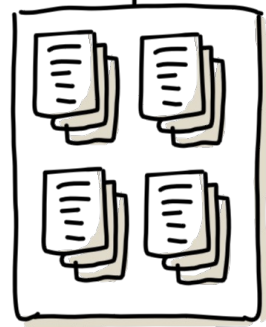
COMMON



DATABASE



DWH DATABASE

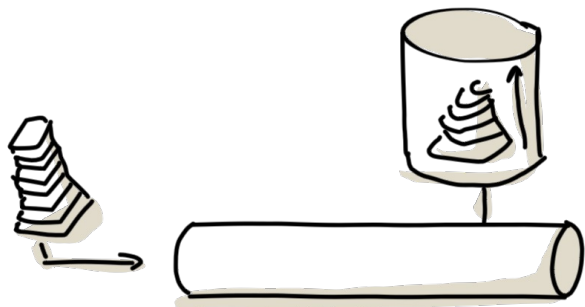


DATA LAKE

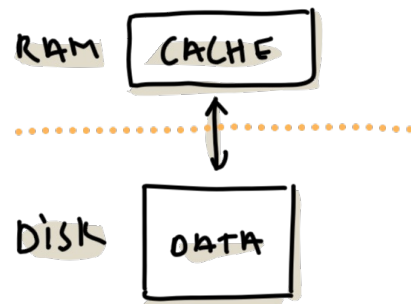


DATA LAKEHOUSE

SPECIFIC



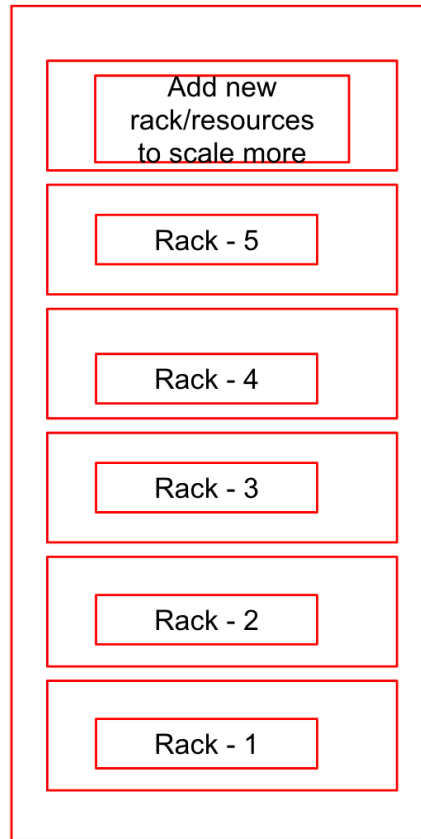
STREAMING CACHE



IN MEMORY DATABASE



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



Host 1
192.168.1.1

Vertical Scaling

To scale more, Add more RAM, CPU, Memory to the **one existing machine**

Horizontal Scaling

To scale more: Add more machines to existing **group of distributed system**

Host 1
192.168.1.1

Host 2
192.168.1.2

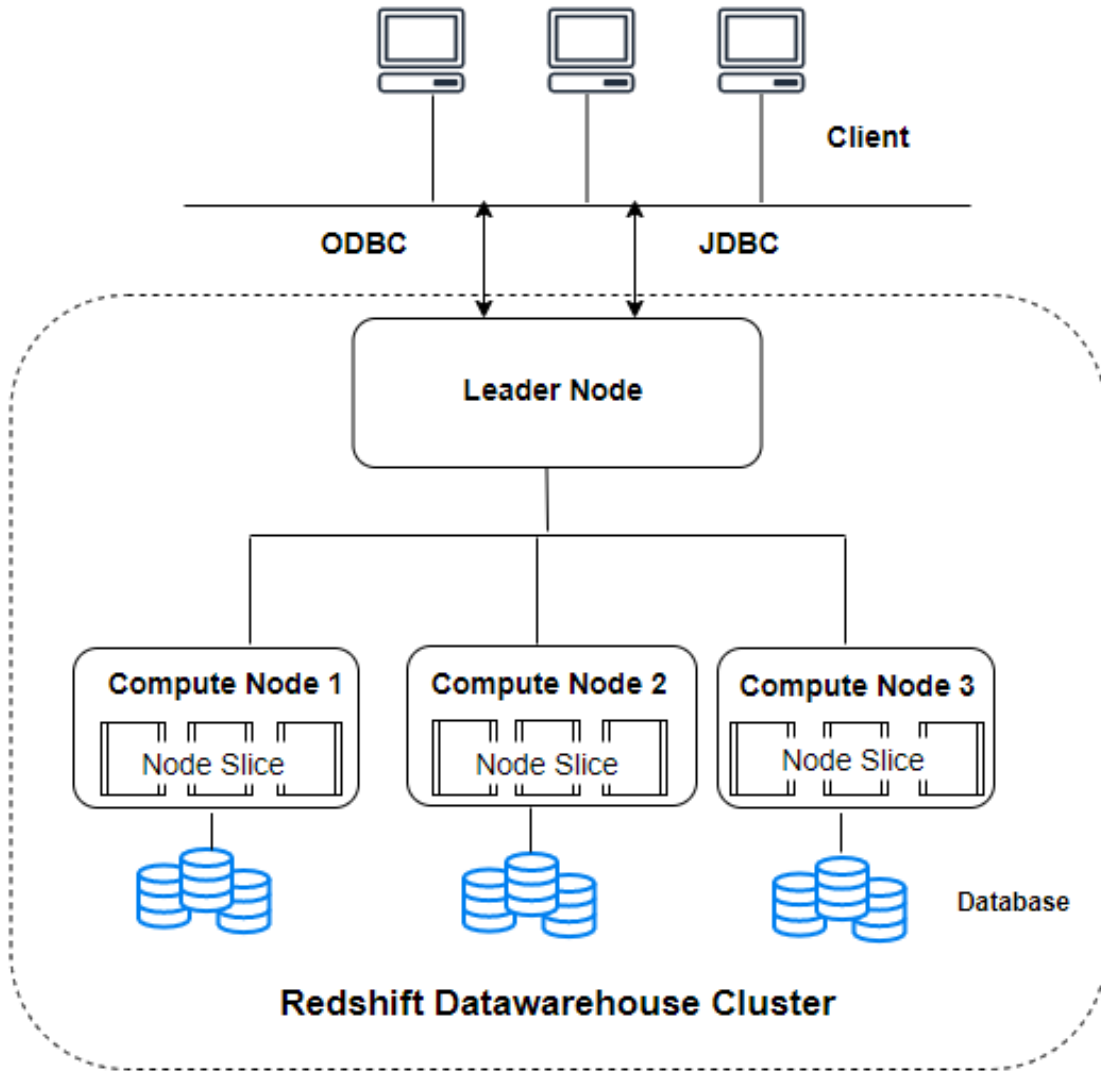
Host 3
192.168.1.3

Host x
192.168.1.x

Add x+1
host to
scale out

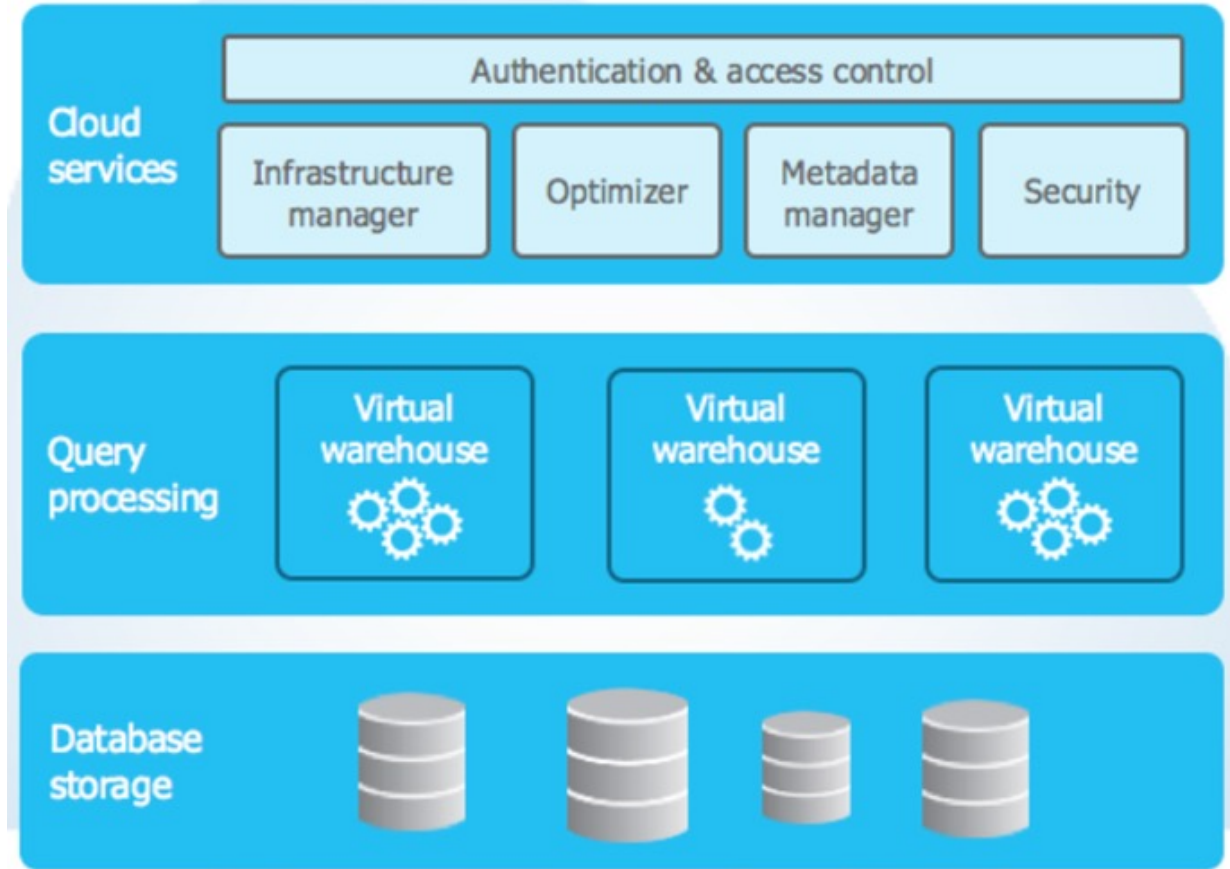


Example: Amazon Redshift



Scaling = Refconfiguring total databases

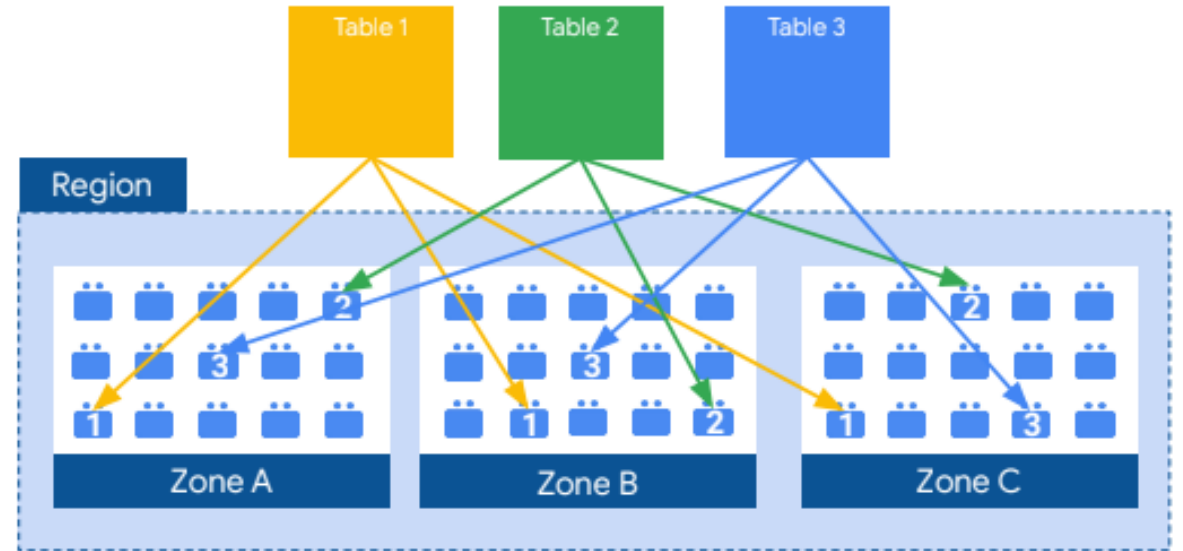
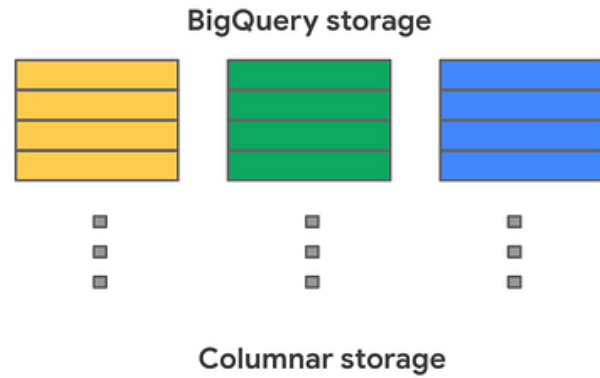
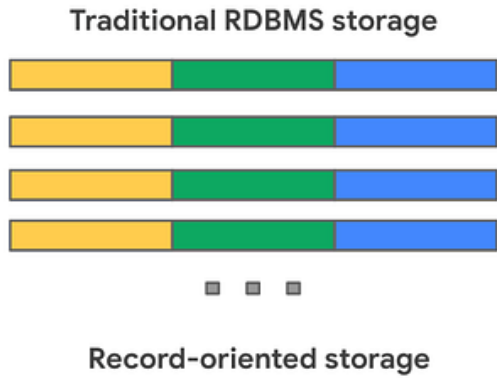
Example: Snowflake



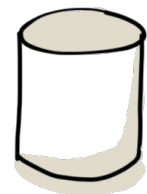
Independent Scaling



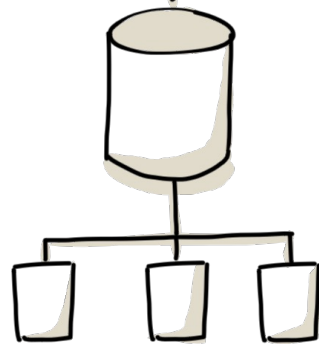
Example: Google BigQuery



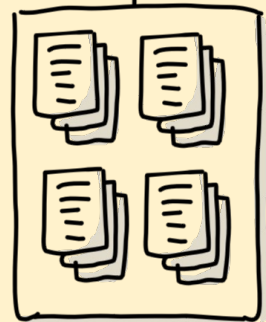
COMMON



DATABASE



DWH DATABASE

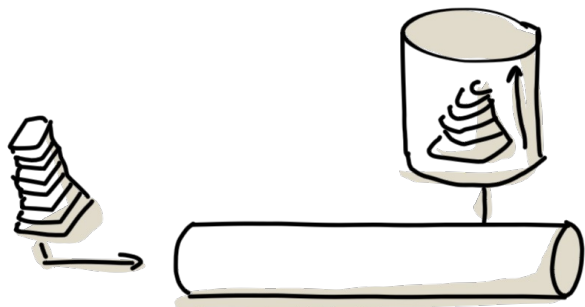


DATA LAKE

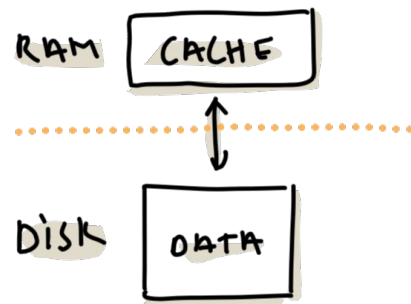


DATA LAKEHOUSE

SPECIFIC



STREAMING CACHE



IN MEMORY DATABASE

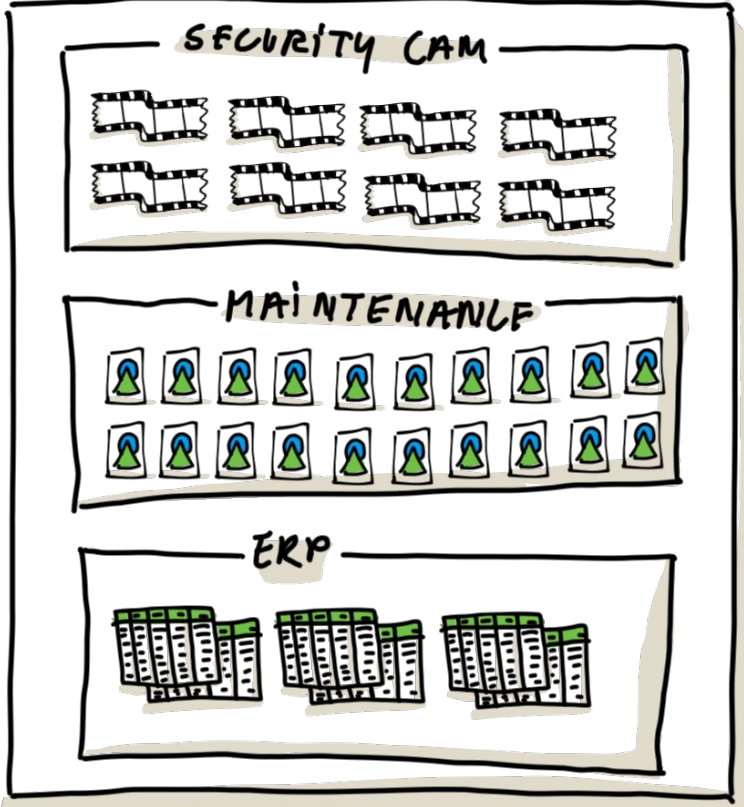


Data Lake

DATA SOURCES

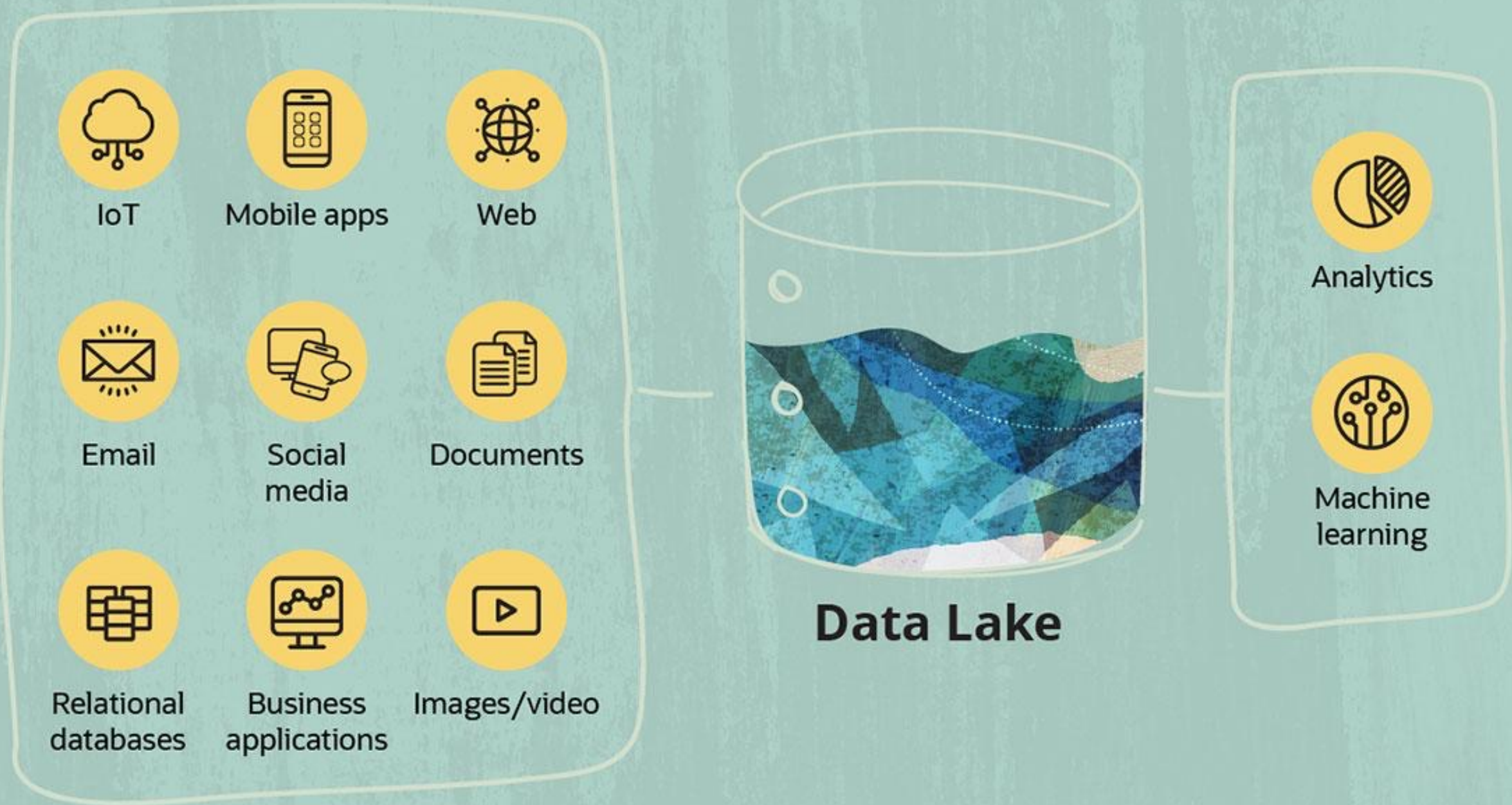


DATA LAKE

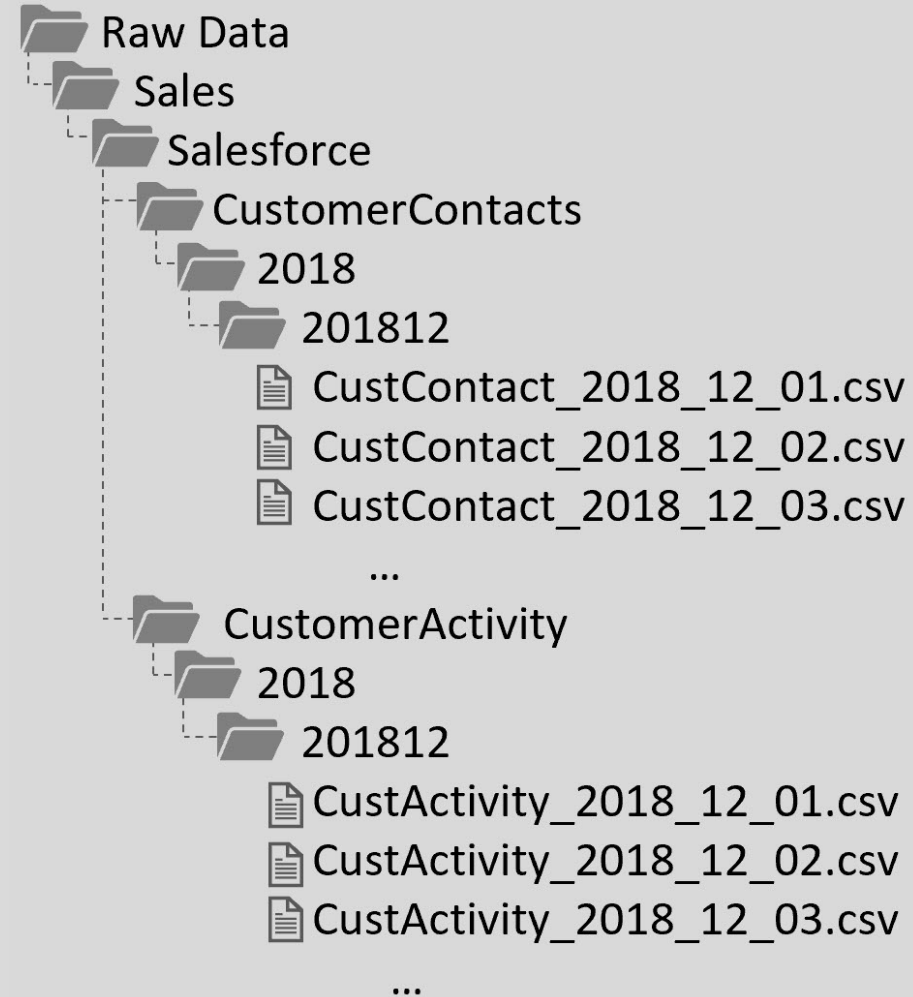


DATA PRODUCTS

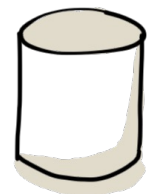




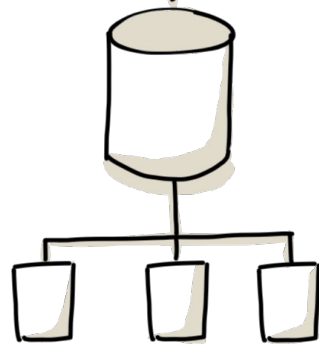
Data Lake: A **Structured** File Repository



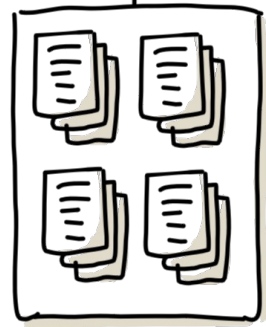
COMMON



DATABASE



DWH DATABASE

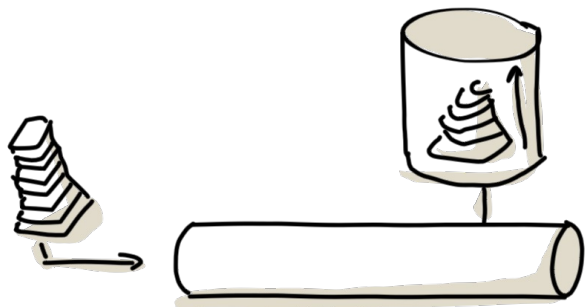


DATA LAKE

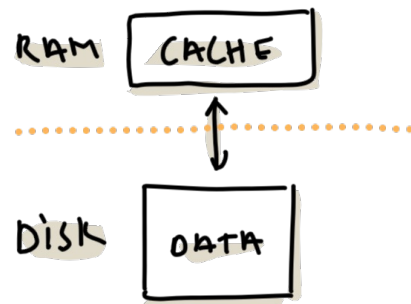


DATA LAKEHOUSE

SPECIFIC



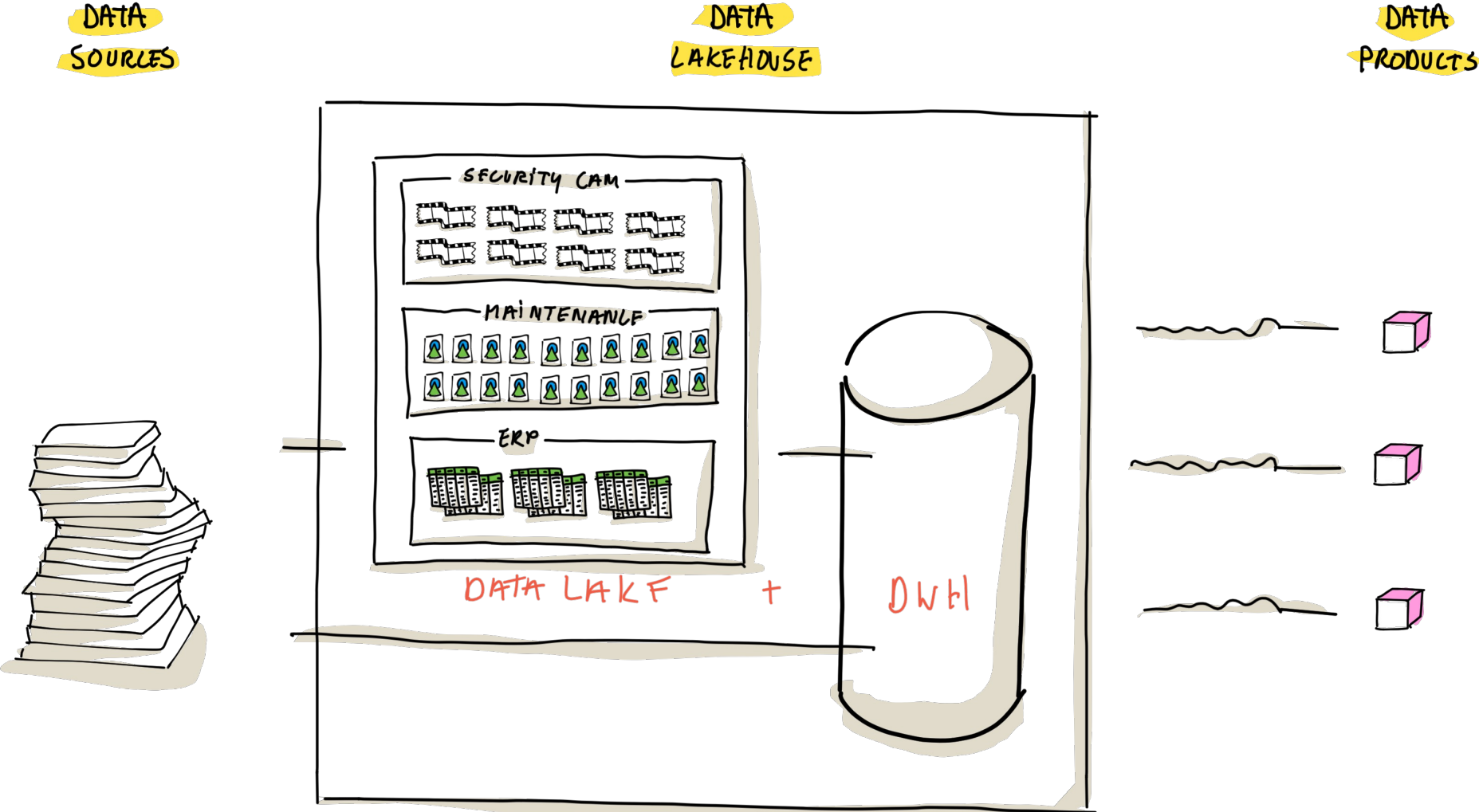
STREAMING CACHE



IN MEMORY DATABASE

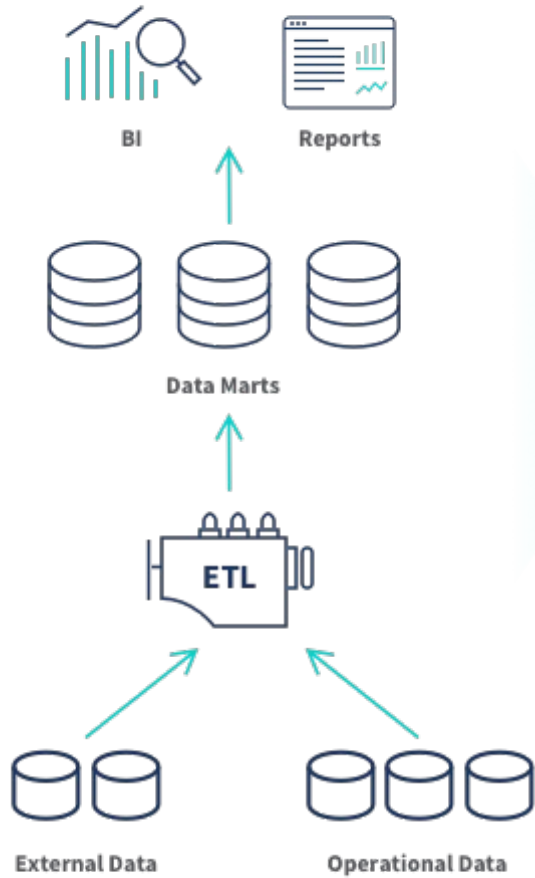


Data Lakehouse



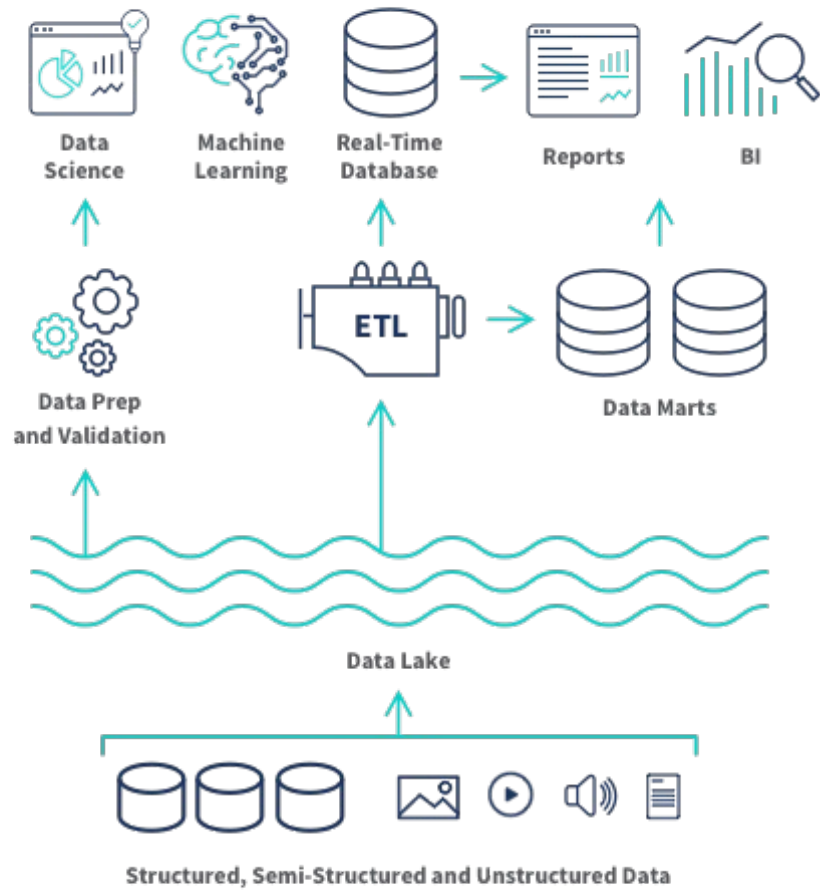
LATE 1980'S

Data Warehouse



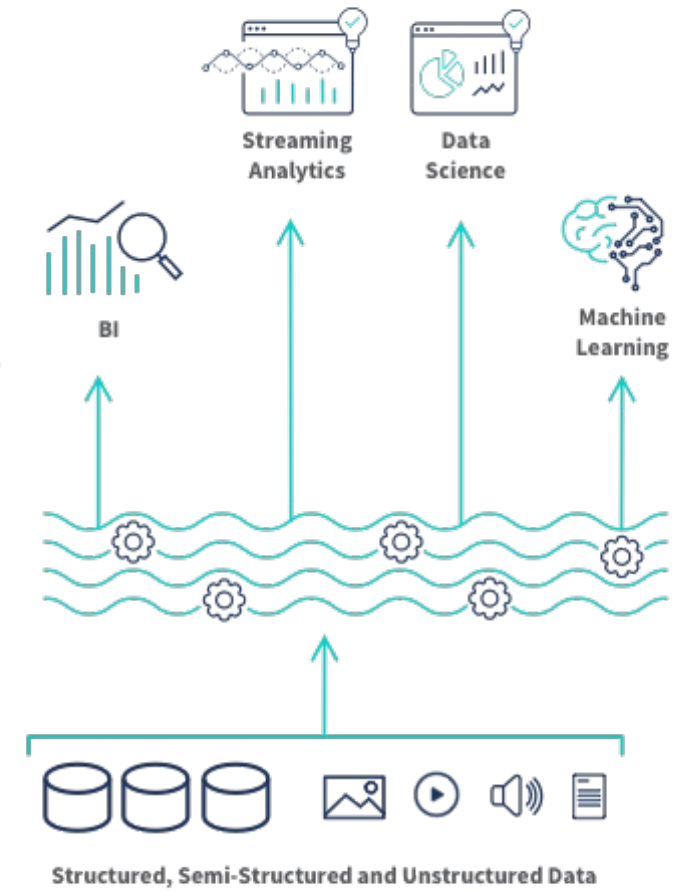
2011

Data Lake

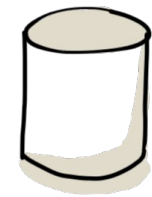


2020

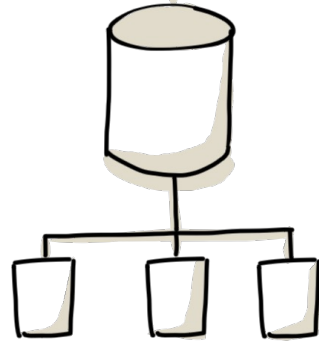
Lakehouse



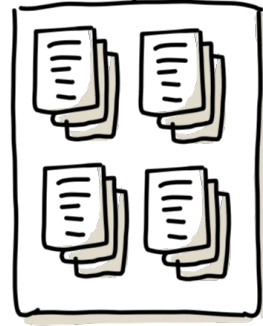
COMMON



DATABASE



DWH
DATABASE

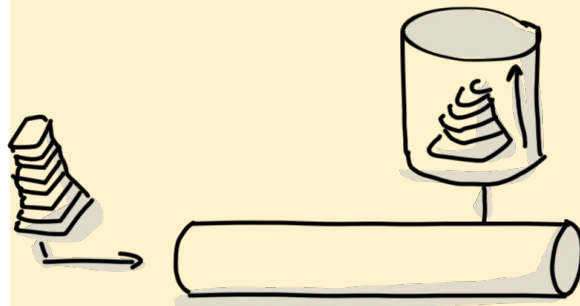


DATA
LAKE

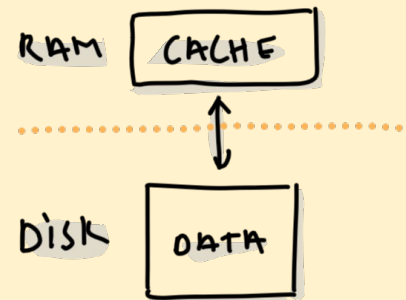


DATA
LAKEHOUSE

SPECIFIC



STREAMING
CACHE

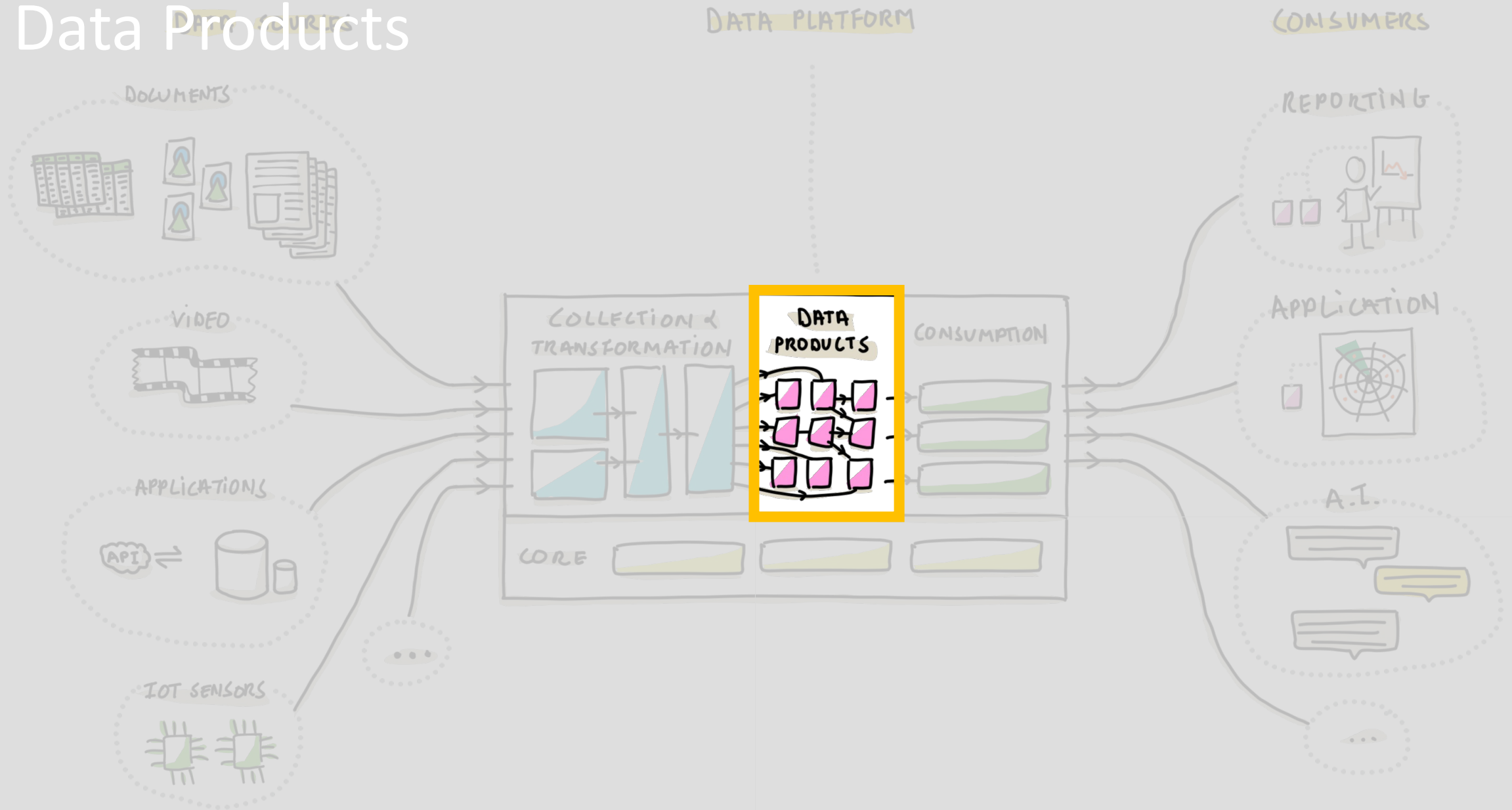


IN MEMORY
DATABASE



	Database	DWH Database	Data Lake	Data Lakehouse
Cost	+++	+++	+	++
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)	++++	++	+++(+)

