DATA STRATEGIE

Jan Meskens

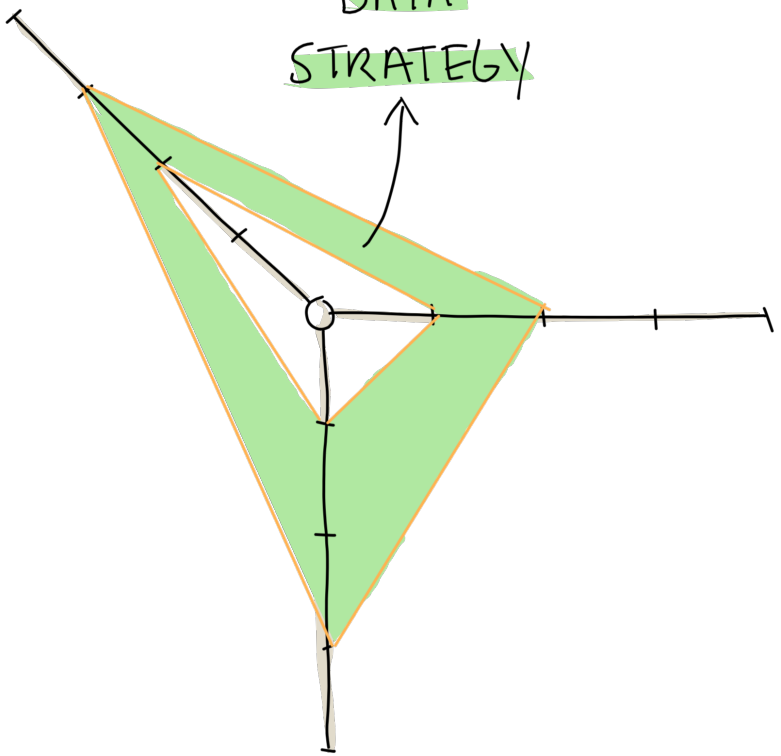
11 – 2024

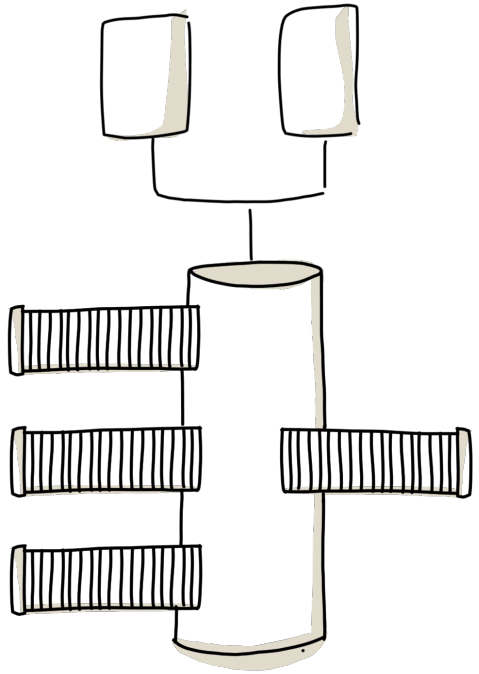
7.

SOLUTIONS

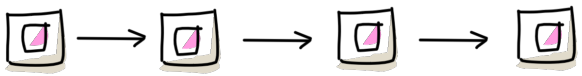
PEOPLE

DATA
STRATEGY





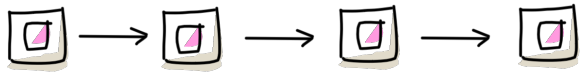
TECHNOLOGY



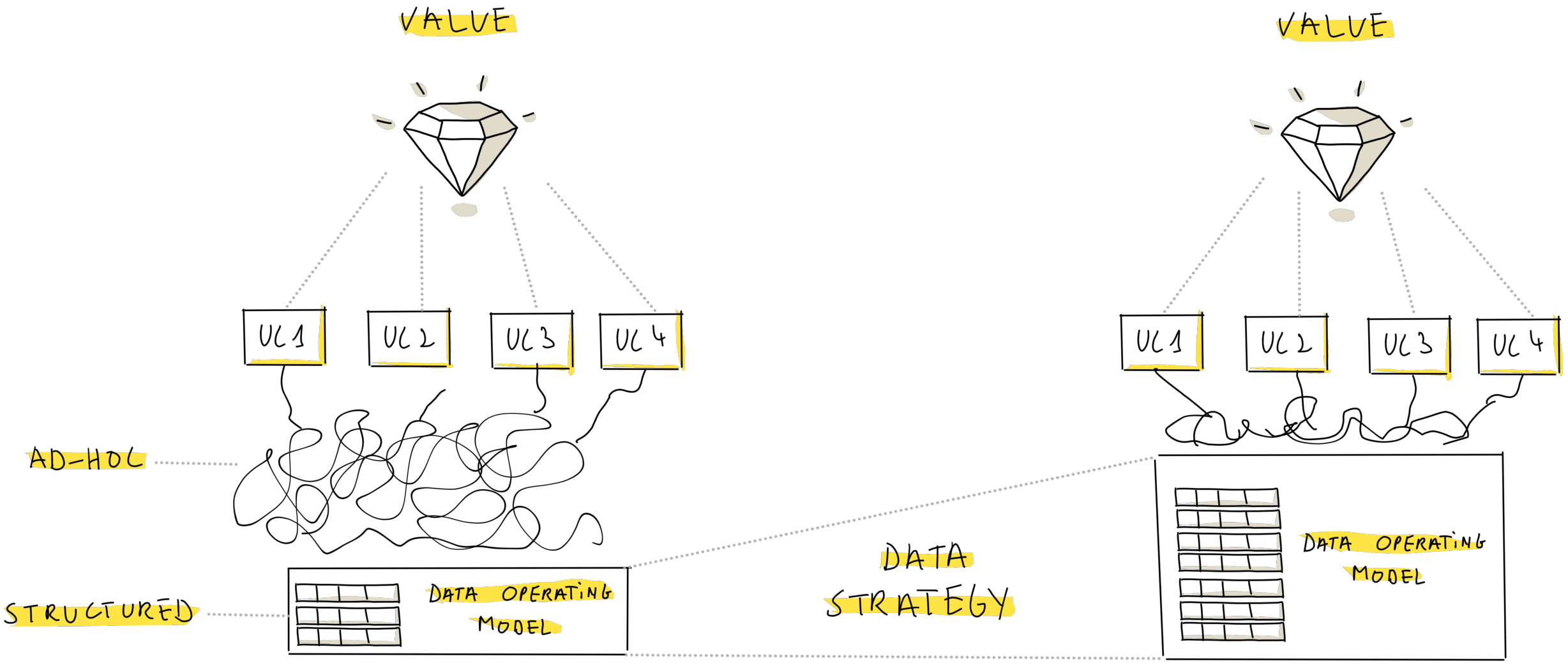
PROCESS

PEOPLE

Which **capabilities** (on People / Process / Technology level) do we miss to realize the **intended business value** by means of the **proposed use cases**?

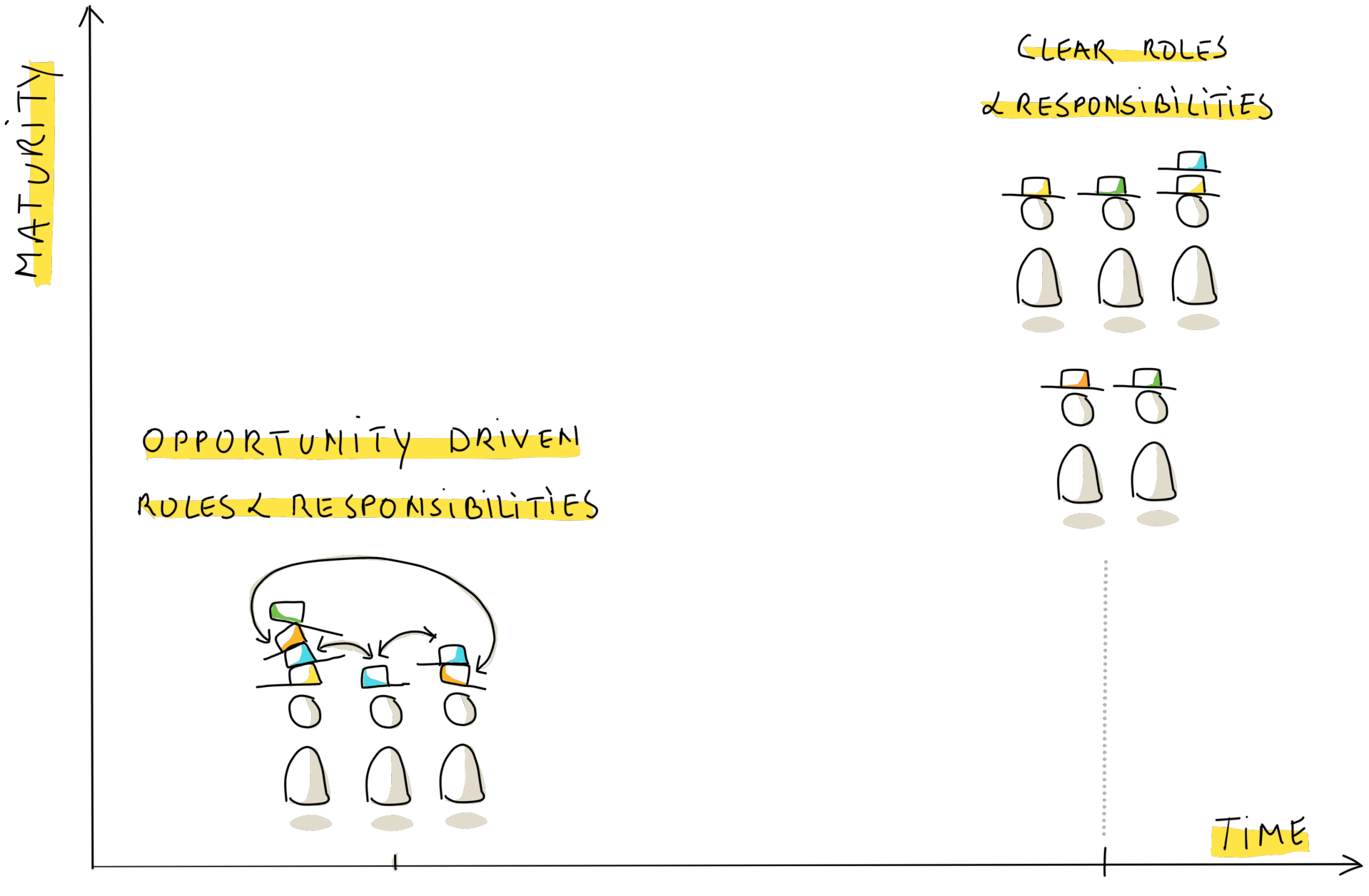


PROCESS


VALUE
UC 1
UC 2
UC 3
UC 4
AD-HOC
STRUCTURED
DATA OPERATING MODEL
DATA STRATEGY
VALUE
UC 1
UC 2
UC 3
UC 4
DATA OPERATING MODEL