2. Data Products



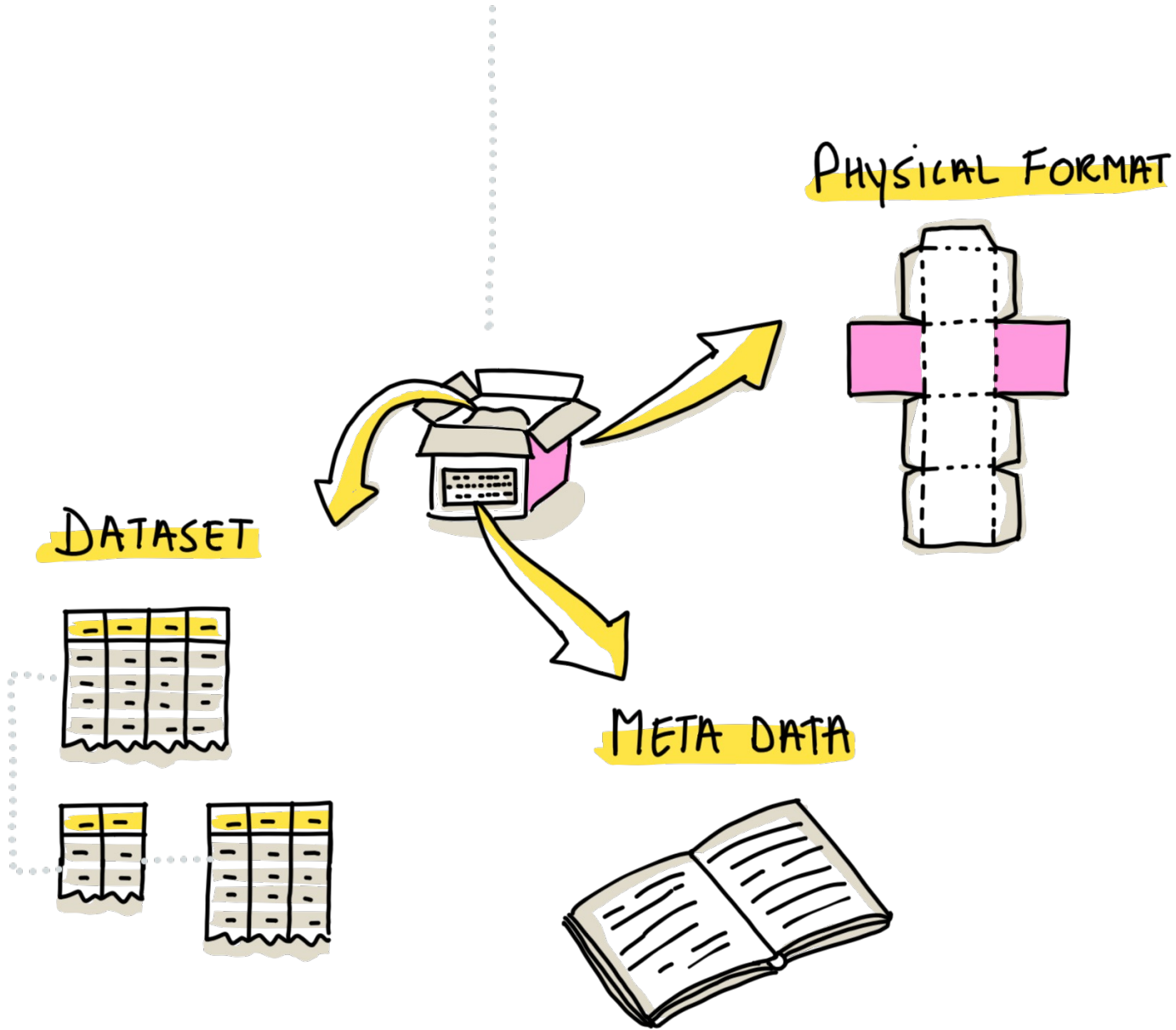
“Operational Plane”

“Analytical Plane”

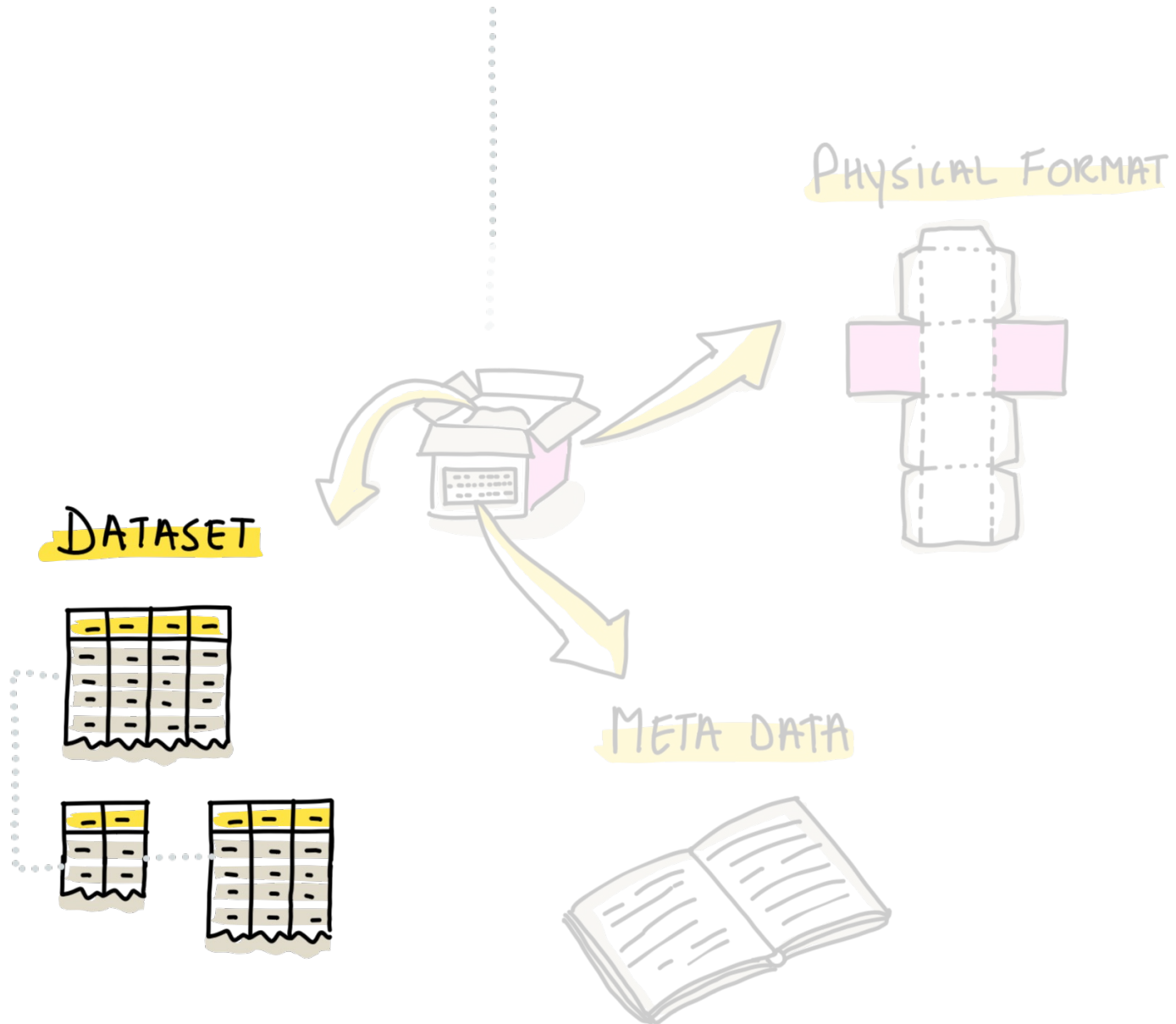
“Operational / Analytical Plane”



DATA PRODUCT



DATA PRODUCT



Dataset (Result from data transformations)

COLUMN NAME or
VARIABLE NAME

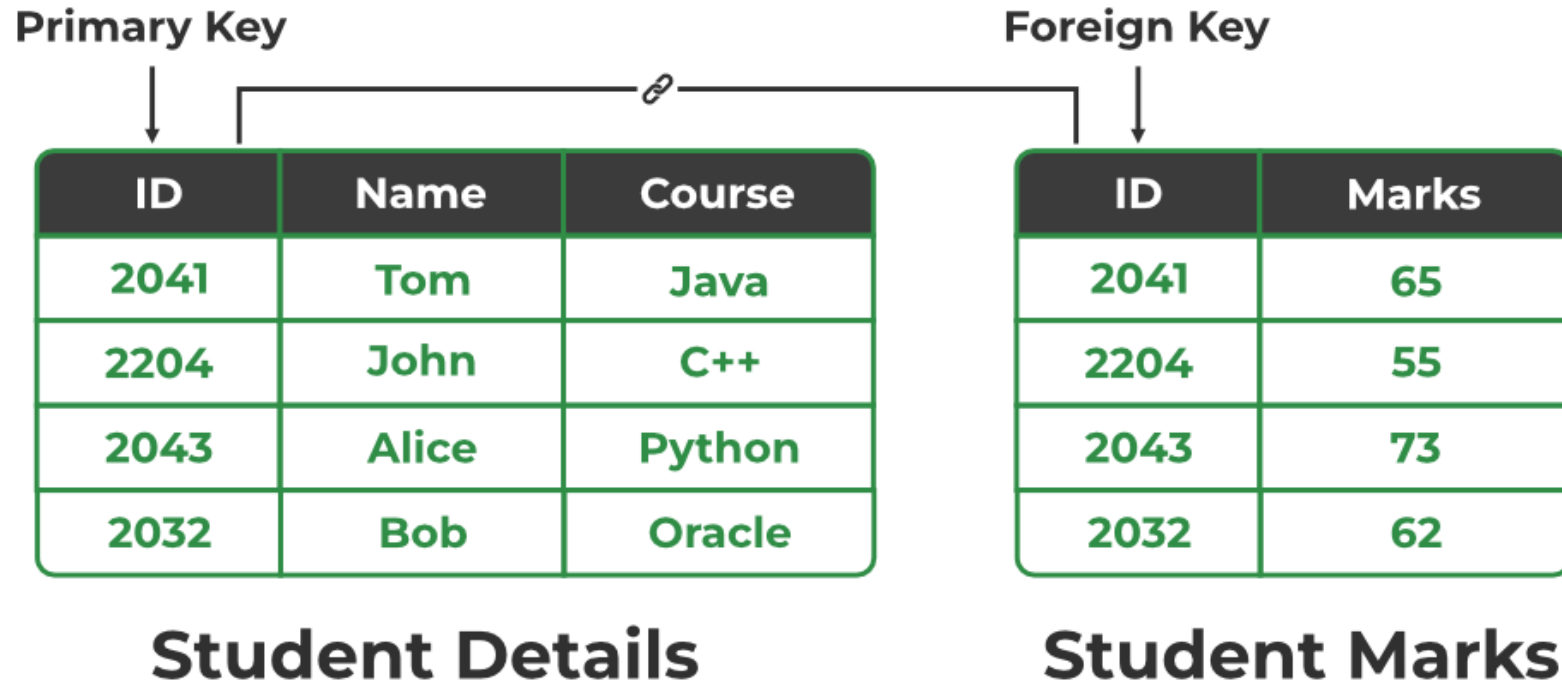
COLUMN or
VARIABLE

	A	B	C	D	E	F	G	H	I	J	K
	Additional_Number_of_Scoring	Review_Date	Average_Score	Hotel_Name	Reviewer_Nationality	Review_Total_Negative_Word_Counts	Total_Number_of_Reviews	Review_Total_Positive_Word_Counts	Total_Number_of_Reviews_Reviewer_Has_Given	Reviewer_Score	days_since_review
1		194	3/8/2017	Hotel Arena	Russia	0	1403	11	7	2.9	0 days
2		194	3/8/2017	Hotel Arena	Ireland	0	1403	105	7	7.5	0 days
3		194	2/11/2017	Hotel Arena	Australia	42	1403	21	9	7.1	3 days
4		194	7/31/2017	Hotel Arena	United Kingdom	210	1403	26	1	3.8	3 days
5		194	7/24/2017	Hotel Arena	New Zealand	140	1403	8	3	6.7	10 days
6		194	7/24/2017	Hotel Arena	Poland	17	1403	20	1	6.7	10 days
7		194	7/17/2017	Hotel Arena	United Kingdom	33	1403	18	6	4.6	17 days
8		194	7/17/2017	Hotel Arena	United Kingdom	11	1403	19	1		10 17 days
9		194	9/7/2017	Hotel Arena	Belgium	34	1403	0	3	6.5	25 days
10		194	8/7/2017	Hotel Arena	Norway	15	1403	50	1	7.9	26 days
11		194	7/7/2017	Hotel Arena	United Kingdom	5	1403	101	2		10 27 days
12		194	6/7/2017	Hotel Arena	France	75	1403	4	12	5.8	28 days
13		194	6/7/2017	Hotel Arena	United Kingdom	28	1403	6	7	4.6	28 days
14		194	4/7/2017	Hotel Arena	Italy	0	1403	59	6	9.2	30 days
15		194	4/7/2017	Hotel Arena	Canada	35	1403	15	1	8.8	30 days
16		194	3/7/2017	Hotel Arena	Italy	0	1403	82	26		10 31 days
17		194	3/7/2017	Hotel Arena	United Kingdom	38	1403	14	8	6.3	31 days
18		194	6/30/2017	Hotel Arena	Ireland	59	1403	64	2	7.5	34 days
19		194	6/29/2017	Hotel Arena	Netherlands	0	1403	33	4	7.1	35 days
20		194	6/20/2017	Hotel Arena	Australia	73	1403	48	16	7.5	44 days
21		194	6/19/2017	Hotel Arena	United Kingdom	40	1403	17	1	6.1	45 days
22		194	12/6/2017	Hotel Arena	France	92	1403	75	12	3.8	52 days
23		194	10/6/2017	Hotel Arena	Germany	364	1403	37	21	5.4	54 days
24		194	5/29/2017	Hotel Arena	Ireland	22	1403	51	1	9.6	66 days
25		194	5/27/2017	Hotel Arena	United Kingdom	3	1403	130	2	9.6	68 days
26		194	5/25/2017	Hotel Arena	United States of America	57	1403	41	4	8.1	70 days

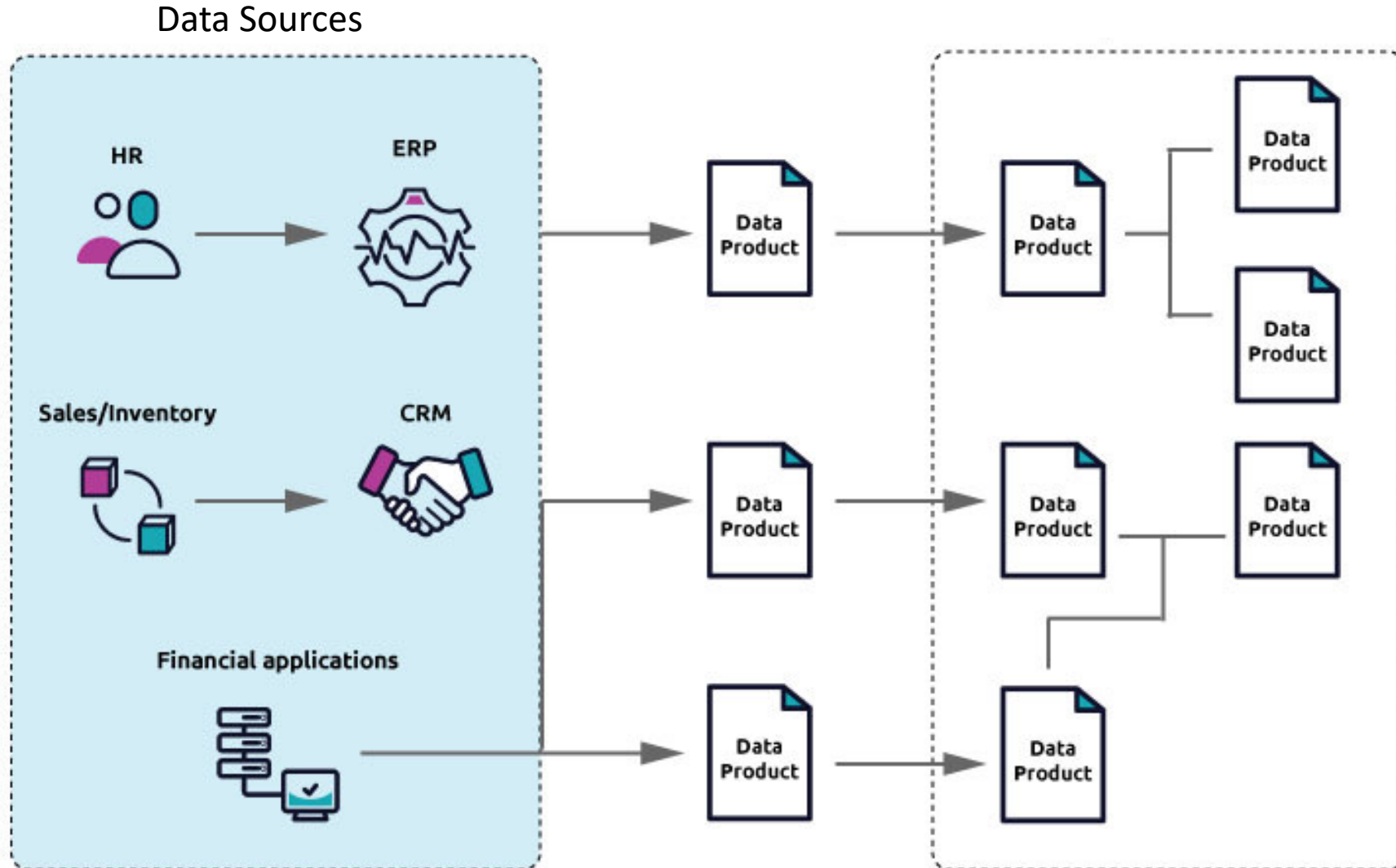
TABLE



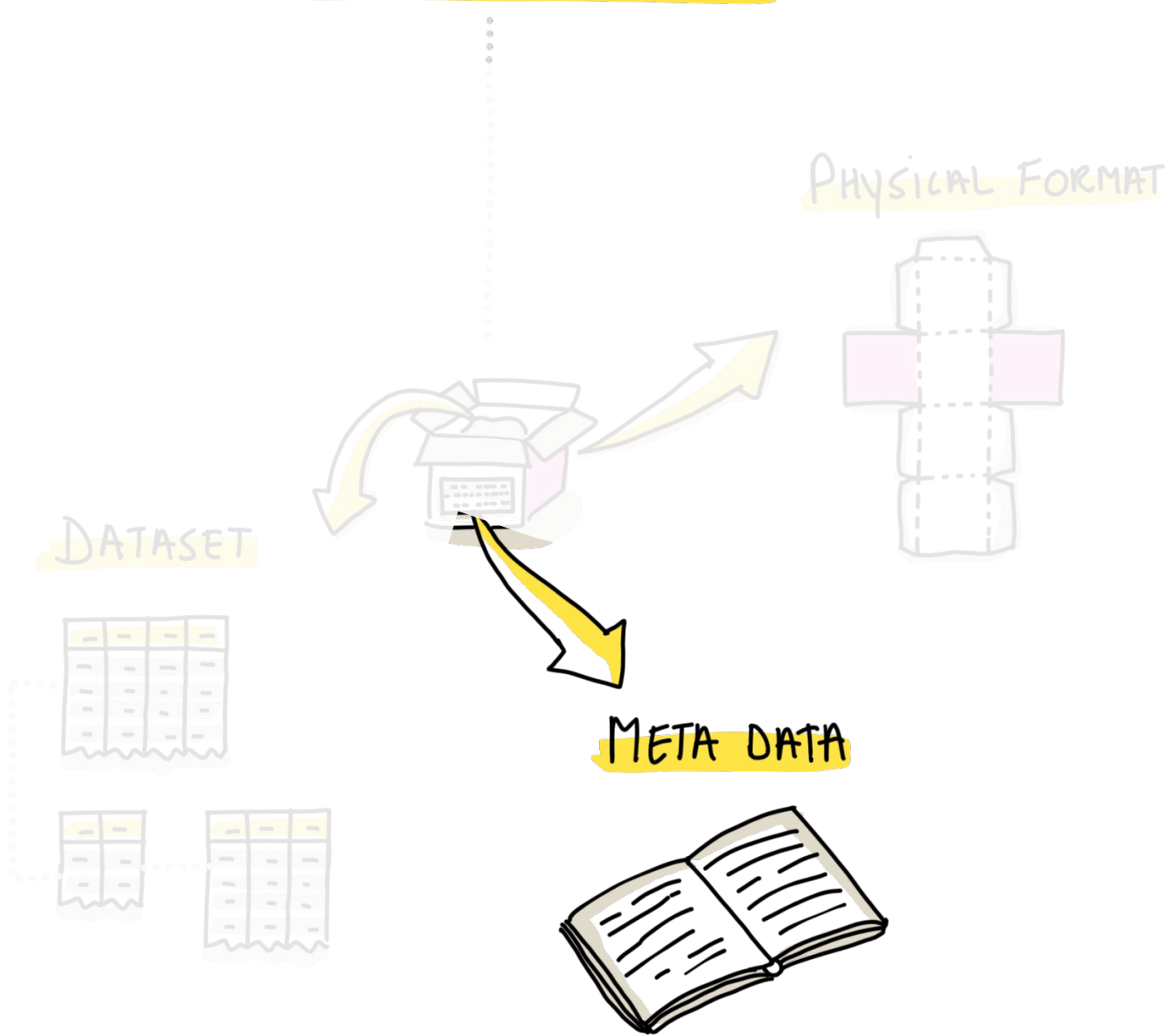
Multi-table datasets



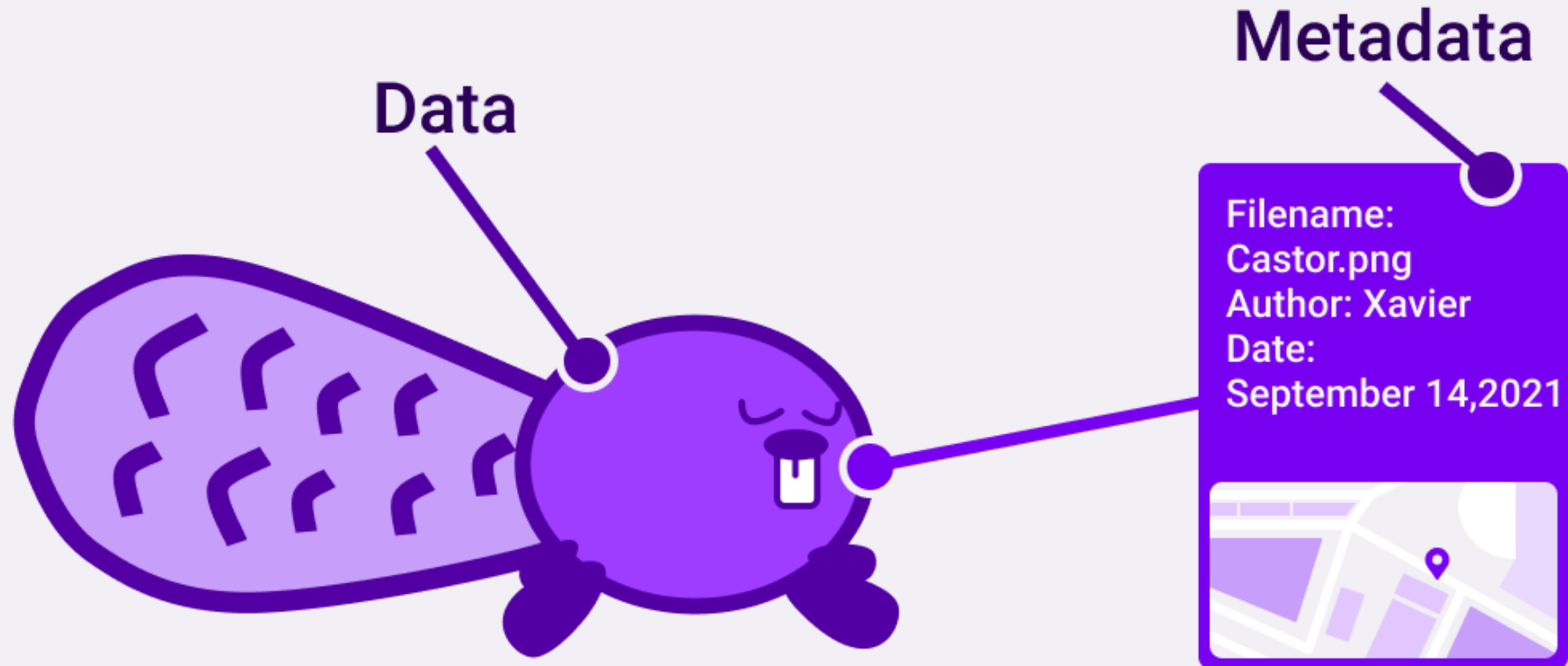
Data product datasets can link to each other



DATA PRODUCT



Metadata = “Data about Data”



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

employee_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	Bemdt	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
employee_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 position title, e.g. Secretary
department_id	int	Employee department. Ref: Department
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

Tool: DBT (Data Build Tool)

Data Movement Tracking



Data Source:

warehouse.customers

customers_snapshot

stg_eltool__customers

warehouse.state

stg_eltool__state

dim_customers

warehouse.orders

stg_eltool__orders

fct_orders

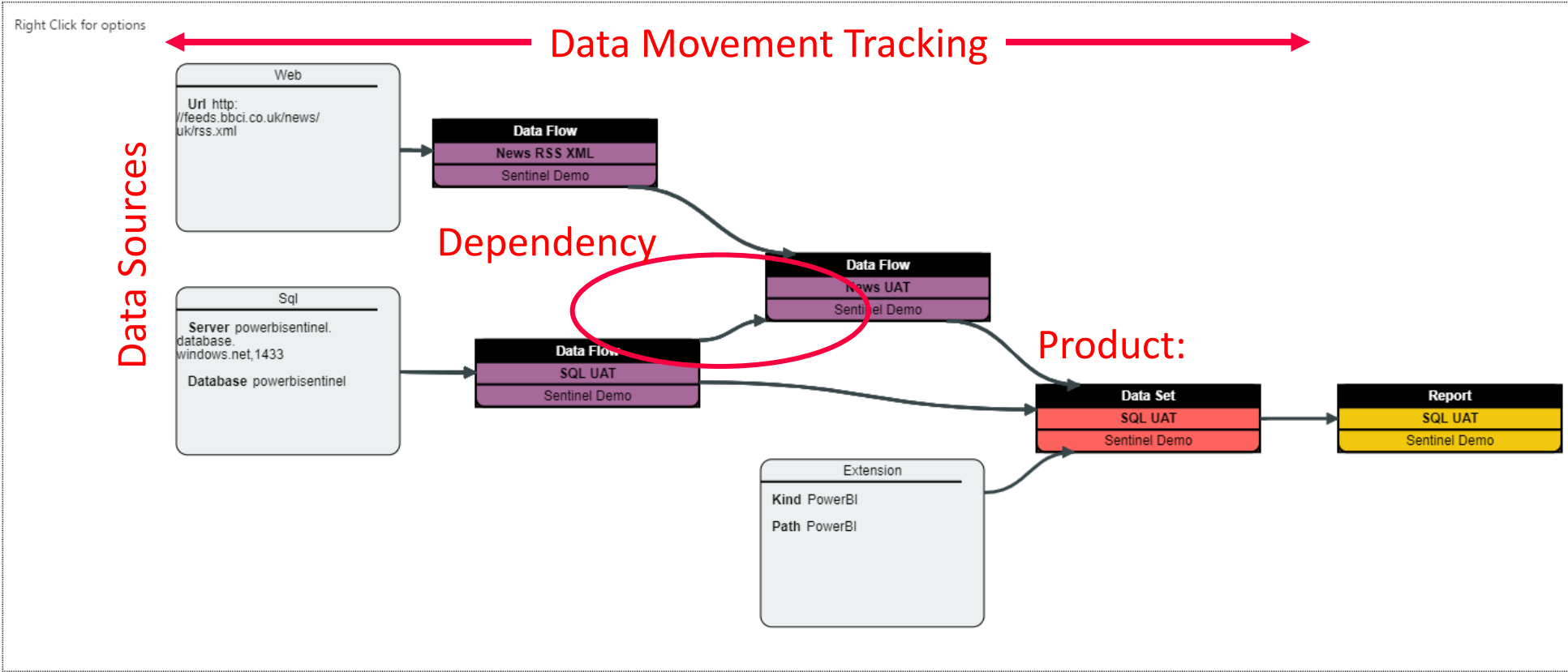
Product:

customer_orders

Dependency: stg_eltool_orders depends on warehouse.orders

Example: E2E Visible Data Lineage

- Home
- Configure
- Backup (history)
- Documentation
- Lineage Explorer
- Change Tracking
- Column Search
- Data Source Explorer
- Risk Assessment Report
- Subscription
- Contact Us
- FAQ
- Submit an Idea
- Sign out



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

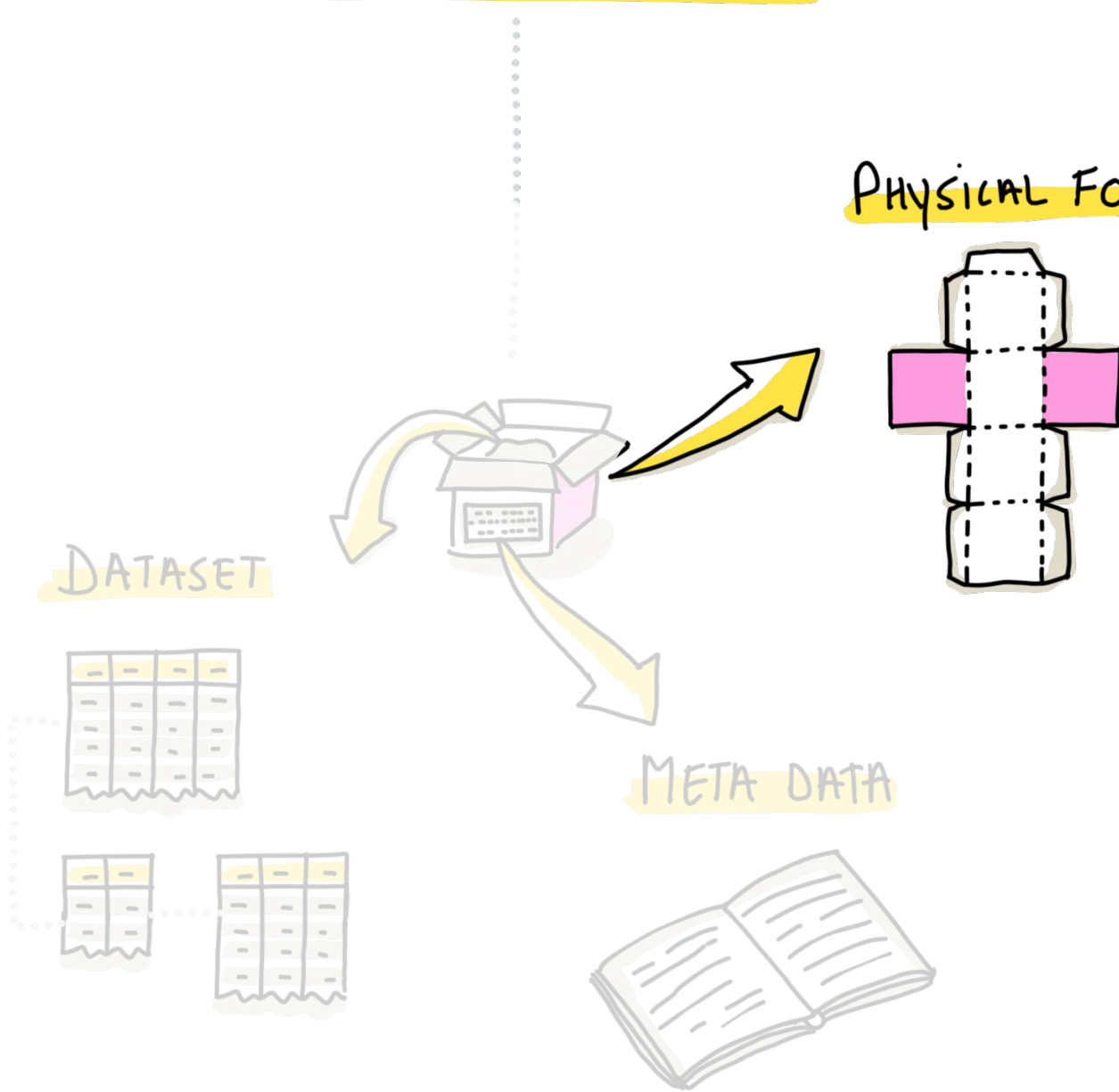
The screenshot displays a web application for a Business Glossary. The interface is divided into several sections:

- Left Sidebar:** Contains navigation icons for 'Assets', 'Glossary', and 'Insights'. The 'Glossary' section is active, showing a search bar and a list of glossaries including 'Aisle', 'Concepts', 'Consumer Product Goods', 'COVID-19', 'Example Glossary', 'Instacart', 'KPIs', and 'Metrics'. Under 'Metrics', 'Customer Acquisition Cost' is selected.
- Top Header:** Shows 'instadart' logo, 'Glossary' title, '+ New' button, a help icon, and the user 'Deakins'.
- Main Content Area:**
 - Header:** 'Customer Acquisition Cost' with a checkmark and a 'TERM' icon.
 - Overview:** A tabbed interface with 'Overview' selected and 'Linked Assets' as an alternative view.
 - Readme:** A section with an 'Edit' button. It contains:
 - Simple method:** A paragraph explaining the method, followed by the formula $CAC = \frac{MCC}{CA}$ and a list of definitions for CAC, MCC, and CA.
 - Complex method:** A paragraph explaining the method, followed by the formula $CAC = \frac{MCC + W + S + PS + O}{CA}$ and a list of definitions for CAC, MCC, W, S, and O.
- Right Panel:** A detailed view of the term with a red border around the top section. It includes:
 - Owners:** A list with a '+', a user icon, and the name 'chris'.
 - Classification:** A list with a '+', a shield icon, and the label 'Confidential'.
 - Certificate:** A green highlighted box containing a checkmark icon, the label 'Verified', and 'chris 3 months ago'.
 - Categories:** A text box stating 'This term does not belong to any category'.
 - Related Terms:** A list with a '+', a document icon, and the text 'Average Selling Price', 'Churn Rate', and 'Customer Lifetime Va...'.
 - Custom Metadata:** A list with a dropdown arrow, a user icon, and the text 'Great Expectations', 'PO number', and 'Airflow ETL Details'.
- Far Right Sidebar:** A vertical list of icons representing various data and resource types: 'Overview', 'Activity', 'Resources', 'Request', 'Property', 'Great ...', 'Data F...', 'PO number', 'Data Q...', 'Data C...', 'Airflo...', and 'Priority'.

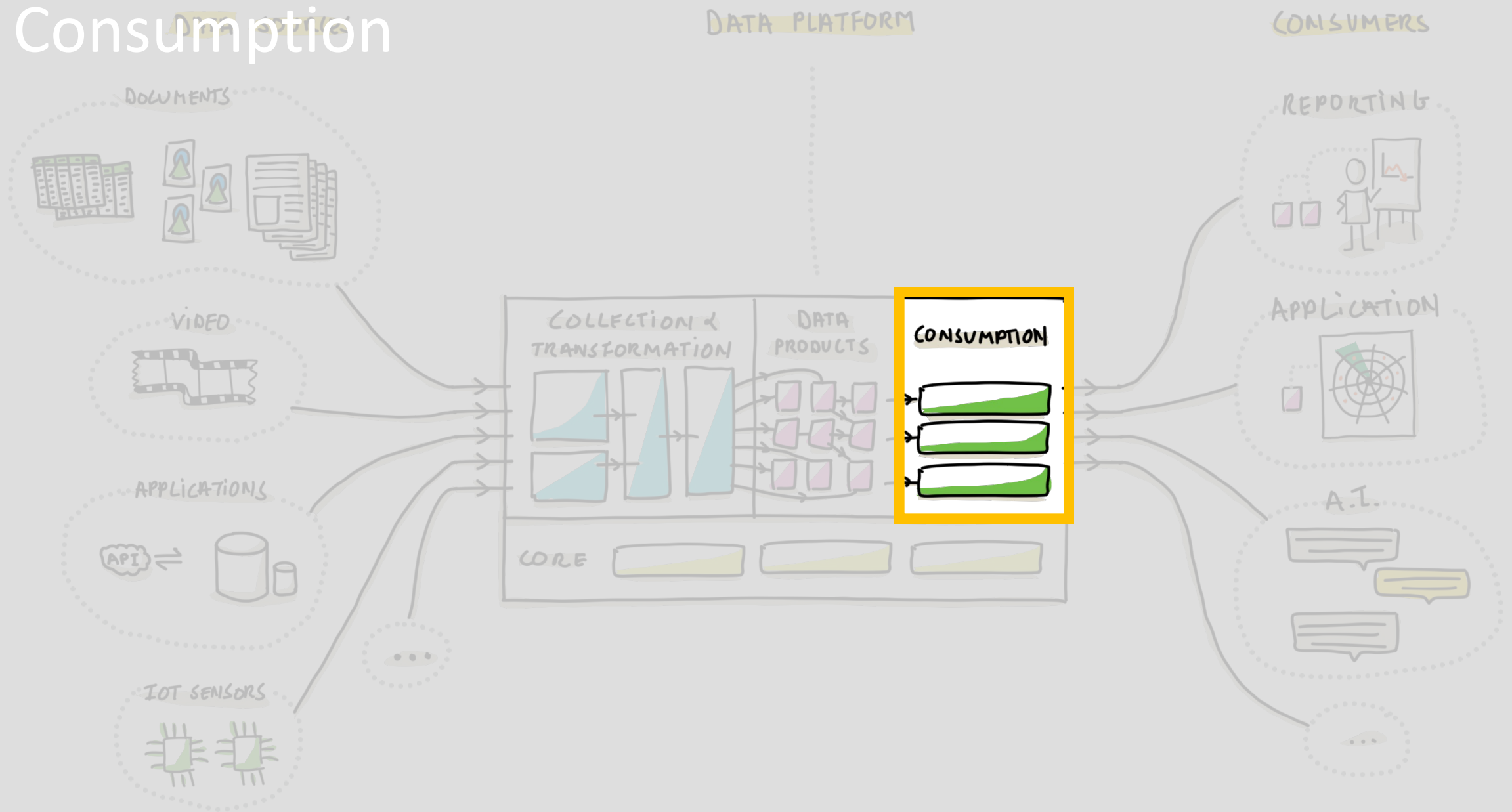


DATA PRODUCT

PHYSICAL FORMAT



3. Consumption



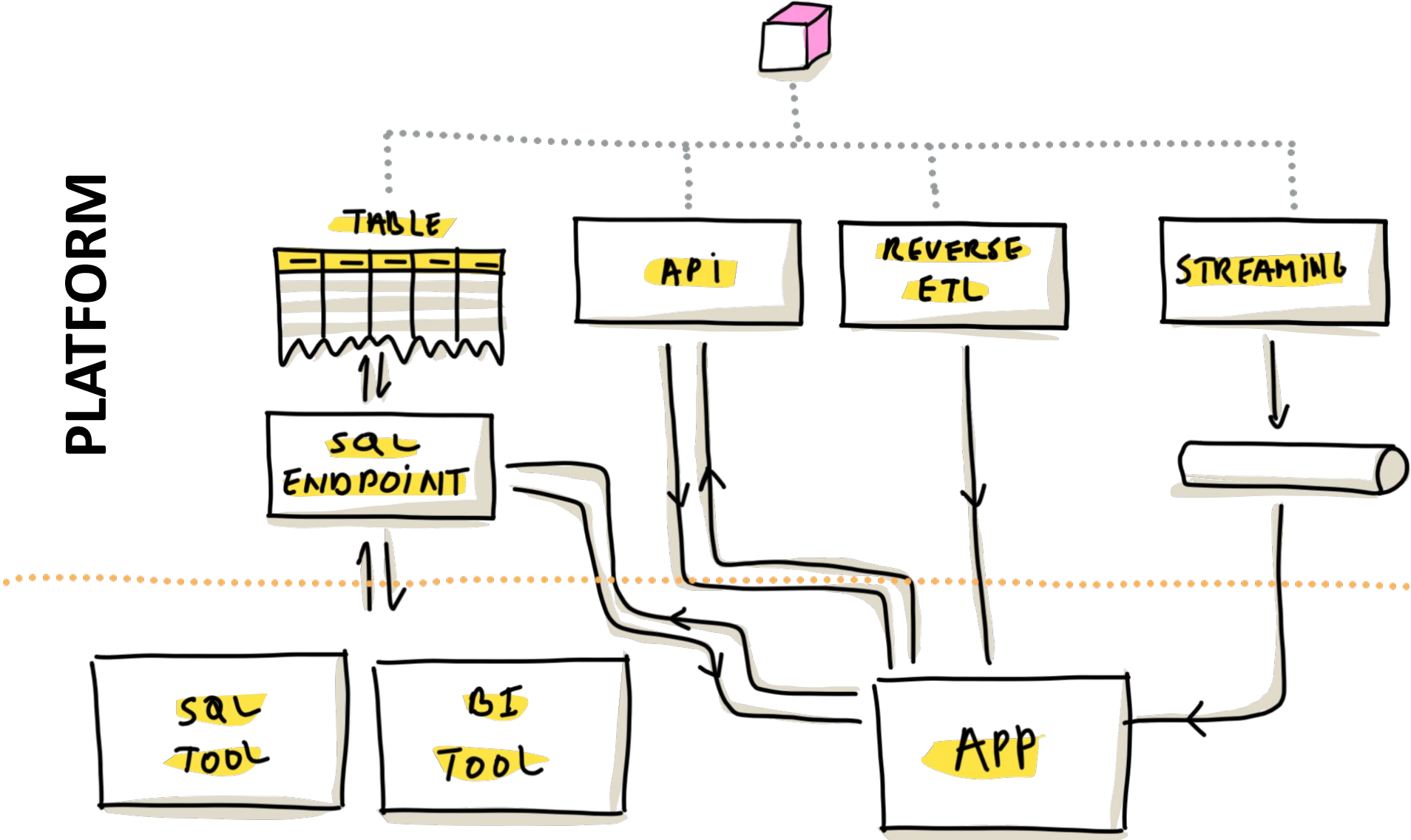
“Operational Plane”

“Analytical Plane”

“Operational / Analytical Plane”



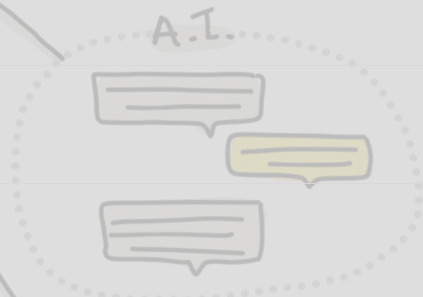
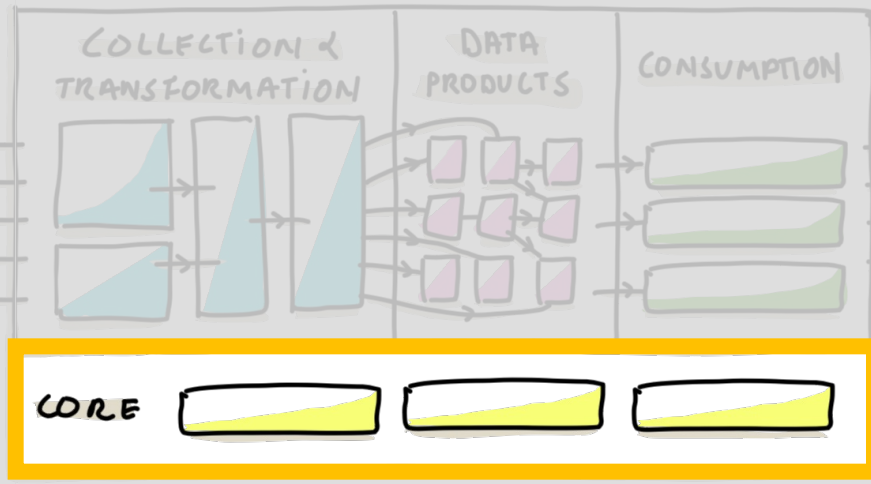
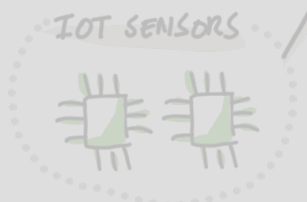
Common Consumption Patterns



4. Core DATA SOURCES

DATA PLATFORM

CONSUMERS



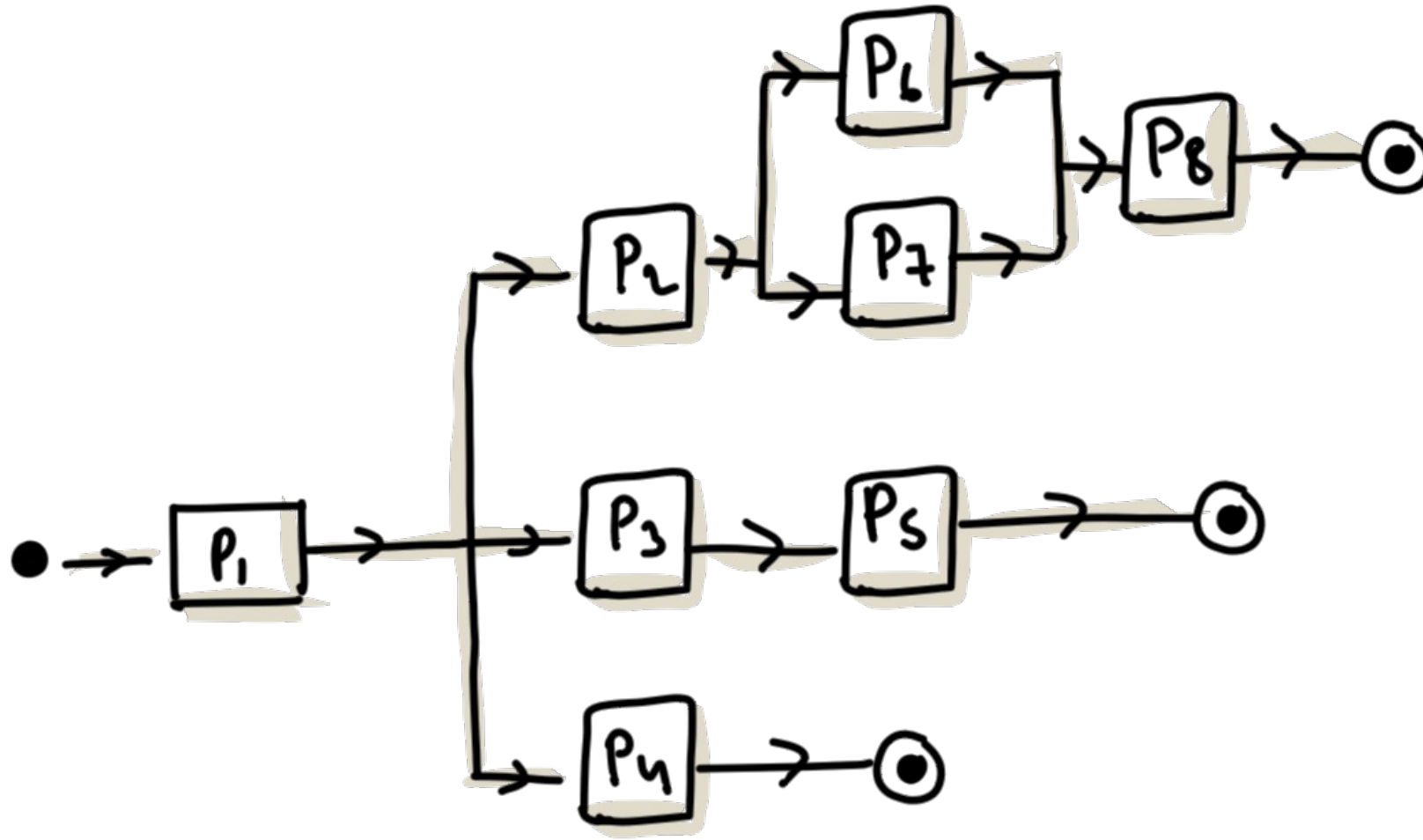
“Operational Plane”

“Analytical Plane”

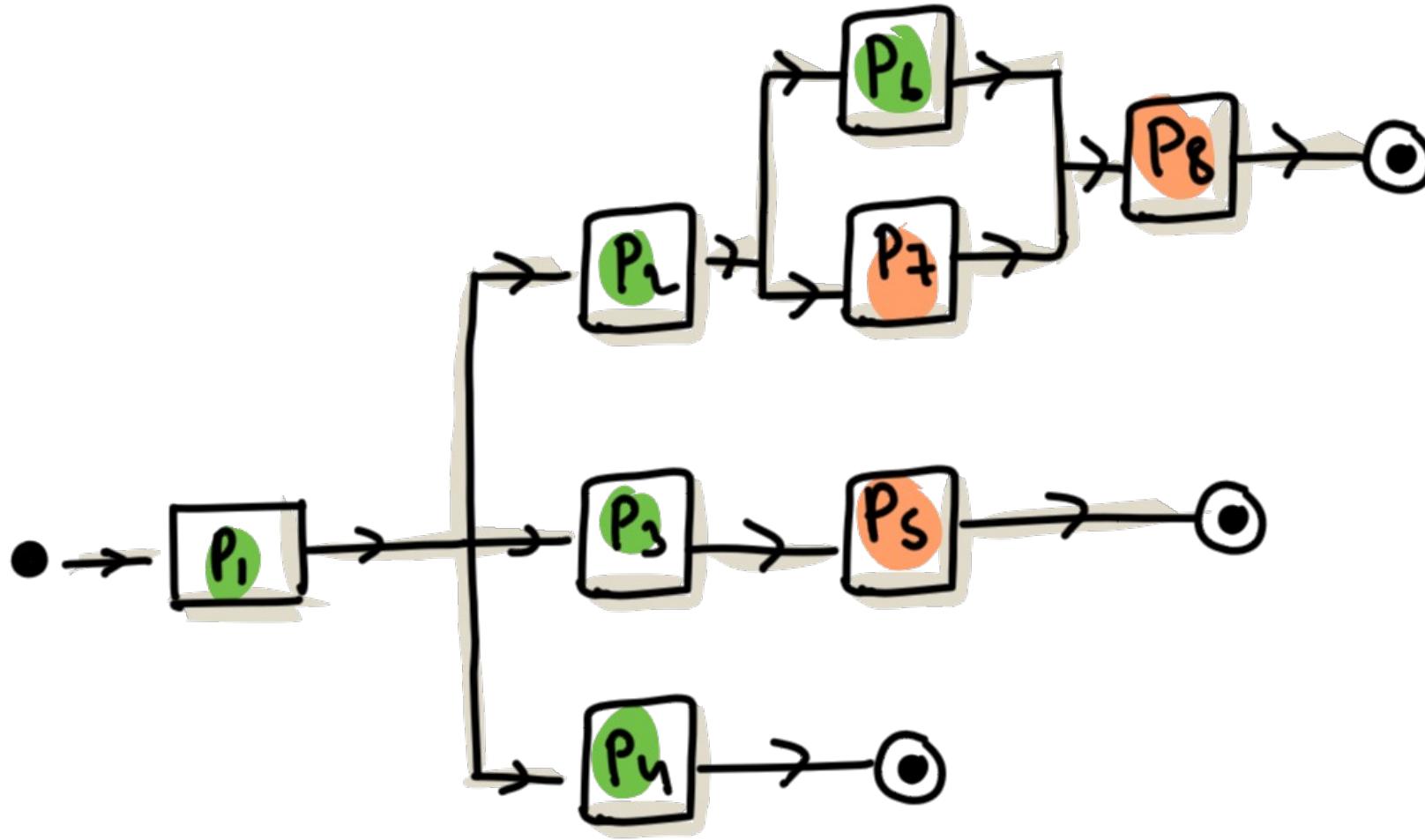
“Operational / Analytical Plane”



Orchestration



Orchestration



Example: Apache Airflow



DAG: example_task_group

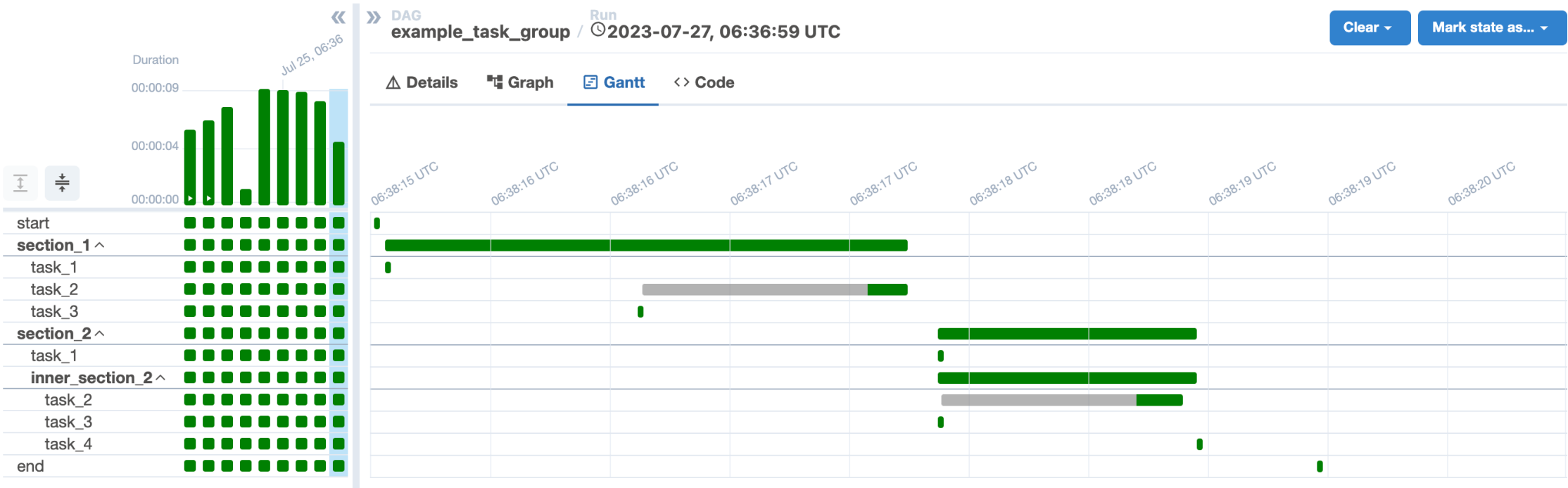
Schedule: 1 day, 0:00:00 Next Run: 2023-07-27, 06:36:59

Grid Graph Calendar Task Duration Task Tries Landing Times Gantt Details Code Audit Log

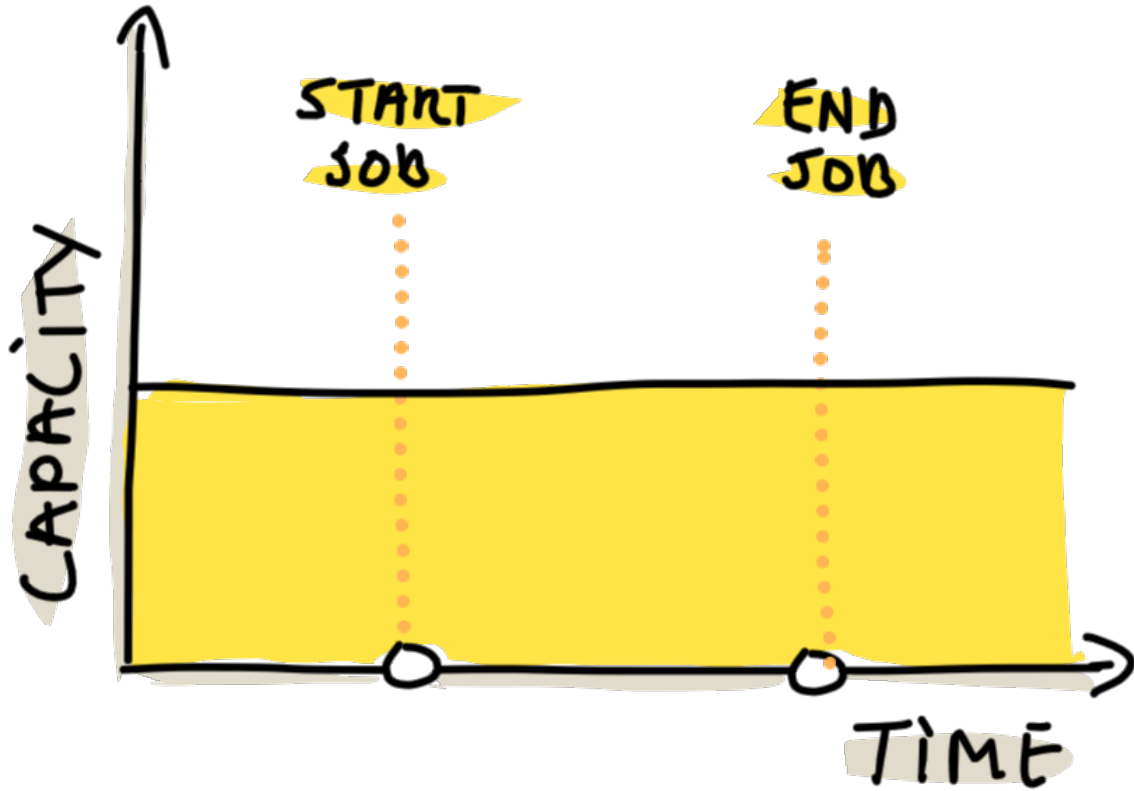
07/28/2023, 06:09:10 AM 25 All Run Types All Run States Clear Filters Auto-refresh

Press **shift** + **/** for Shortcuts

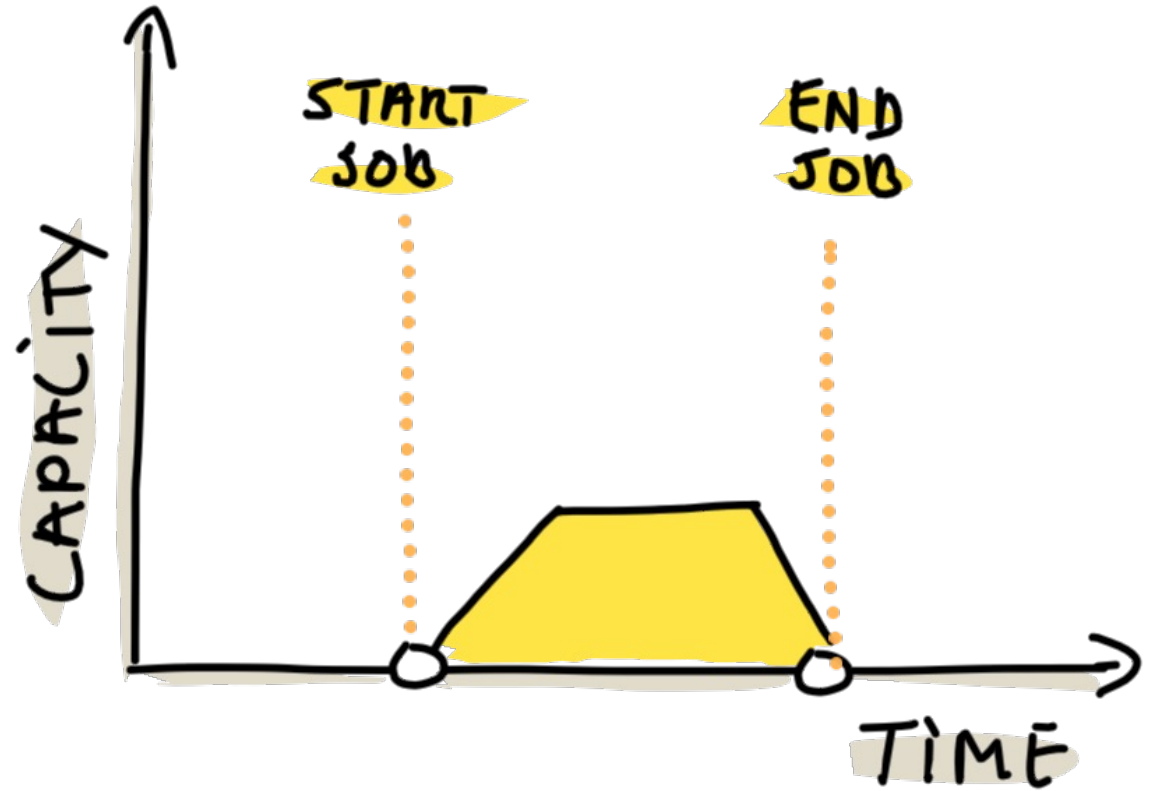
- deferred
- failed
- queued
- removed
- restarting
- running
- scheduled
- shutdown
- skipped
- success
- up_for_reschedule
- up_for_retry
- upstream_failed
- no_status



Infrastructure Management (Compute)

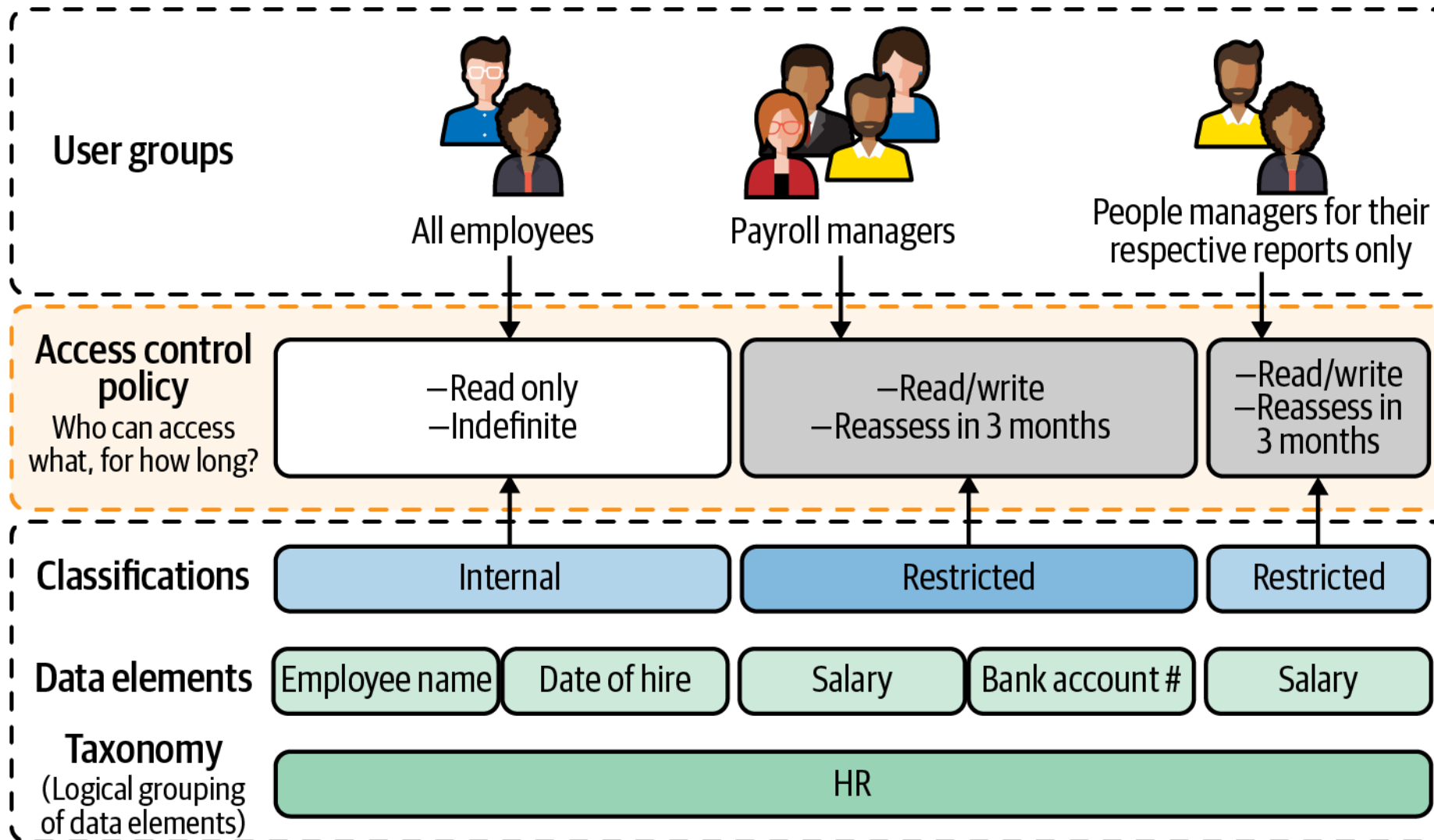


Static



Dynamic

Security & Access Control



Monitoring

Search anything (⌘ J)

TRIAL EXPIRES IN 13 DAYS

SCHEDULE CALL

Jerome Williamson
willi.jerome@green.com

PIPELINES

ACTIVATE

TRANSFORM

DESTINATIONS

DOCS

LIVE CHAT

v1.38

Overview

Transformation

Schema Mapper

Load Status

Activity Log

#475

mysql-source-new
MySQL · Ingests every 15 minutes

redshift-destination
Amazon Redshift · Loads every 15 minutes

ACTIVE PAUSE

Pipeline Activity

1h 12h 24h

Ingestion: 08.3K (421.64 epm)

Transformations: 07.1K (421.64 epm)

Schema Mapper: 04.2K (421.64 epm)

Load: 08.3K (421.64 epm)

Jobs

Events Ingested

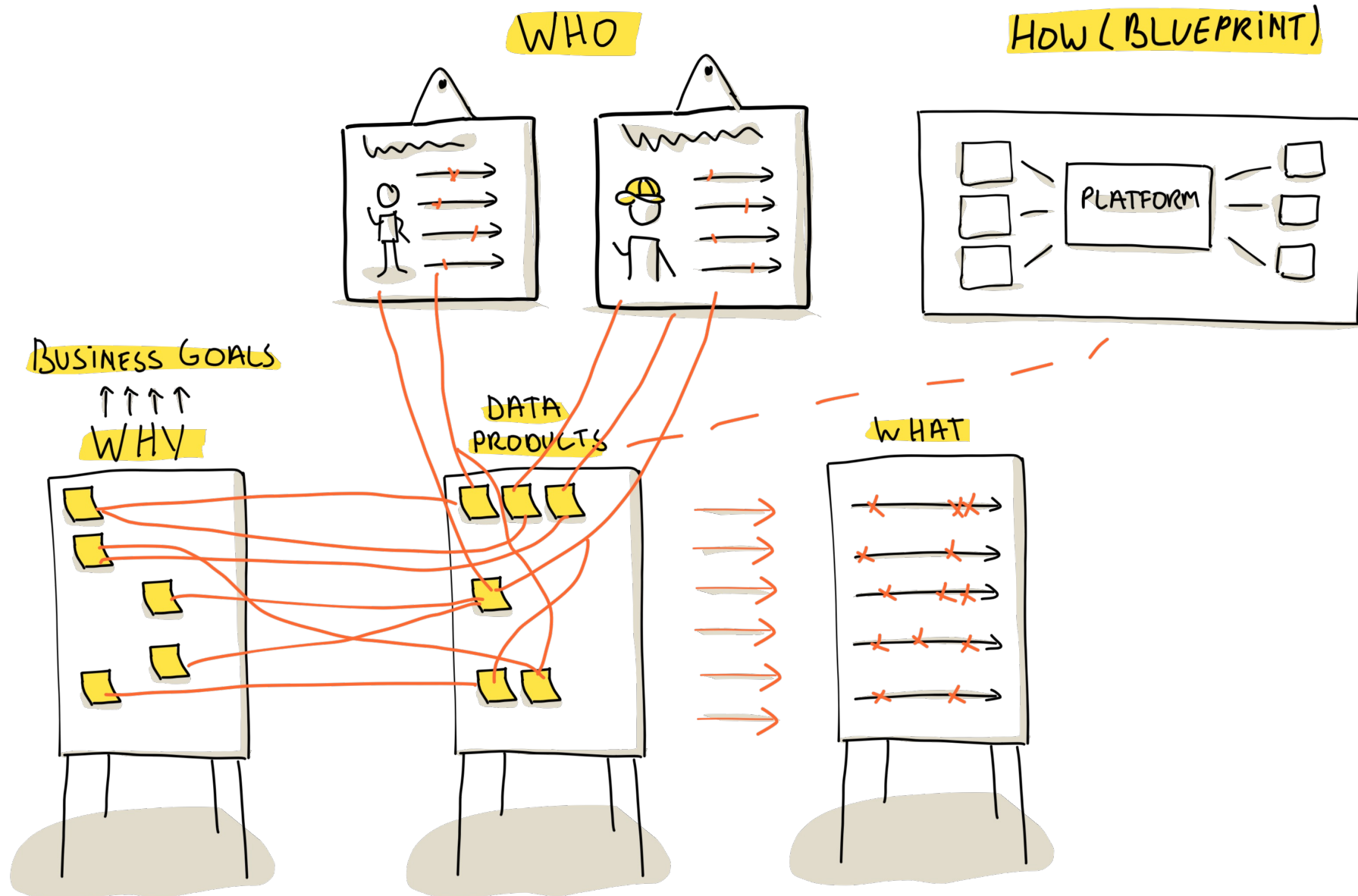
4.8M Events not Loaded

1-10 of 53

<input type="checkbox"/> object-name_new_updated Historical Load Running	0			QUEUED Not Synced Yet
<input type="checkbox"/> employee_records_updated Position: Sep 26, 2018 3:16:59 PM (UTC)	2.13M		1.28M Events not Loaded	PAUSED Last Synced: 6 Minutes Ago
<input type="checkbox"/> new_customer_data_generated Historical Load Running · Position: Sep 26, 2...	3.43M		3.28M Events not Loaded	ACTIVE Last Synced: 6 Minutes Ago
<input type="checkbox"/> food_categories_new Position: Sep 26, 2018 3:16:59 PM (UTC)	4.21M		18.29K Events not Loaded	FAILED Last Synced: 6 Minutes Ago
<input type="checkbox"/> register_file_loads Position: Sep 26, 2018 3:16:59 PM (UTC)	2.43M		1.38M Events not Loaded	ACTIVE Last Synced: 6 Minutes Ago
<input type="checkbox"/> new_customer_data_generated				ACTIVE



EXERCISE: HIGH LEVEL BLUEPRINT?



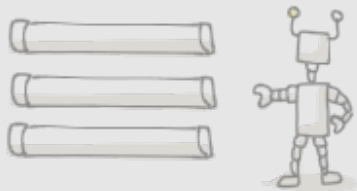
DATA PLATFORM



INFRASTRUCTURE



TECHNOLOGY



DATA GOVERNANCE

PEOPLE



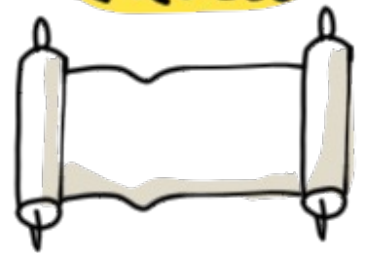
ORGANIZATION



PROCESSES



POLICIES, STANDARDS & RULES



DATA GOVERNANCE: **TABLE OF CONTENTS**

1. WHY

2. WHAT

A. POLICIES, STANDARDS & RULES

B. PROCESSES

C. PEOPLE

D. ORGANIZATION



DATA GOVERNANCE: **TABLE OF CONTENTS**

1. WHY

2. WHAT

A. POLICIES, STANDARDS & RULES

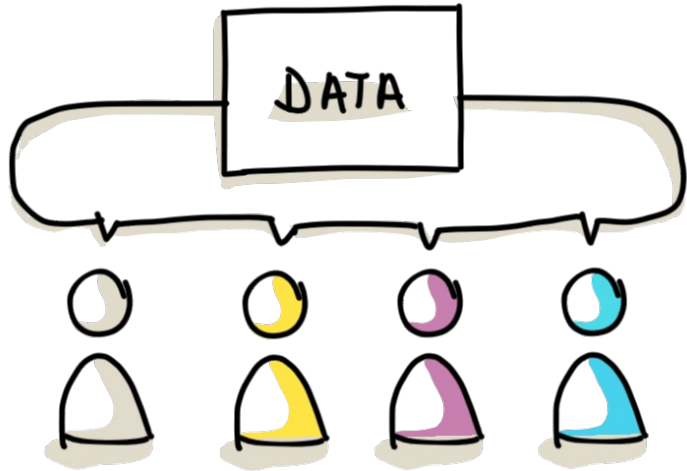
B. PROCESSES

C. PEOPLE

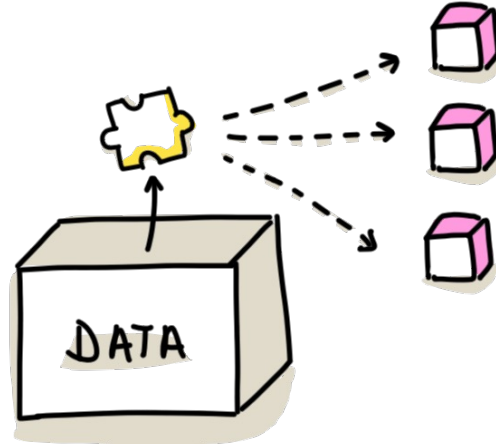
D. ORGANIZATION



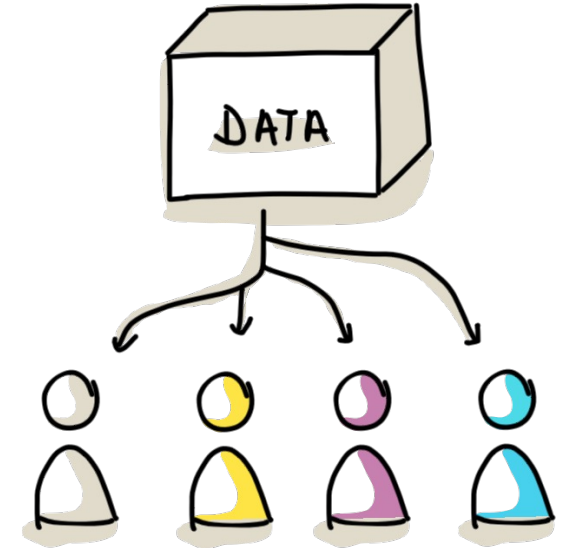
#1 One Common Data Language



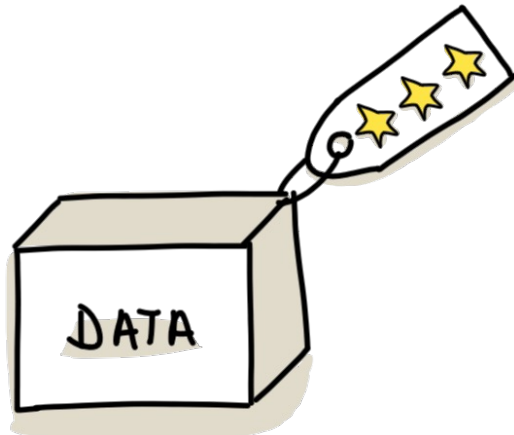
#2 Improved Data Reusability



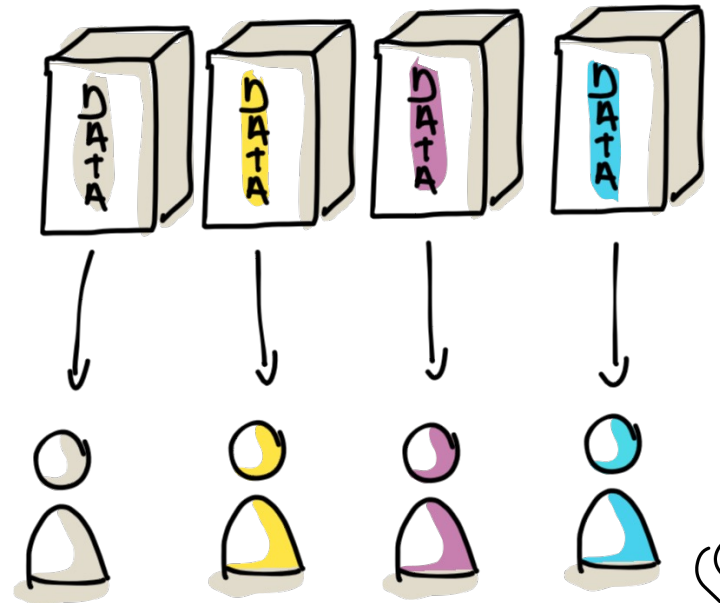
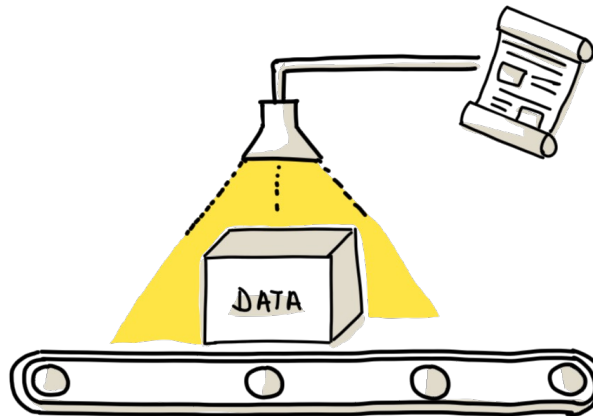
#3 Single Source of Truth



#4 Increased Confidence in Data



#5 Better Data Management Mechanisms



DATA GOVERNANCE: **TABLE OF CONTENTS**

1. WHY

2. WHAT

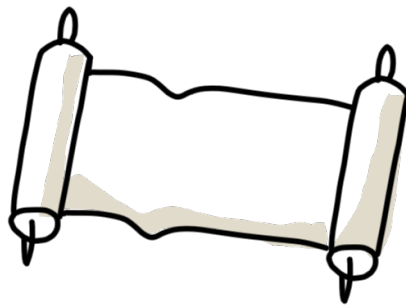
A. POLICIES, STANDARDS & RULES

B. PROCESSES

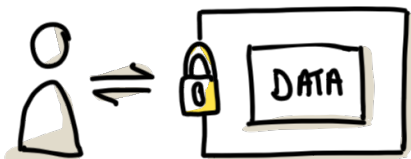
C. PEOPLE

D. ORGANIZATION





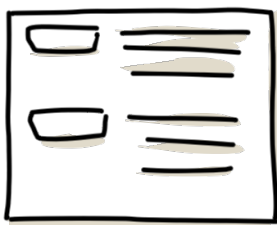
POLICIES, STANDARDS & RULES



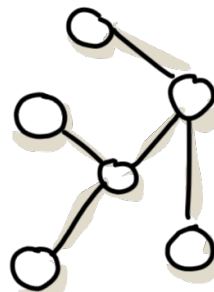
CLASSIFICATION & SECURITY



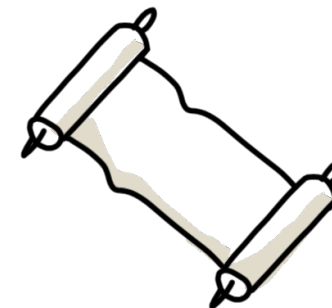
STORAGE ZONES



DEFINITIONS | META-DATA



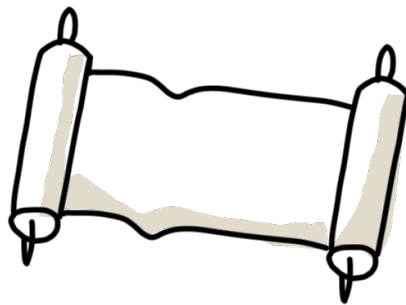
DATA MODEL



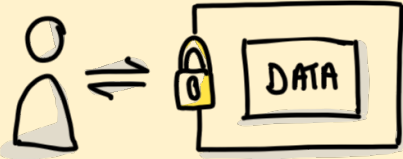
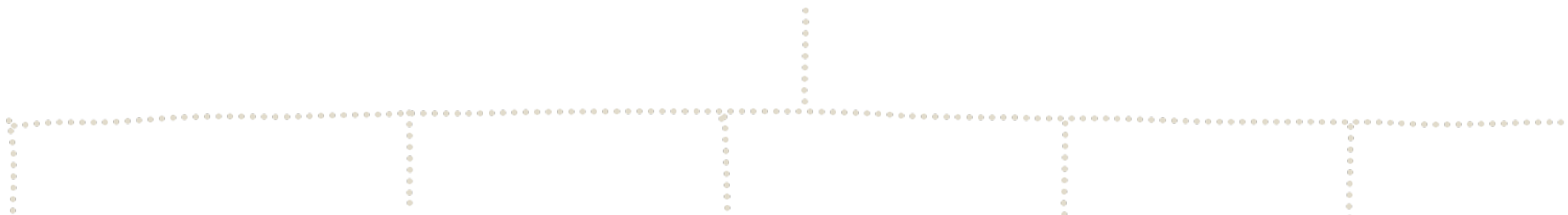
REGULATIONS

...