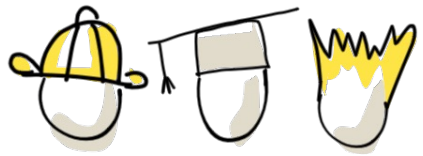
7.1

PEOPLE & ORGANIZATION

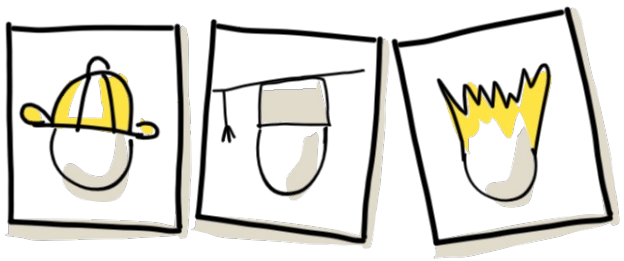

MATURITY
CLEAR ROLES
& RESPONSIBILITIES
OPPORTUNITY DRIVEN
ROLES & RESPONSIBILITIES
TIME

7.1

PEOPLE & ORGANIZATION

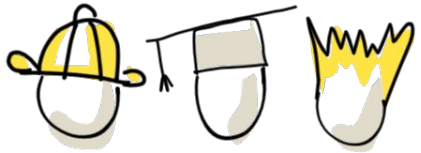


PEOPLE

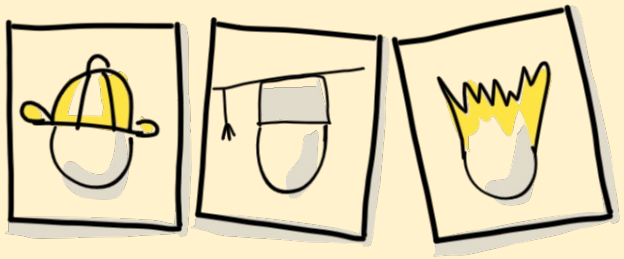


ROLES & RESPONSIBILITIES

DATA LITERACY

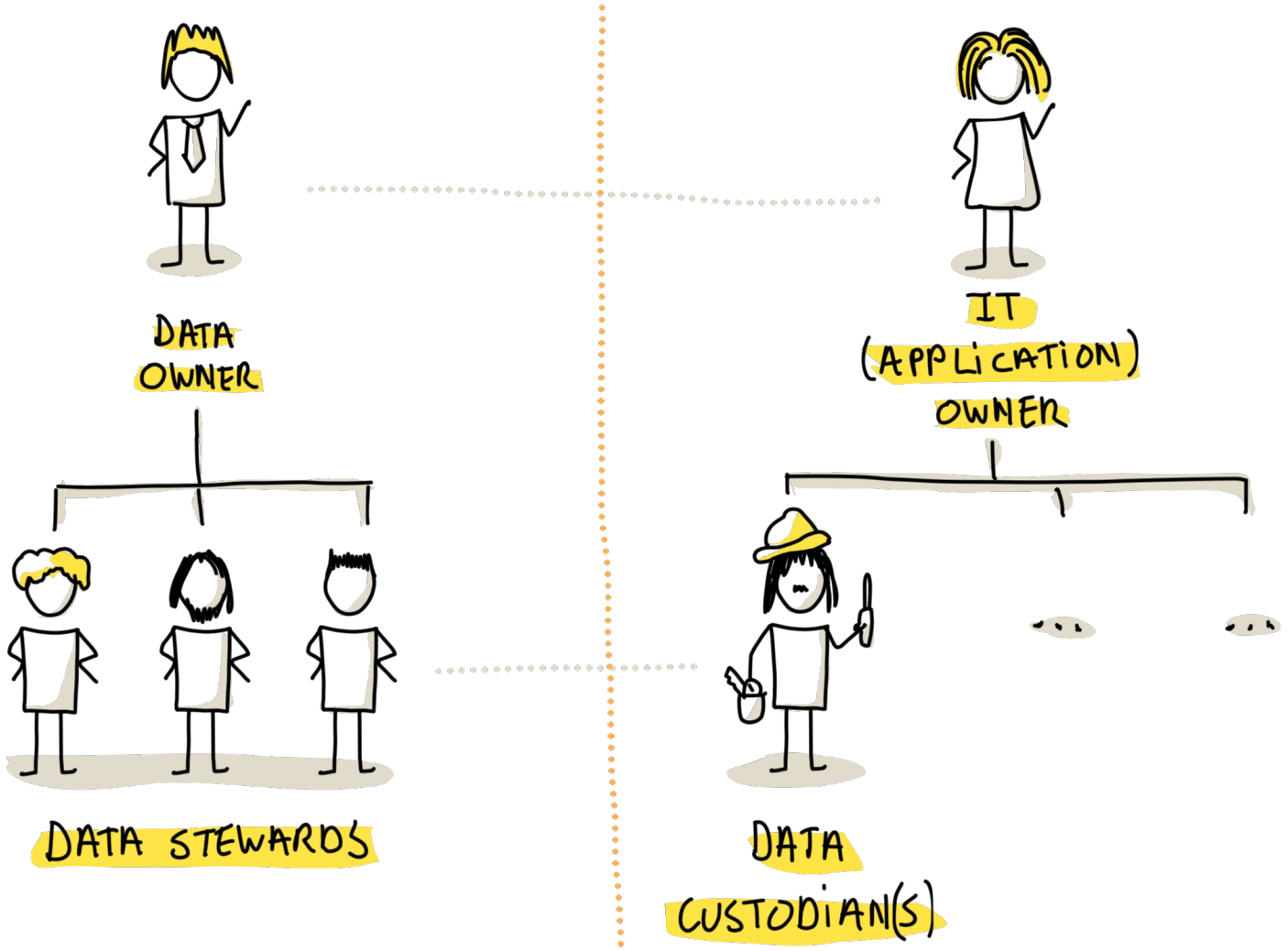


PEOPLE



ROLES & RESPONSIBILITIES

DATA LITERACY



Data Owner

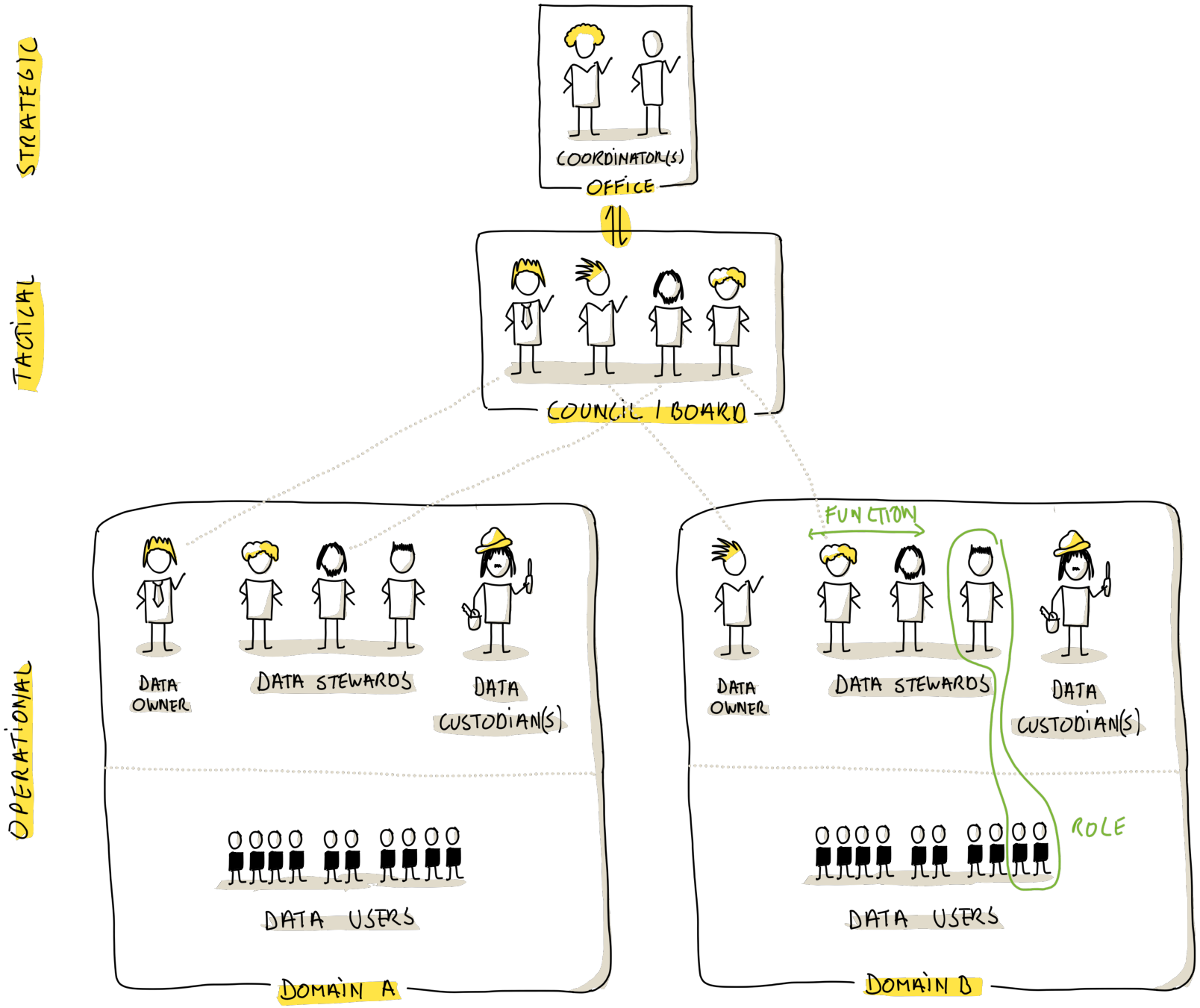
Accountable for the classification, protection, use, and quality of one or more data sets within an organization.

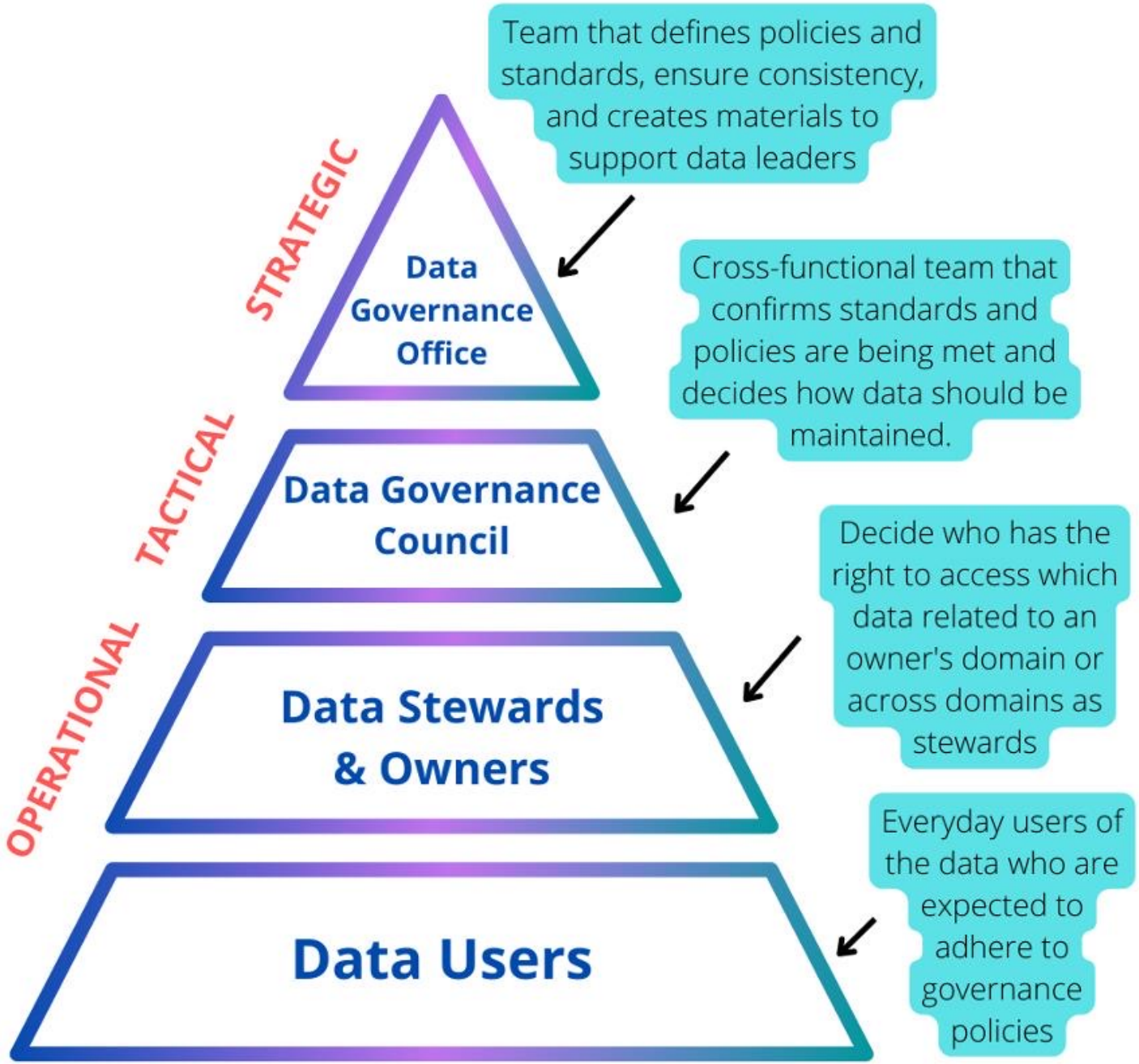
Data Steward

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.


STRATEGIC
TACTICAL
OPERATIONAL
COORDINATOR(S)
OFFICE
COUNCIL / BOARD
DATA OWNER
DATA STEWARDS
DATA CUSTODIAN(S)
DATA USERS
DOMAIN A
FUNCTION
DATA OWNER
DATA STEWARDS
DATA CUSTODIAN(S)
ROLE
DATA USERS
DOMAIN B



[Planning Data Governance: Approach & Best Practices]

Roles / Bodies	data domain owner	data owner	data governance coordinator	data stewards
DG board	Accountable for all data in his domain			
DG council		Takes care of the data, end-2-end oversight	Promoter of data governance, coordinates actions	
DG office				Data management activities

[5 Pillars of a Data Governance Framework]

Jan Meskens • You
Data Strategy Consultant | Speaking, sketching and writing about th...
6h •

What are the most common roles & responsibilities in Data? 👷 👩‍⚕️ 👮 👧 👷‍♂️

As I started drawing the landscape of data-related professions, I've identified a constellation of key roles. But I'm certain there's more to uncover!

So far, I've mapped out:

-Data Owners
-Data Stewards
-Data Custodians
-Data Analysts
-Data Engineers
-AI Engineers
-Data Scientists
-BI Analysts

This is just the beginning.

What other roles am I missing?

Kristof Bouckaert • 1st
Data & Analytics Solution Manager
5h •••

Data architect ? AI architect?

Like · 3 | Reply

Jérôme Huberty • 2nd
Global Data Stewardship Manager @Umicore
1h (edited) •••

(Business) Process owner ,
Data visualisation an
Data governance off
Master data Manage
Data quality Manage
Data Manager,
Integration tester,
Platform engineer ?

Like · 1 | Reply

Jan Uyttenhove • 1st
Lead data & platform architect
7m •••

Data Product owner?

Like · 1 | Reply

Sophie Angenot • 2nd
QuaData is your Agile Data Governance Coach
5h •••

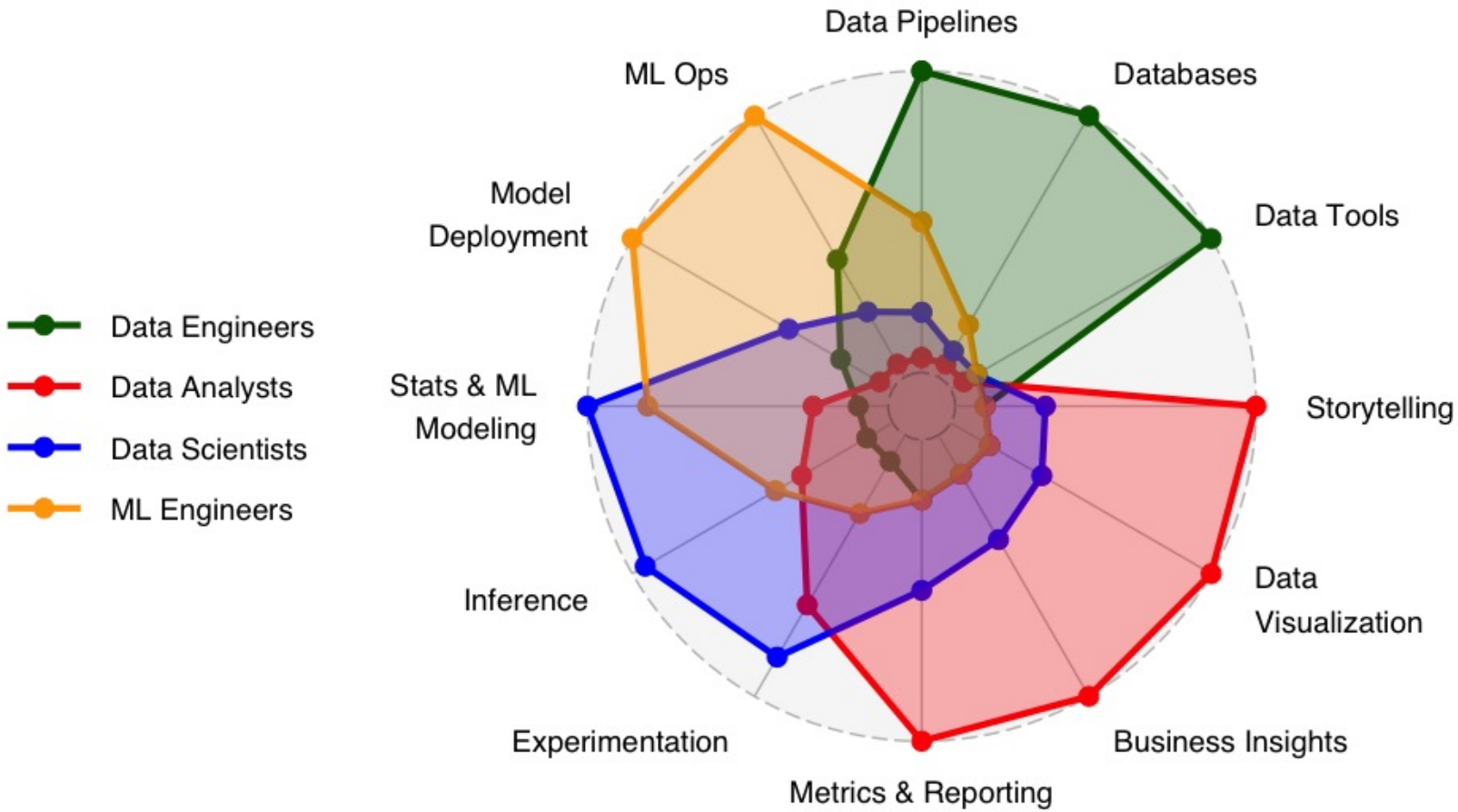
Chief data officer, Data Governance Lead, Data Protection Officer

Like · 4 | Reply

Sevil Oktem Demaerschalk • 2nd
Business Analyst | Process Analysis |Data Governance
4h •••

Data validators and data producers

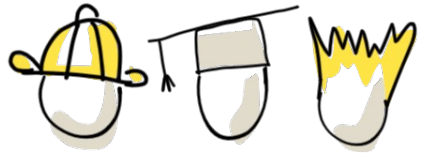
Like · 1 | Reply



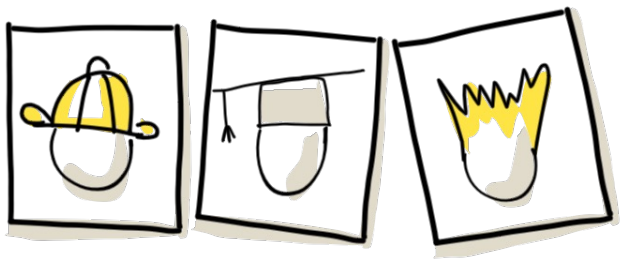
[A Guide to Data Roles]

	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

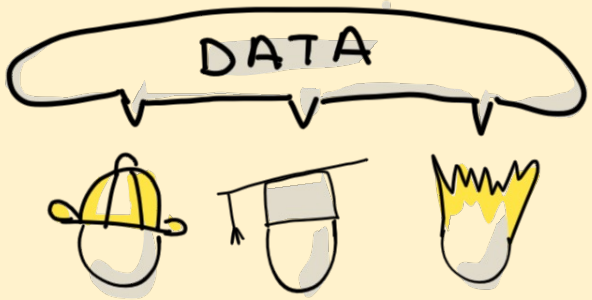
[Data Science Roles]



PEOPLE



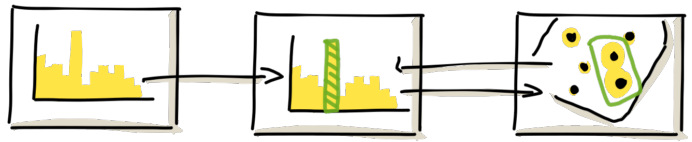
ROLES & RESPONSIBILITIES



DATA LITERACY

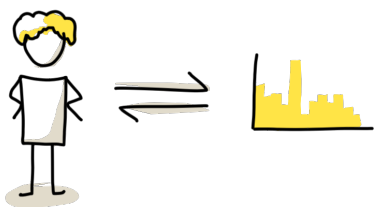
Data in the hands of a few data experts can be powerful, but data at the fingertips of many is what will be truly transformational.