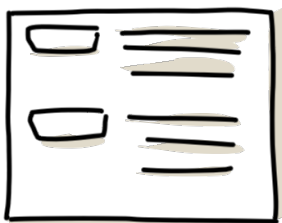
POLICIES, STANDARDS & RULES



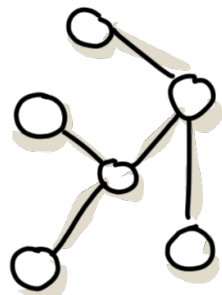
CLASSIFICATION & SECURITY



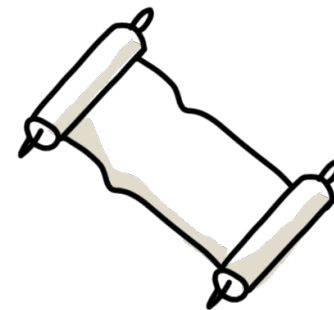
STORAGE | ZONES



DEFINITIONS | META-DATA



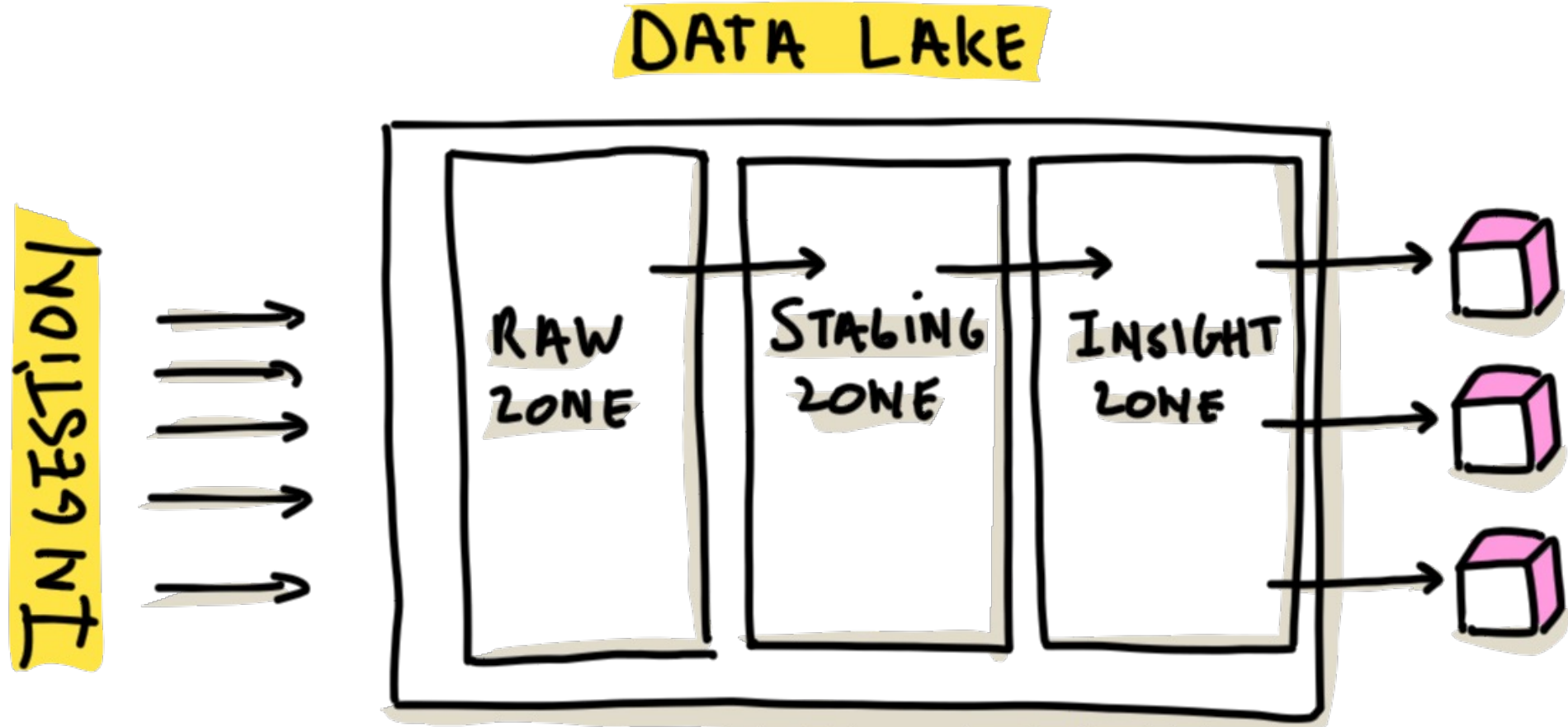
DATA MODEL



REGULATIONS



Example Data Policy



Example Data Policy

	RAW ZONE	STAGING ZONE	INSIGHT ZONE
TYPES OF DATA	<ul style="list-style-type: none">• Any kind of data including unstructured data• Examples: videos, tekst files, .csv files, ...	<ul style="list-style-type: none">• Known and structured data• Data from multiple sources is likely to be joined here• Data engineers prep and cleanse data	<ul style="list-style-type: none">• Known, enriched, integrated and cleaned data• Privacy controles like removing personal data
ACCESS	<ul style="list-style-type: none">• Very restricted access• Likely a handful of people or just an admin	<ul style="list-style-type: none">• More access• Mostly data engineers	<ul style="list-style-type: none">• Highest level of access• Most if not all data analysts/scientists



Example Data Policy

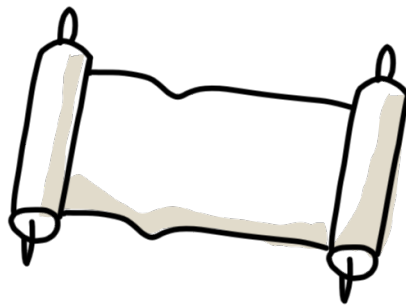
Raw zone	Staging zone	Insight zone
Plain text customer name	Hashed customer name	Redacted customer name
Anderson, Dan	Anderson, #####	*****
Buchanan, Cynthia	Buchanan, #####	*****
Drexel, Frieda	Drexel, #####	*****
Harris, Javiar	Harris, #####	*****

Plain text = raw data

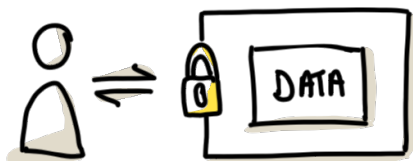
Hashed = Fixed length string of letters / numbers

Redacted = obscured information that's personally identifiable, confidential or sensitive.





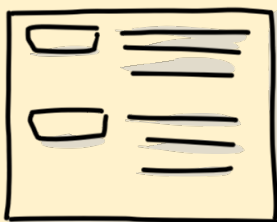
POICIES, STANDARDS & RULES



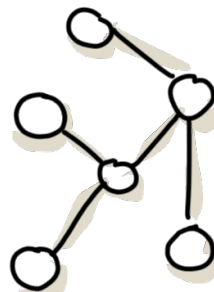
CLASSIFICATION & SECURITY



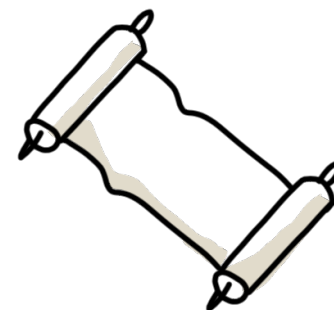
STORAGE | ZONES



DEFINITIONS | META-DATA



DATA MODEL



REGULATIONS



“Every data element that is exposed to end-users should be defined and documented in the Data Dictionary Meta-Data”

Column	Description	SQL Server Datatype	Oracle Datatype	DB2 Datatype	Teradata Datatype	Sybase Datatype	MySQL Datatype
DAY_ID	Day the action was started.	DATE	TIMESTAMP	DATE	DATE	DATE	DATE
HOUR_ID	Hour the action was started.	TINYINT	NUMBER(3)	SMALLINT	BYTEINT	TINYINT	TINYINT
MINUTE_ID	Minute the action was started.	SMALLINT	NUMBER(5)	SMALLINT	SMALLINT	SMALLINT	SMALLINT
SERVERID	GUID of the Intelligence Server processing the request.	CHAR(32)	CHAR(32)	CHAR(32)	CHAR(32)	CHAR(32)	CHAR(32)
SERVERMACHINE	Name of the Intelligence Server processing the request.	VARCHAR(255)	VARCHAR2(255)	VARCHAR(255)	VARCHAR(255)	VARCHAR(255)	VARCHAR(255)



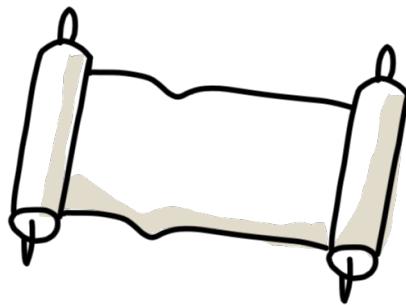
“Every report should have an ‘about section’
describing the necessary meta-data”

The screenshot shows a web application interface for a report titled 'Cruise general'. The left sidebar contains a menu with 'Pagina's' and several options: 'Cruise PAX', 'Cruise vessels', 'Cruise frequency', and 'ABOUT' (which is highlighted). The main content area displays the 'ABOUT' section with the following metadata:

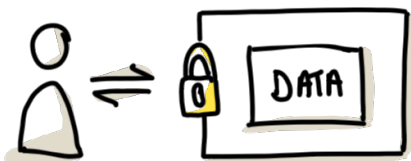
- Doel:** Zicht geven op evolutie cruise passagiers en aanlopen
- Auteur:** [Redacted]
- Owner rapport:** [Redacted]
- Brondata:** BeheersControleDWH
- Naam PBI dataset:** MasterDS_Tab_Scheepvaart
- Updatefrequentie:** dagelijks

The interface also includes a search bar at the top right and a toolbar with options like 'Bestand', 'Exporteren', 'Delen', 'Chatten in Teams', 'Deze gegevens verkennen', and 'Inzichten'.





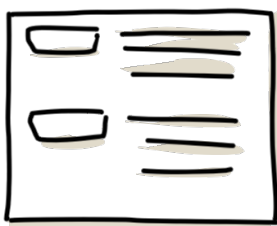
POLICIES, STANDARDS & RULES



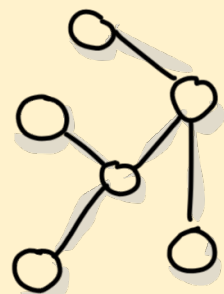
CLASSIFICATION & SECURITY



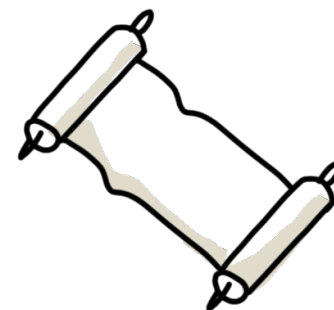
STORAGE | ZONES



DEFINITIONS | META-DATA



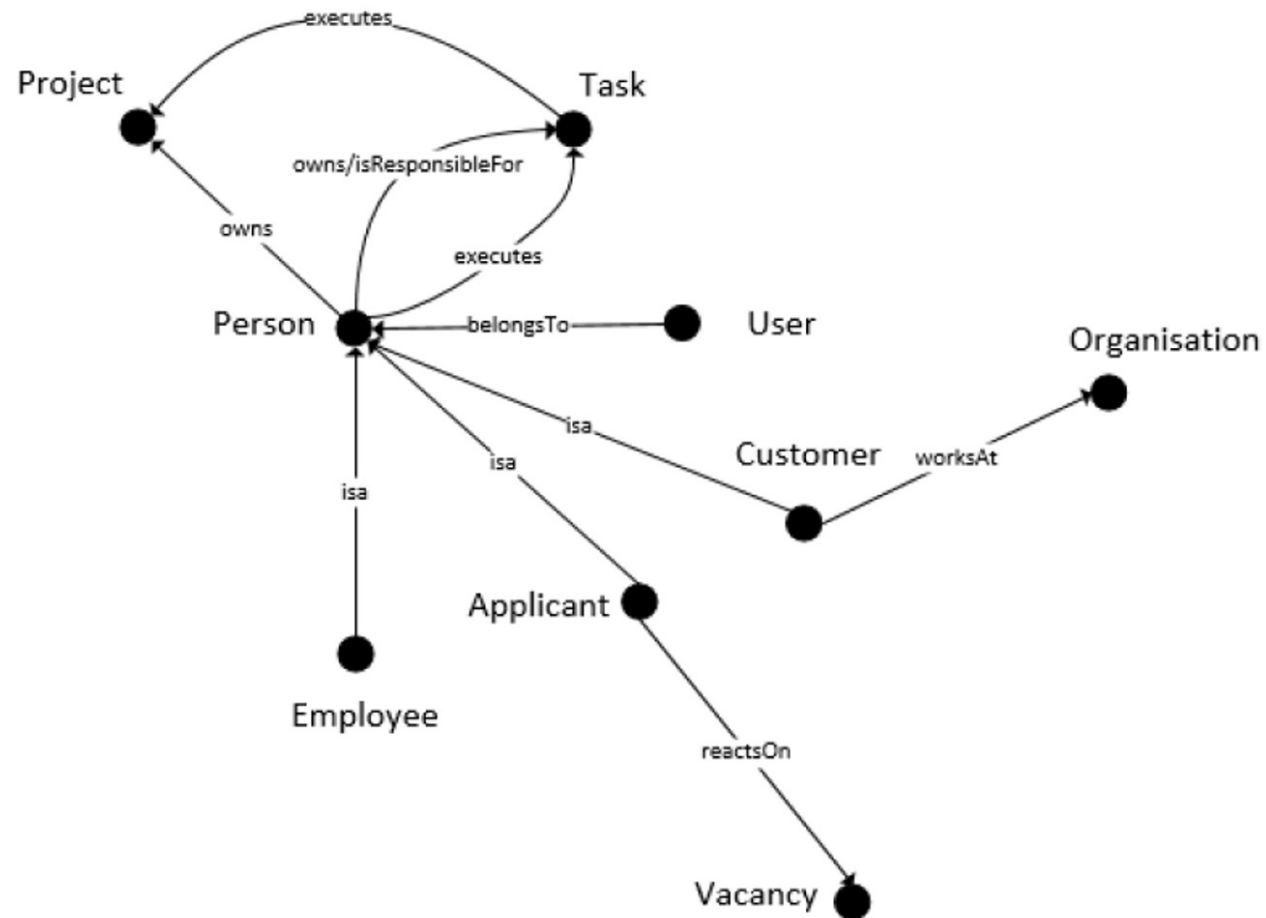
DATA MODEL

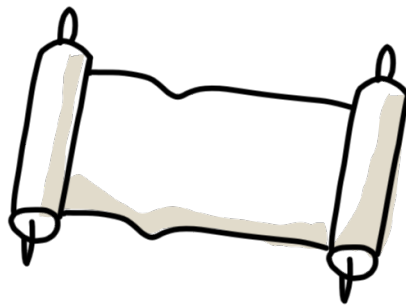


REGULATIONS

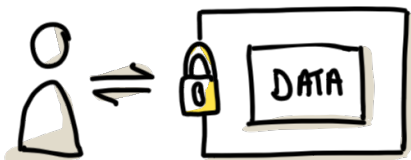


“Every data element that is exposed to end-users should be part of the overall company semantic data model”





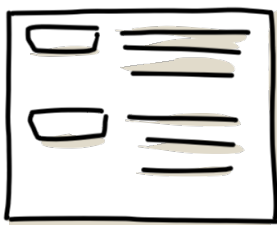
POLICIES, STANDARDS & RULES



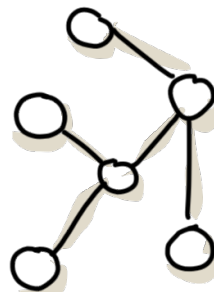
CLASSIFICATION & SECURITY



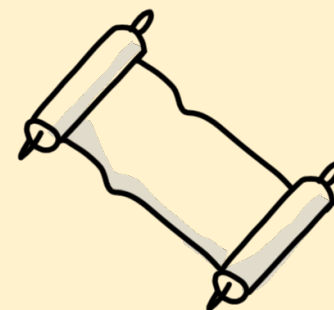
STORAGE ZONES



DEFINITIONS | META-DATA



DATA MODEL



REGULATIONS



Bigger Responsibility, Bigger Repercussions

Fines of up to 4% of turnover
Organizations in breach of GDPR can be fined up to 4% of annual global turnover or €20 Million.



Breach notification within 72 hrs
Breaches must be reported within 72 hours of first having become aware of the breach.



Increased territorial scope
Applies to any company processing personal data of EU citizens, regardless of location.



Privacy by design
Data protection from the onset of the designing of systems, rather than a retrospective addition.



Consent matters
Explicit consent must be provided in an intelligible and easily accessible form.



Right to be forgotten
Entitles the data subject to have the data controller erase his/ her personal data (and potentially third parties, too).



Right to access and portability
Users can inquire whether and how their personal data is being processed.



Mandatory data protection officers
Appointed in certain cases, to facilitate the company's need to demonstrate GDPR compliance.



Bigger Responsibility, Bigger Repercussions

Fines of up to 4% of turnover
Organizations in breach of GDPR can be fined up to 4% of annual global turnover or €20 Million.



Breach notification within 72 hrs
Breaches must be reported within 72 hours of first having become aware of the breach.



Increased territorial scope
Applies to any company processing personal data of EU citizens, regardless of location.



Consent matters
Explicit consent must be provided in an intelligible and easily accessible form.



Right to access and portability
Users can inquire whether and how their personal data is being processed.



Privacy by design
Data protection from the onset of the designing of systems, rather than a retrospective addition.








Right to be forgotten
Entitles the data subject to have the data controller erase his/ her personal data (and potentially third parties, too).



Mandatory data protection officers
Appointed in certain cases, to facilitate the company's need to demonstrate GDPR compliance.



CSRD: A Unified Reporting Framework

	Framework	Objectives / Disclosures	Audience
	Global Reporting Initiative (GRI)	Helps companies to report on issues related to economic, environmental, and social impacts of business operations. Primarily used to inform stakeholders on current sustainability performance.	<ul style="list-style-type: none"> All stakeholders Industry specific guidance (E.g., Oil & Gas, Coal and Agriculture)
	Sustainability Accounting Standards Board (SASB)	Industry-specific standards that help companies identify topics that are likely to impact their financial or operational performance. These are built around 5 dimensions; environment, social capital, human capital, business model & innovation, and leadership & governance.	<ul style="list-style-type: none"> Investors Industry specific guidance
	Climate Disclosure Standards Board (CDSB)	A framework for reporting on climate and natural capital information.	<ul style="list-style-type: none"> All stakeholders
	Carbon Disclosure Project (CDP)	Disclosure through questionnaires, not standards, of a company's environmental impact. It is focused on climate change, water security, forests, and supply chain topics, which allows to benchmark against peers.	<ul style="list-style-type: none"> All stakeholders Industry specific guidance
	International Integrated Reporting Council (IIRC)	Provides guidance for integrated reporting. This allows companies to combine financial data and ESG data, for better performance data & analysis.	<ul style="list-style-type: none"> Investors No Industry – specific guidance



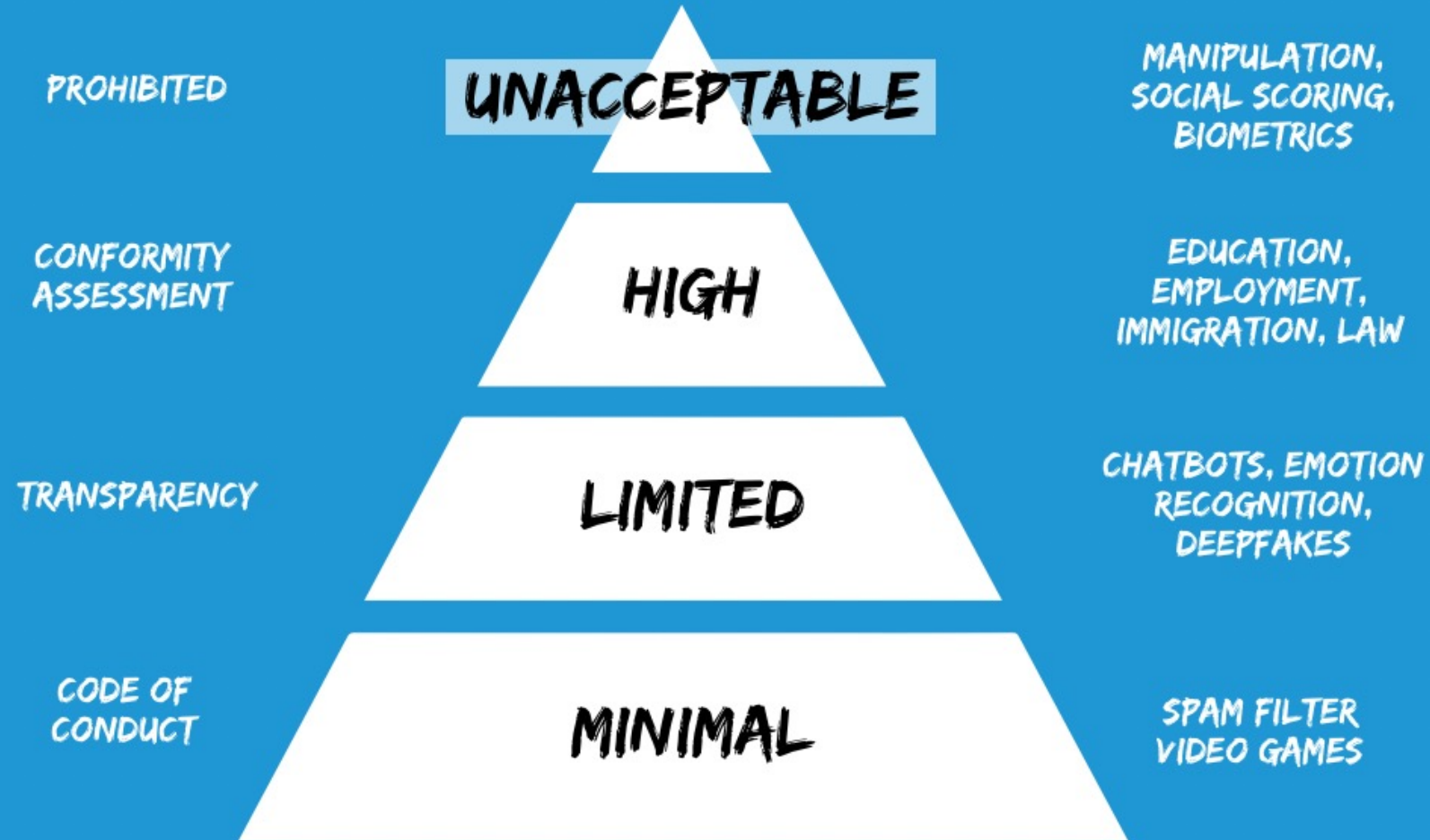
Institution	Objectives / Disclosures
European Financial Reporting Advisory Group (EFRAG)	Under CSRD, EFRAG will standardize NFRD compliant sustainability reporting standards.



CSRD Compliant



THE 4 TYPES OF RISKS IN THE **E.U. AI ACT**



DATA GOVERNANCE: **TABLE OF CONTENTS**

1. WHY

2. WHAT

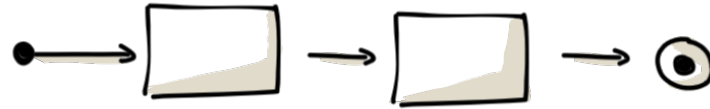
A. POLICIES, STANDARDS & RULES

B. PROCESSES

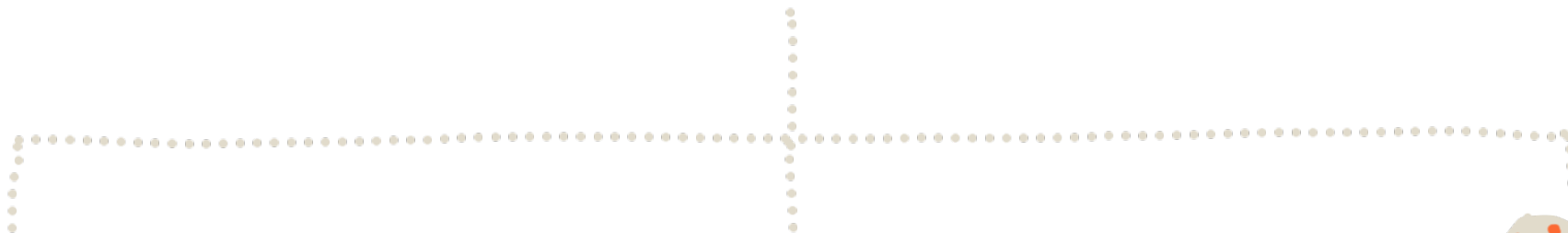
C. PEOPLE

D. ORGANIZATION





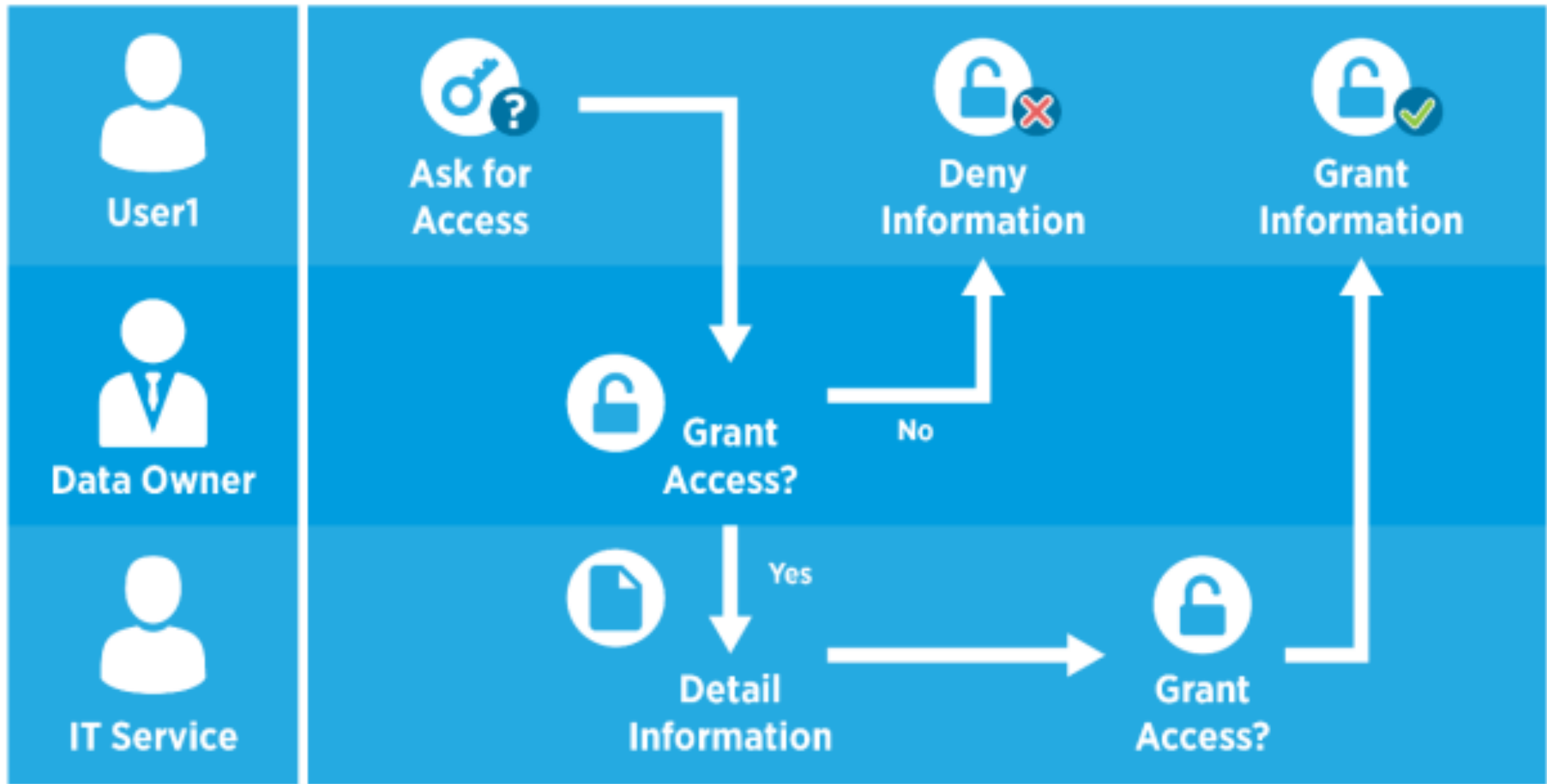
PROCESSES



GETTING DATA ACCESS

HOW TO REQUEST NEW FEATURES?

DATA QUALITY



Step 3

What are the data objects you need access to?

Data object	Permissions
SALES Schema	Read X
PURCHASING Schema	Read X



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

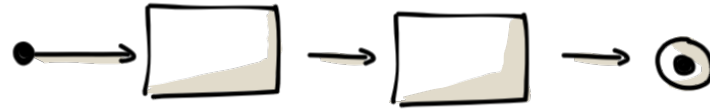


FINANCE
16 data objects

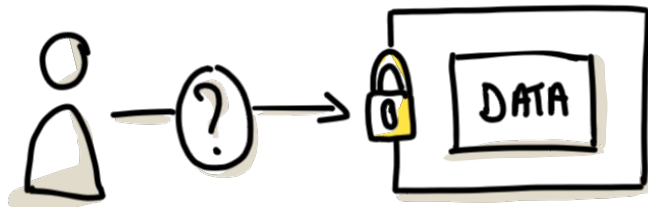


Jin Doe
jin@raito.io

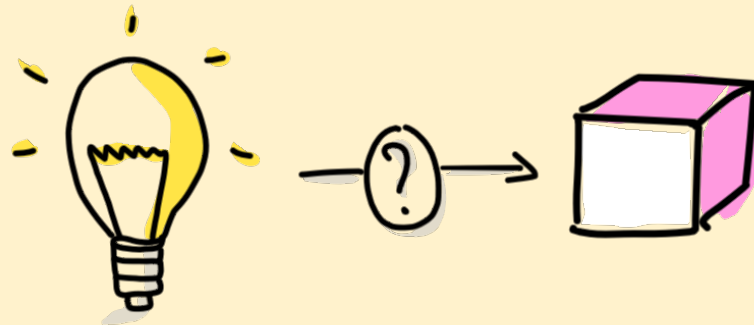
Implement



PROCESSES



GETTING DATA ACCESS

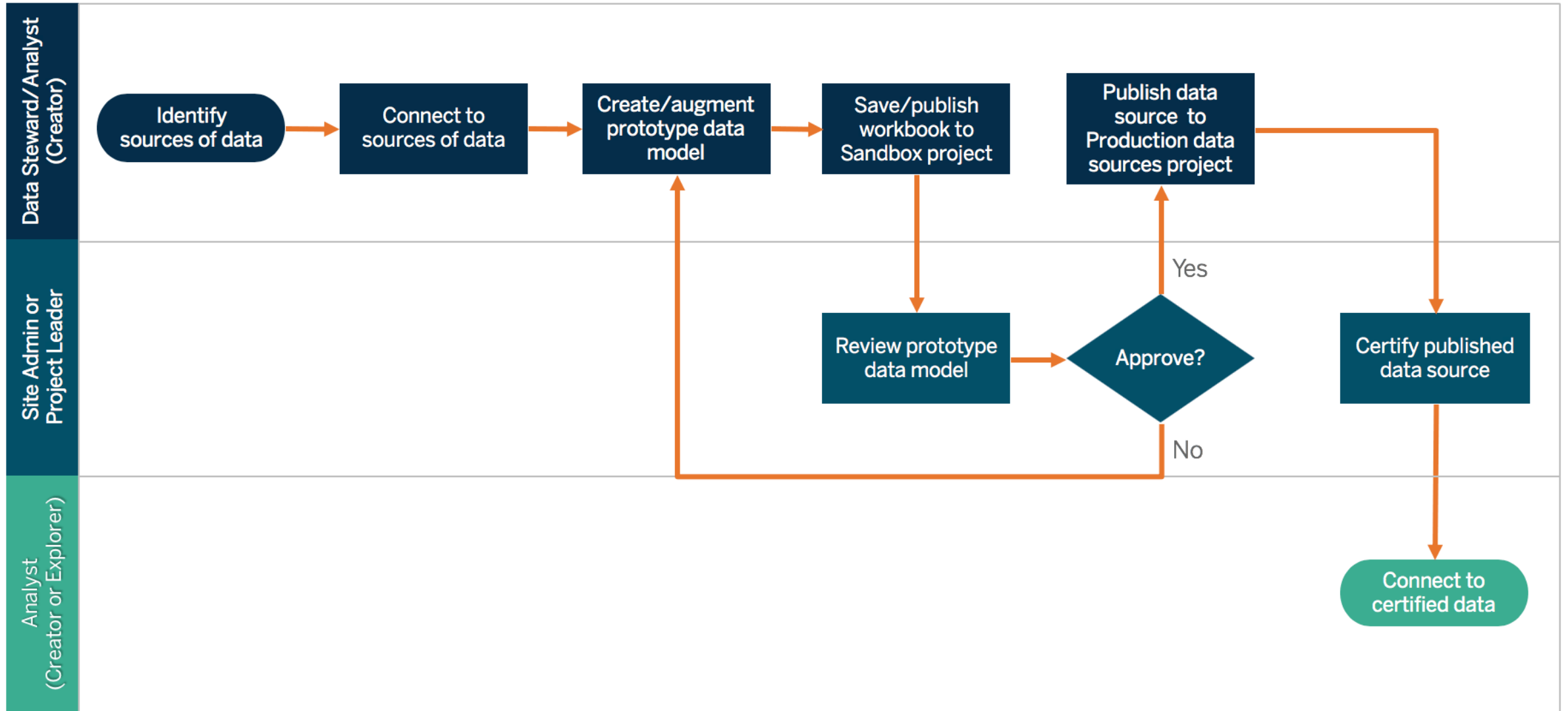


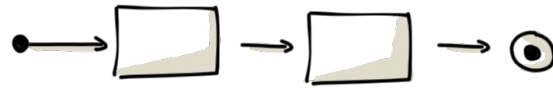
HOW TO REQUEST NEW FEATURES?



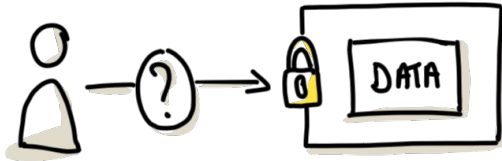
DATA QUALITY

Self Service Process

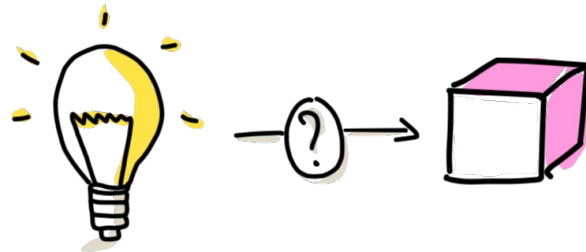




PROCESSES



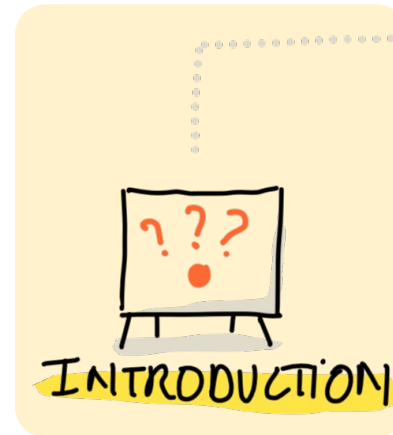
GETTING DATA ACCESS



HOW TO REQUEST NEW FEATURES?



DATA QUALITY



INTRODUCTION



PROLESS



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-G. Lewis		NC	www.burlington.com		
6	Pyramid Construction Inc.	427-4427	2 Place Juss			www.pyramid.com		
7	Dickenson plc	785-241-6200	1301 Hoch l		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 & Gas	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								

Annotations:

- Not Standardized**: Points to the WEBSITE column for rows 2, 3, and 4.
- Not Complete**: Points to the BILLINGCITY column for row 5.
- Not Vaild**: Points to the PHONE column for row 12.
- Not Consistent**: Points to the BILLINGSTATE column for rows 11 and 12.



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.

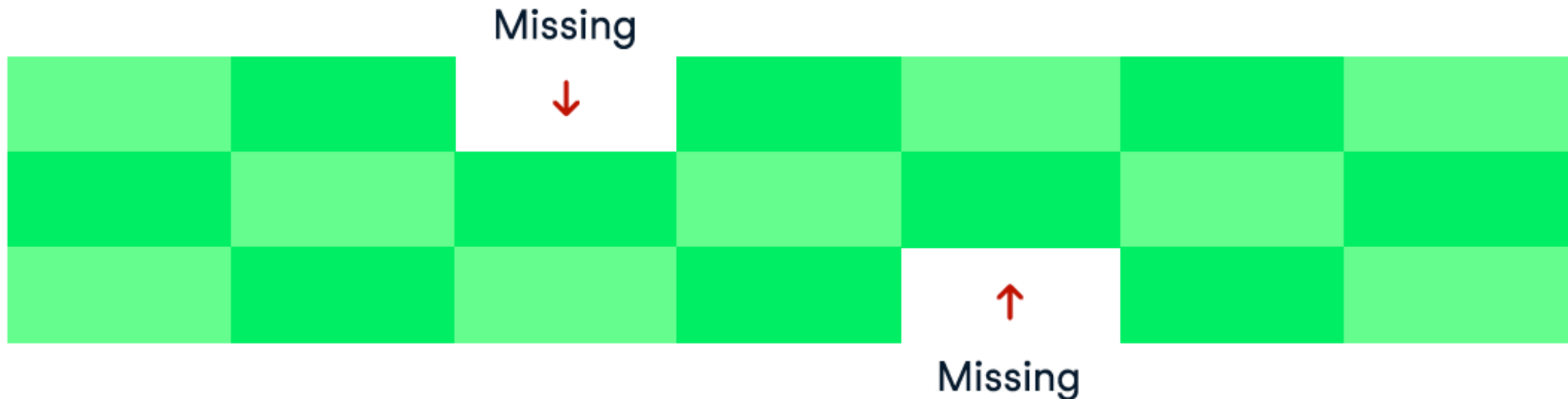
Completeness – Validity – Accuracy – Uniqueness
Timeliness – Consistency





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.



Completeness Example

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

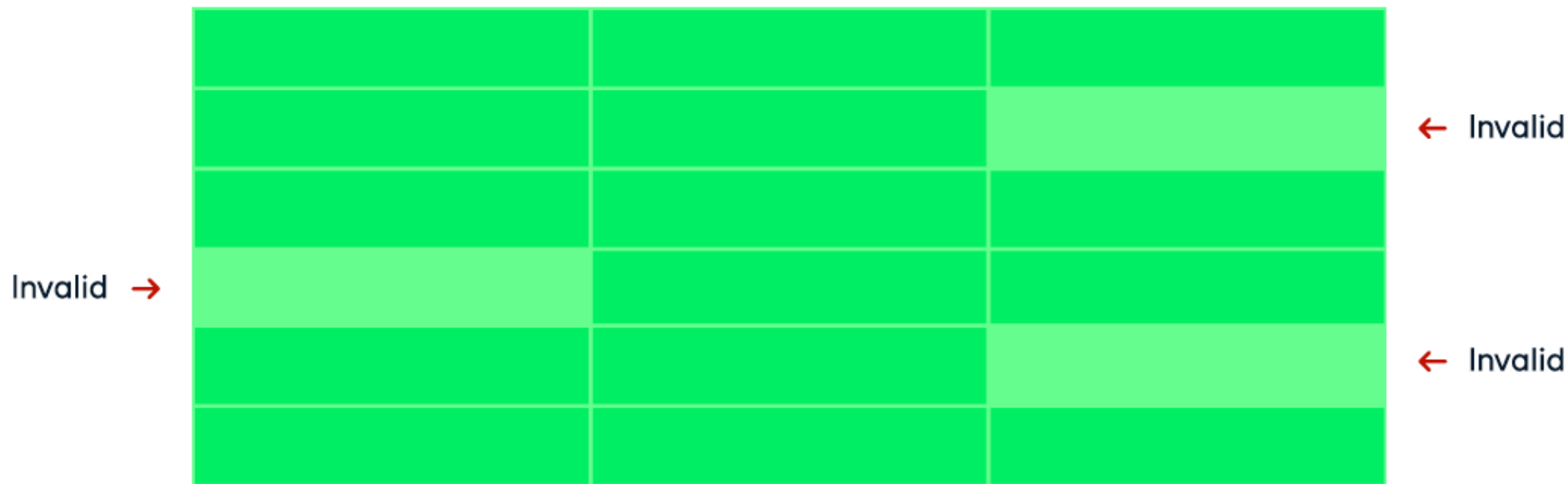
CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert 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





Validity

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



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





Accuracy

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

Verified Source Document

Orange	Orange	Orange
Green	Green	Green
Blue	Blue	Blue
Purple	Purple	Purple

Downstream Table

Orange	Orange	Orange
Green	Green X	Green
Blue	Blue	Blue
Purple	Purple	Purple



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.

Tax Form

Name: Ava Shiffer Birthdate: 10/30/1990

Address: 910 Quality St

City: Washington State: DC

Zip: 20008



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



2032 SW 35th Street



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.



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





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



Timeliness Example

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

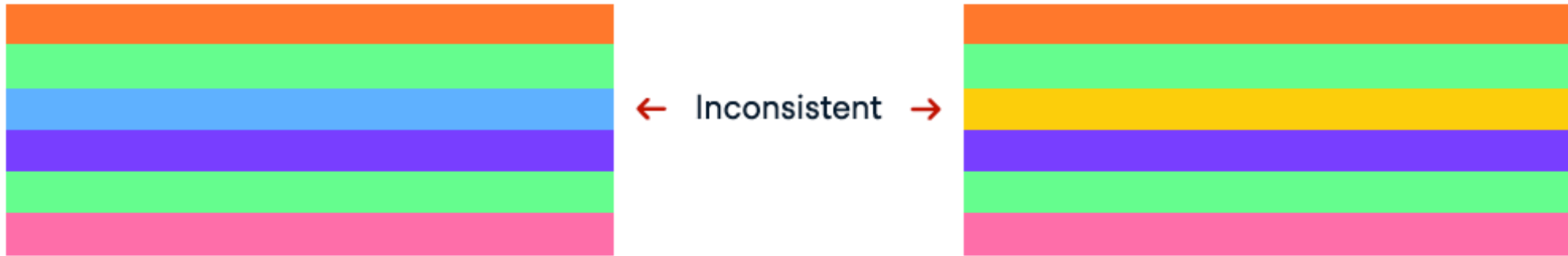


CustomerID	CustomerName
100000192	01-01-2023 11:07 am
100000198	01-01-2023 11:07 am
100000120	01-01-2023 11:07 am



Consistency

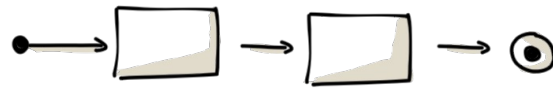
Consistency is a data quality dimension that measures the degree to which data is the same across all instances of the data. Consistency can be measured by setting a threshold for how much difference there can be between two datasets.



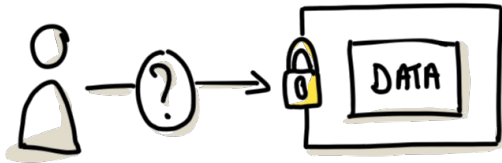
Consistency Example

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

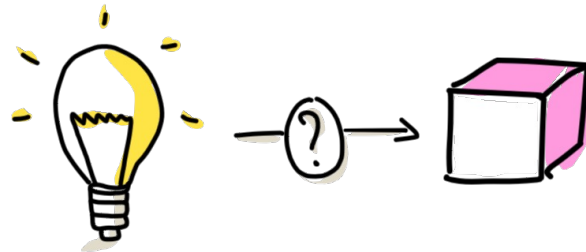




PROCESSES



GETTING DATA ACCESS



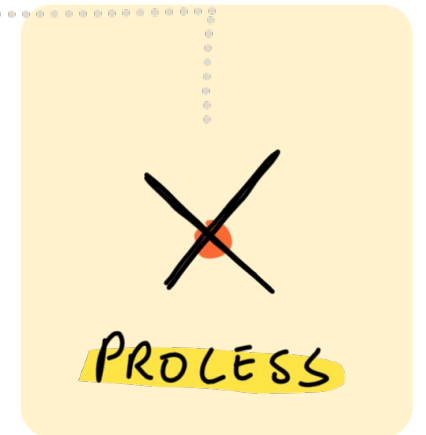
HOW TO REQUEST NEW FEATURES?



DATA QUALITY



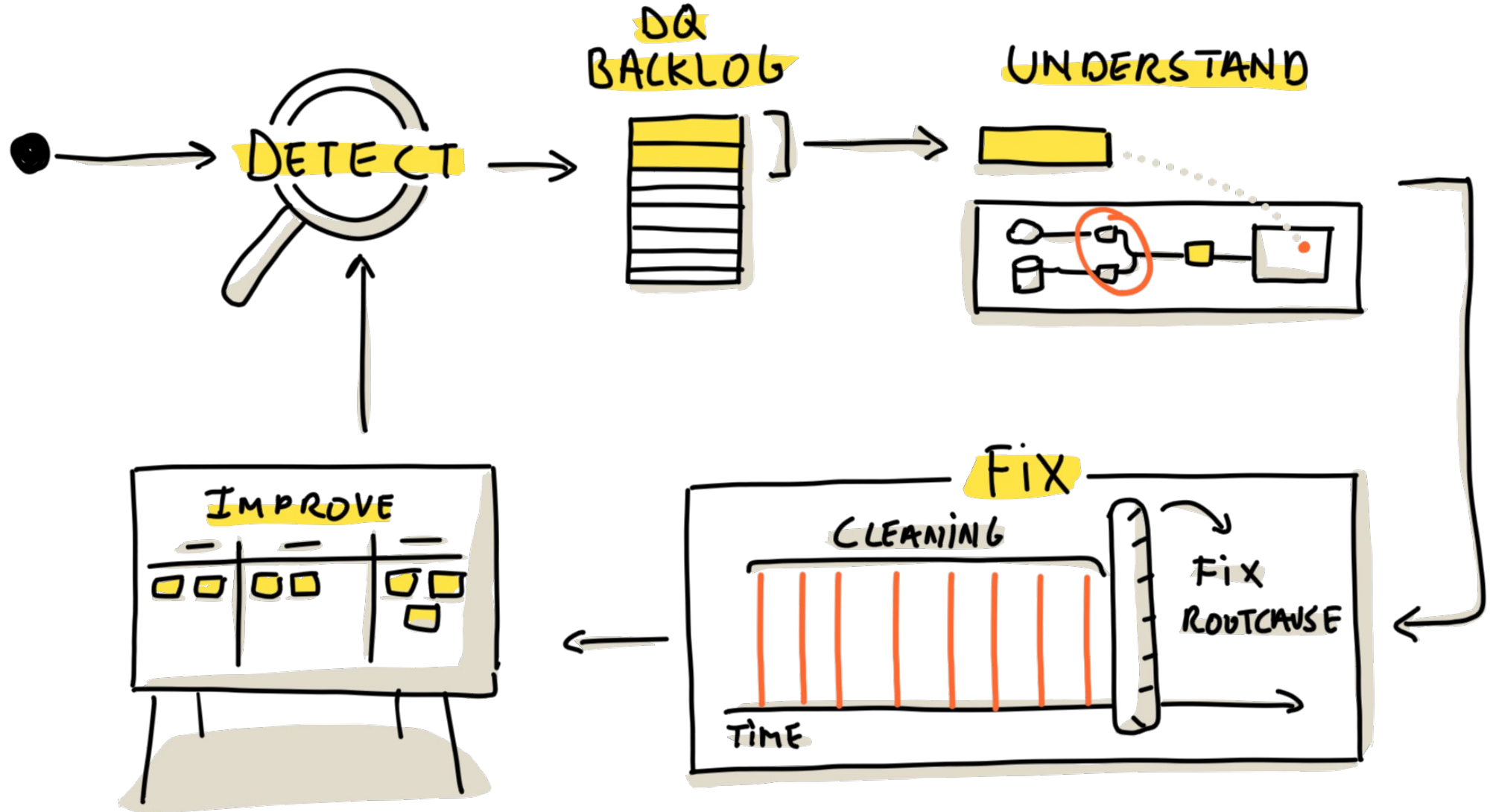
INTRODUCTION



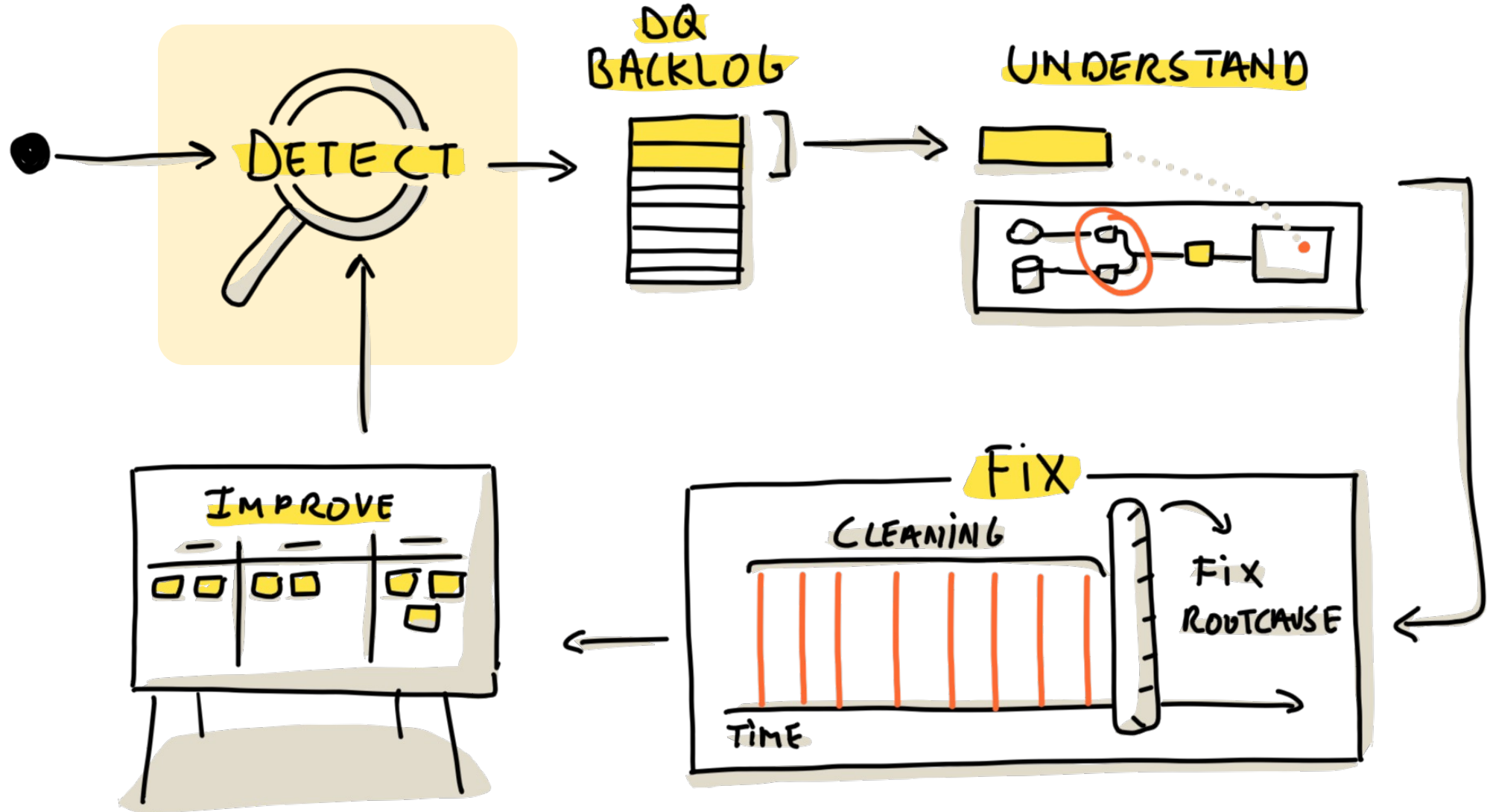
PROLESS



Data Quality Process



Data Quality Process



DQ Tests

SQL customer_id.assert.sql

```
1 -- check if customers contains null values
2 SELECT customer_id as customer_id,
3        customer_name as customer_name
4 FROM customers
4 WHERE customer_id IS NULL
```

run

if 0 row returned:

Assertion passed ✓

if >=1 row(s) returned:

Assertion failed ✗



DQ Monitoring



Home > Knowledge Catalog > Sources > MDM
party_full

Use In

Overview Profile Data Quality Data Preview Lineage Relationships **2999 Records** **7 Attributes** **Profiled 2 mins**

Filter attributes, values, masks

Name	Terms	Insights	Top 3 Values	Mask Analysis
<u>src_primary_key</u>		3 Duplicates	3% NNN 0% 145 0% 146	3% LLL 47% DDD 50% DDDD
<u>src_name</u>	<u>Last Name</u>	3 Duplicates	24% Null 3% Green 2% Kazmer	6% LLLL 5% LLLLL Show All +29
<u>src_sin</u>	<u>Social Insurance Number</u>	NULL 24%	24% Null 0% 103792776 0% SIN: 999670052	24% LLL: DDDDDDDI 18% DDDDDDDDD Show All +22
<u>src_card</u>	<u>Credit Card Number</u>	7 Exceptions	2% ##### 1% ##### 0% #####	98% DDDDDDDDDDDDI 2% LLLL





Home > Knowledge Catalog

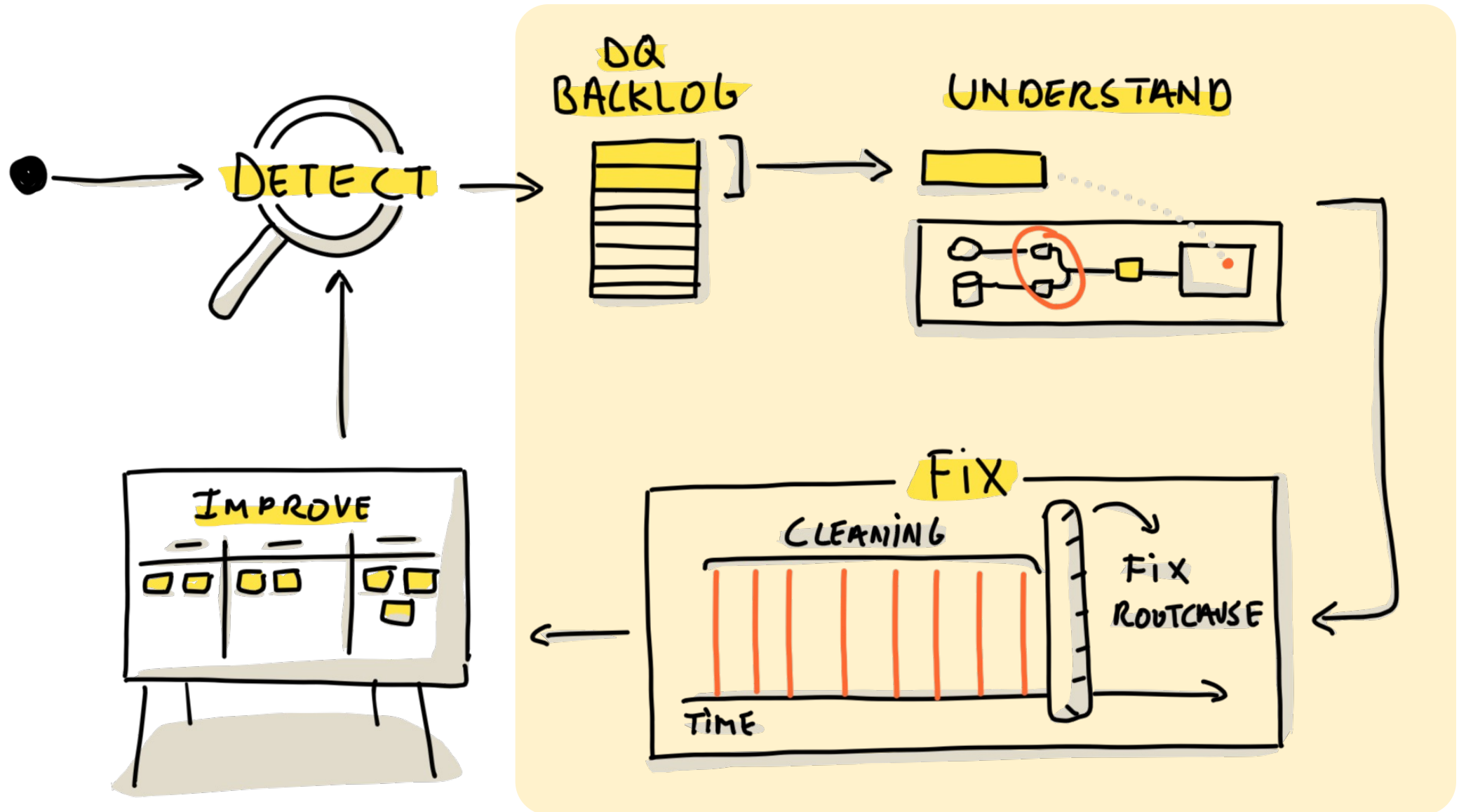
Data Assets

Filter by name, owner, creation date...

<input type="checkbox"/>	Name	Terms	Data Quality	# R
<input type="checkbox"/>	<u>src_person</u>	PII Employee Enum	<div style="width: 100%;"><div style="width: 70%; background-color: green;"></div><div style="width: 30%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>Master customer</u>	PII Customer	<div style="width: 100%;"><div style="width: 20%; background-color: green;"></div><div style="width: 80%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>Customers 2019</u>	PII Customer	<div style="width: 100%;"><div style="width: 95%; background-color: green;"></div><div style="width: 5%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>comp</u>	Account	<div style="width: 100%;"><div style="width: 90%; background-color: green;"></div><div style="width: 10%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>Customer campaigns</u>	Customer Campaign	<div style="width: 100%;"><div style="width: 95%; background-color: green;"></div><div style="width: 5%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>cstmr</u>	PII Customer	<div style="width: 100%;"><div style="width: 95%; background-color: green;"></div><div style="width: 5%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>employees_2020</u>	PII Employee	<div style="width: 100%;"><div style="width: 70%; background-color: green;"></div><div style="width: 30%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>Master address</u>	Address	<div style="width: 100%;"><div style="width: 20%; background-color: green;"></div><div style="width: 80%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>cstomers_2019_ext</u>	PII Customer	<div style="width: 100%;"><div style="width: 95%; background-color: green;"></div><div style="width: 5%; background-color: red;"></div></div>	
<input type="checkbox"/>	<u>account_list</u>	PII Account	<div style="width: 100%;"><div style="width: 90%; background-color: green;"></div><div style="width: 10%; background-color: red;"></div></div>	



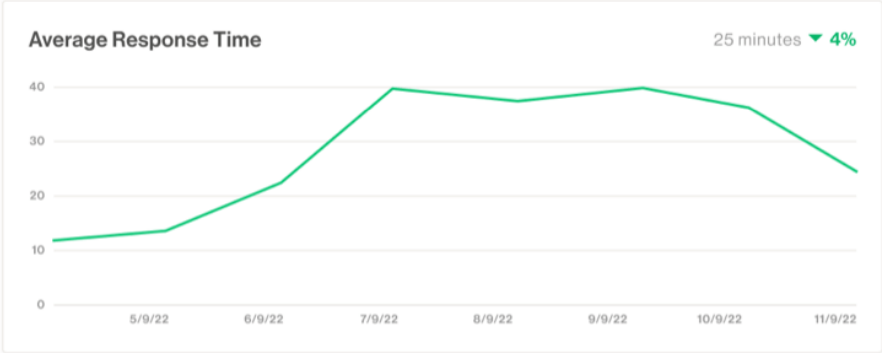
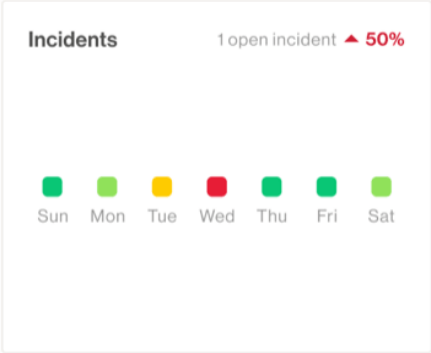
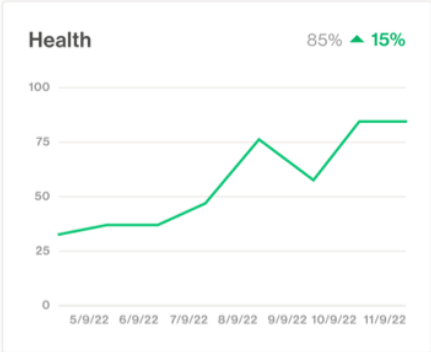
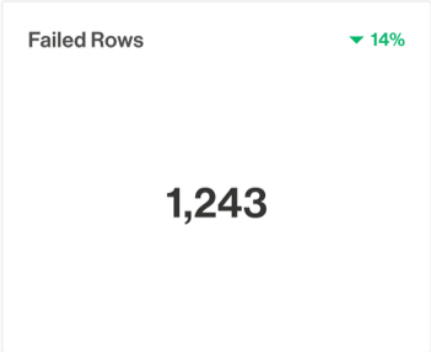
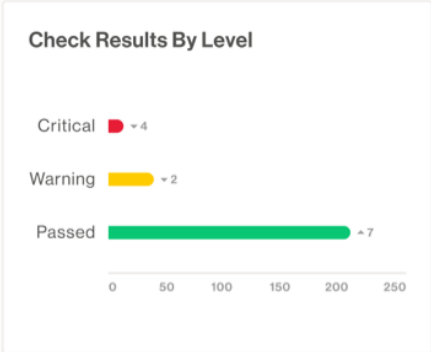
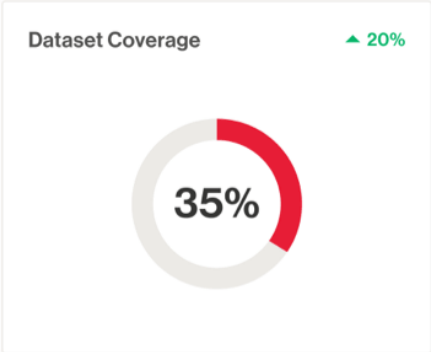
Data Quality Process



Dashboard

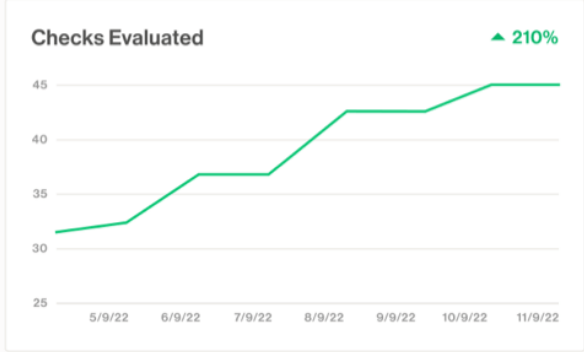
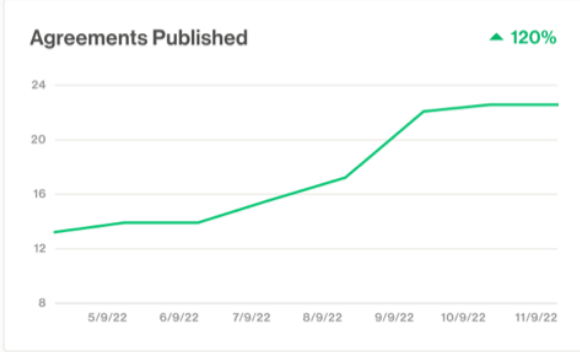
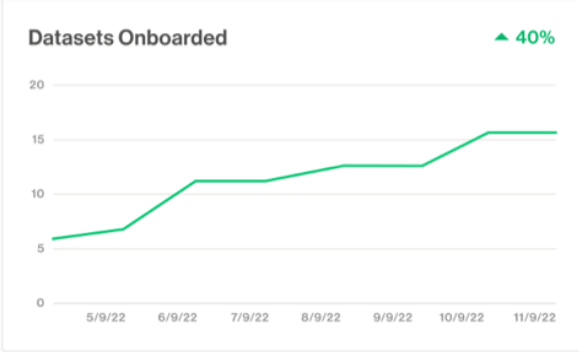
Last 7 days

Finance



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



DATA GOVERNANCE: **TABLE OF CONTENTS**

1. WHY

2. WHAT

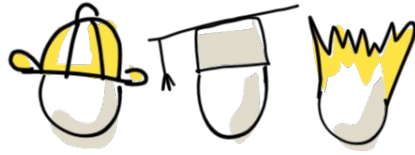
A. POLICIES, STANDARDS & RULES

B. PROCESSES

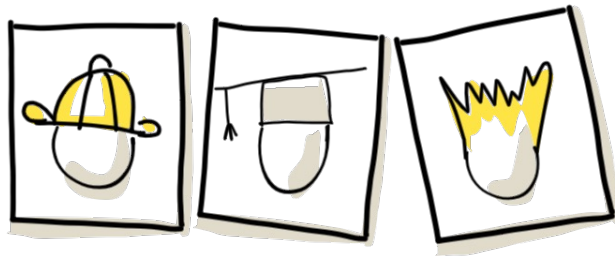
C. PEOPLE

D. ORGANIZATION

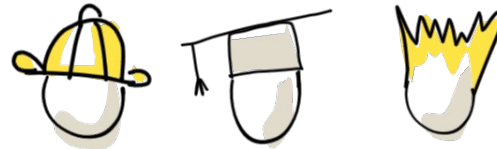




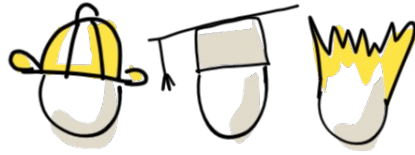
PEOPLE



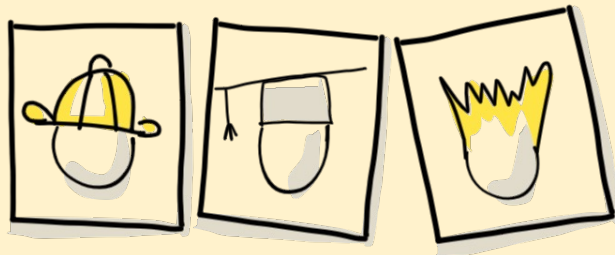
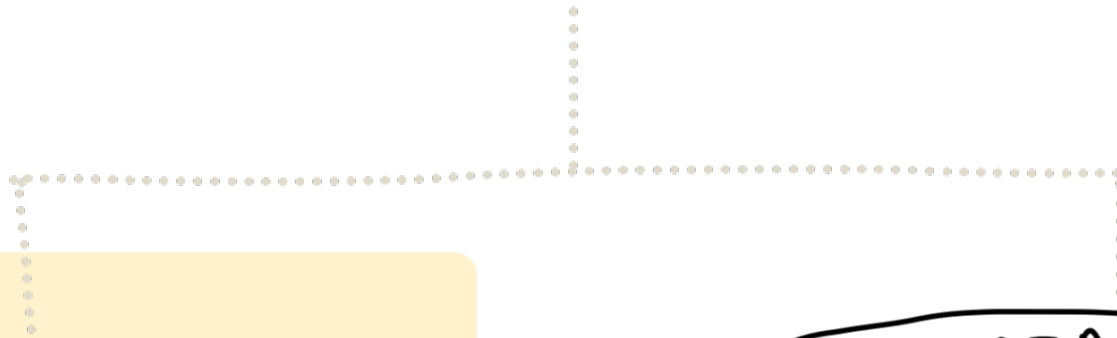
ROLES &
RESPONSIBILITIES



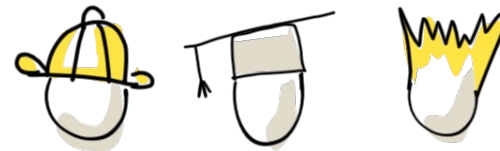
DATA LITERACY



PEOPLE

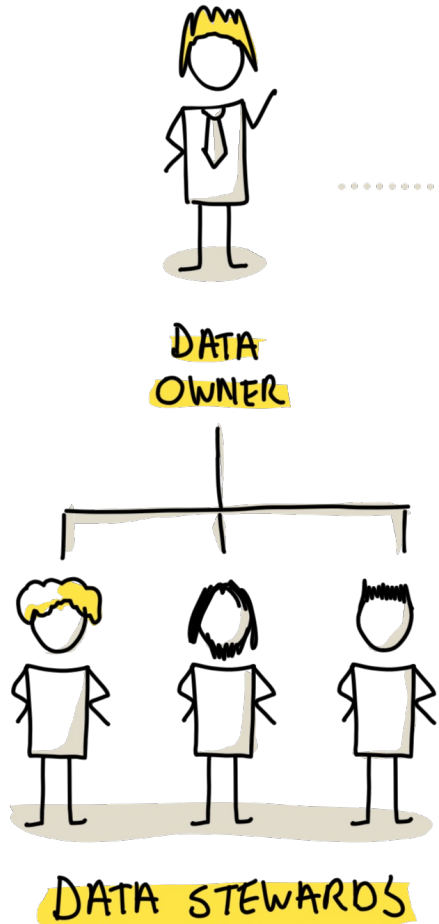


ROLES &
RESPONSIBILITIES

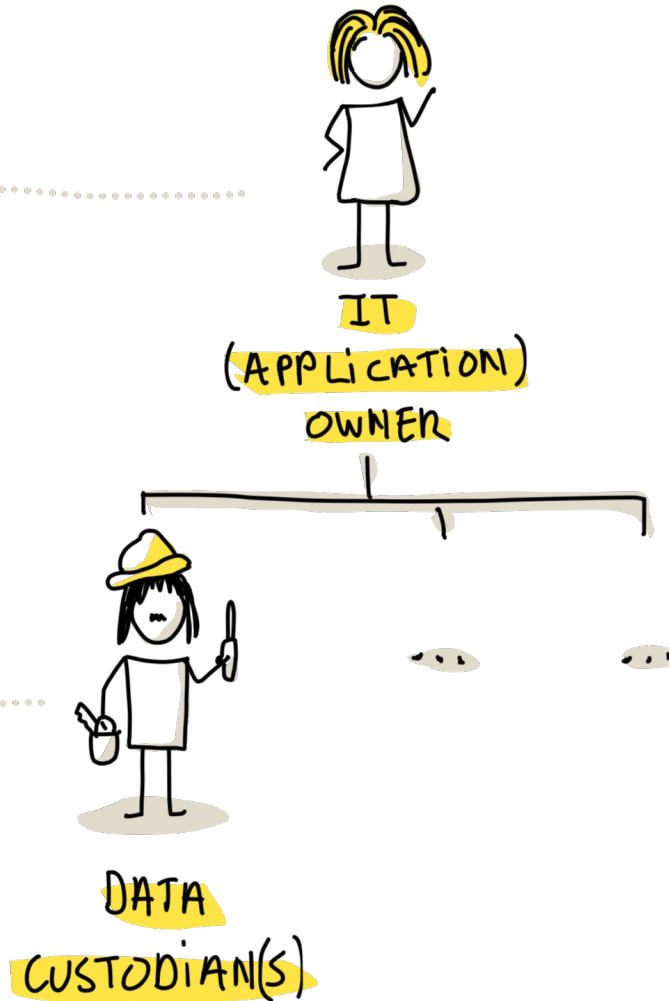


DATA LITERACY

Business



IT



Data Owner

Accountable for the classification, protection, use, and quality of one or more data sets within an organization.

Data Steward

A subject expert with a thorough understanding of a particular data set. Responsible for ensuring the classification, protection, use, and quality of that data, in line with the standards set by the Data Owner.

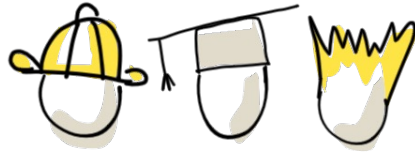
Data Custodian

Responsible for technical data changes based on requirements specified by the Data Owner.

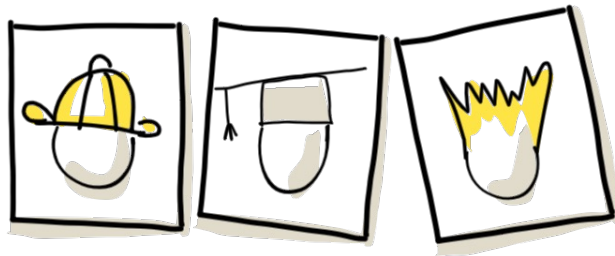


	Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Data Scientist	Mid	High	Mid	Top	High	Lower
Data Engineer	Mid	High	Top	Lower	Lower	Mid
Data Analyst	Mid	Top	Mid	Mid	Lower	Lower
ML Engineer	Mid	Lower	Mid	Mid	High	Top
Product Owner	Top	Mid	Lower	Lower	Top	Lower
Project Manager	High	Lower	Lower	Lower	Mid	Mid

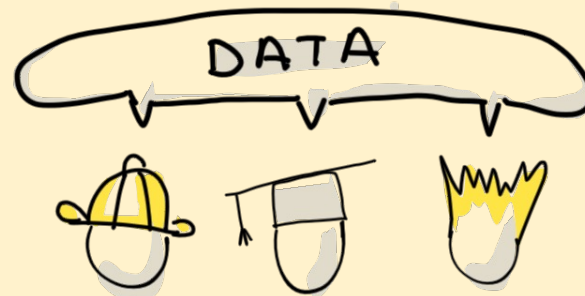




PEOPLE



ROLES &
RESPONSIBILITIES



DATA LITERACY

Data in the hands of a few data experts can be powerful, but data at the fingertips of many is what will be truly transformational.



Data Literacy Skill Levels

			
Reading Dashboard User	Creating insights by filtering existing dashboards	Conclusions	Implications Causations
Writing Dashboards 	Editing Existing Dashboard	Build new workbooks and perform ad-hoc analyses	Create new impactful and actionable Dashboards (for own and team)
Writing Data Models 	Write simple queries and edit existing queries (for ad-hoc analyses)	Create Data Models and know where to find the right data	Create new data model Explore existing data models for compatibility
Speaking	Presenting Explain needs	Discuss Ask the right questions	Convince and Convey Strategic plans based on data

COMMUNITY of PRACTICE



Professional Development
April 30th, 2013
CTLT, UBC



CREATING JOINT AGENDAS



DATA GOVERNANCE: **TABLE OF CONTENTS**

1. WHY

2. WHAT

A. POLICIES, STANDARDS & RULES

B. PROCESSES

C. PEOPLE

D. ORGANIZATION



CENTRALIZED

(Data Mesh) DECENTRALIZED

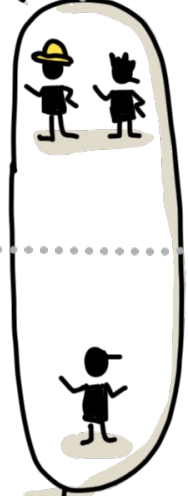
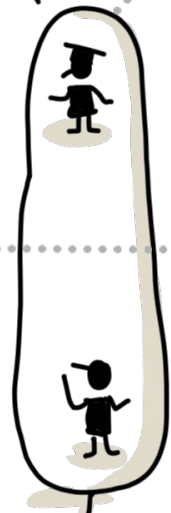
DATA
PLATFORM

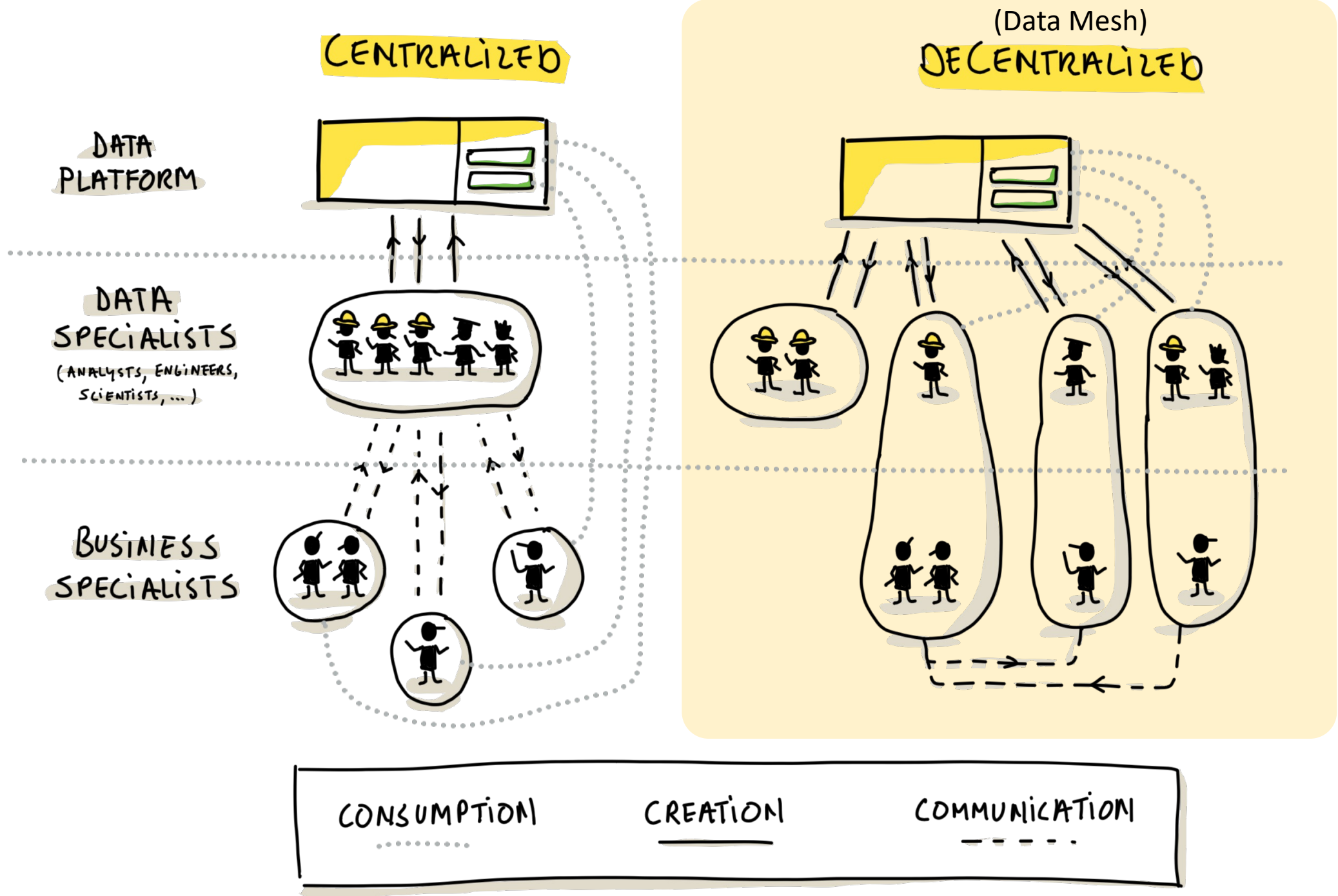


DATA
SPECIALISTS
(ANALYSTS, ENGINEERS,
SCIENTISTS, ...)