[Brent Dykes]



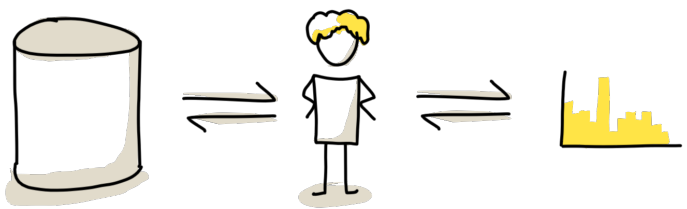
COMMUNICATING WITH DATA

DATA LITERACY



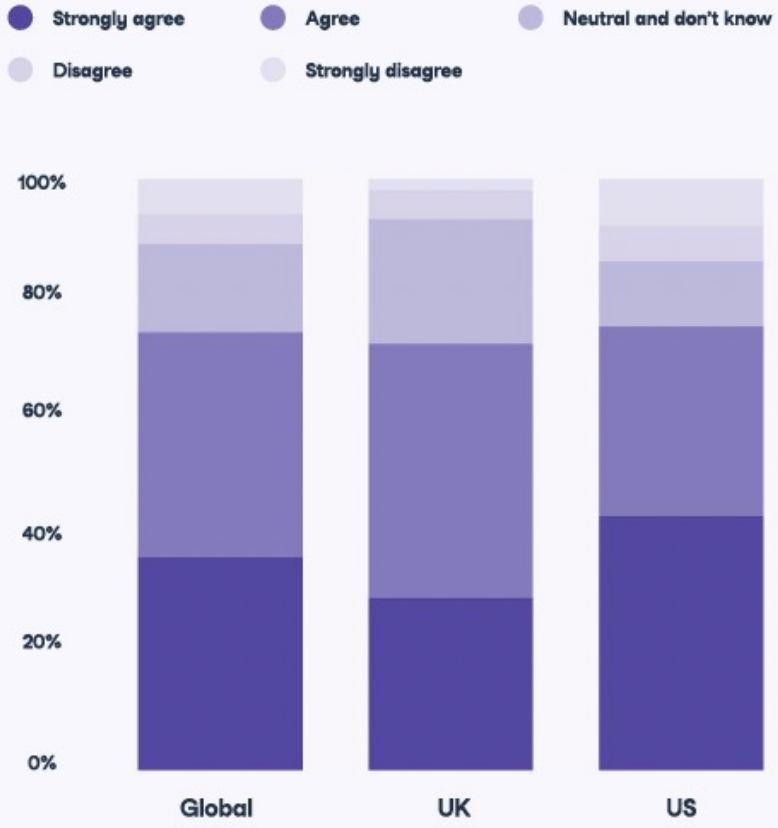
READING DATA

WORKING WITH DATA



NEARLY THREE QUARTERS OF LEADERS BELIEVE THOSE WITH SUFFICIENT DATA LITERACY SKILLS OUTPERFORM THOSE WITH INADEQUATE DATA SKILLS

Question asked: *"Do you agree or disagree with the following statement: The people in my organization with sufficient data literacy skills outperform those with inadequate data literacy skills"*



LEADERS POINT TO INACCURATE, AND SLOW DECISION-MAKING, AS THE BIGGEST RISKS FOR INADEQUATE DATA SKILLS ON THEIR TEAMS

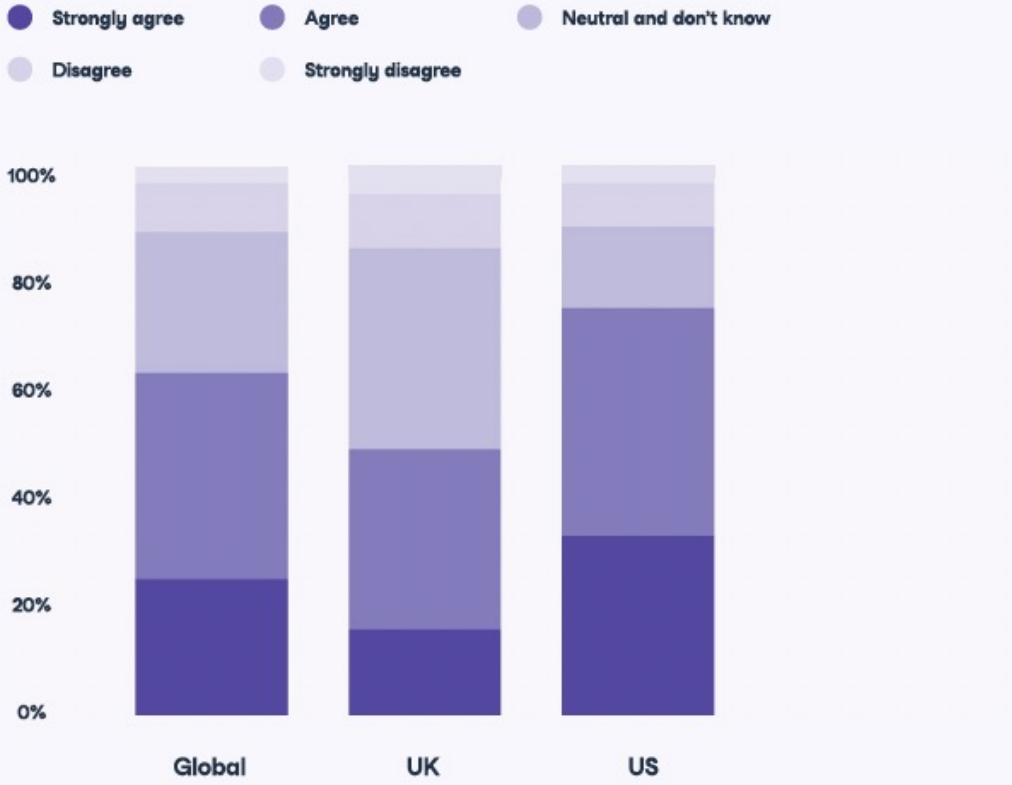
Question asked: *"What risks is your department or team facing if your people do not have adequate data skills?"*

EMPLOYEES WITH ADEQUATE DATA SKILLS OUTPERFORM THOSE WITH INSUFFICIENT DATA SKILLS ON A VARIETY OF DIMENSIONS

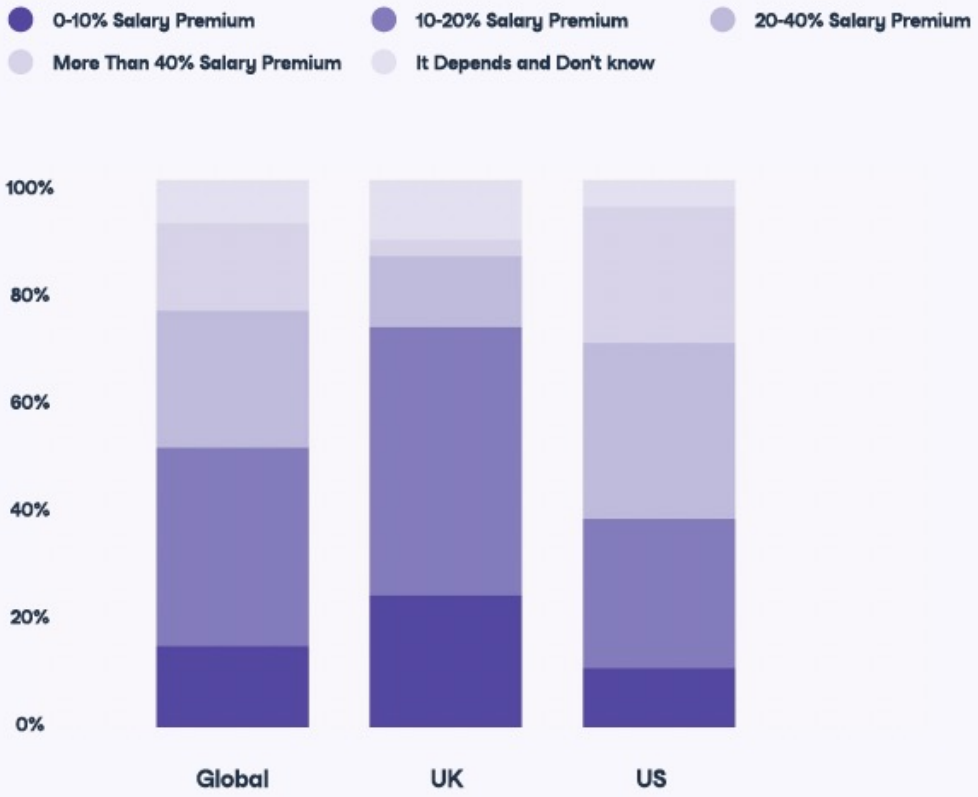
Question asked: *"What value do data literate employees provide over those with insufficient data skills? (Rank them by importance)"*

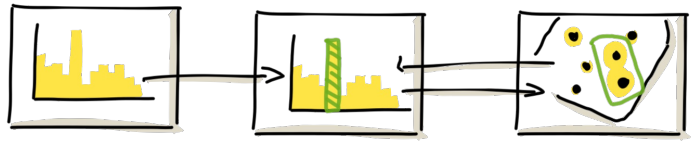
LEADERS ARE WILLING TO PAY A PREMIUM FOR EMPLOYEES WITH STRONG DATA LITERACY SKILLS

Question asked: *"Do you agree or disagree with the following statement: "When hiring someone new, I'm willing to pay a higher salary to a candidate who has good data literacy skills over a candidate who does not"*



Question asked: *"If you answered yes to the previous question, what salary premium are you willing to pay to a candidate with high data literacy skills?"*





COMMUNICATING WITH DATA

DATA LITERACY

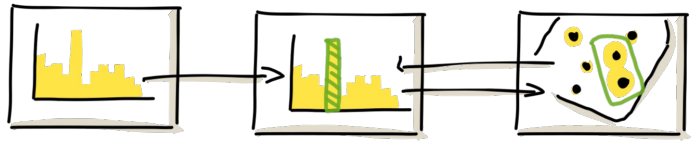
READING DATA

WORKING WITH DATA

bol.
de winkel van ons allemaal

Data Literacy Skill Levels

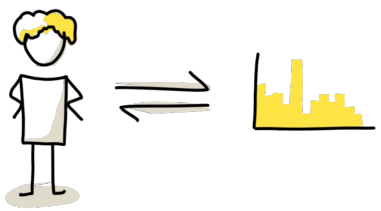
	★☆☆	★★☆	★★★
Reading Dashboard User	Creating insights by filtering existing dashboards	Conclusions	Implications Causations
Writing Dashboards	Editing Existing Dashboard	Build new workbooks and perform ad-hoc analyses	Create new impactful and actionable Dashboards (for own and team)
Writing Data Models	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



COMMUNICATING WITH DATA

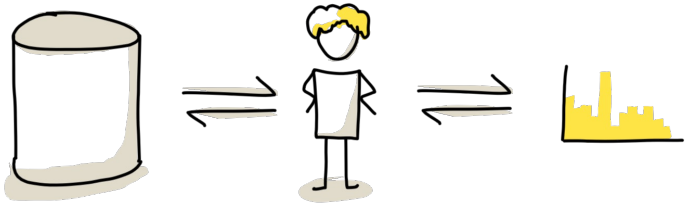
DATA LITERACY

- [ ] AWARENESS CAMPAIGNS
- [ ] INFORMAL COP
- [ ] FORMAL TRAININGS
- [ ] LOWER THE DATA BARIER



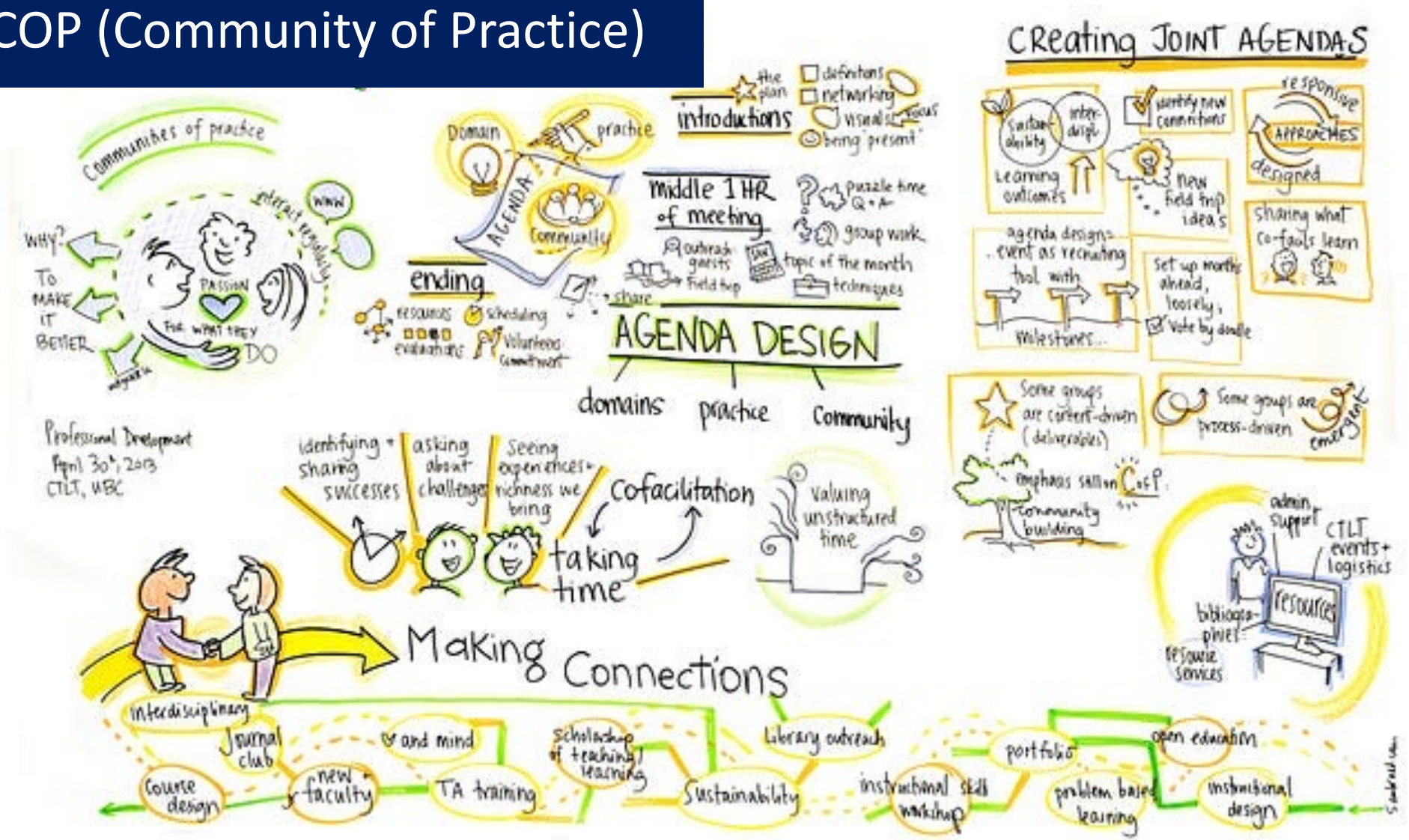
READING DATA

WORKING WITH DATA



Awareness Campaigns

Informal COP (Community of Practice)