BUSINESS
SPECIALISTS





SOURCES

TO INGEST



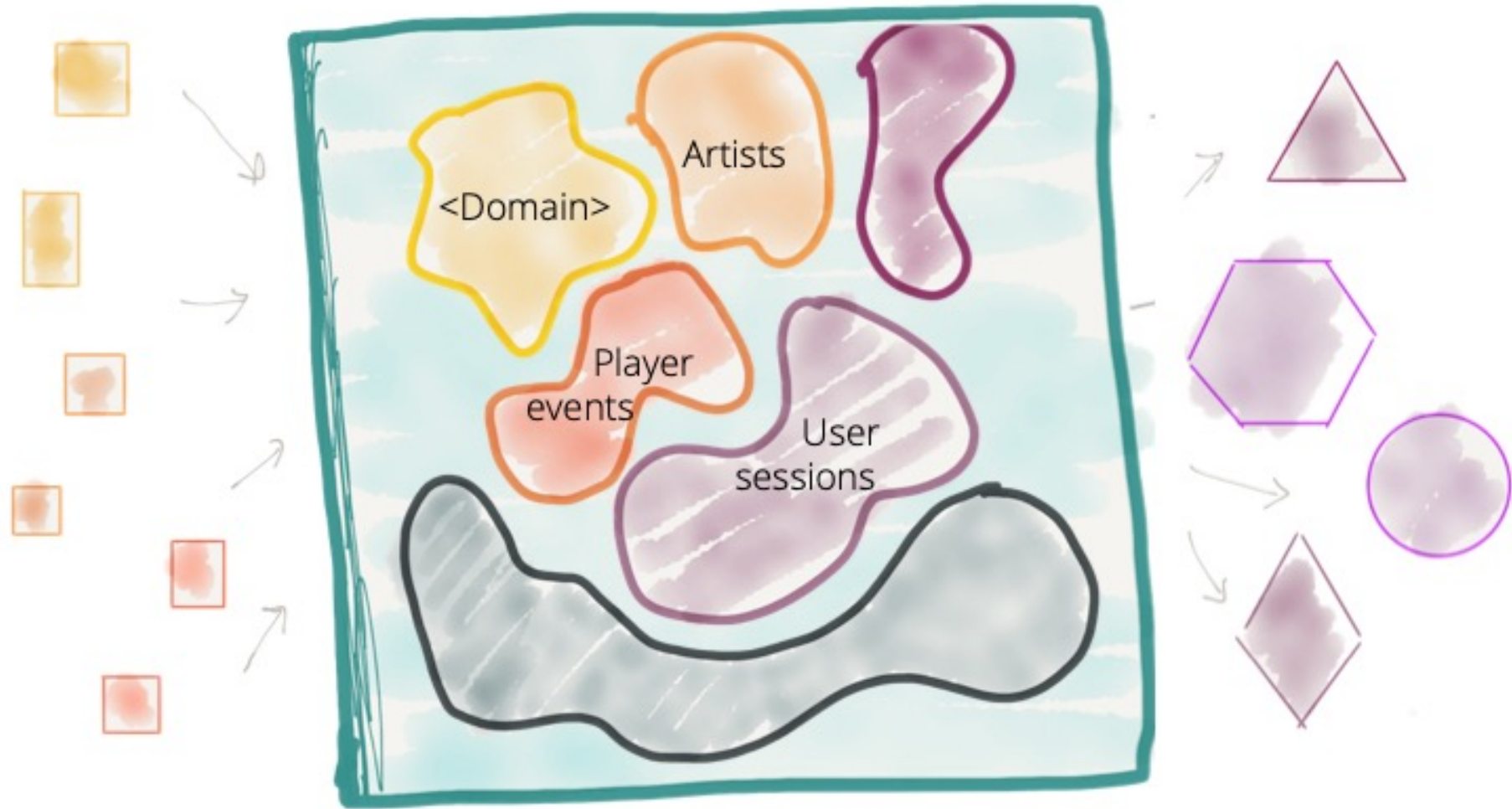
BIG DATA PLATFORM

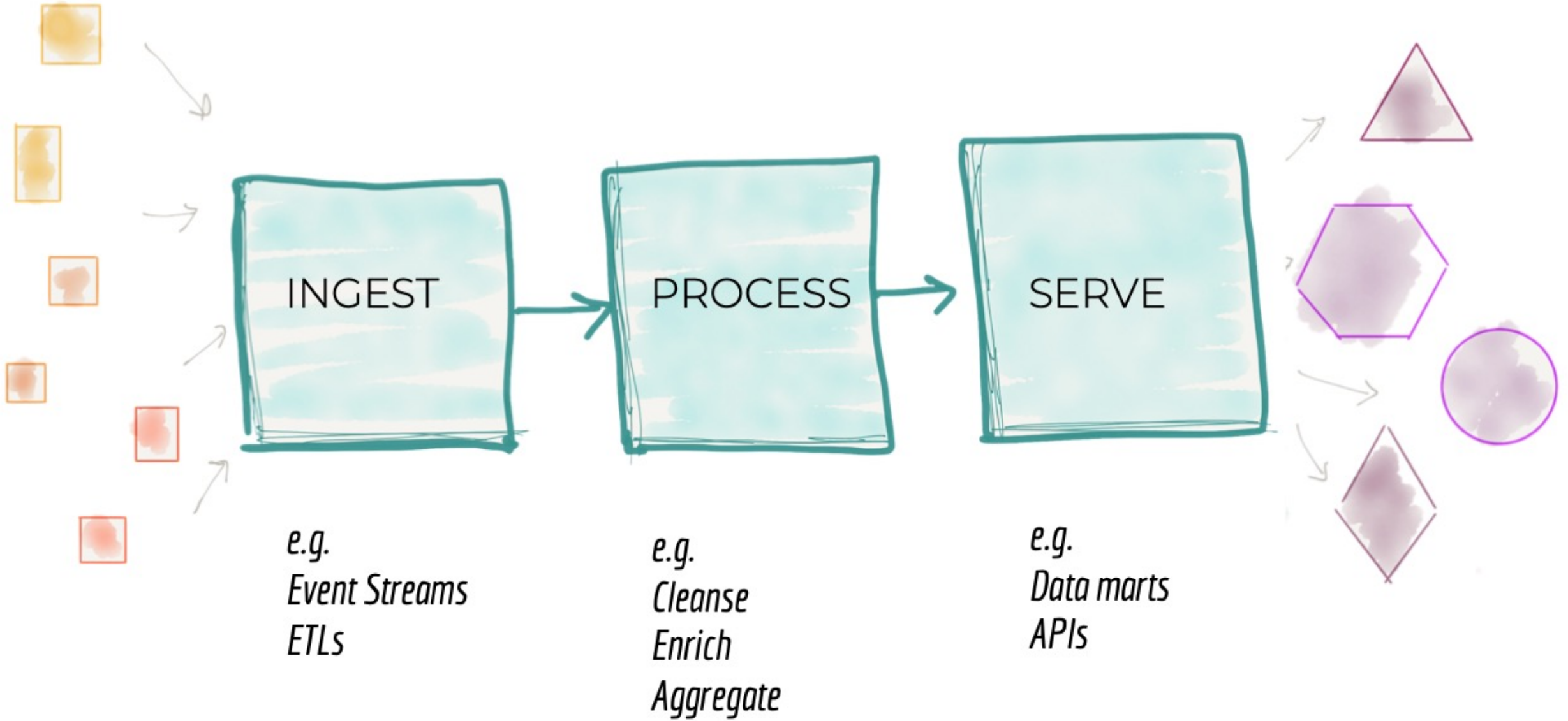


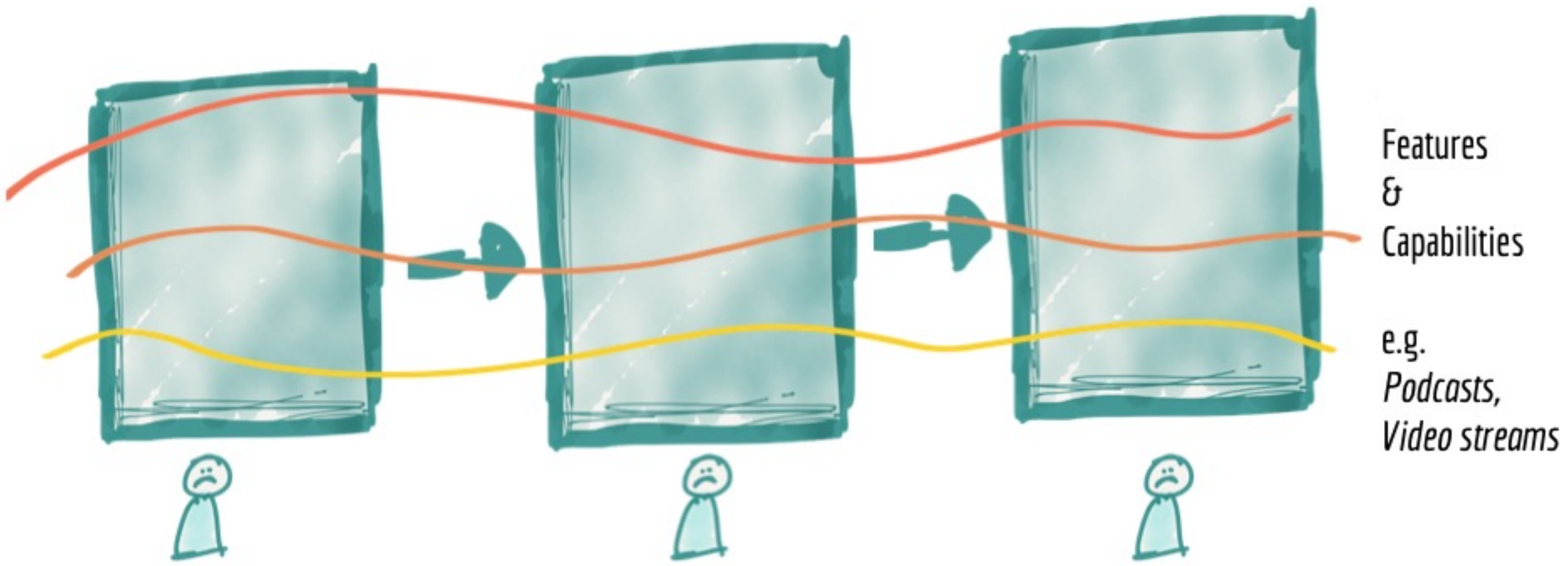
CONSUMERS

TO SERVE





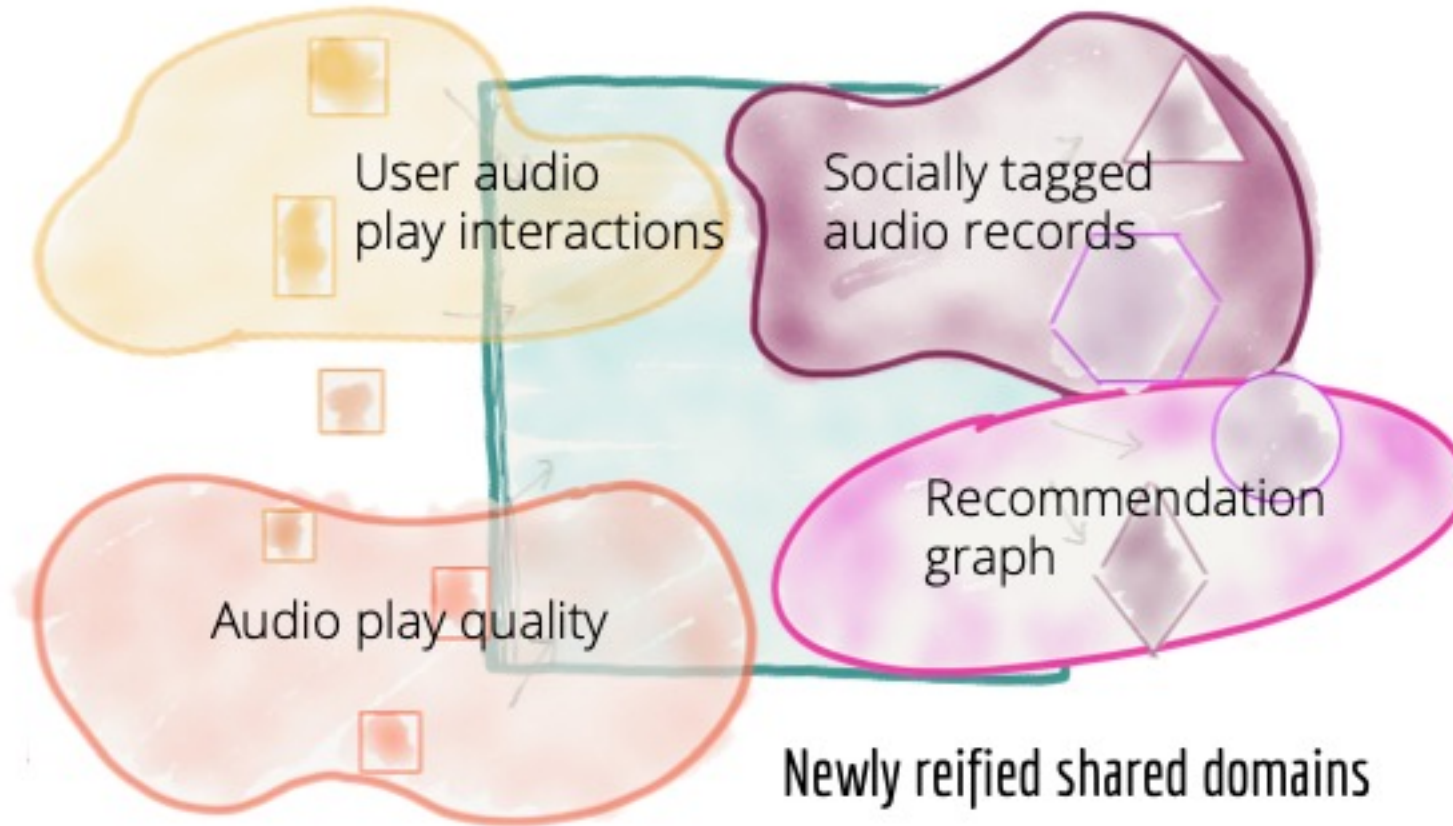




ANOTHER APPROACH

Domains aligned with the source

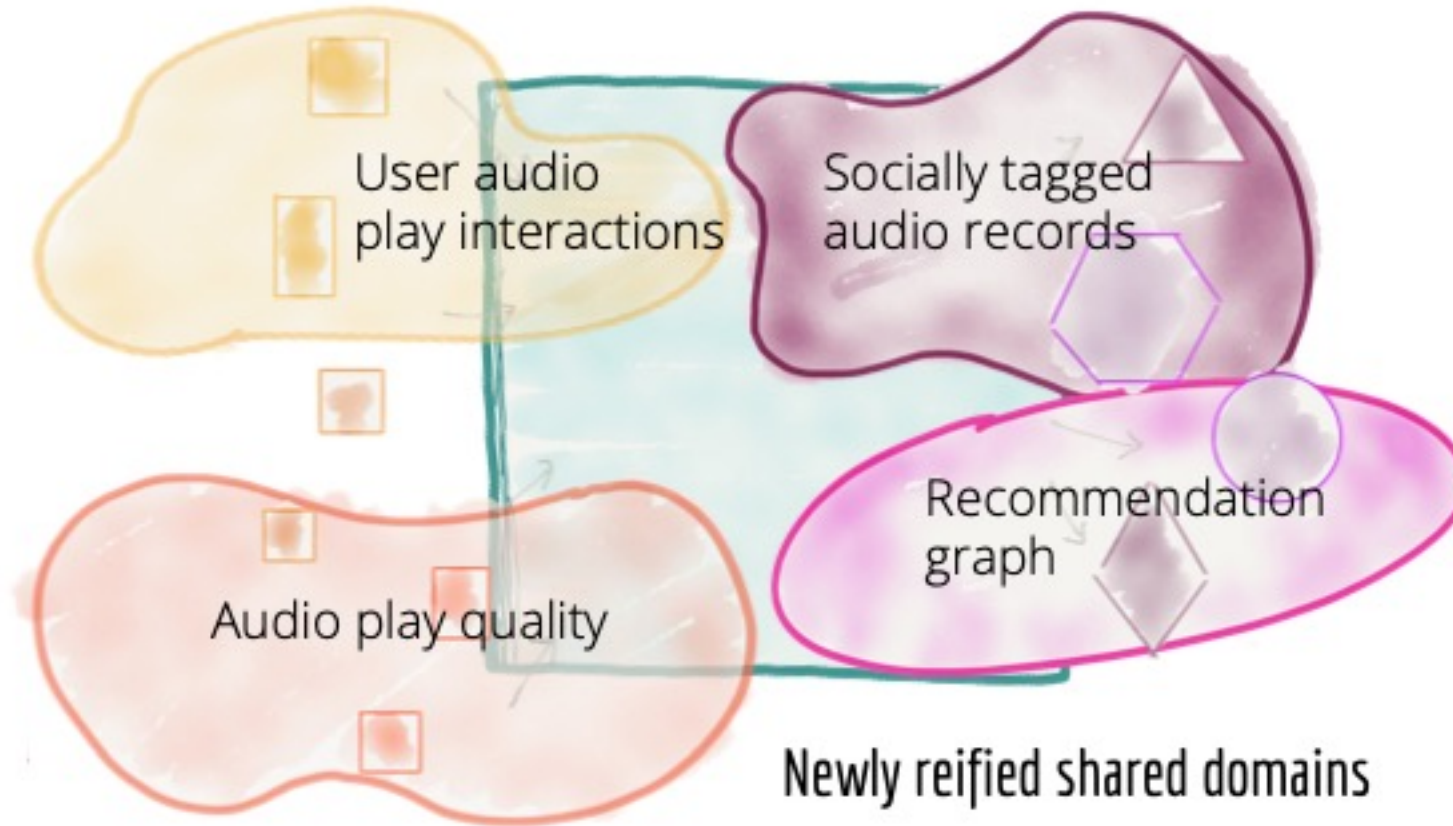
Domains aligned with the consumption



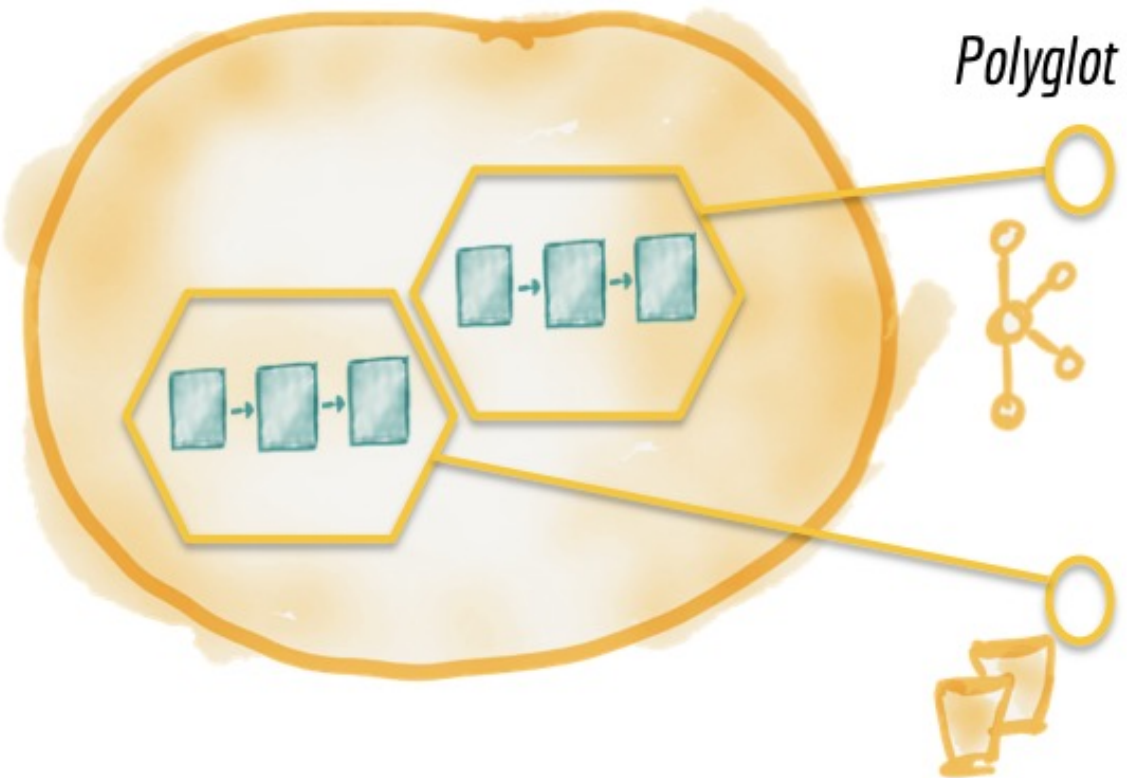
ANOTHER APPROACH

Domains aligned with the source

Domains aligned with the consumption



Domain



Polyglot Data Products

DISCOVERABLE



ADDRESSABLE



TRUSTWORTHY
(DEFINED & MONITORED SLOs)



SELF-DESCRIBING



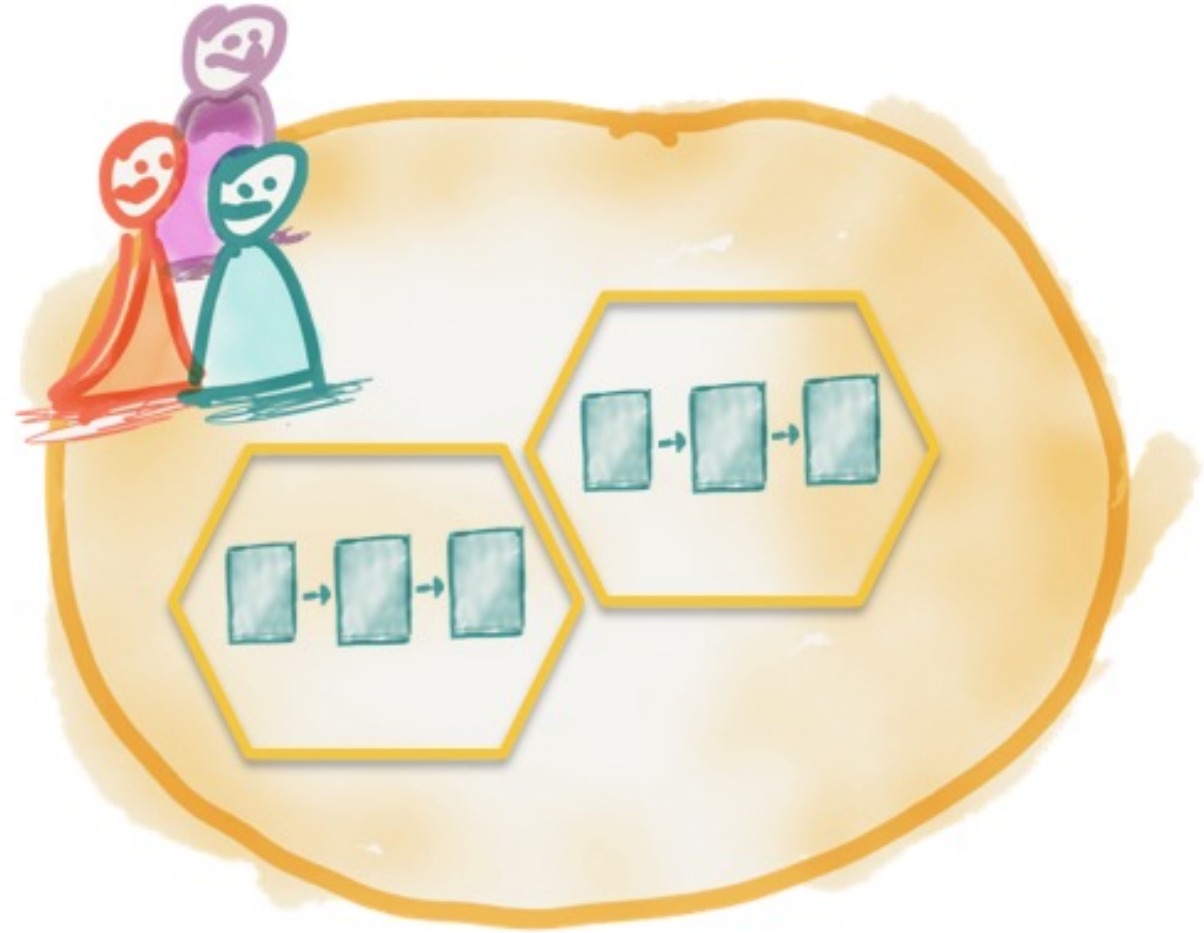
INTER OPERABLE
(GOVERNED BY OPEN STANDARDS)



SECURE

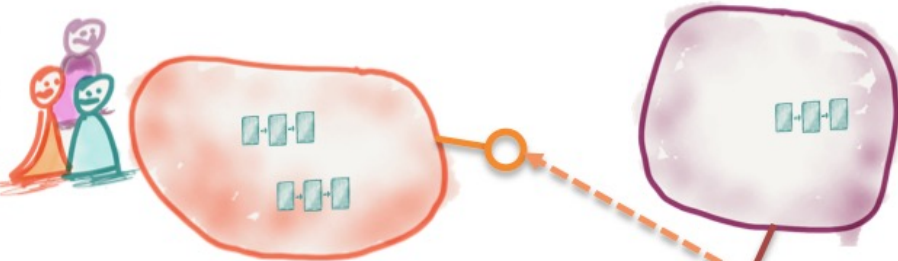


Data Engineer
Domain Data Product Owner
Software Developer
Infra Engineer
...

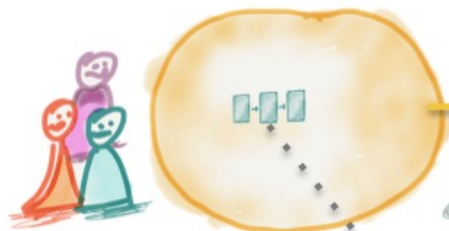


Global Governance & Open Standards
(enable interoperability)

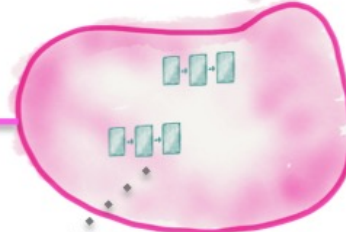
Domain data owned & served by cross-functional team



New Data Domains can be created correlating data from other domains

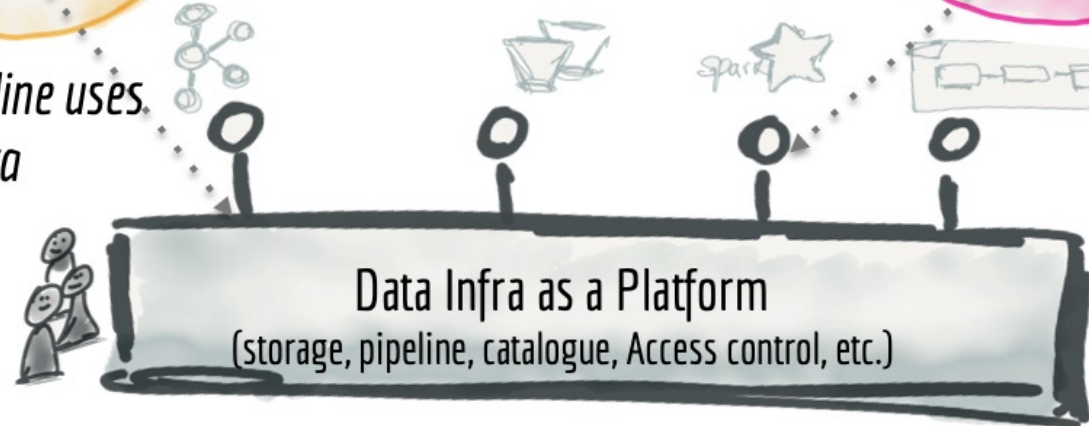


Domain's internal data pipeline uses common self-serve data infra



Domain oriented data served as a product for any other domain to use

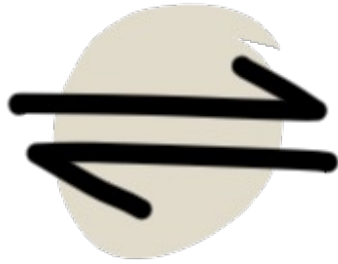
Data infra engineers



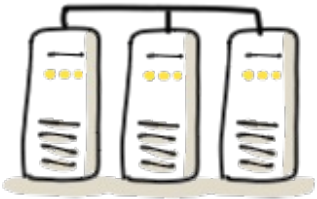
Data Infra as a Platform
(storage, pipeline, catalogue, Access control, etc.)



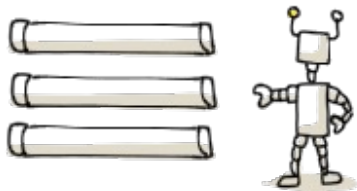
DATA PLATFORM



INFRASTRUCTURE



TECHNOLOGY



DATA GOVERNANCE

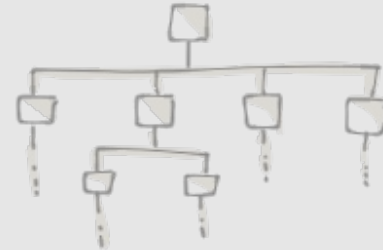
PEOPLE



PROCESSES



ORGANIZATION

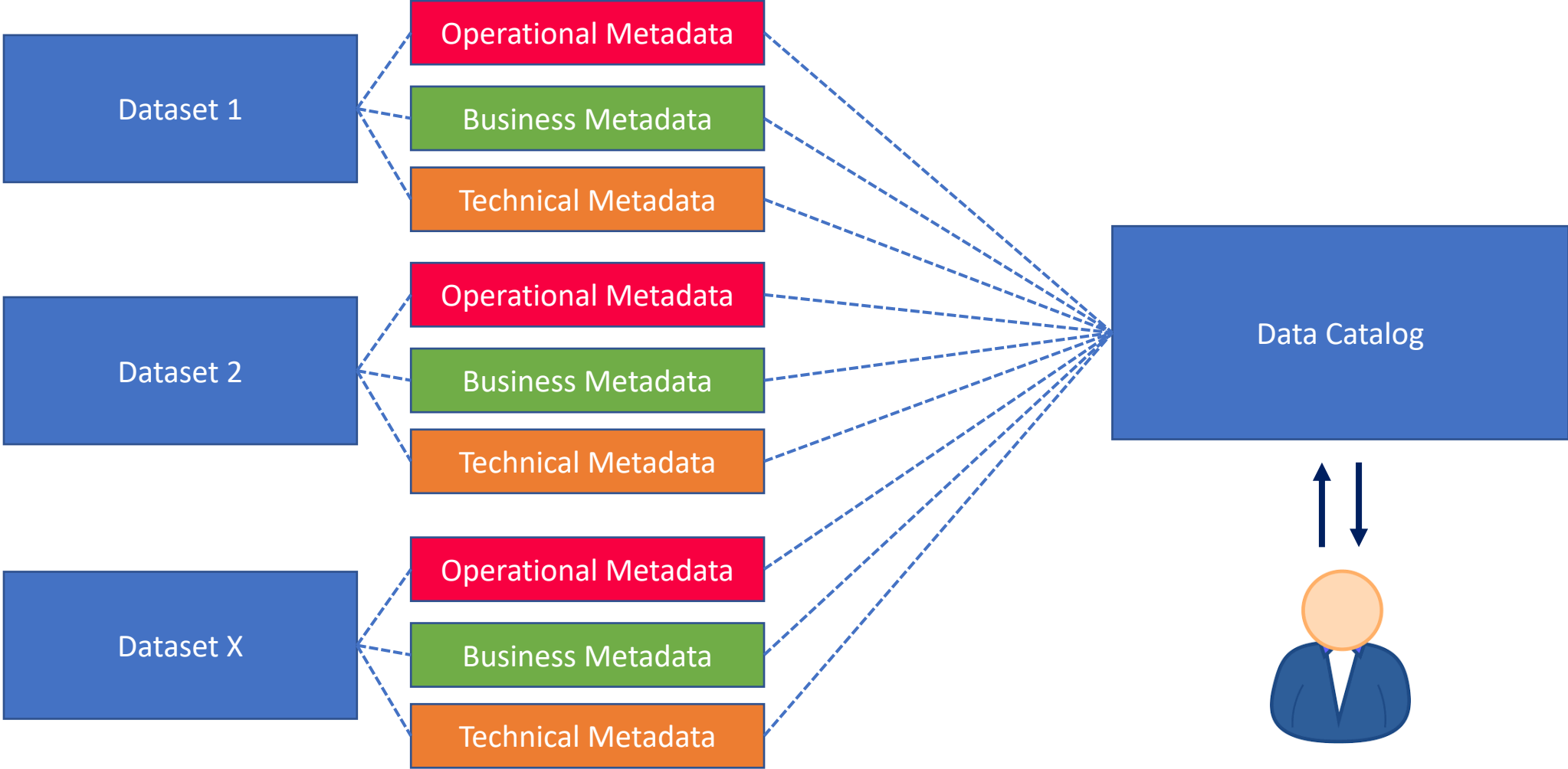


POLICIES, STANDARDS & RULES



Data Catalog

Bringing Data Governance and Technology together



A Data Catalog helps analysts and other data users to find the data that they need.

A Data catalog combines:

- A collection of metadata
- Search tools
- Data Management tools
- An inventory of available data
- Information about data quality





Key filters All filters

5107 Search results for: "*"

Views [Grid icon] [List icon]

Followed

Asset type Dataset (2341) Report (2541) Glossary Item (225)

Validation status Unassigned (2945) Validated (87) Not validated (25) Under review (2049) Out of use (1)

Owner Unassigned (2241) 1. SDG team : No poverty (2067) 0. SDG team : cross SDGs (2) Data Citizens (1) Key users APAC - Finance (1) Local stewards APAC (3) [Show all](#)

Source Unassigned (2121) BW4 Barry (683) Power BI (65) SAP BW on HANA (566) SAP BW/4HANA (3) Tableau test (1666)

Duration
Report | Validated

Duration description

(SDG) funding - Power BI dataset
Dataset | Validated

This data set contains information on obtained funding for SDG

1.0 Organisatie context
Glossary Item | Validated

Deze groep bevat alle objecten als demo voor de requirements: A LG001 tot en met ALG006 Opmerkingen: ALG005: Versiebeheer zou ondersteund moeten worden. => Op vandaag nog niet bes...

1.5 DC: Connectie datacatalog
Glossary Item | Validated

Deze groep bevat alle objecten als demo voor de requirements: DC001 tot en met DC003 de gebruikte dataset is DC001: Data set 1 - Personeel

Activity Status
Glossary Item | Validated

Indicates the status of a project. The following statuses are used: Pipeline/identification: the project is currently under consideration Implementation: the project is currently ongoing Completion: th...

Allocation Volume
Glossary Item | Validated

Total allocation volume

Average Order Value
Glossary Item | Validated

Average value of each individual sales order

Azure Marketplace follow up
Report | Validated

This report allows us to follow up customers onboarding via Azure Marketplace

BVP
Glossary Item | Validated

Business Glossary - define KPIs, data definitions and business terms
Glossary Item | Validated

Business Rule 1
Glossary Item | Validated

explanation

Cash Conversion Cycle
Glossary Item | Validated

The CCC is a metric that expresses the length of time it takes for a company to convert resource inputs into cash flows.. The lower the better. This indicator is used to optimize financial resources...

Cash To Cash
Glossary Item | Validated

Measures the number of days between the payment for raw materials and the payment of the final product sold to the customer. T

Channel Performance
Glossary Item | Validated

Claim payment
Glossary Item | Validated

Claim Payment (CP) means an amount payable to you under the Policy to compensate you for the credit losses you have sustaine



Find data assets across your projects and organizations

 Search

Popular Tables

Most queried BigQuery tables and views in the past 30 days

airlines

travel-bookings-236421.trip_ez.airlines

analytics_travel_agency

travel-bookings-236421.trip_ez.analytics_travel_agency

analytics_travel_agency_hotel_flight

travel-bookings-236421.trip_ez.analytics_travel_agency_hotel_flight

bookings

travel-bookings-236421.trip_ez.bookings

bookings_2019

travel-bookings-236421.trip_ez.bookings_2019

car_brand_most_used

travel-bookings-236421.trip_ez.car_brand_most_used

city_most_visited_year

travel-bookings-236421.trip_ez.city_most_visited_year

customers

travel-bookings-236421.trip_ez.customers

flight

Explore data assets

Explore all the data assets available to you



All BigQuery resources



Datasets



Data streams and Pub/Sub topics



Tables and Views

Tag Template

Tag your data using predefined templates



[Explore tag templates](#)



[Create a tag template](#)

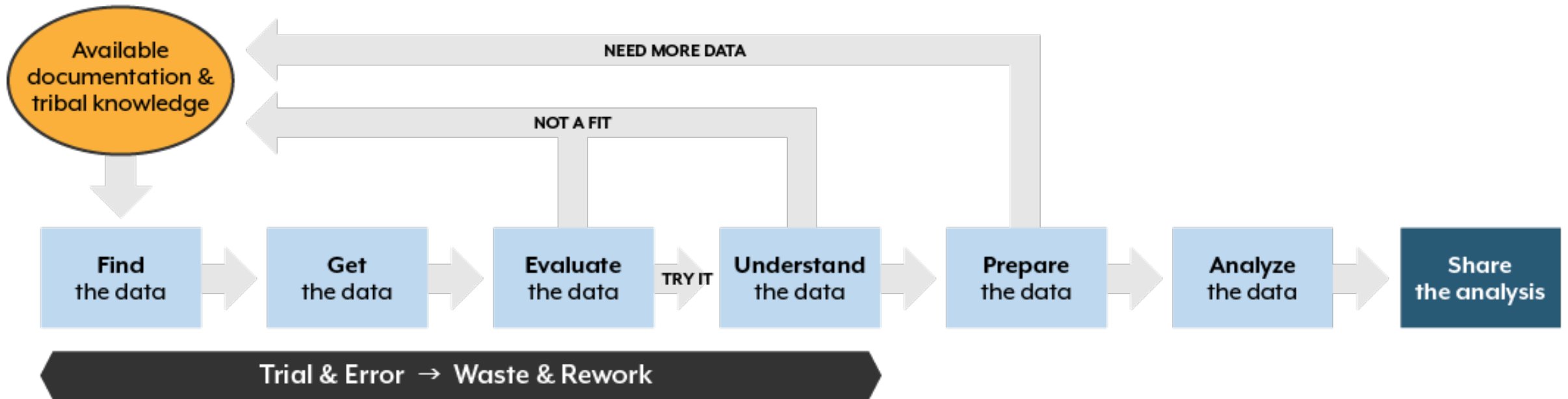
Search tips

To match exactly (type and system), use "=" as the search separator. To match within text, use "*" as the search operator

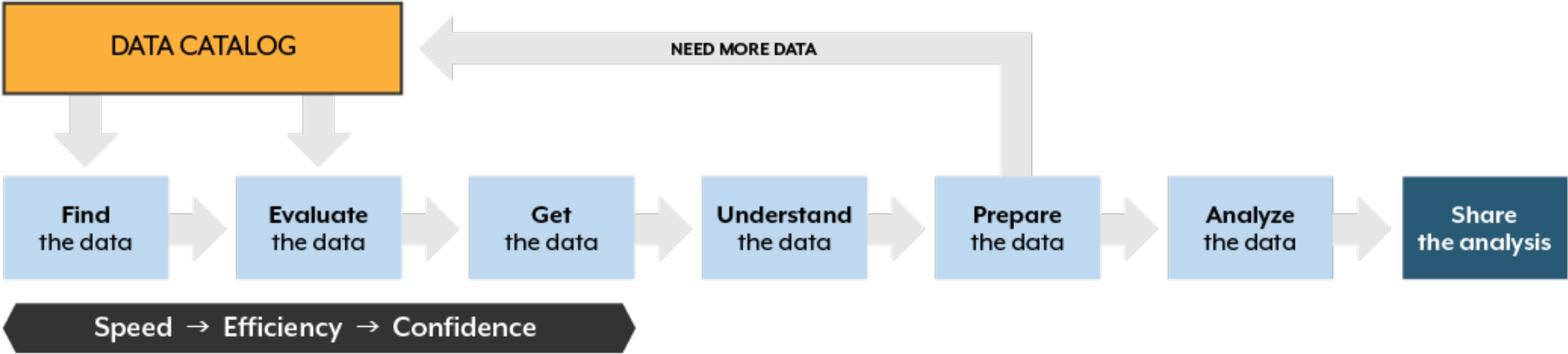
Search by	Example
tag	tag.travel-bookings-236421.my-template-id.somefield.somevalue



Without Data Catalog



With Data Catalog



WRAP UP

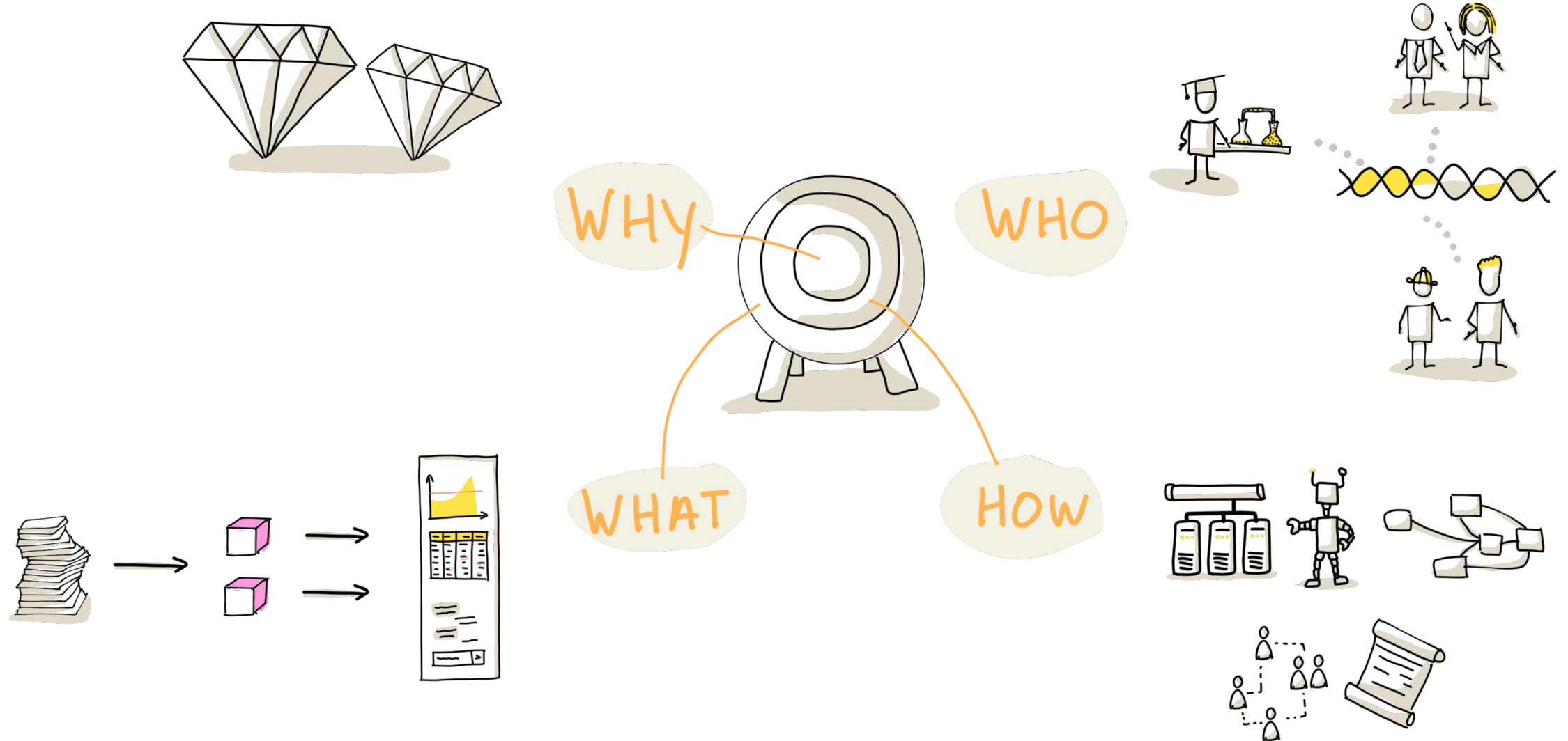
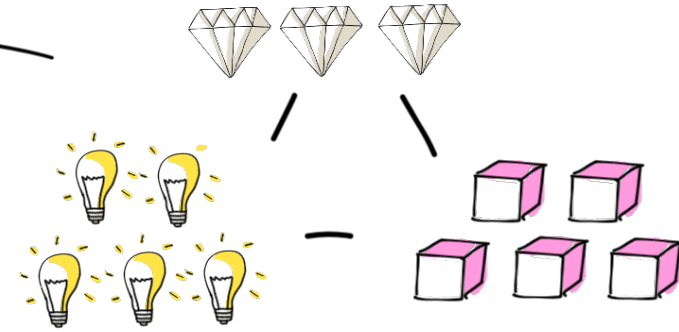
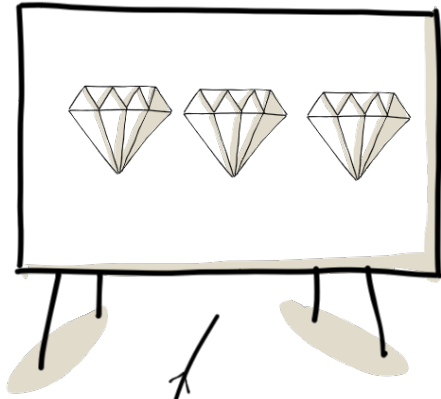


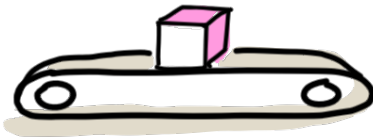
TABLE OF CONTENTS

BUSINESS STRATEGY



DATA PRODUCT

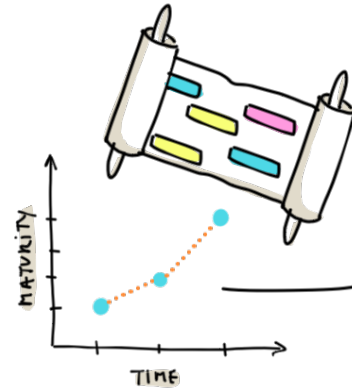
WHY-WHAT-WHO-HOW



DELIVERY

	1	2	3	4	5
CRIT 1		●			
CRIT 2			●		
CRIT 3	●				
CRIT 4				●	

DATA MATURITY

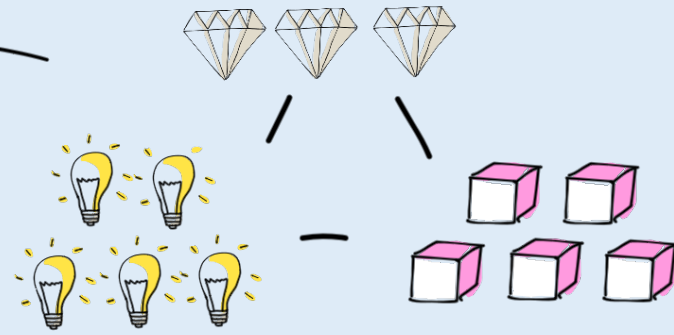
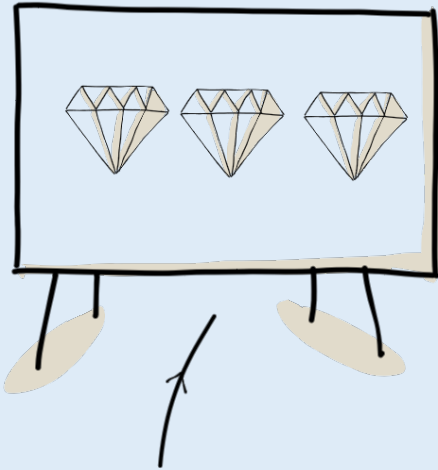


STRATEGY + ROADMAP



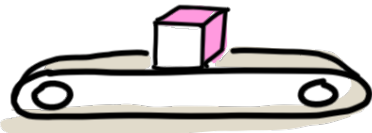
TABLE OF CONTENTS

BUSINESS STRATEGY



DATA PRODUCT

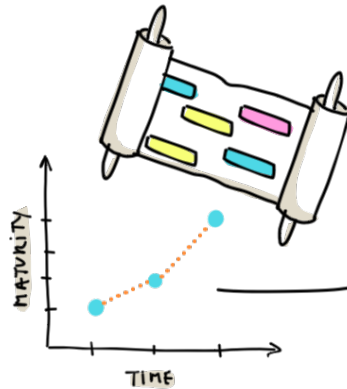
WHY-WHAT-WHO-HOW



DELIVERY

	1	2	3	4	5
CRIT 1		●			
CRIT 2			●		
CRIT 3	●				
CRIT 4				●	

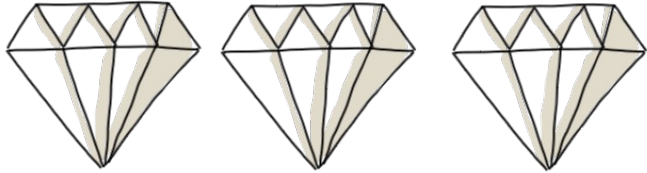
DATA MATURITY



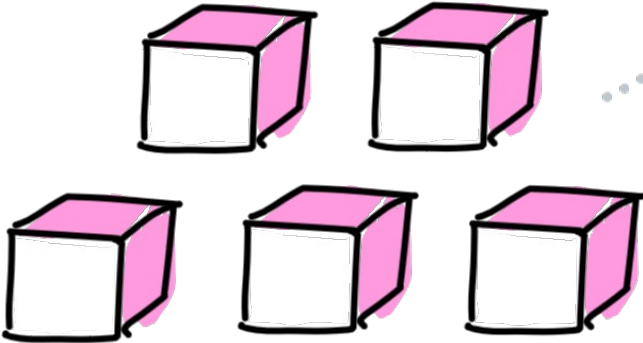
STRATEGY + ROADMAP



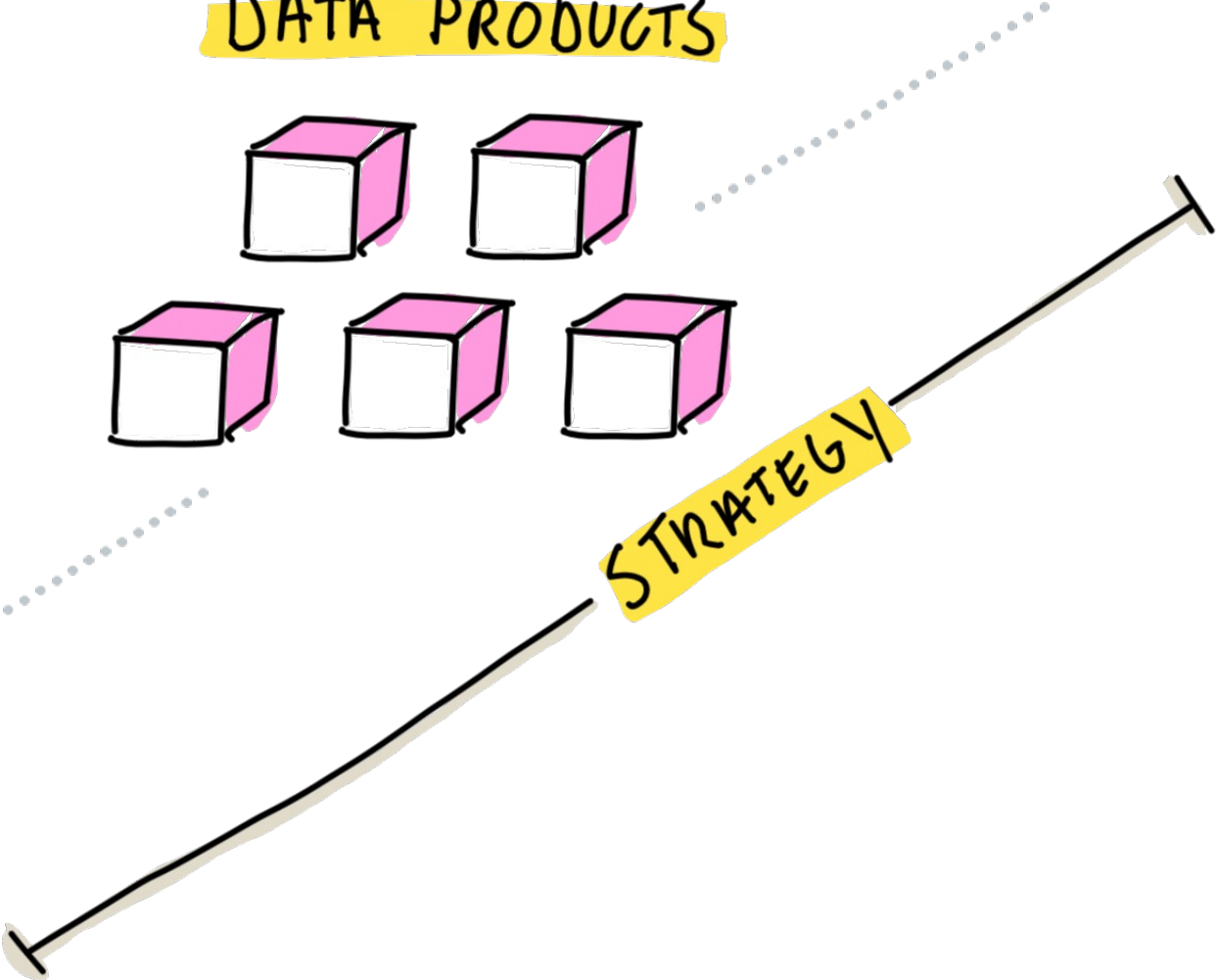
BUSINESS GOALS



DATA PRODUCTS



DATA PRODUCT IDEAS



BUSINESS GOALS



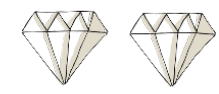
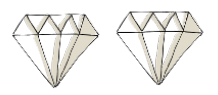
BUSINESS STRATEGY

DIGITAL STRATEGY

DATA STRATEGY

MARKETING STRATEGY

...

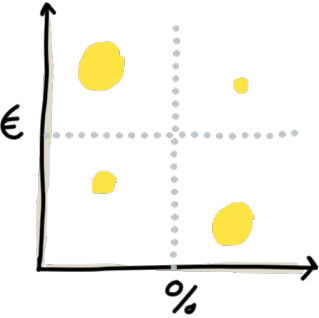


BUSINESS STRATEGY



"REDUCE CUSTOMER CHURN $\leq 5\%$ "

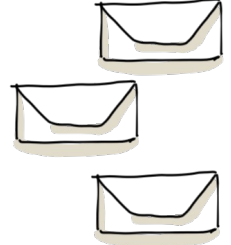
RETENTION



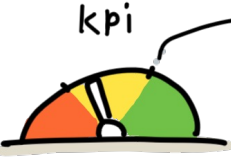
⇔

Q1	Q2	Q3	Q4
■	■	■	■
	■		■
	■		■

MARKETING



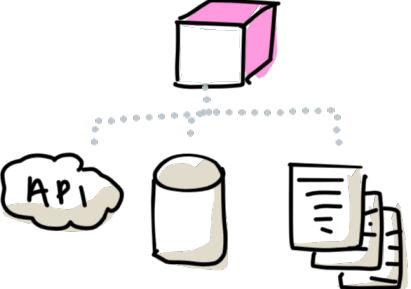
FINANCE



CUSTOMER SERVICE



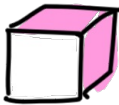
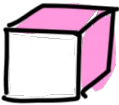
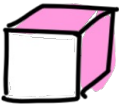
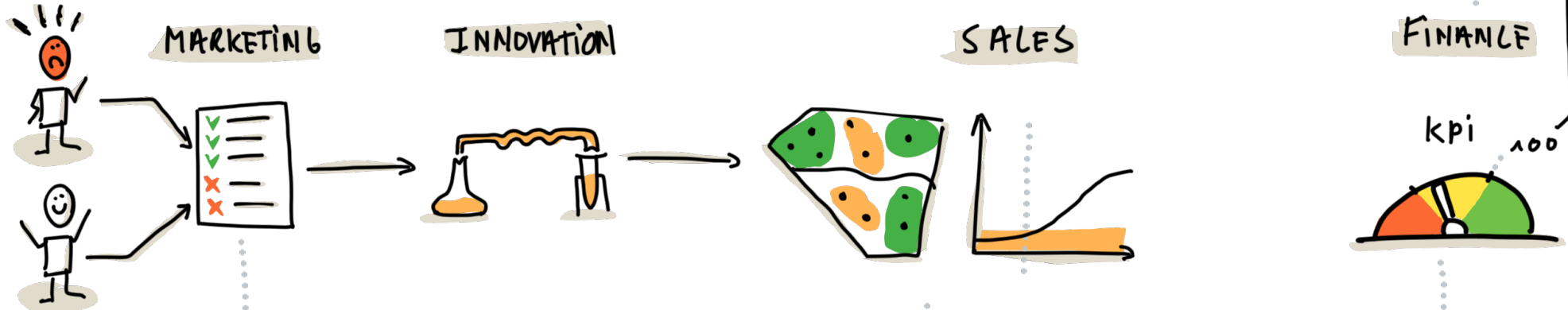
DATA



BUSINESS STRATEGY



"LAUNCH A NEW PRODUCT WITH 100 CUSTOMERS"



DATA



EXERCISE: RELATE DATA PRODUCT(S) TO YOUR BUSINESS GOALS

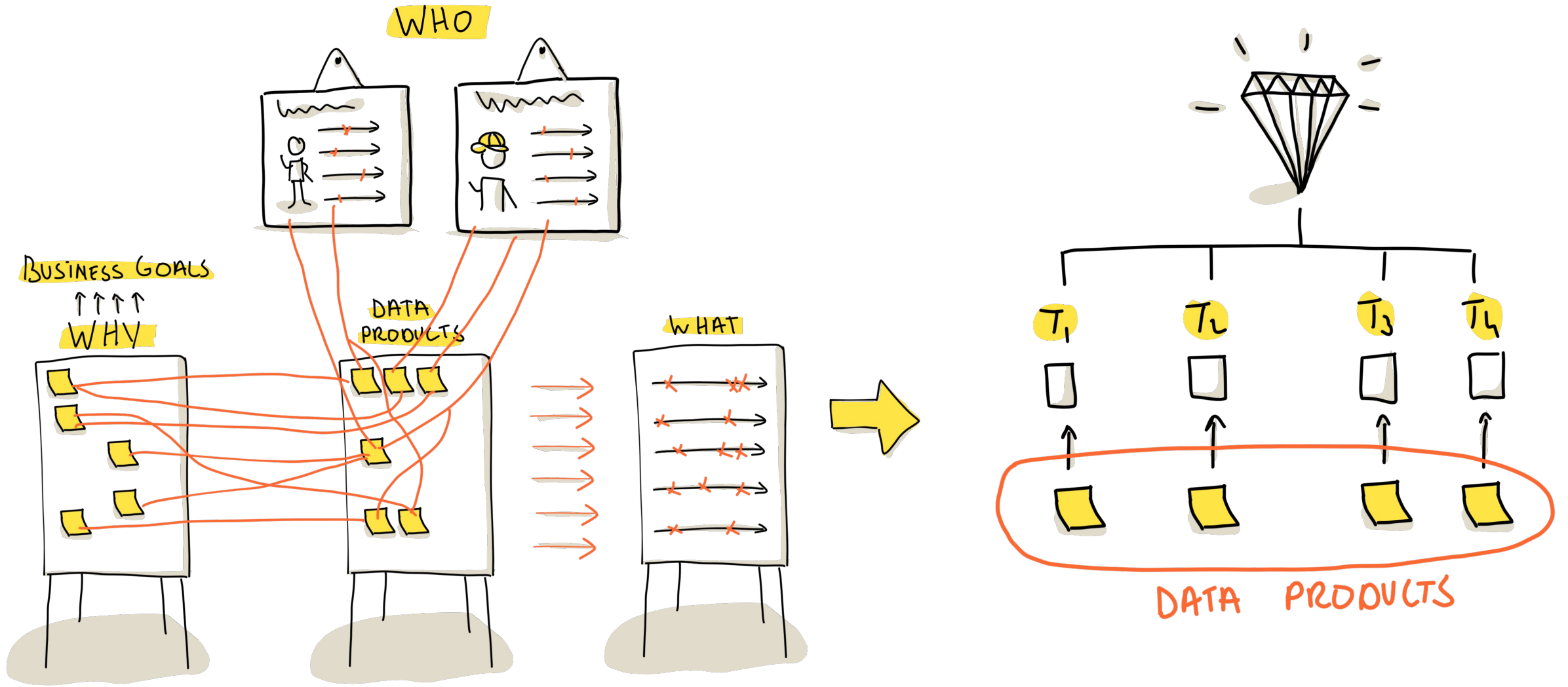
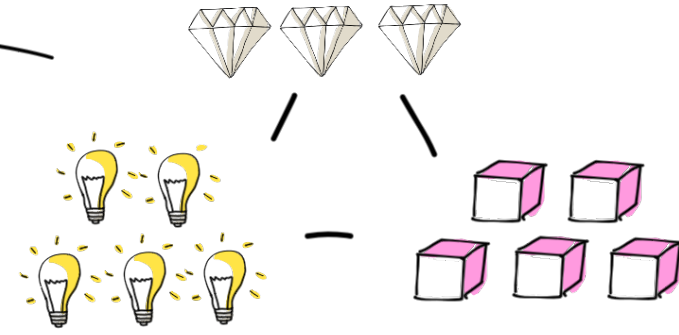
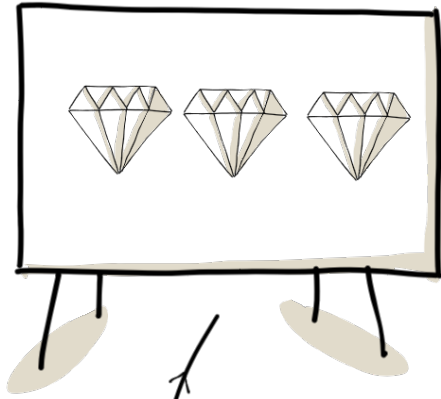


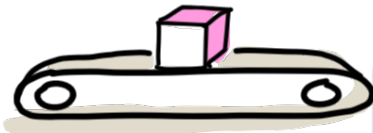
TABLE OF CONTENTS

BUSINESS STRATEGY



DATA PRODUCT

WHY-WHAT-WHO-HOW



DELIVERY

MATURITY

TIME

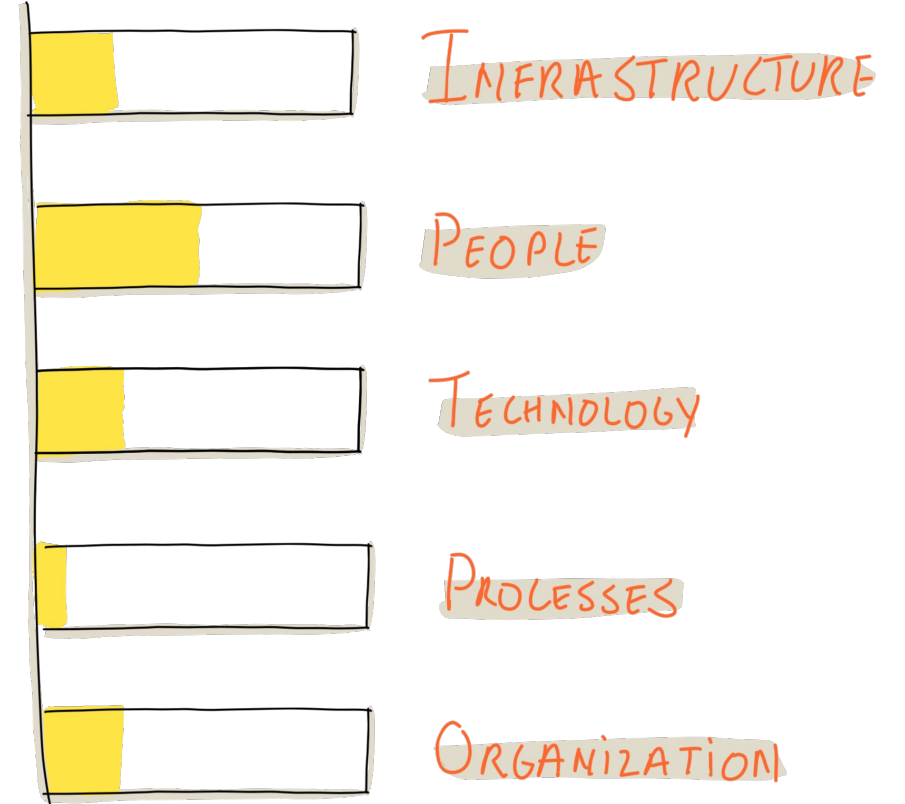
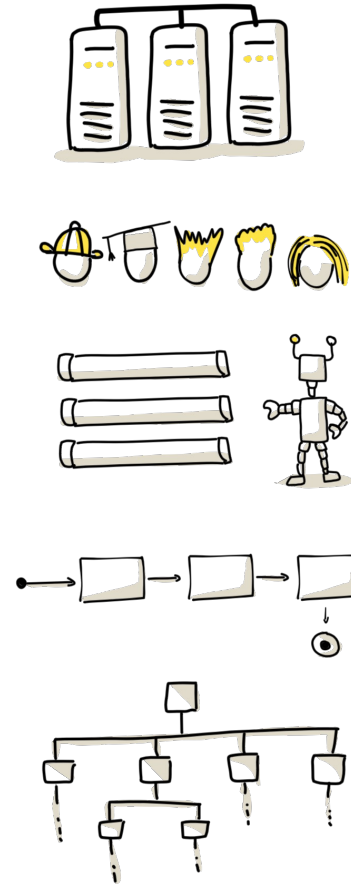
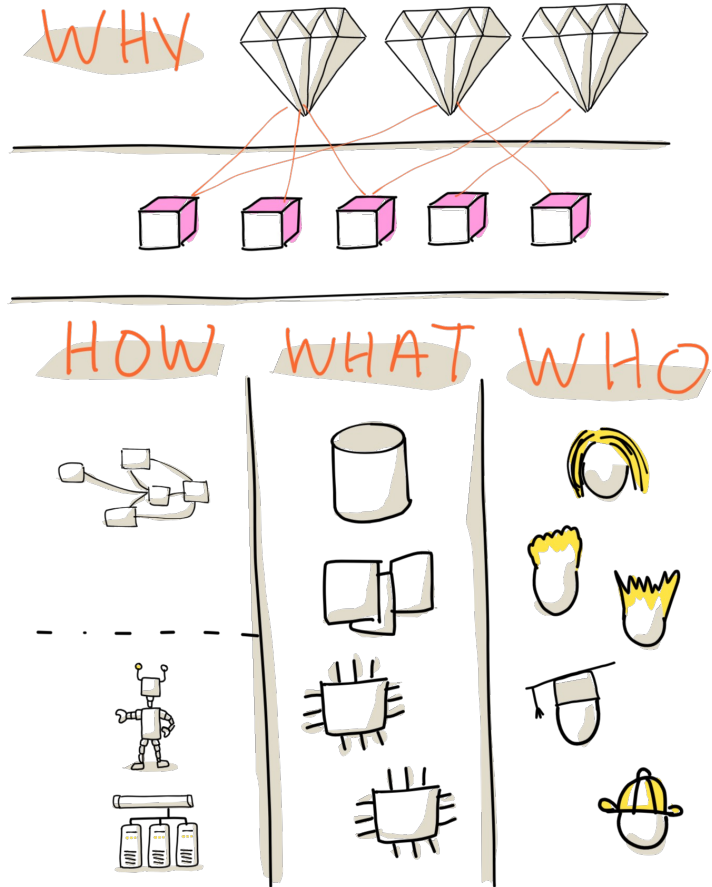
STRATEGY + ROADMAP

	1	2	3	4	5
CRIT 1		●			
CRIT 2			●		
CRIT 3	●				
CRIT 4				●	






DATA MATURITY



From Product Thinking to Strategic Thinking

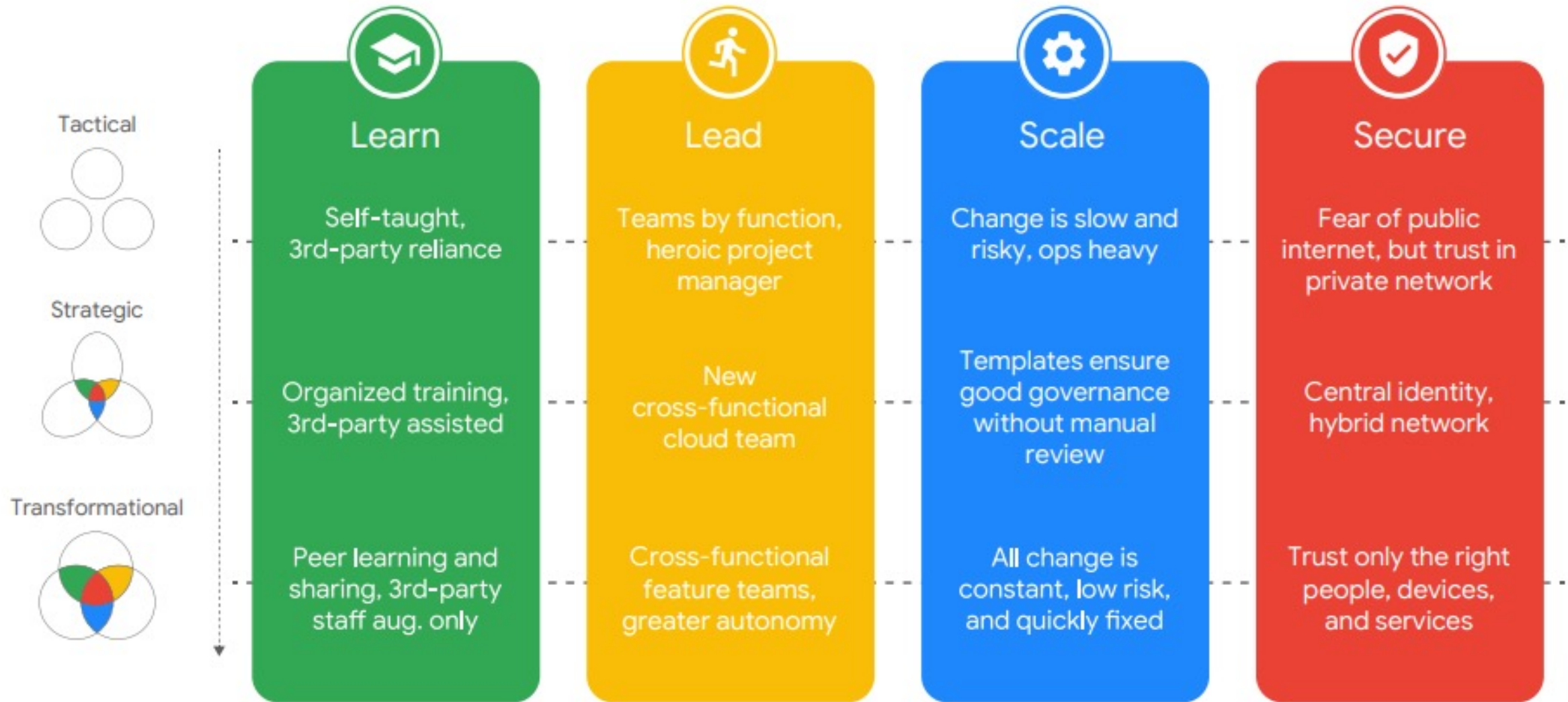


A blueprint for data literacy

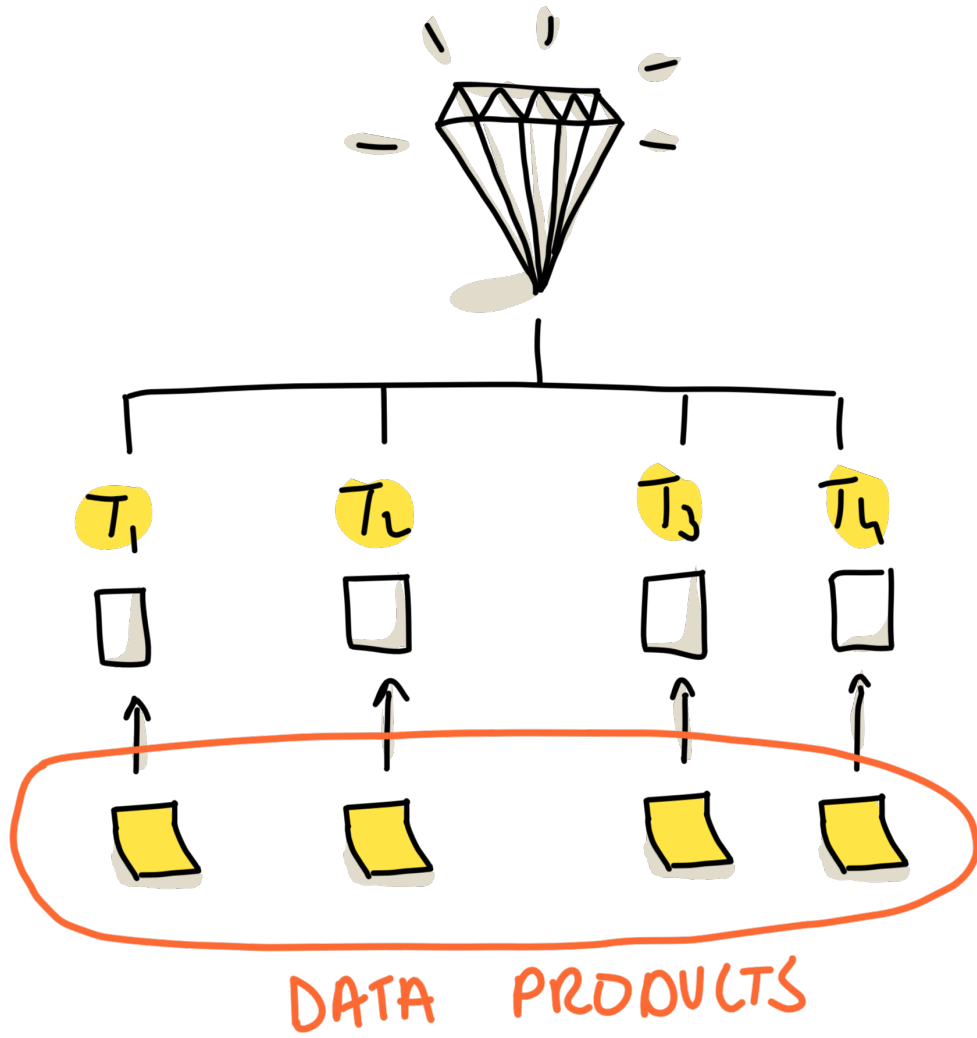
	 Infrastructure	 People	 Tools	 Organization	 Processes
Data Reactive	<p>Data is collected on ad-hoc basis with disparate tools, and there is no trusted, centralized data storage</p> <ul style="list-style-type: none"> ✓ Develop a data infrastructure strategy 	<p>No one has the skills to work with data — the organization does not have a data culture</p> <ul style="list-style-type: none"> ✓ Prove value with proof of concept ✓ Build executive support ✓ Center learning around data strategy ✓ Invest in data infrastructure talent 	<p>Variety of ad-hoc legacy tools to work with data that are rarely leveraged</p> <ul style="list-style-type: none"> ✓ Prioritize infrastructure, and people 	<p>There is no data team, and no data strategy to support it</p> <ul style="list-style-type: none"> ✓ Prioritize infrastructure, and people 	<p>Any data work is done on ad-hoc basis and there are no processes for working with data</p> <ul style="list-style-type: none"> ✓ Prioritize infrastructure, and people
Data Scaling	<p>Only a few key experts understand how data is accessed in the organization. There is no organization-wide access to, or trust in data</p> <ul style="list-style-type: none"> ✓ Centralize data storage ✓ Establish data governance and quality policy ✓ Define data access for high-impact teams 	<p>Minimal data culture with very few people believing in the importance of data or having the skills to work with data</p> <ul style="list-style-type: none"> ✓ Reward change agents and champions ✓ Define and outline data culture ✓ Set the stage for organization-wide upskilling 	<p>Mostly legacy tools with ad-hoc use of modern tooling</p> <ul style="list-style-type: none"> ✓ Provide access to inclusive modern tooling ✓ Align tooling with infrastructure strategy 	<p>Data strategy in place — with no centralized data team or embedded expertise set in motion</p> <ul style="list-style-type: none"> ✓ Define data team organizational model 	<p>Very few, limited data processes exist in siloed teams</p> <ul style="list-style-type: none"> ✓ Define data team processes with other business units
Data Progressive	<p>Data is accessible, and data infrastructure is maturing. However, data is not easily discoverable, compliant, understood, or actionable.</p> <ul style="list-style-type: none"> ✓ Democratize data access with data discoverability and management tools ✓ Strengthen data quality and operationalize data 	<p>Data is strategic, but underutilized throughout the organization. Organization-wide data literacy is lacking, and data upskilling is still limited.</p> <ul style="list-style-type: none"> ✓ Roll out organization-wide data upskilling fit for all data personas ✓ Assess, track, & reward skill development ✓ Start innovating with data 	<p>Modern tooling accessible, however limited data democratization hinders value.</p> <ul style="list-style-type: none"> ✓ Develop frameworks to democratize data and lower barrier to entry to working with tools 	<p>Data team set in place, however impact is limited to requests and analysis</p> <ul style="list-style-type: none"> ✓ Develop a hybrid model of embedded and centralized, to drive data strategy and expand value 	<p>Mature data processes for high data competency teams only</p> <ul style="list-style-type: none"> ✓ Develop scalable data processes through organization by centralizing shared insights, promoting collaboration, and lowering barrier to entry
Data Literate	<p>Data is collected, discoverable, reliable, understood, compliant, and actionable throughout the organization</p> <ul style="list-style-type: none"> ✓ Innovate and automate infrastructure processes ✓ Monitor data products in production 	<p>Everyone has the skills necessary to work with, and understand data. Continuous learning is part of the data culture.</p> <ul style="list-style-type: none"> ✓ Keep learning central to organization success ✓ Measure skill matrix development through organization 	<p>Modern tooling and frameworks enable higher adoption and easier data driven decision-making</p> <ul style="list-style-type: none"> ✓ Refine frameworks and contribute to open-source community ✓ Invest in collaborative tooling 	<p>Organizational model for scalable data science</p> <ul style="list-style-type: none"> ✓ Refine ad-hoc organizational models and enable further democratization 	<p>Data processes to scale collaboration and efficiency</p> <ul style="list-style-type: none"> ✓ Center collaboration at the heart of data processes