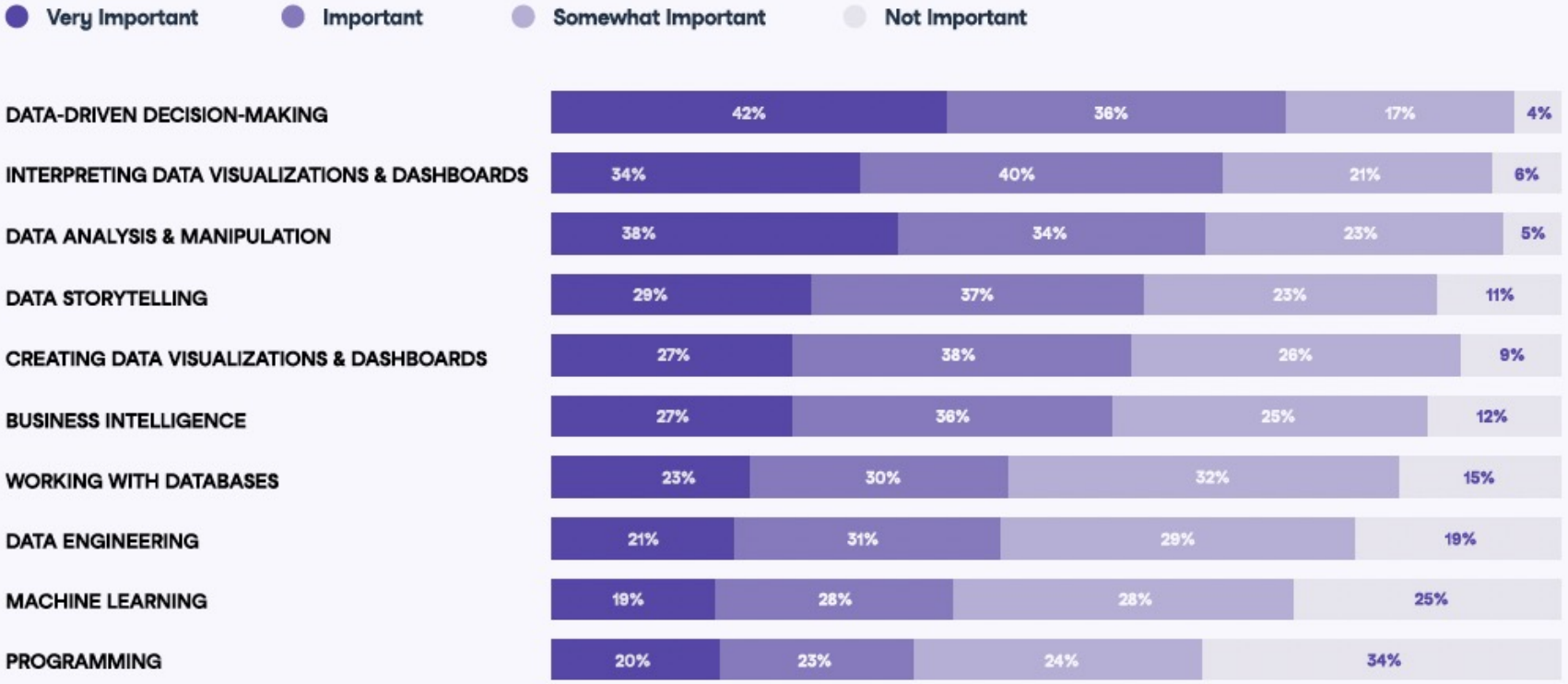


[Unlocking the Wisdom of Crowd Through Community of Practice]

Formal Training

DATA-DRIVEN DECISION-MAKING IS SEEN AS THE MOST IMPORTANT SKILL LEADERS LOOK TO HAVE ON THEIR TEAMS

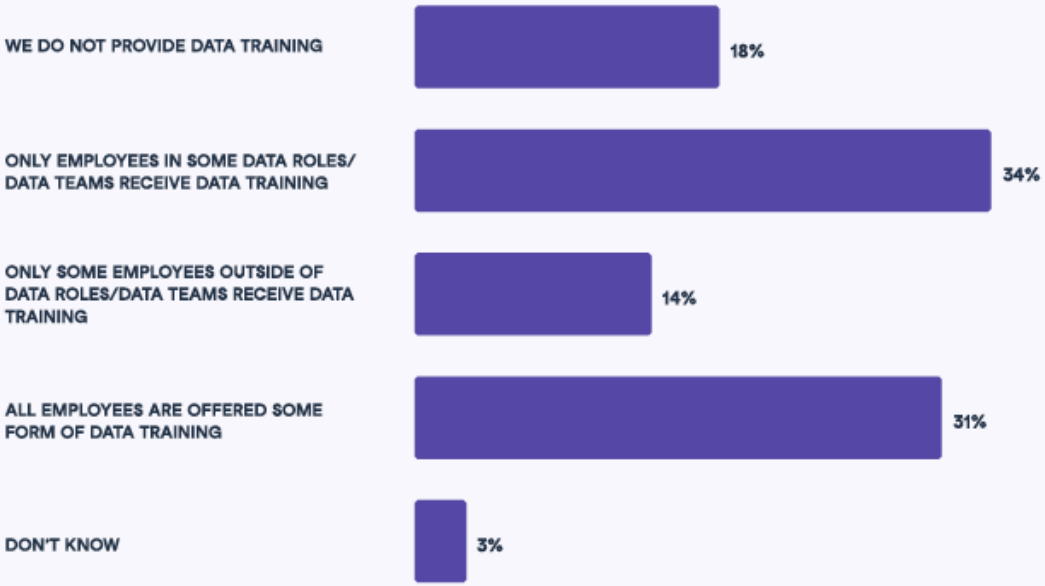
Question asked: *"How important, if at all, are the following data skills for the day-to-day tasks of employees in your organization?"*



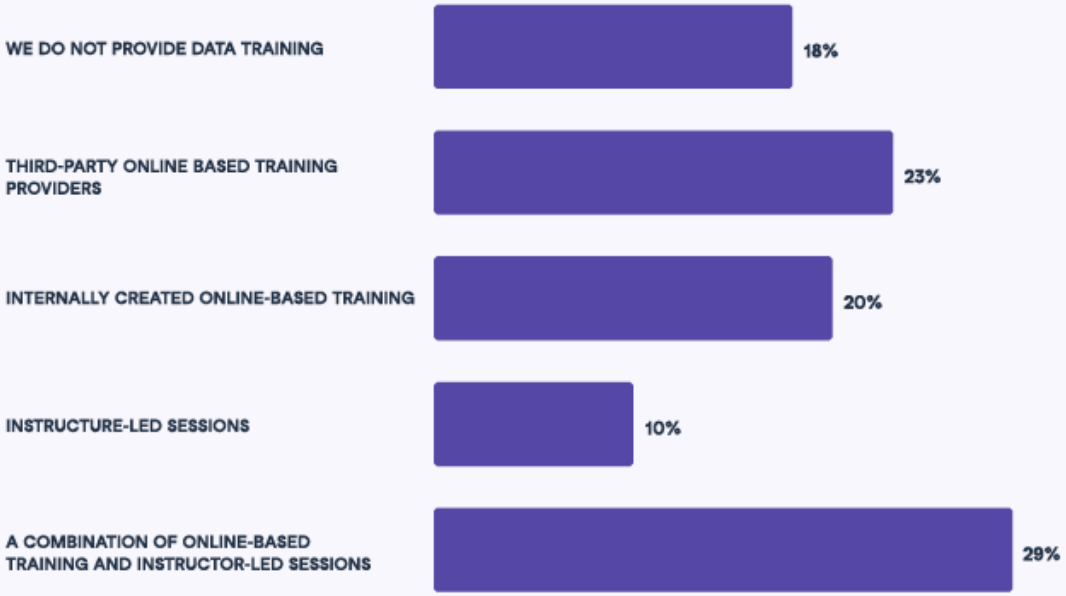
Formal Training

MOST ORGANIZATIONS DON'T HAVE A MATURE TRAINING PROGRAM SET IN PLACE, LEVERAGING A VARIETY OF LEARNING METHODOLOGIES

Question asked: *"What would best describe the state of data training at your organization?"*



Question asked: *"How do you upskill your workforce on data skills? (Single select option)"*



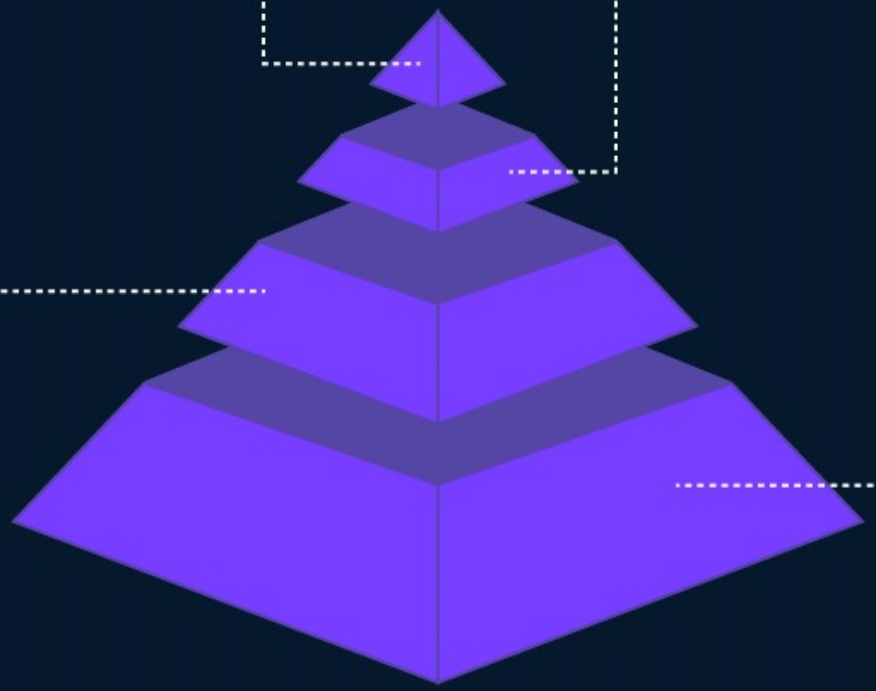
datacamp

Results (ROI)

- **Transformational goals:** Measurement of initial transformational goal set out in the learning program
- **Employee retention:** The learning team can look at churn, billing, and retention for skill academy graduates versus non-graduates

Behavioral Change

- **Tool usage:** If you have an internal data platform, you can work with your engineering teams to measure how learners are engaging with your company's data in real life
- **Data culture participation:** The number of learners who become part of data culture events such as hackathons, tech talks, etc



Learning

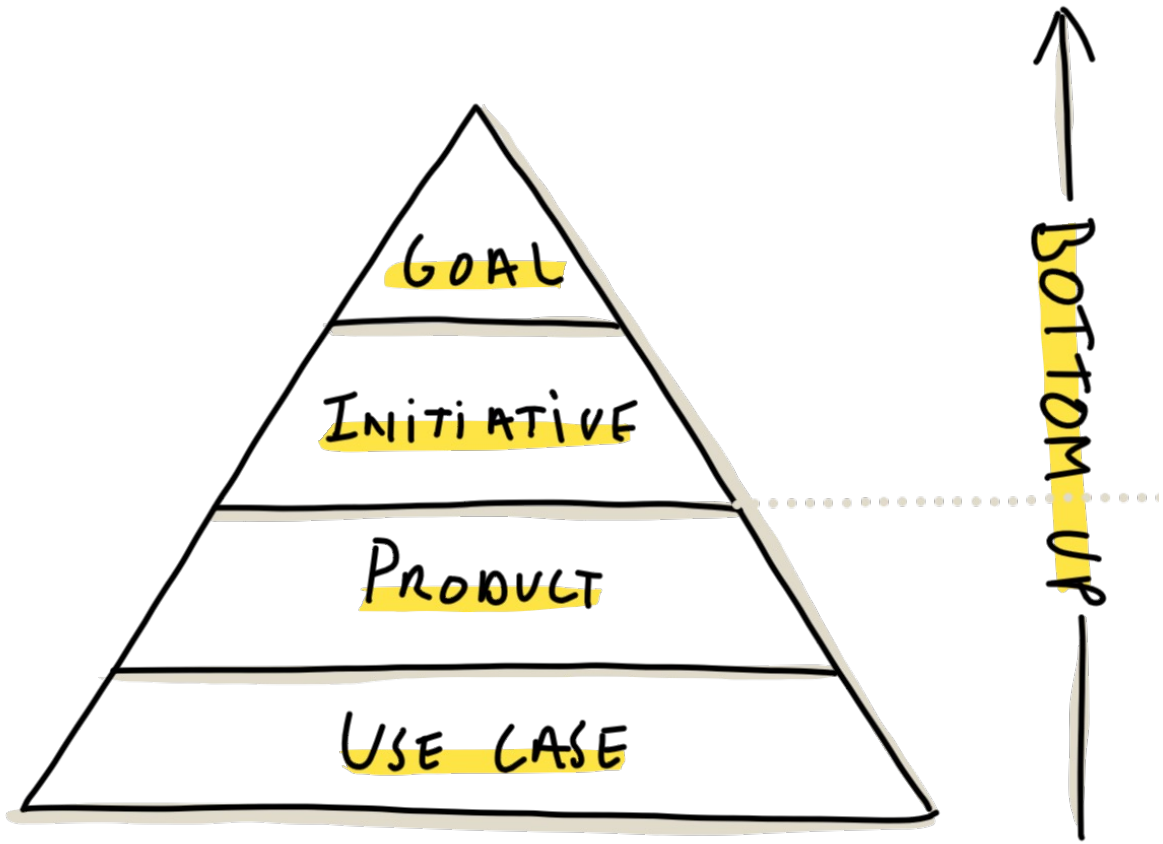
- **Completion rates:** Percentage of employees who have completed the program
- **Assessment evolution:** How your people are ranking on DataCamp assessments throughout the learning program
- **XP points gained on platform:** DataCamp provides XP points for completing lessons, courses, assessments, projects, and more

Reaction

- **Adoption rates:** Number of active users in the program
- **Satisfaction rates:** Percentage of positive feedback on anonymous surveys
- **Email engagement rates:** How are learners interacting with upskilling-related emails and newsletters?

ROLES & RESPONSIBILITIES?

DATA LITERACY



BOTTOM UP

ROLES & RESPONSIBILITIES?

DATA LITERACY

DATA MANAGEMENT STRUCTURE?

DATA
- EXPERTS ?
- OWNERS ?
- STEWARDS ?

BOTTOM UP

ROLES & RESPONSIBILITIES?

DATA LITERACY

DATA MANAGEMENT STRUCTURE?

LITERACY IMPROVEMENT TRACKS?

DATA
- EXPERTS ?
- OWNERS ?
- STEWARDS ?

LITERATE ?

EXERCISE: WHAT IS THE PEOPLE IMPACT + ACTIONS IN YOUR UCS?

BOTTOM UP

ROLES & RESPONSIBILITIES?

DATA LITERACY

DATA MANAGEMENT STRUCTURE?

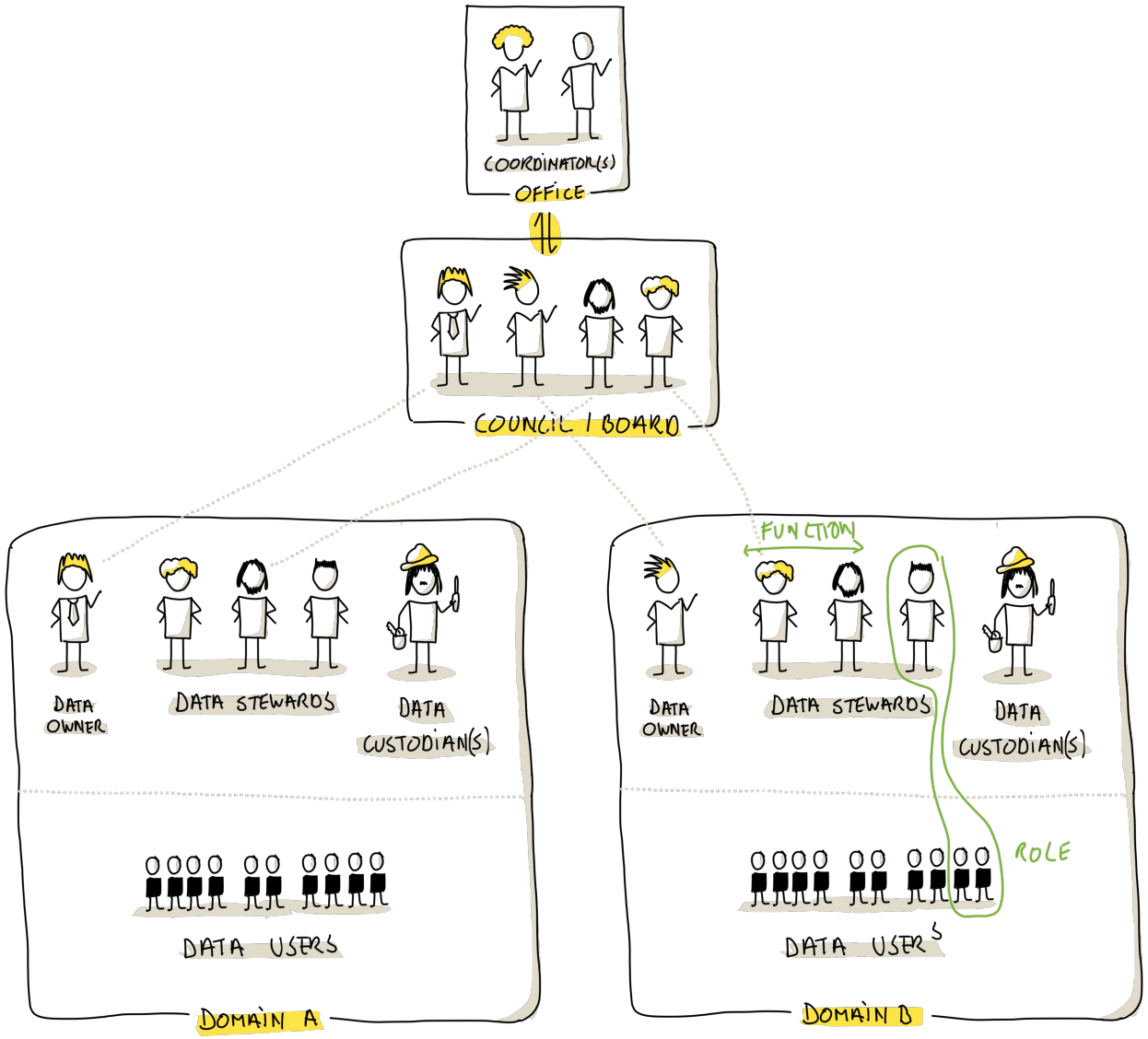
LITERACY IMPROVEMENT TRACKS?

DATA – EXPERTS ?
– OWNERS ?
– STEWARDS ?

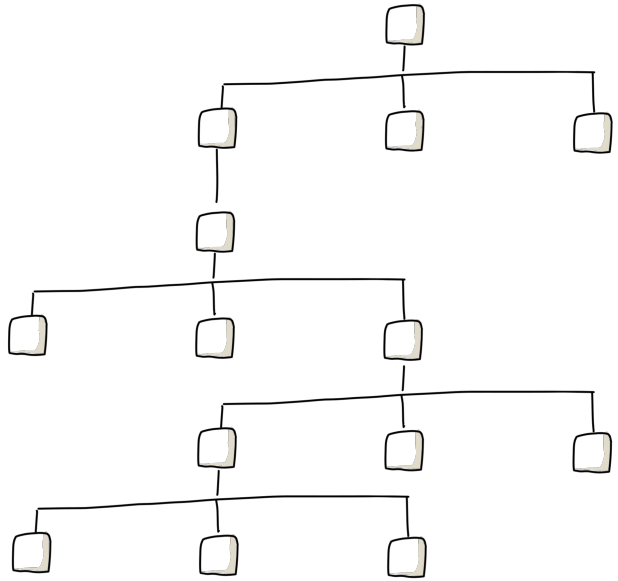
LITERATE ?

7.1

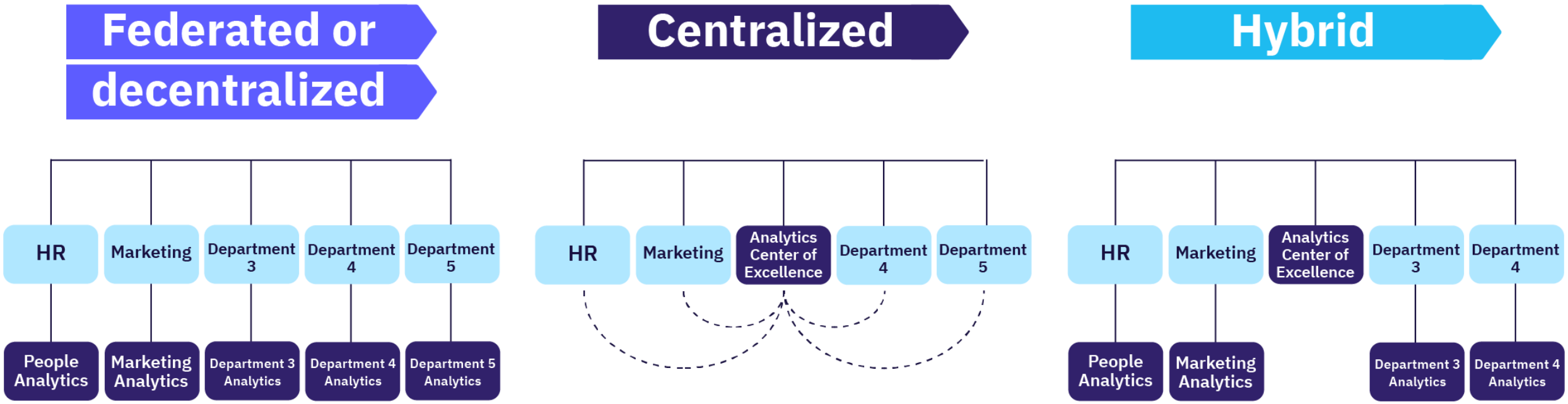
PEOPLE & ORGANIZATION



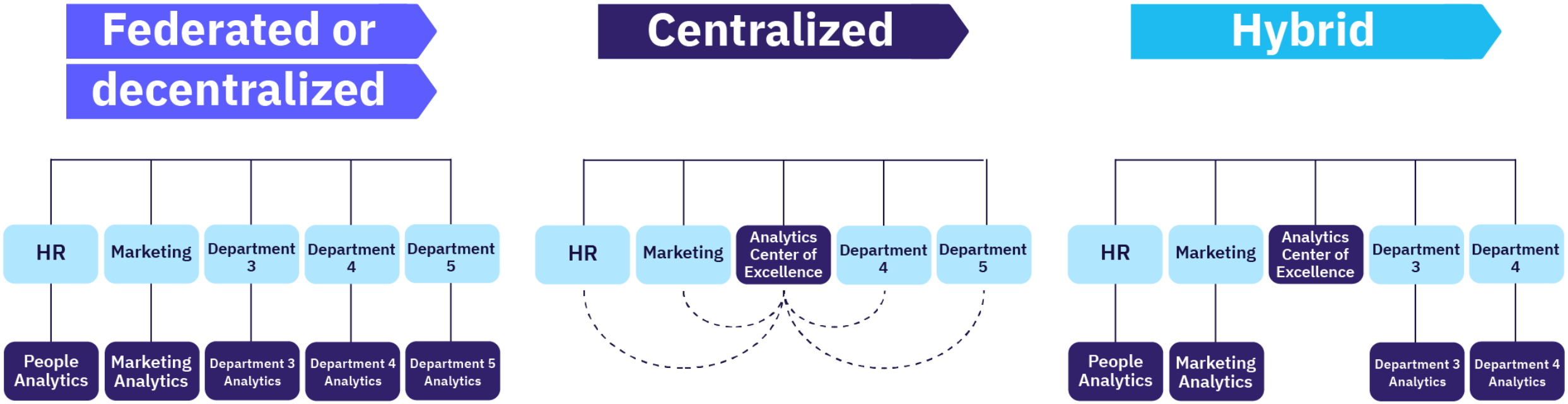
ORGANIZATIONAL MODEL



Different Organization Types

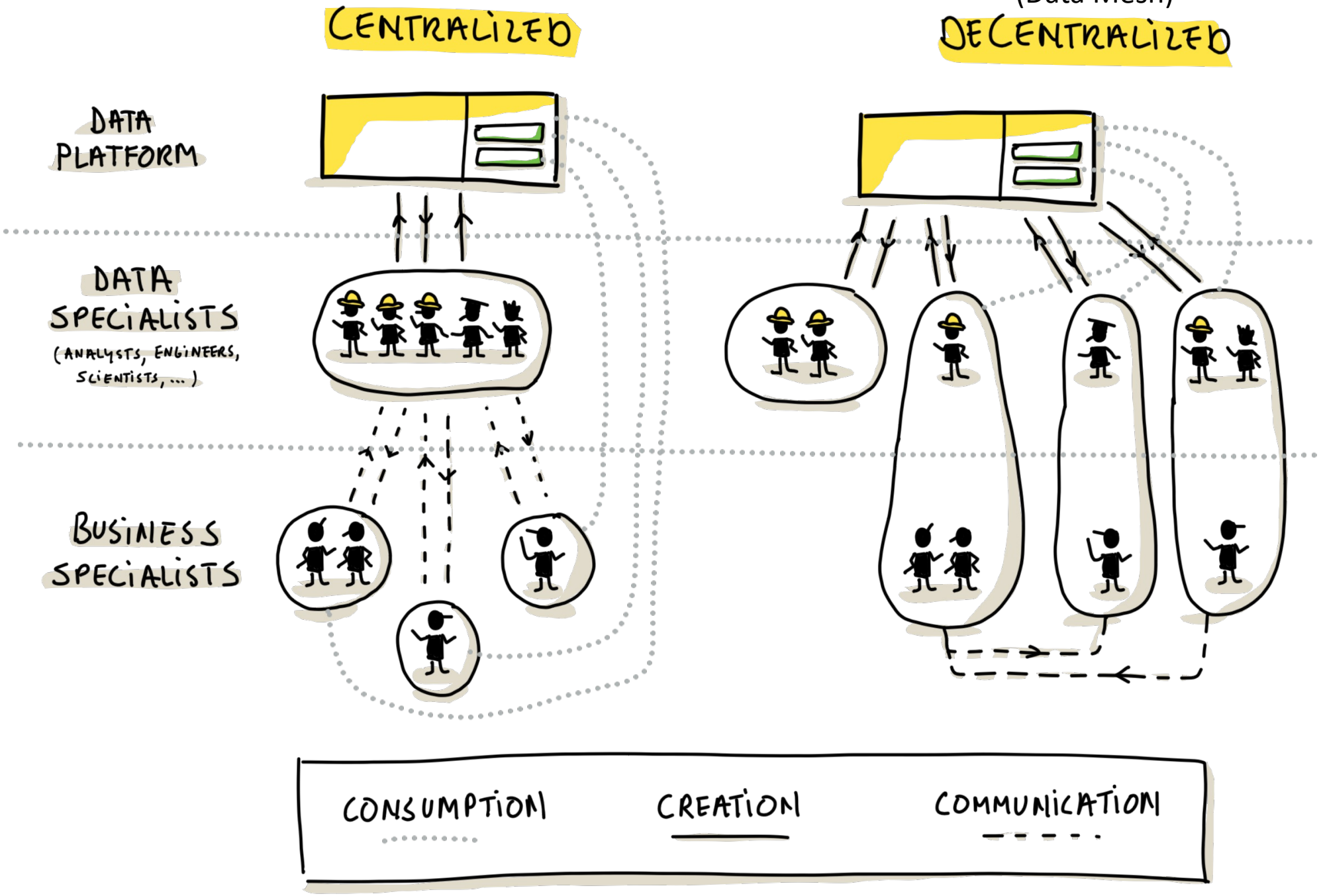


Different Organization Types



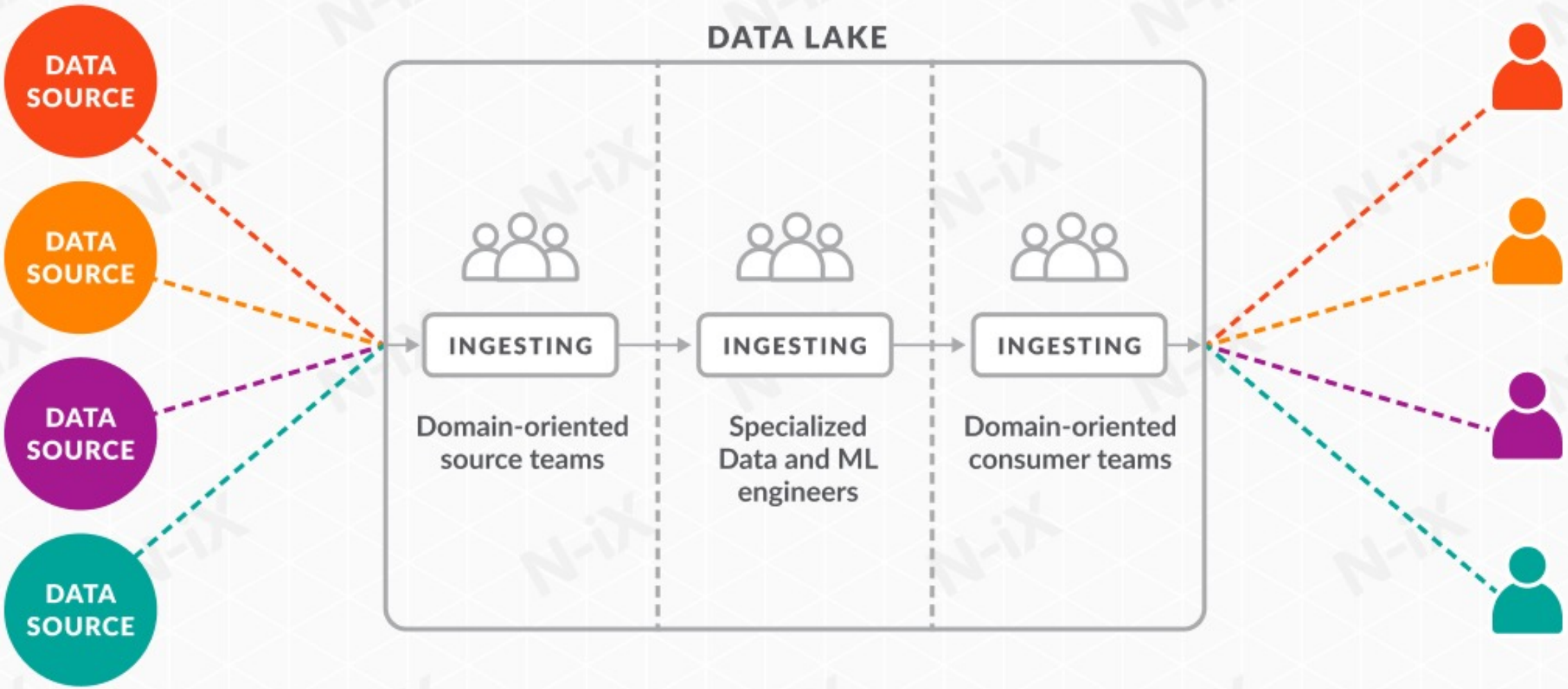
"Data Mesh"

(Data Mesh)


CENTRALIZED
DECENTRALIZED
DATA PLATFORM
DATA SPECIALISTS
(ANALYSTS, ENGINEERS, SCIENTISTS, ...)
BUSINESS SPECIALISTS
CONSUMPTION
CREATION
COMMUNICATION