Level	Structured Processes	Discoverability & Compounding	Analytical Speed & Agility	Breadth & Depth of Impact	Organizational Cohesion
1 Ad Hoc Exploration	Practitioners operate autonomously in a black box	Assets stored locally, emailed around	Limited talent and tools	Ivory tower, no tangible value	Analytics island, purely transactional
2 Repeating, but Limited	Recurring workflows discussed, no enforcement	Assets stored centrally, but lack metadata / permissions	Some tools and talent investment	Static reports in a few business areas	Some collaboration with line managers
3 Defined and Controlled	Formalized process, manually enforced	Assets stored and tagged centrally with metadata and permissions	Ideas rapidly tested with novel methods / tools	Results translated into multiple operational workflows	Analytics are key stakeholders in strategic decisions
4 Optimized and Automated	Best practices codified into infrastructure, transparency for all	All asset versions stored / tagged, searchable, reproducible	Cutting-edge tools, comfortable at the analytical frontier	Data products drive org with robust safeguards	Analytics enmeshed in business; proactively anticipate needs





EXERCISE: ESTIMATE YOUR DATA MATURITY



DATA MATURITY

	1	2	3	4	5
CRIT 1		●			
CRIT 2			●		
CRIT 3	●				
CRIT 4				●	

1. PEOPLE

Ungoverned	Stage 1	Stage 2	Fully Governed
No roles & responsibility	Limited roles & responsibility	Roles & responsibilities defined	Roles & responsibilities defined
No data owners accountable for data	No data owners accountable for data	Some data owners in place	All data owners in place
No data stewards appointed with responsibility for data quality	Some data stewards in place for data quality, but scope too broad, like whole department	Data stewards in place and assigned to data governance working groups for specific data	Data stewards in place assigned to data governance working groups for specific data
No one accountable for data privacy	No one accountable for data privacy	Chief privacy officer accountable for privacy, no tools	Chief privacy officer accountable for privacy with tools
No subject-matter experts (SMEs) identified for data entities	Some SMEs identified, but not engaged	SMEs identified and in data governance working groups	SMEs identified and in data governance working groups

2. PROCESS

Ungoverned	Stage 1	Stage 2	Fully Governed
No common business vocabulary	Common business vocabulary begun in a glossary	Common business vocabulary established	Common business vocabulary complete and maintained
No way to know data location, quality or sensitivity	Data catalog on some systems	Data catalog on all structured data	Data catalog on structured and unstructured data in all systems
No process to monitor data quality, data privacy, or data access security	Some ability to monitor data quality, some ability to monitor privacy, such as queries	Monitoring and stewardship of data quality and data privacy on core systems	Monitoring and stewardship of data quality and data privacy on all systems
No way to know whether a policy violation occurs or process if one occurs	Data access security violation detection in some systems	Data access security violation detection in all systems	Data access security violation detection in all systems
No availability of fully trusted data assets	Development started for a small set of trusted data assets using data fabric software	Several core trusted data assets created using data fabric	Continuous delivery of trusted data assets through enterprise data marketplace

3. POLICIES

Ungoverned	Stage 1	Stage 2	Fully Governed
No data governance classification schemes on confidentiality and retention	Data governance classification scheme for confidentiality	Data governance classification scheme for both confidentiality and retention	Data governance classification scheme for both confidentiality and retention
No policies and rules to govern data quality	Policies and rules to govern data quality begun in common vocabulary in business glossary	Policies and rules to govern data quality defined in common vocabulary in catalog business glossary	Policies and rules to govern data quality defined in common vocabulary in catalog business glossary
No policies and rules to govern data retention	Some policies and rules to govern data retention	Policies and rules to govern data retention consolidated in the data catalog using a classification scheme	Policies and rules to govern data retention consolidated in the data catalog using classification schemes and enforced everywhere

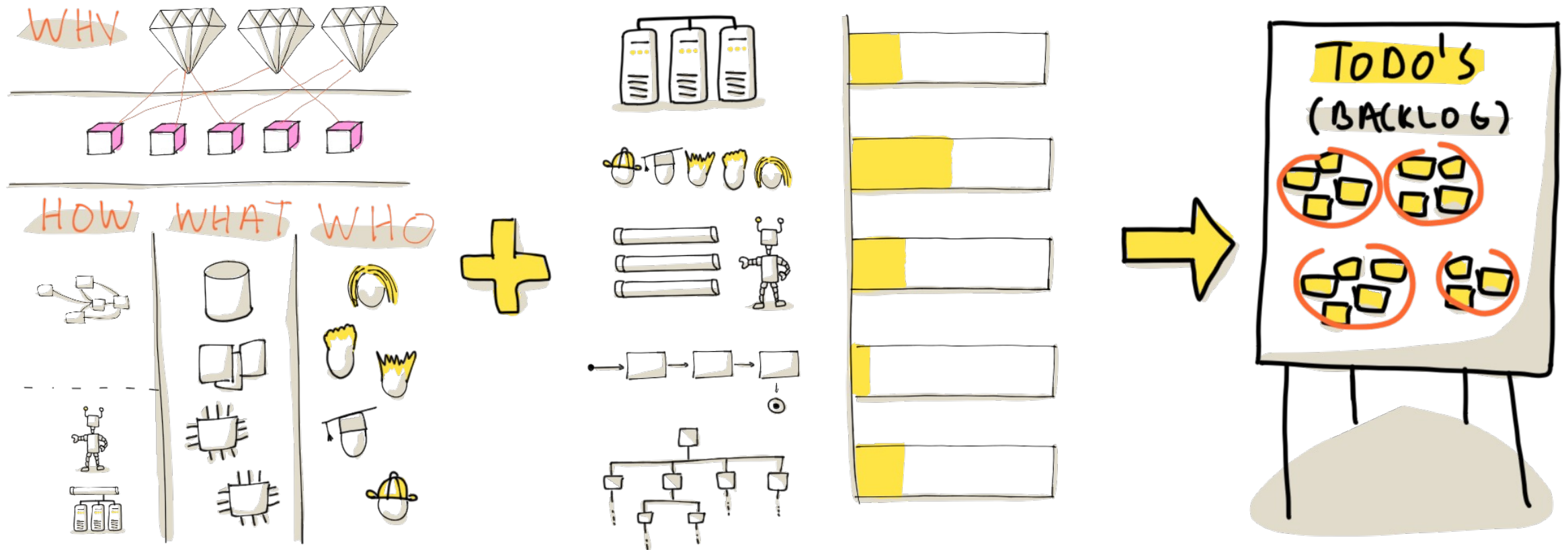


4. TECHNOLOGY

Ungoverned	Stage 1	Stage 2	Fully Governed
No data catalog with auto data discovery, profiling, and sensitive data detection	Data catalog with auto data discovery, profiling, and sensitive data detection purchased	Data catalog with auto data discovery, profiling, and sensitive data detection purchased	Data catalog with auto data discovery, profiling, and sensitive data detection purchased
No metadata lineage	Metadata lineage available in data catalog on trusted assets		
No data access security tool	Data access security in multiple technologies		Data access security enforced in all systems
No data platform software with cloud and datacenter connectivity	Data platform software with cloud and datacenter connectivity and catalog integration purchased		



From Product Thinking to Strategic Thinking to a Data Roadmap



EXERCISE: DATA BACKLOG

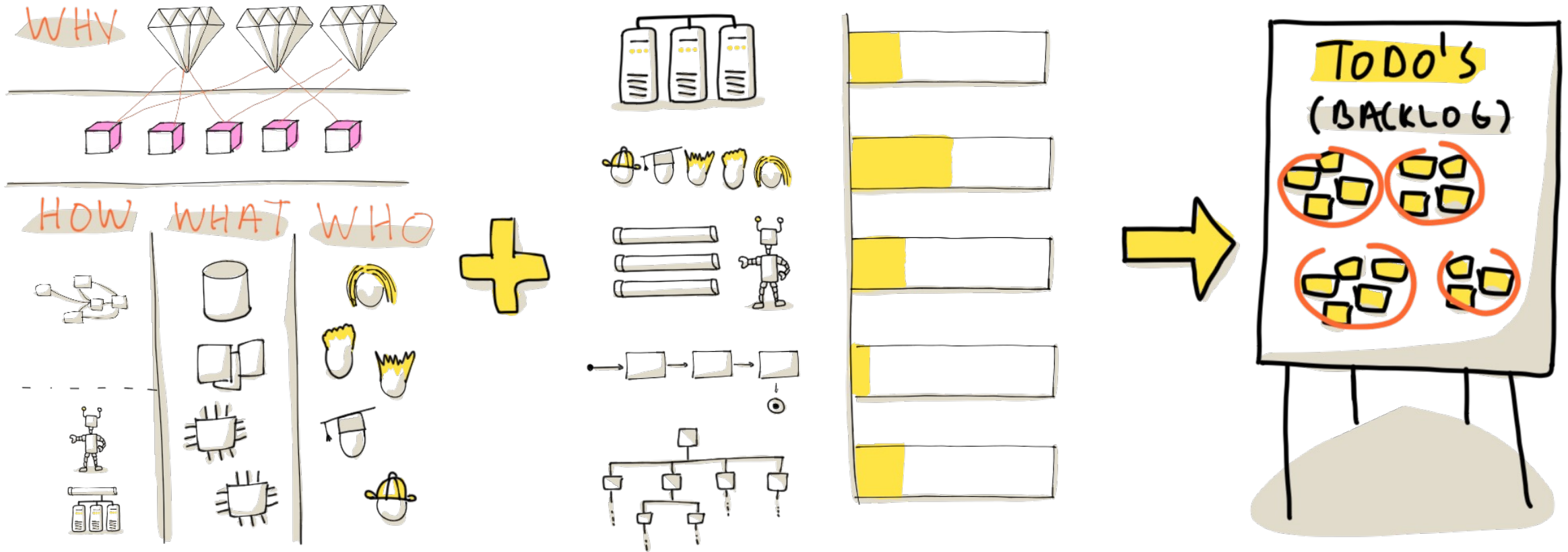
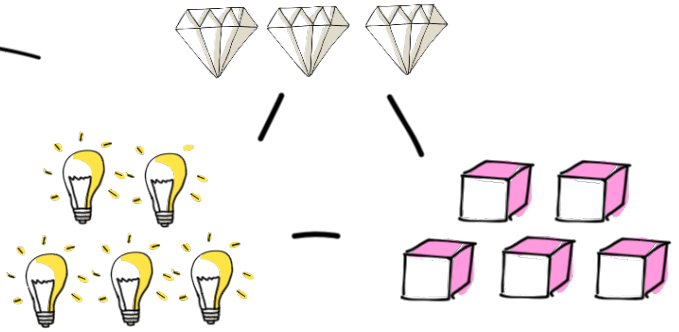
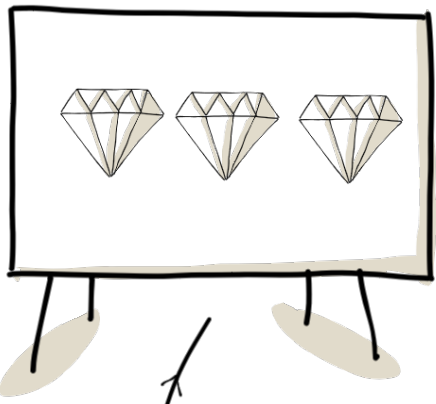


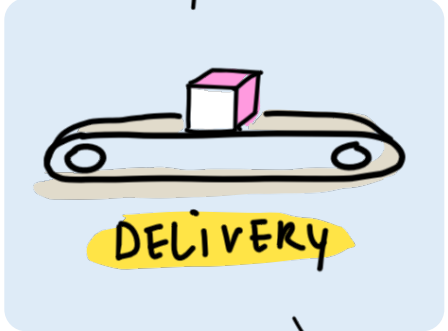
TABLE OF CONTENTS

BUSINESS STRATEGY



DATA PRODUCT

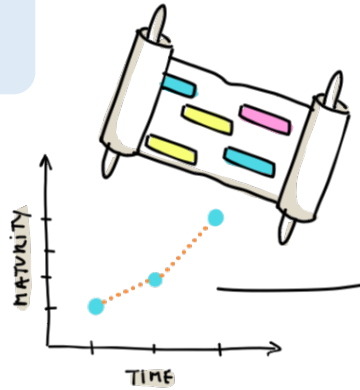
WHY-WHAT-WHO-HOW



DELIVERY

	1	2	3	4	5
CRIT 1		●			
CRIT 2			●		
CRIT 3	●				
CRIT 4				●	

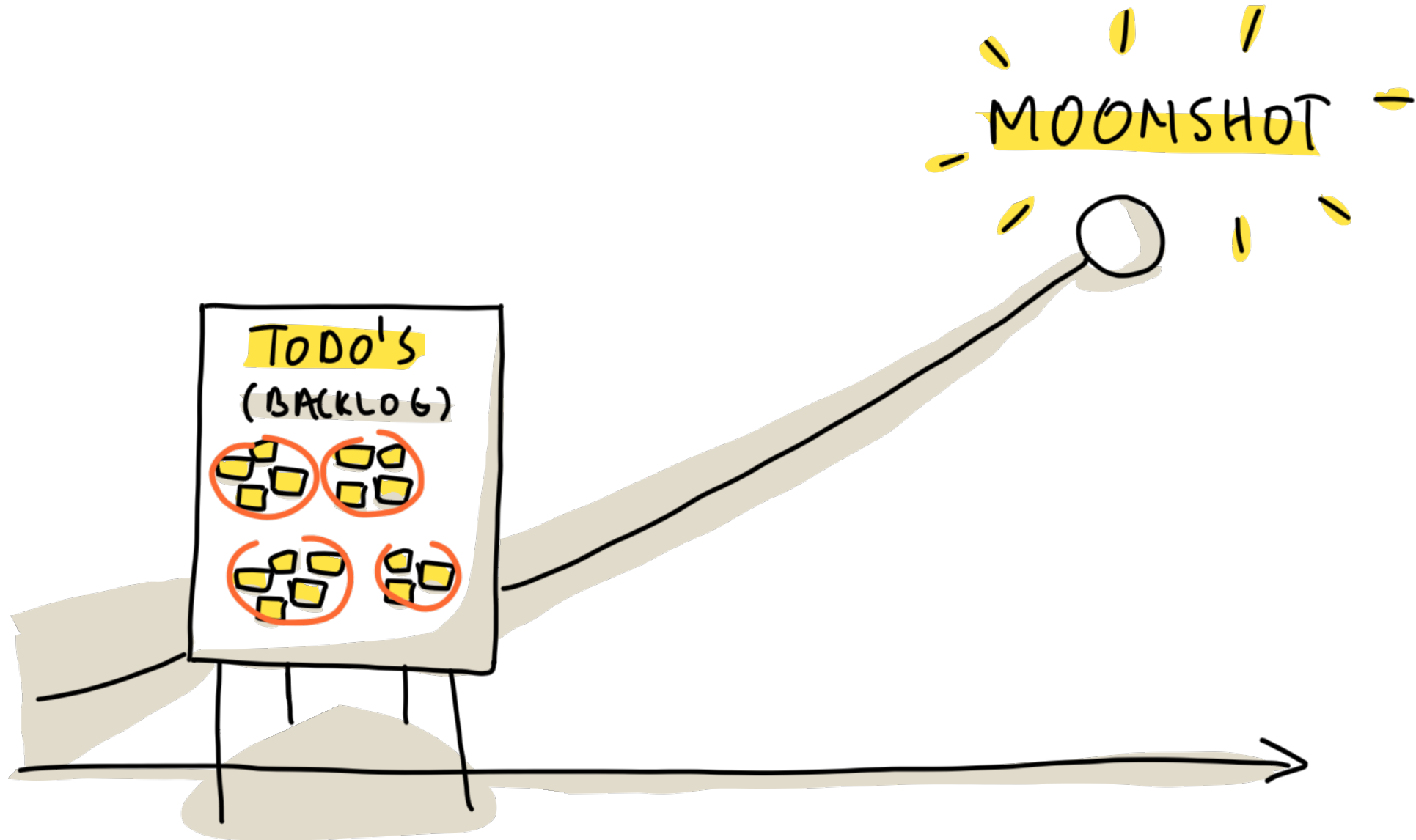
DATA MATURITY

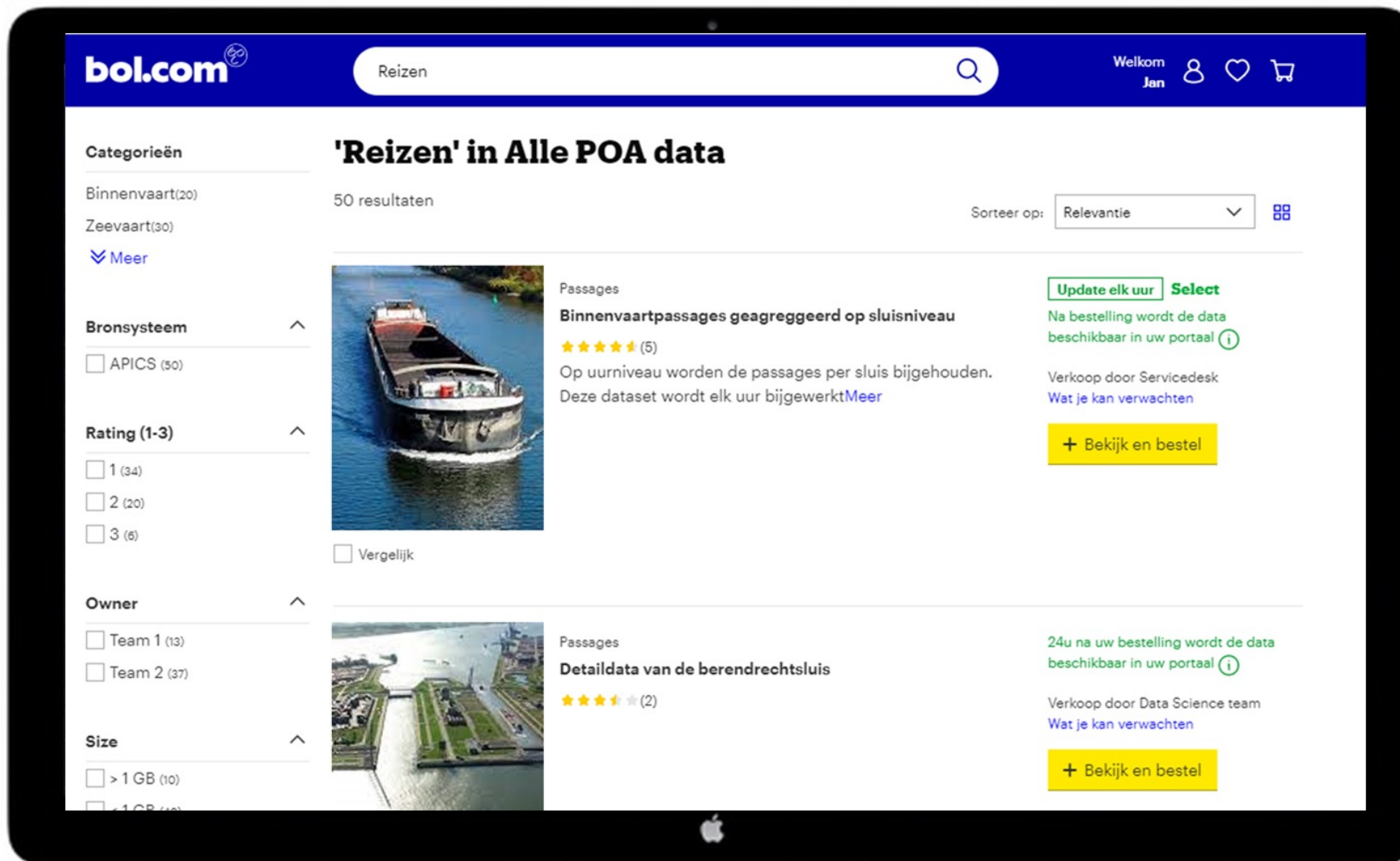


STRATEGY + ROADMAP



Make it Sellable





Ambitie:

Het dataplatform wordt **POA's Bol.Com shop voor data!**

OUR DATA MISSION

We will...

...define centrally governed **data handling standards**

...introduce cross-functional **MDM**, based on a **single source of truth**

...ensure good **Data Quality** through measurement and improvement initiatives

...work with **business people** to turn data into information

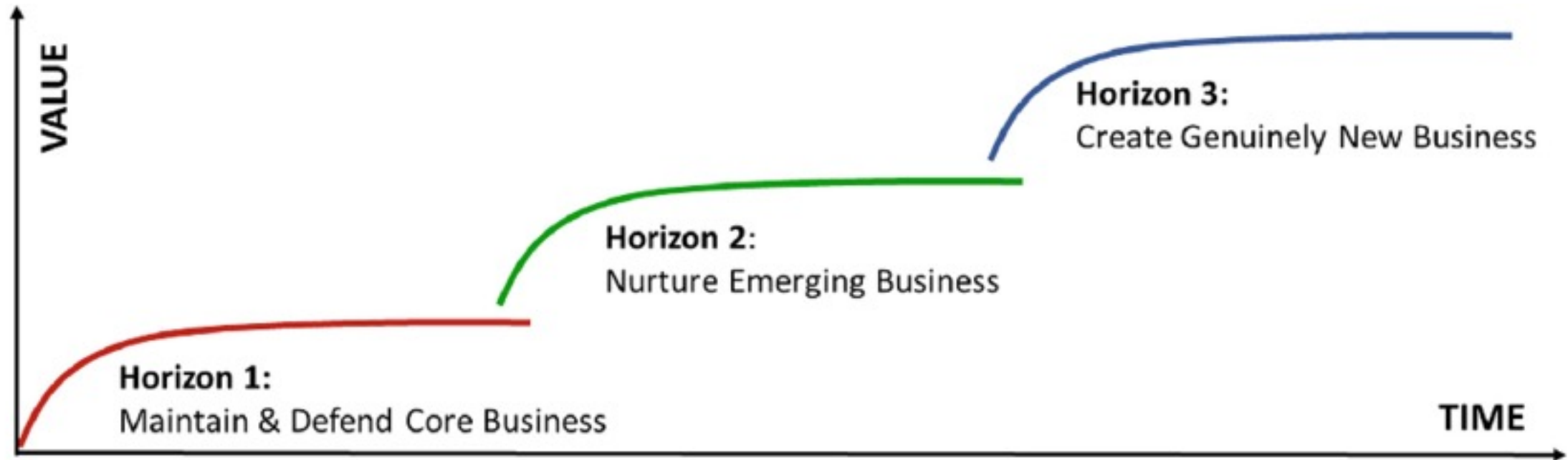
...have all of this supported by the right **toolset**

...implement adequate **ethical standards** in dealing with data

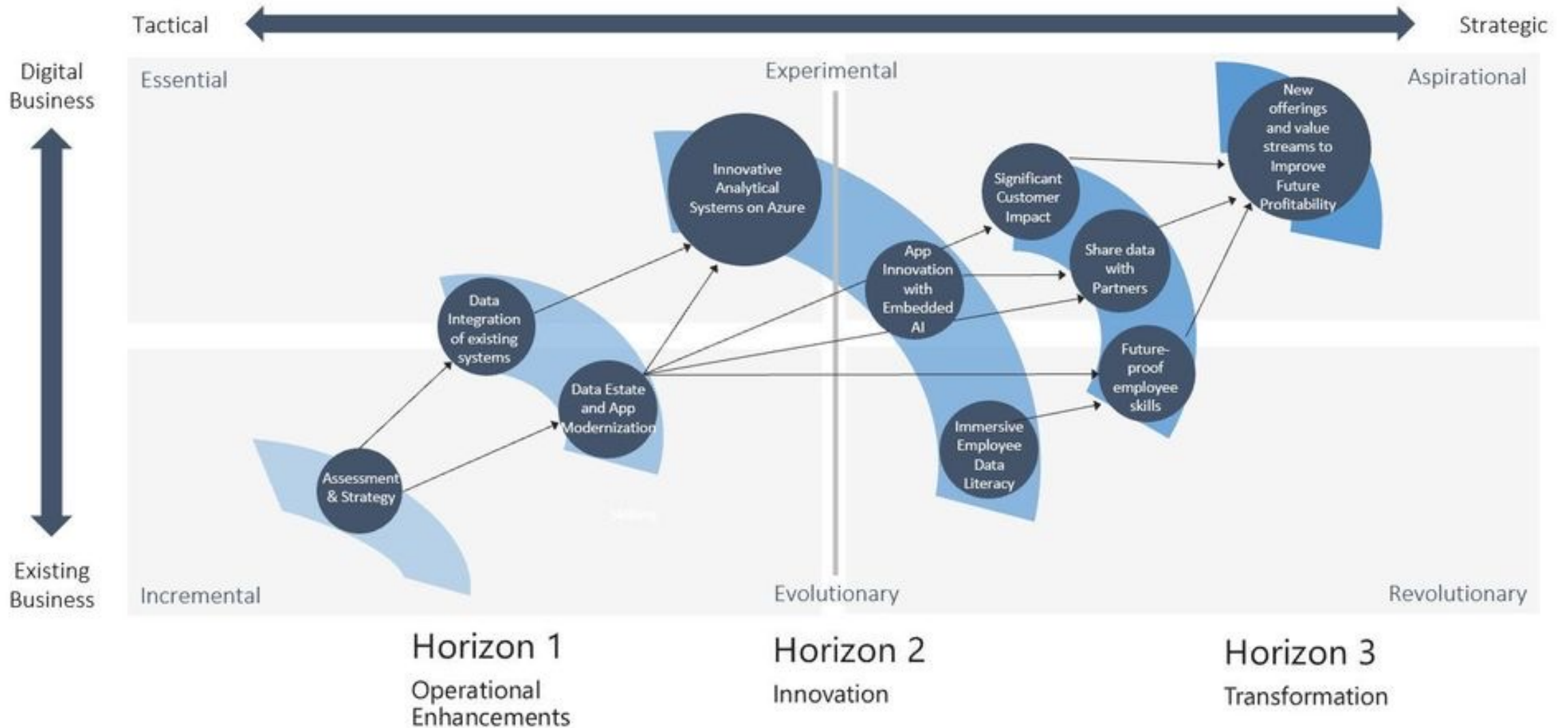
...**train** and **connect** all entities in data matters



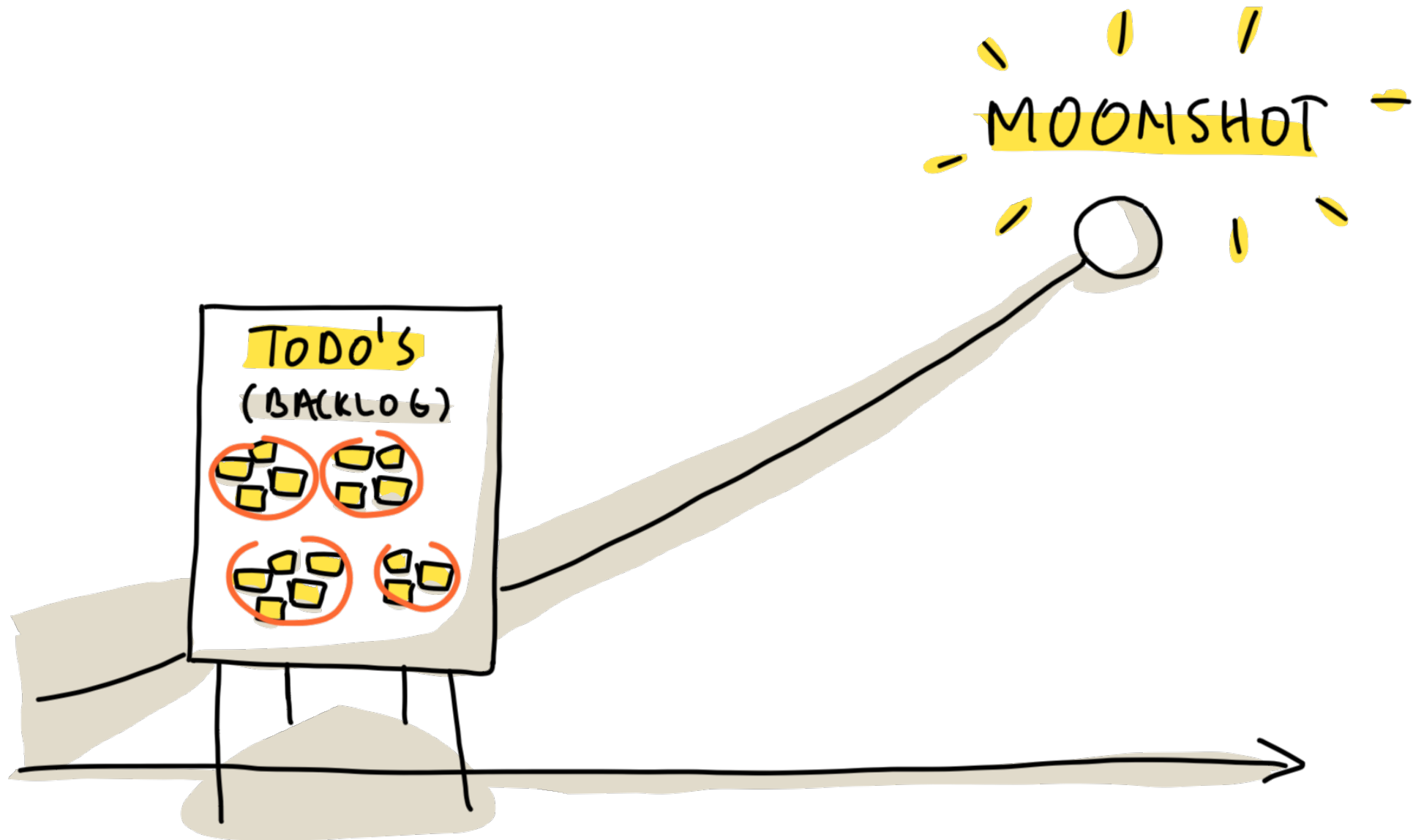
Three Horizons Framework



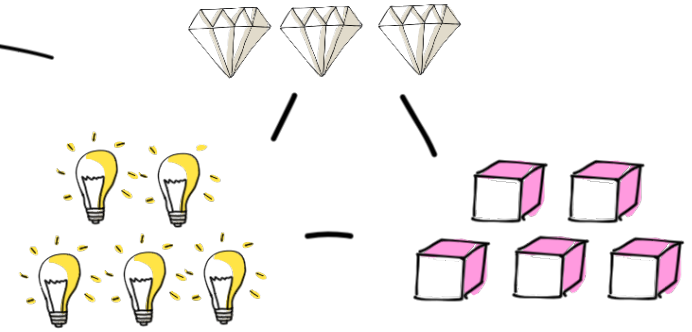
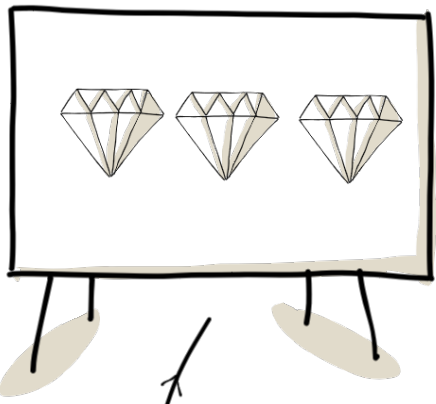
Horizons of Transformation with Azure Data & AI



EXERCISE: CAN YOU TRY TO MAKE A "MOONSHOT"?

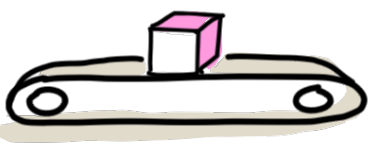


BUSINESS STRATEGY



DATA PRODUCT

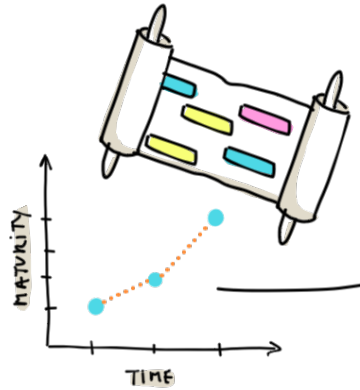
WHY-WHAT-WHO-HOW



DELIVERY

	1	2	3	4	5
CRIT 1		●			
CRIT 2			●		
CRIT 3	●				
CRIT 4				●	

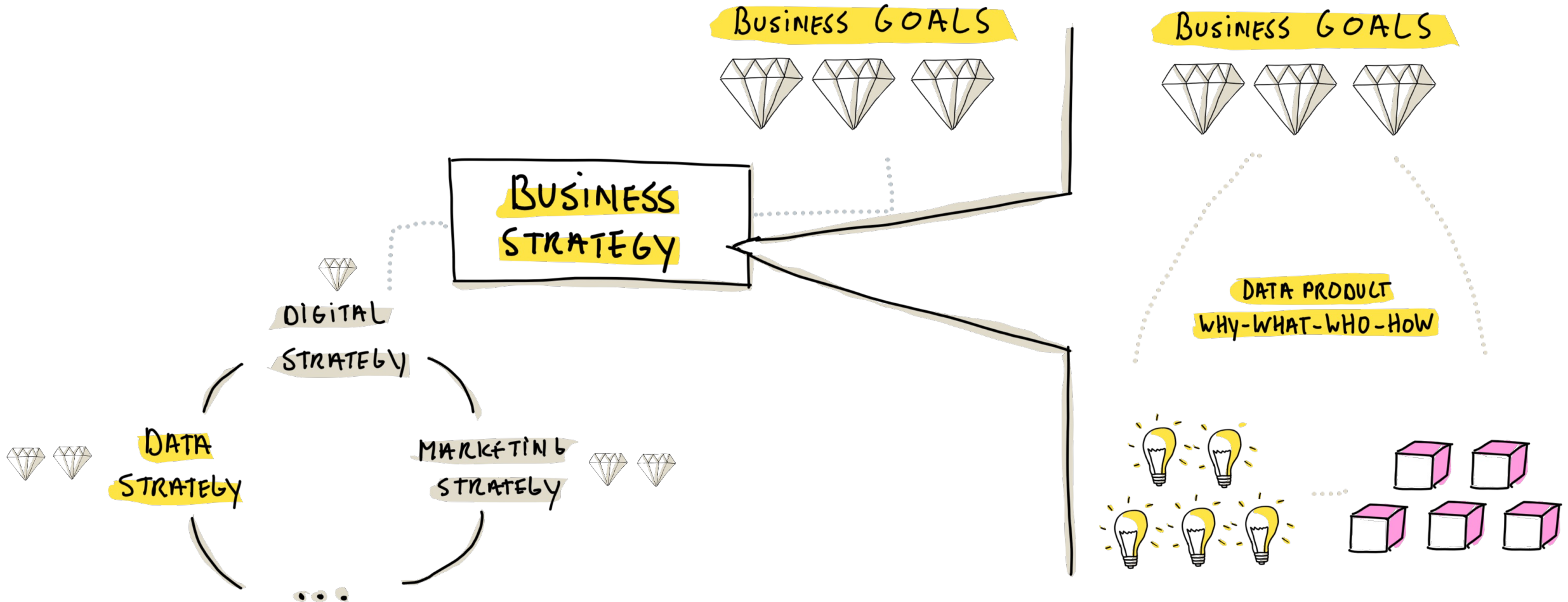
DATA MATURITY



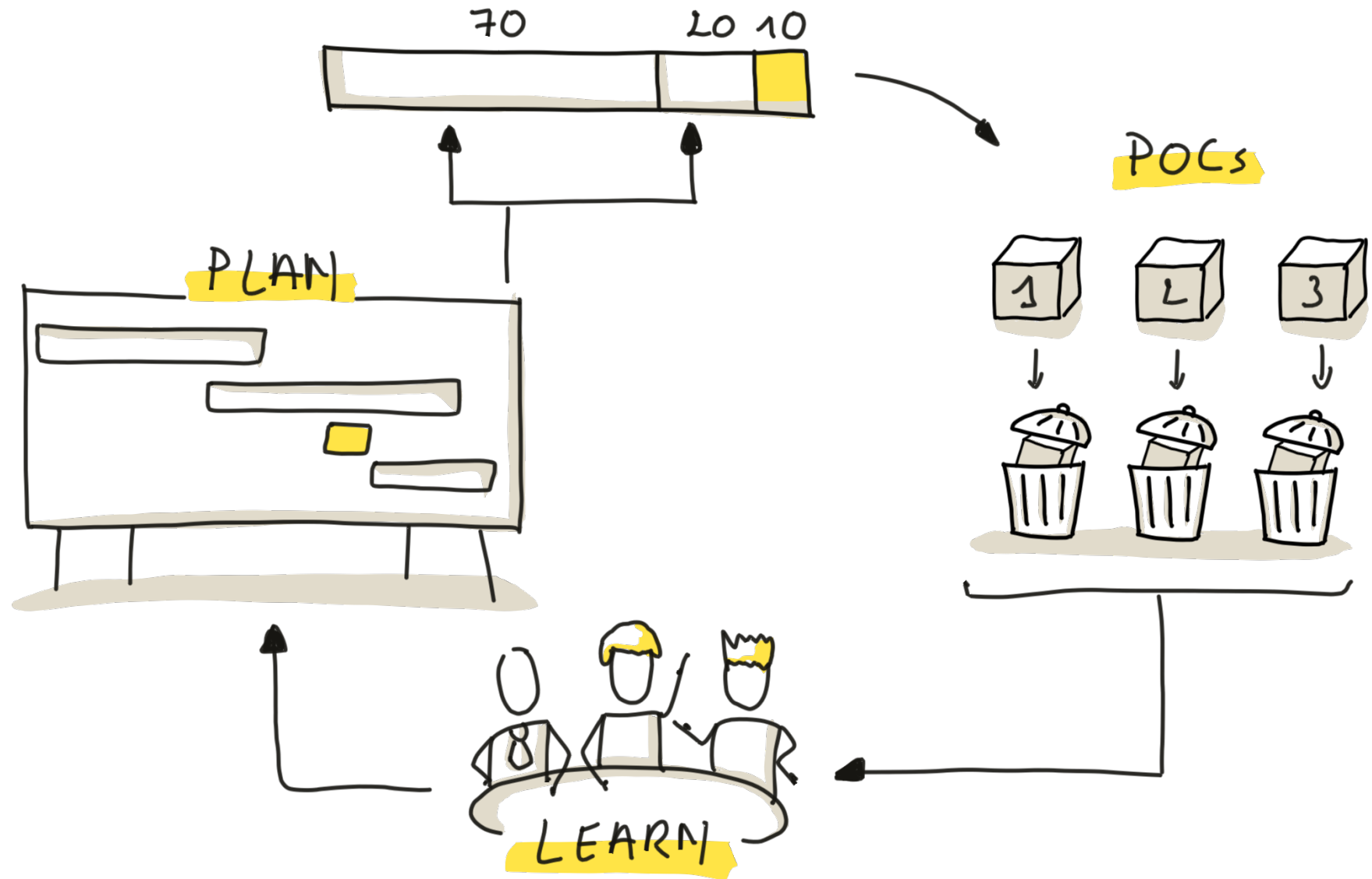
STRATEGY + ROADMAP



Prioritize Pragmatism over Bureaucracy



Create Room for Innovation



Eat Your Own Dogfood



"LAUNCH A NEW PRODUCT
WITH 100 CUSTOMERS"



kpi



"REDUCE CUSTOMER
CHURN $\leq 5\%$ "



kpi



Culture eats Strategy for Breakfast