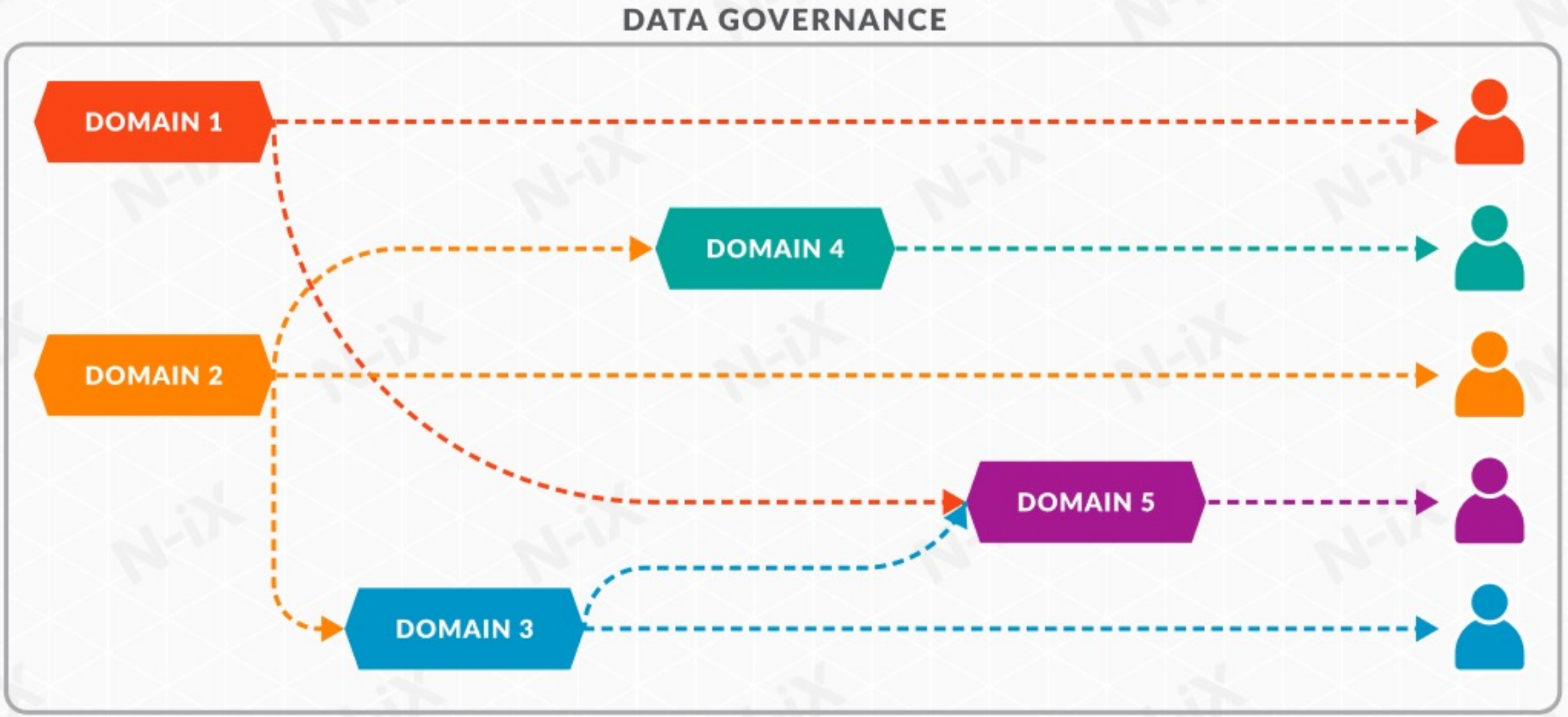
Simplified diagram of a typical data lake infrastructure

N-iX



Simplified diagram of a typical data mesh infrastructure

N-iX



Think twice about a data mesh if…

1. Your organization is too small
2. Your Data Maturity is low
3. You are not used to run multi-year transformation projects
4. You don’t have a domain oriented organization

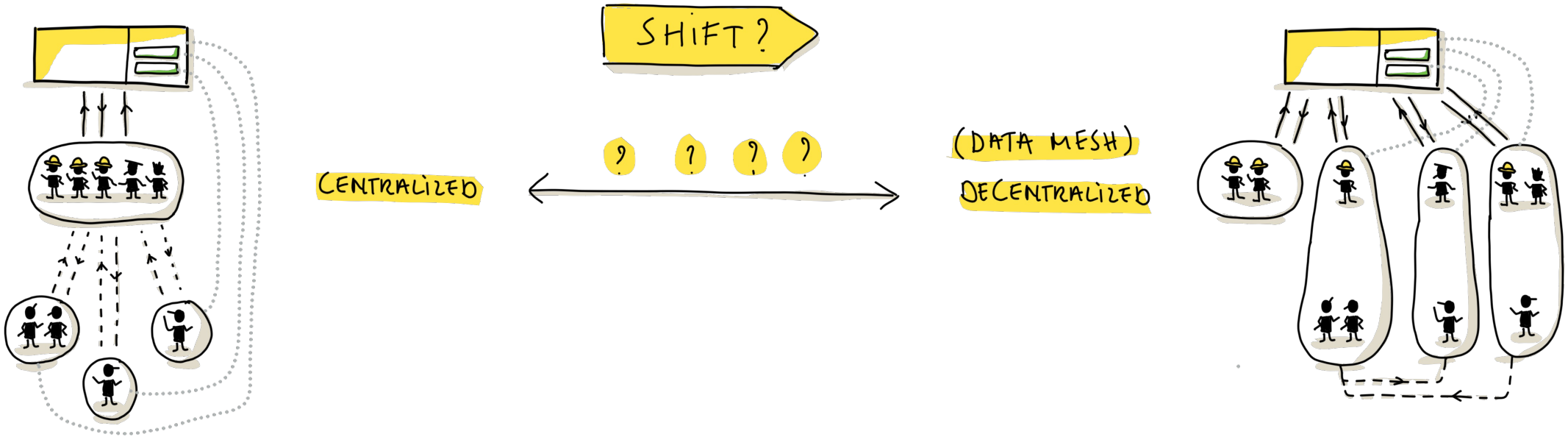
Data Mesh: *My View*

ADVANTAGES	DISADVANTAGES
Agility	Duplication of effort
Freedom to Innovate per Domain	Harder to enforce Governance, Security, Policy, …
Stimulation of Data & Data Product Ownership	Inter-domain communication and coordination
Accountability	Domain Data Dependencies
Domain Knowledge	Incoherent workflows (per domain view)
	Hard to increase Data Skills within Domain teams

[inspired on ‘Use Data Ops With Your Data Mesh to Prevent Data Mush’]

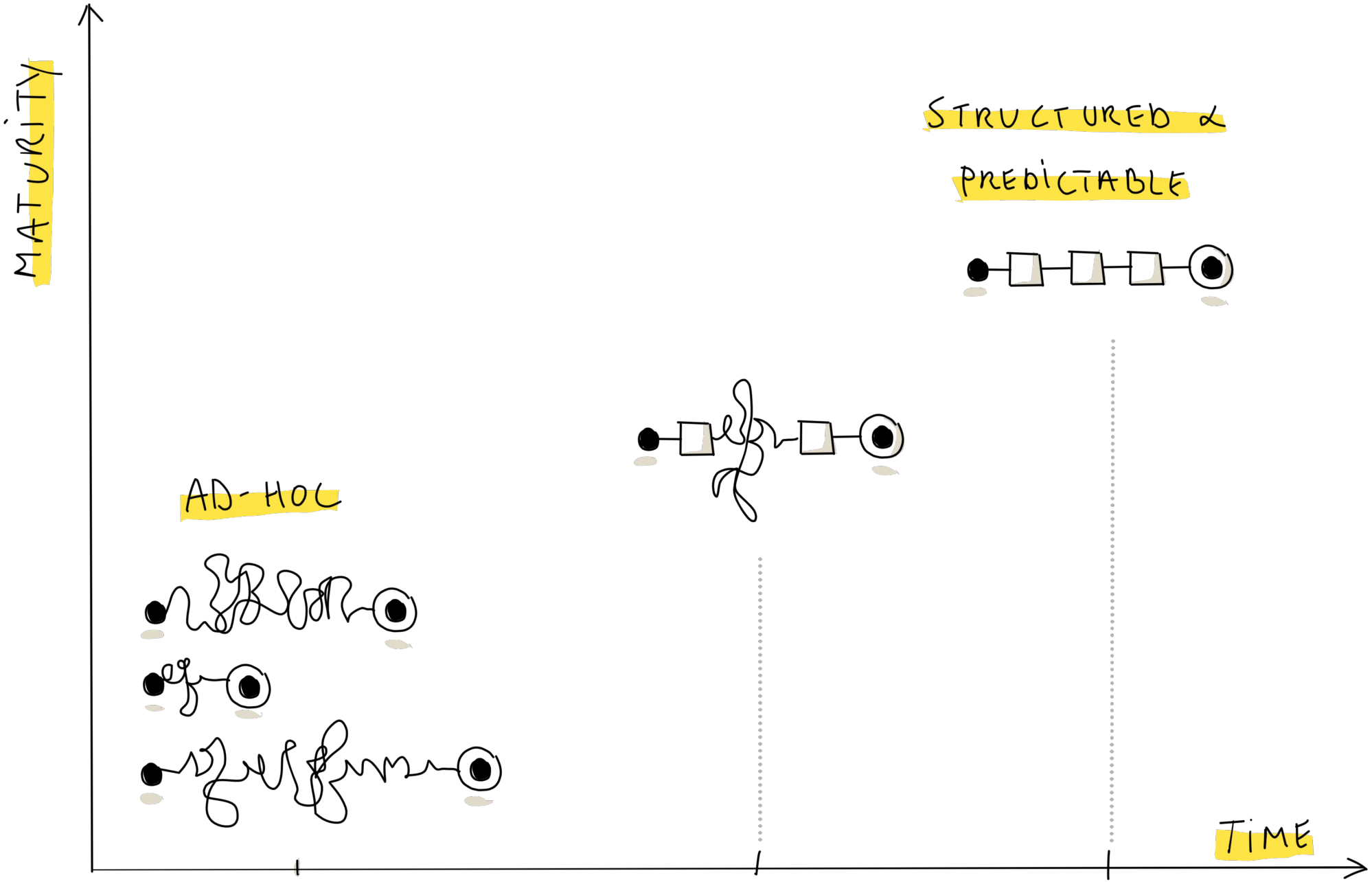
DISCUSSION: ORGANIZATION IMPACT RELEVANT?

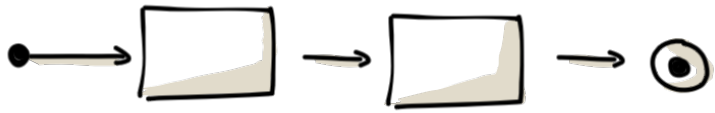
- Which type of organization do you have today (centralized – hybrid – decentralized)?
- How can an organizational change help you to deliver your use cases?



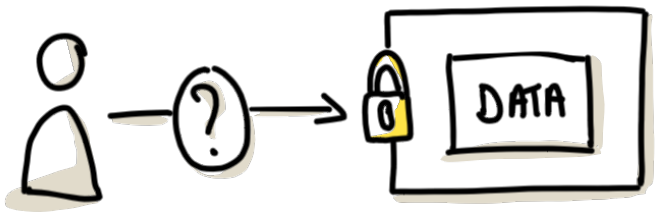
7.2

PROCESSES

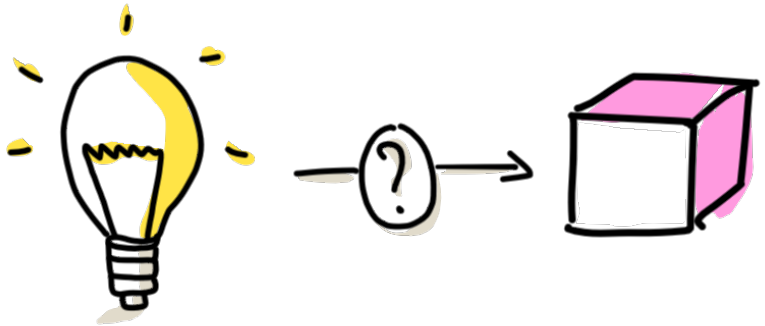

MATURITY
STRUCTURED & PREDICTABLE
AD-HOC
TIME



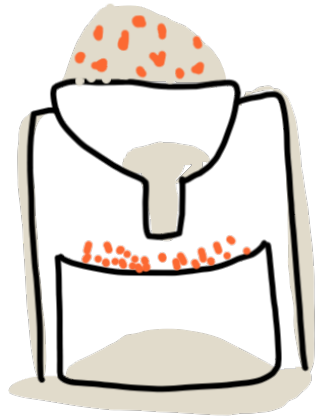
PROCESSES



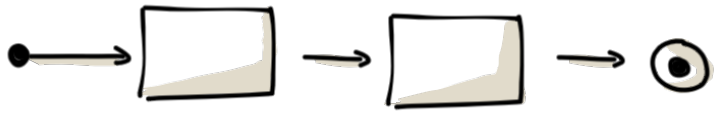
GETTING DATA ACCESS



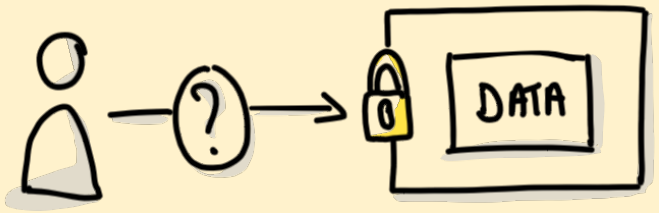
HOW TO REQUEST NEW FEATURES ?



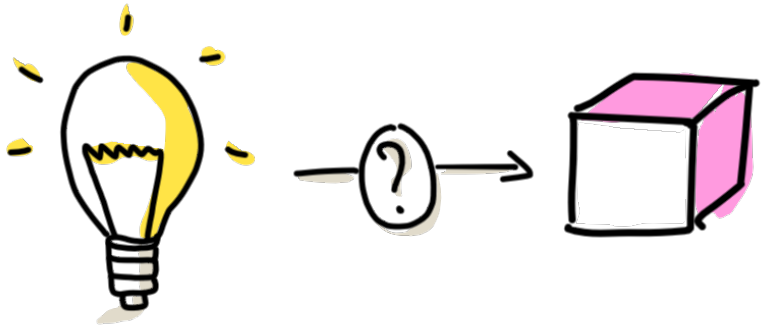
DATA QUALITY



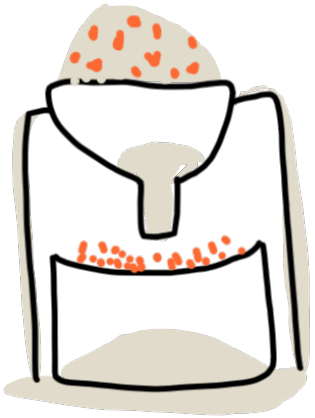
PROCESSES



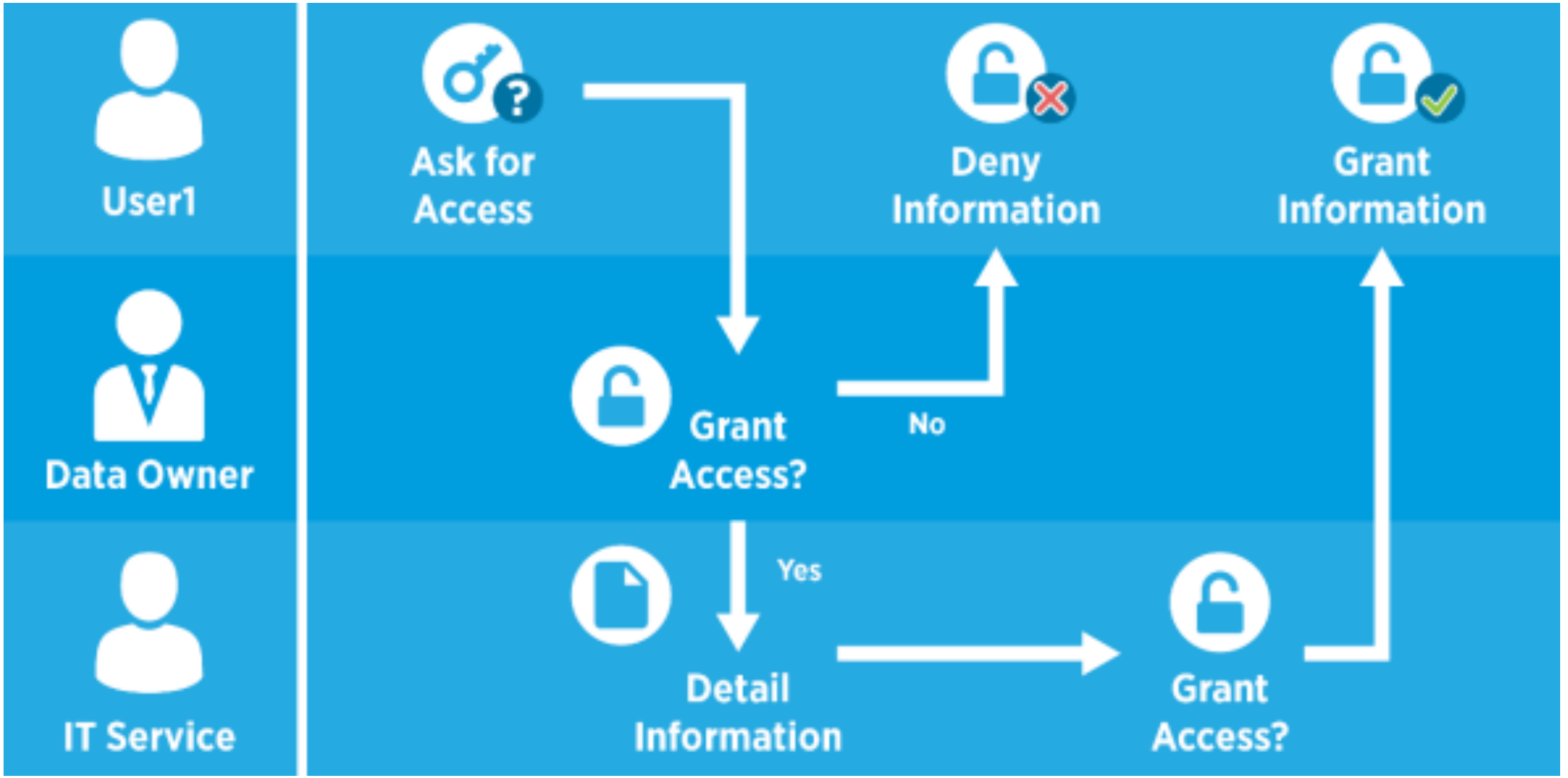
GETTING DATA ACCESS



HOW TO REQUEST NEW FEATURES ?



DATA QUALITY

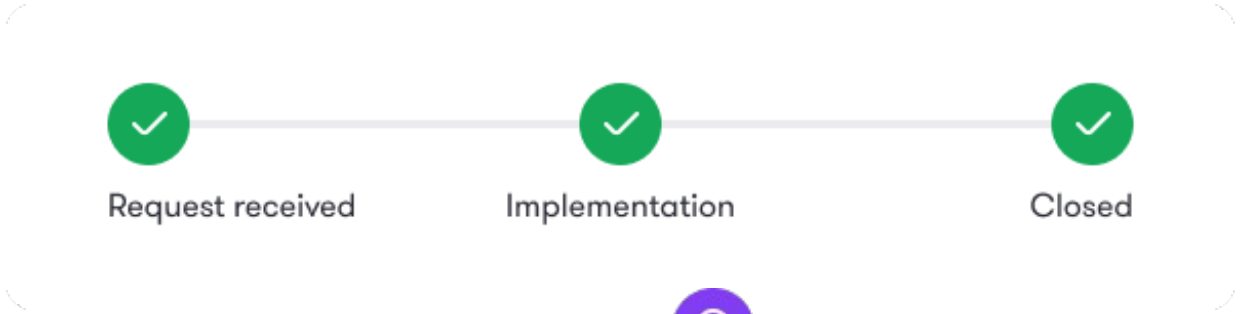

User1
Data Owner
IT Service
Ask for Access
Grant Access?
No
Deny Information
Yes
Detail Information
Grant Access?
Grant Information

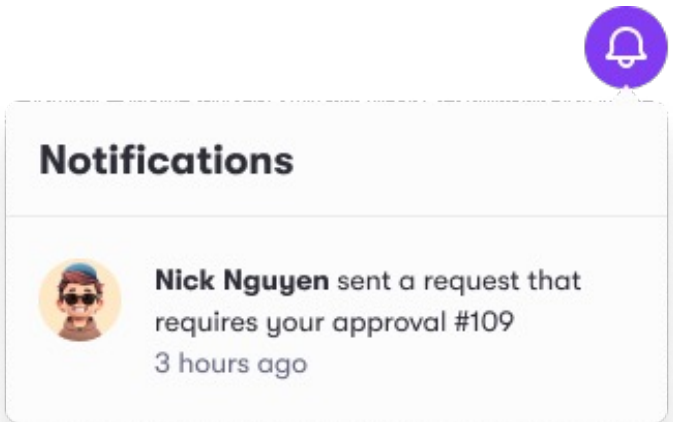
raito

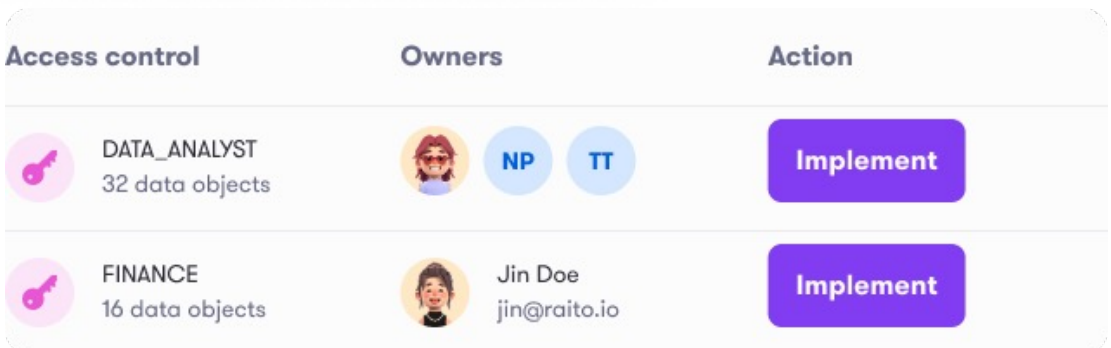
Step 3

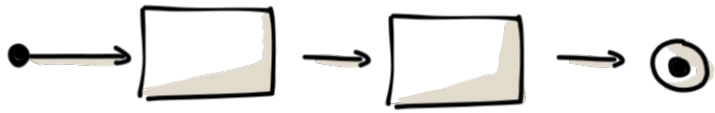
What are the data objects you need access to?

Data object	Permissions
SALES Schema	Read ×
PURCHASING Schema	Read ×

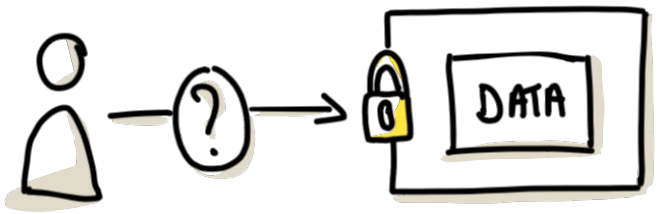




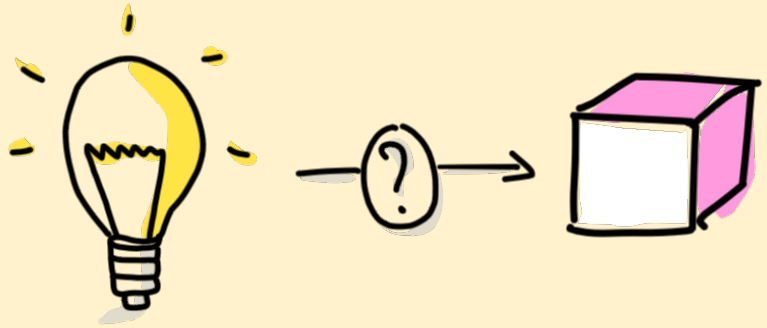




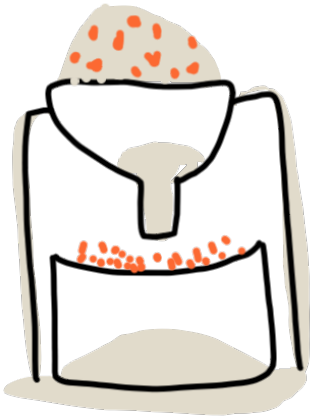
PROCESSES



GETTING DATA ACCESS

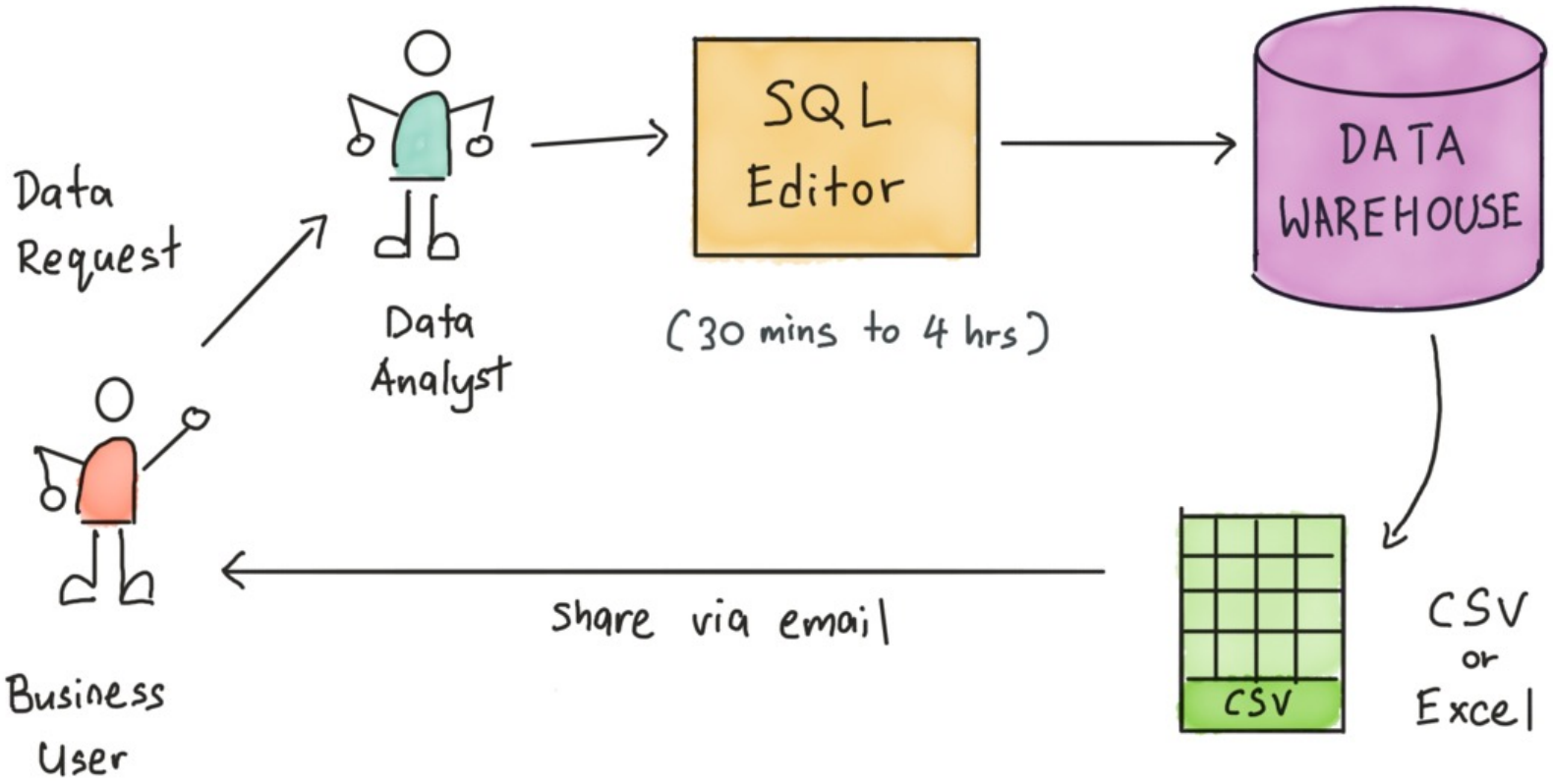


HOW TO REQUEST NEW FEATURES ?



DATA QUALITY

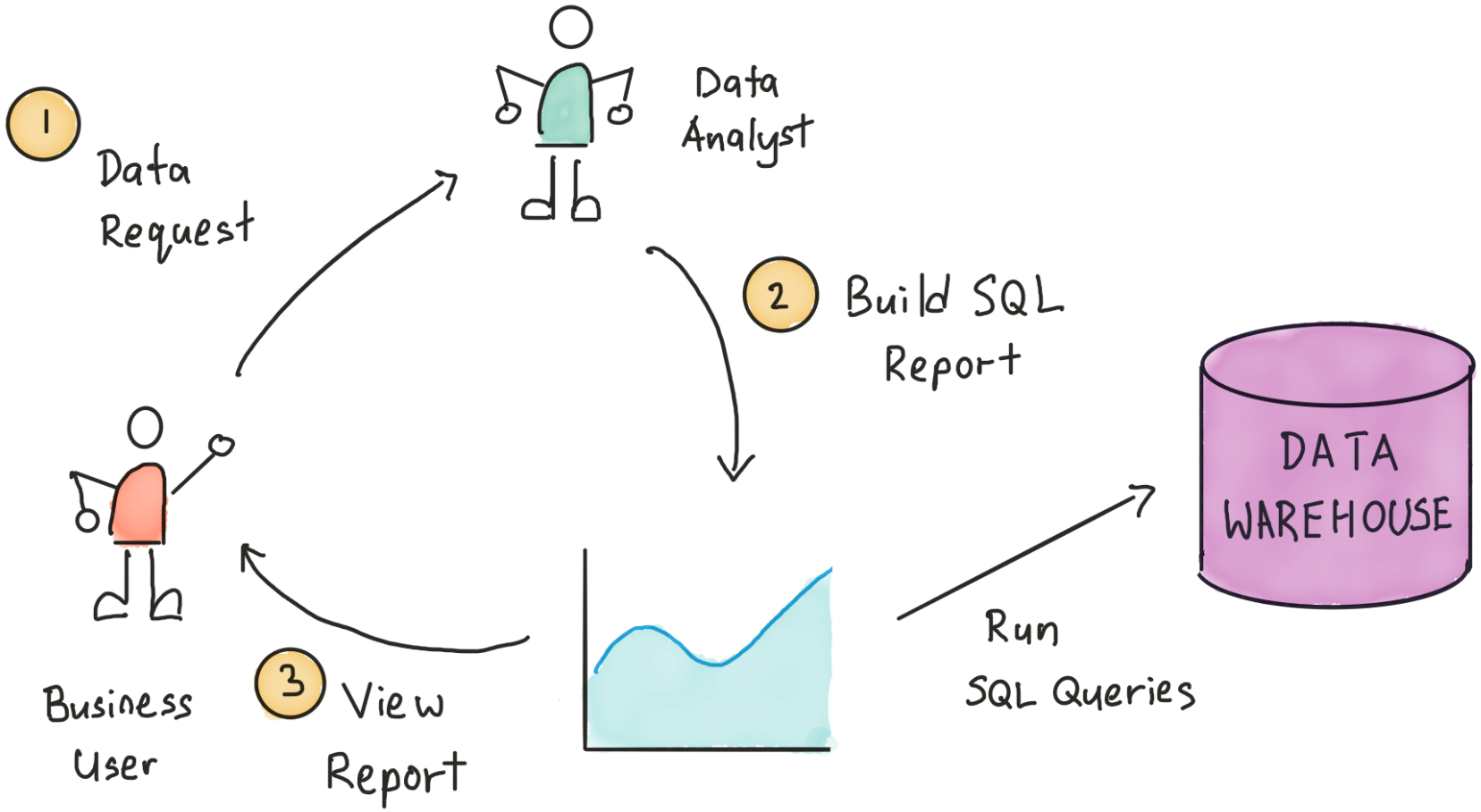
1 – Ad-Hoc Query Process



Slightly different input: repeat the whole process!

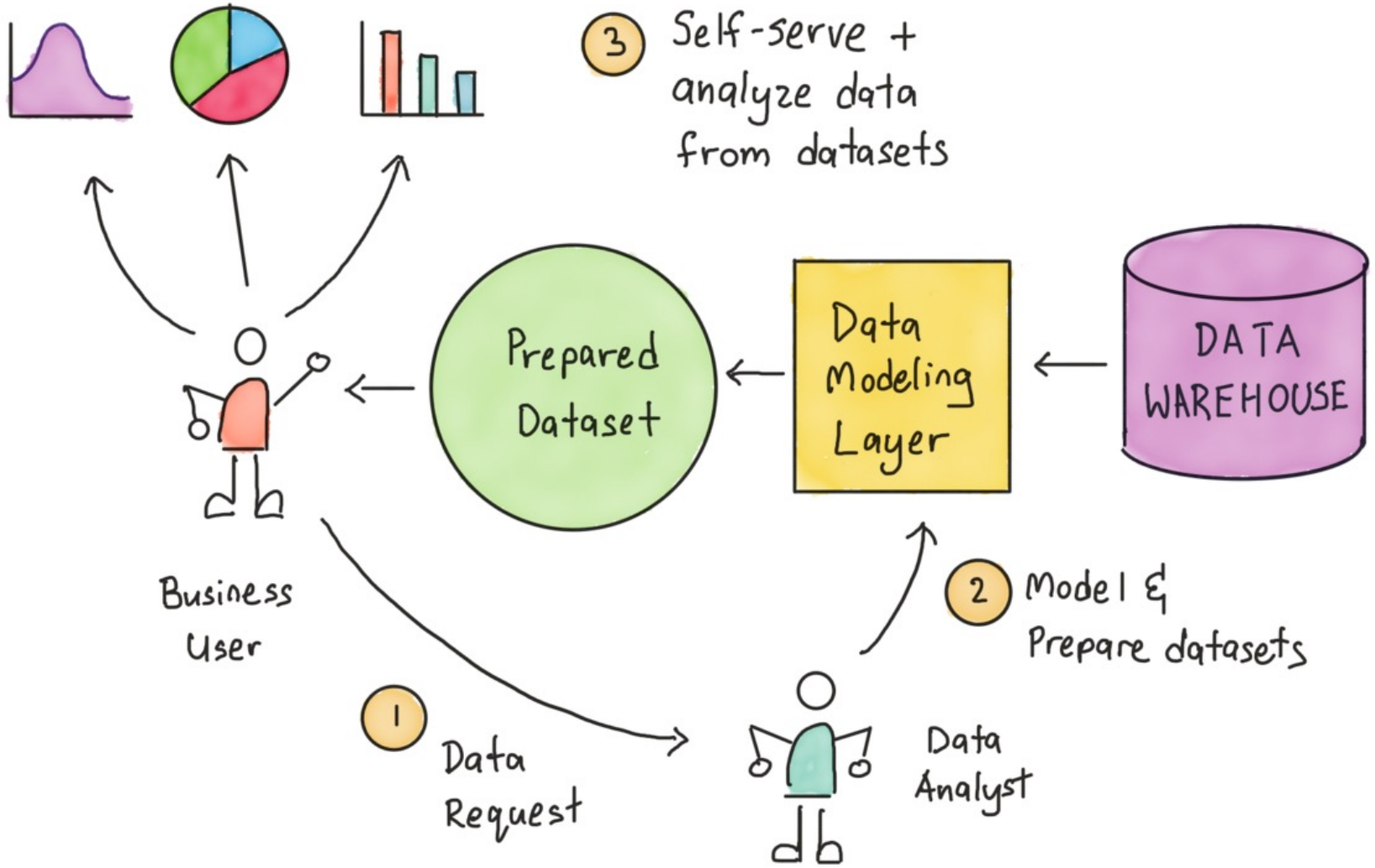
[Self Service Analytics]

2 – Corporate Reports & Dashboards



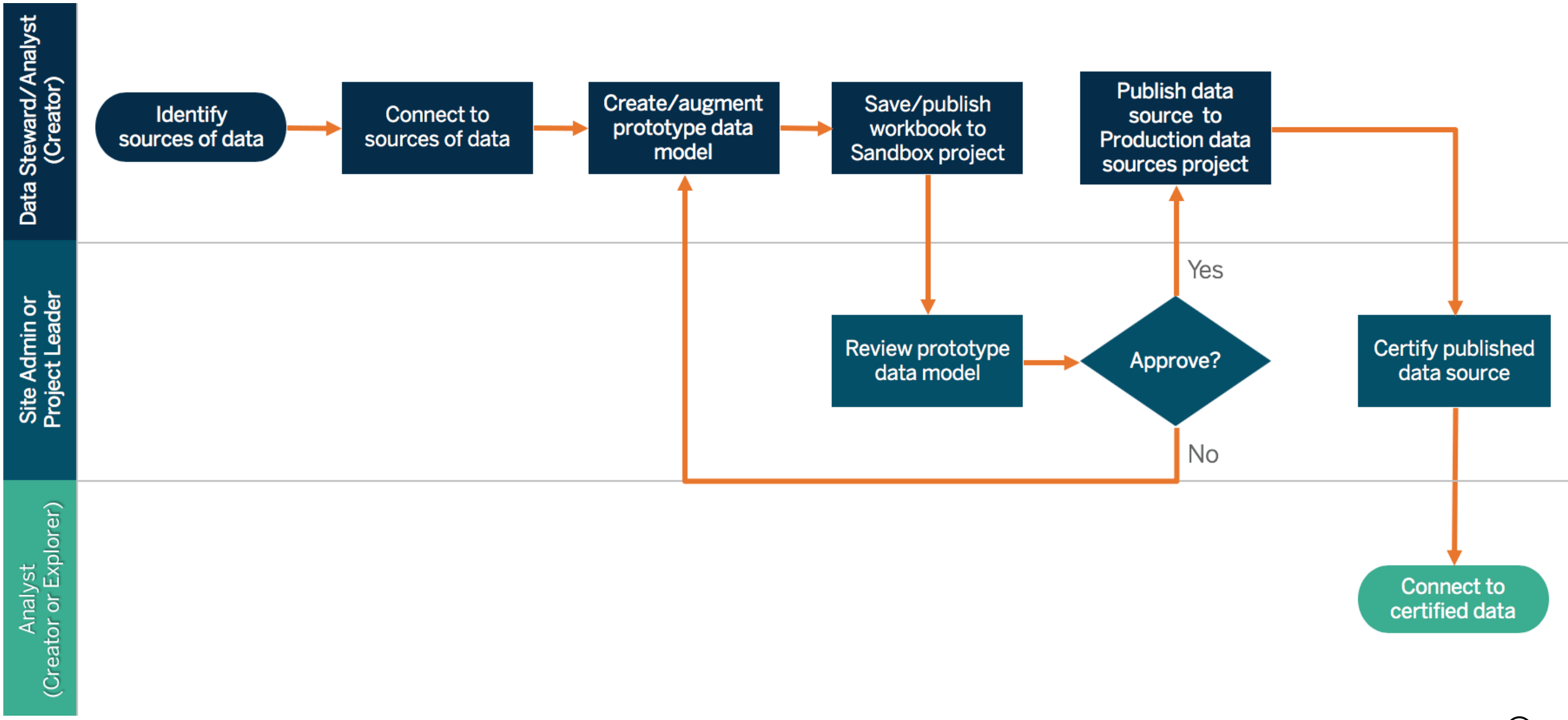
[Self Service Analytics]

3 – Self-Service Reporting



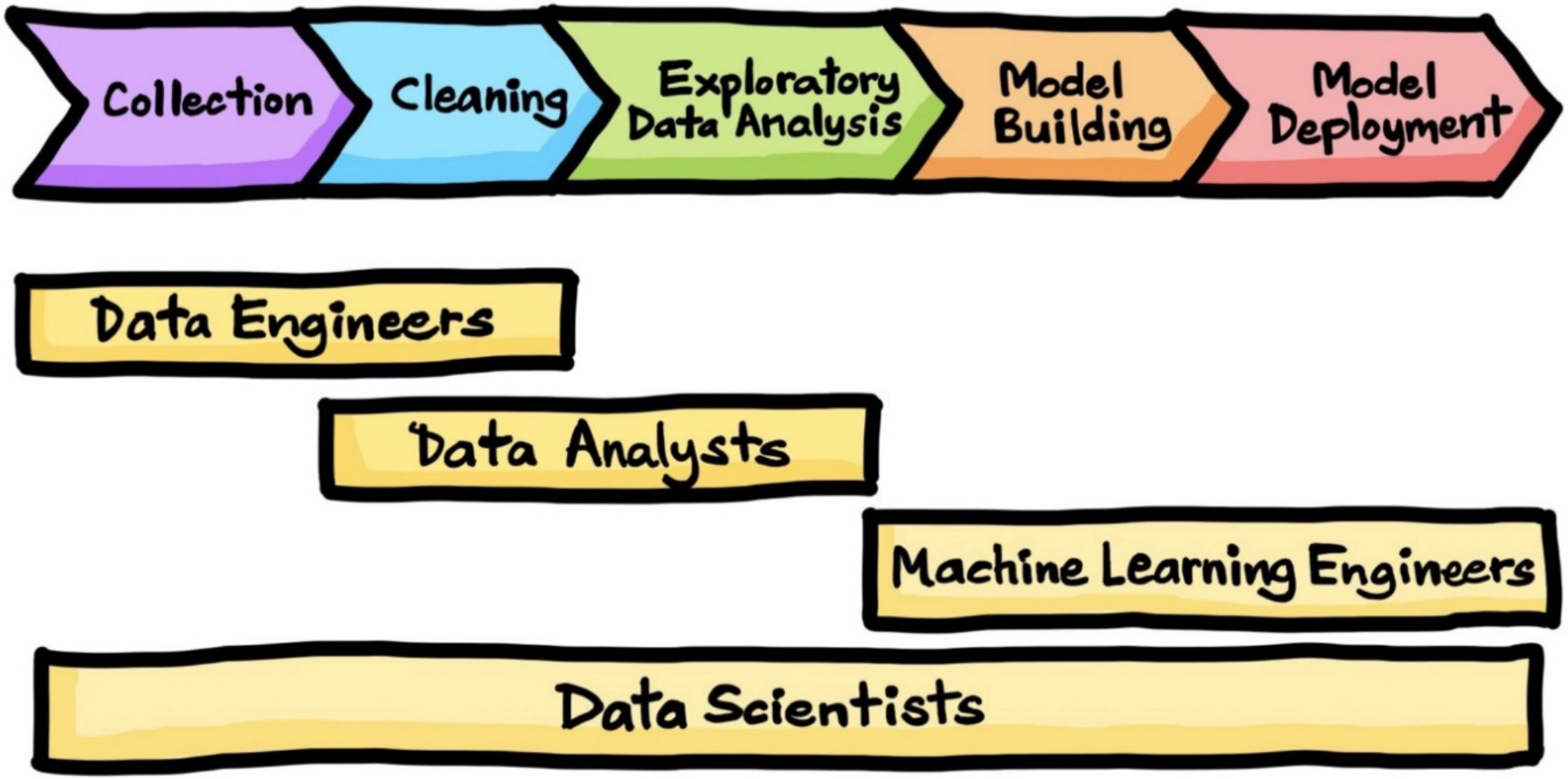
[Self Service Analytics]

Self Service Process



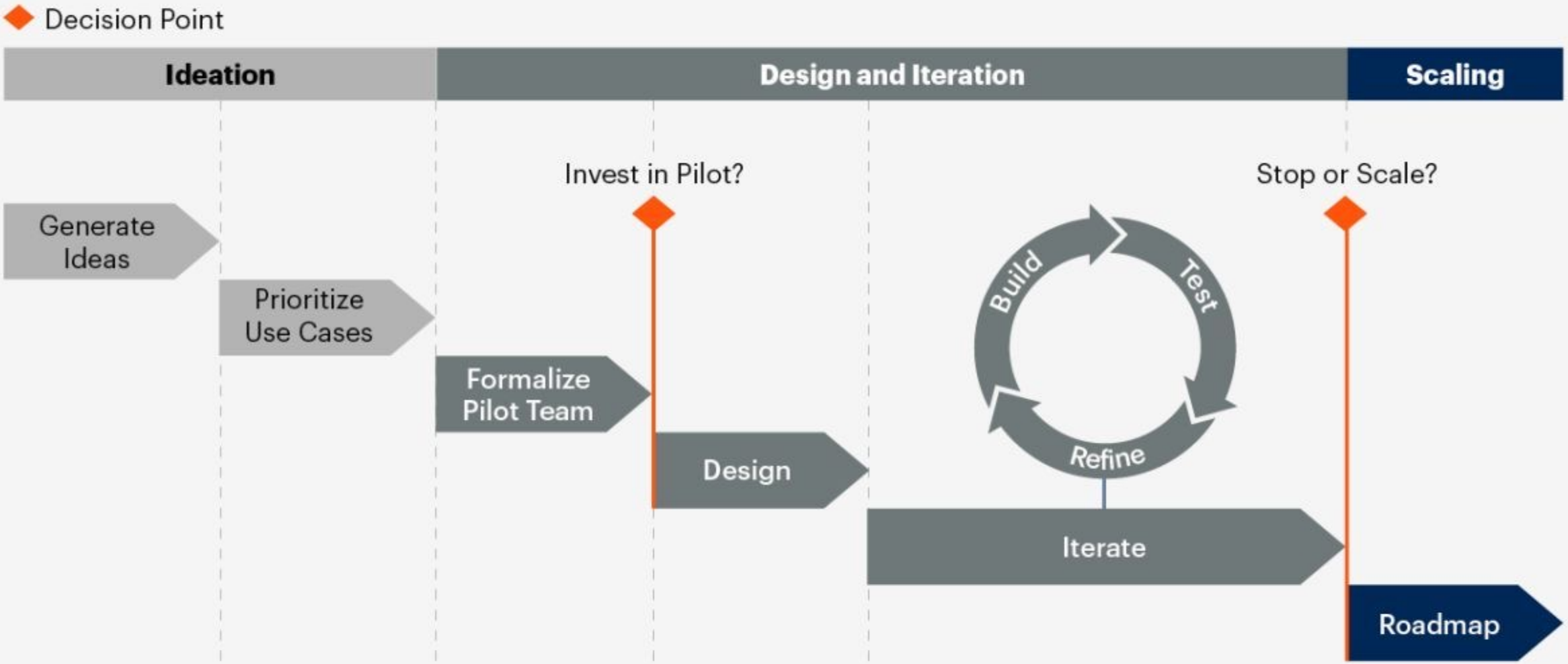
[What is Data Governance]

Data Science Processes



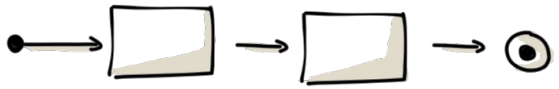
[Data Science Process: A Beginners Guide]

Generative AI Pilot Phases and Decision Points

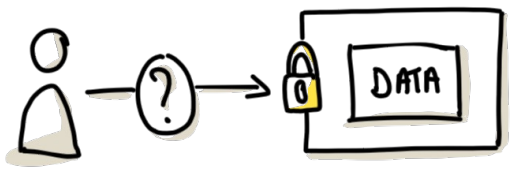


Source: Gartner

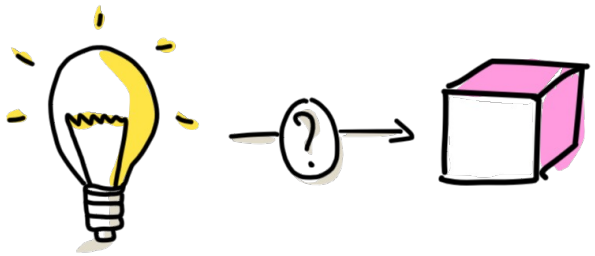
Gartner®



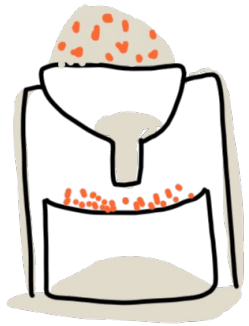
PROCESSES



GETTING DATA ACCESS



HOW TO REQUEST NEW FEATURES ?

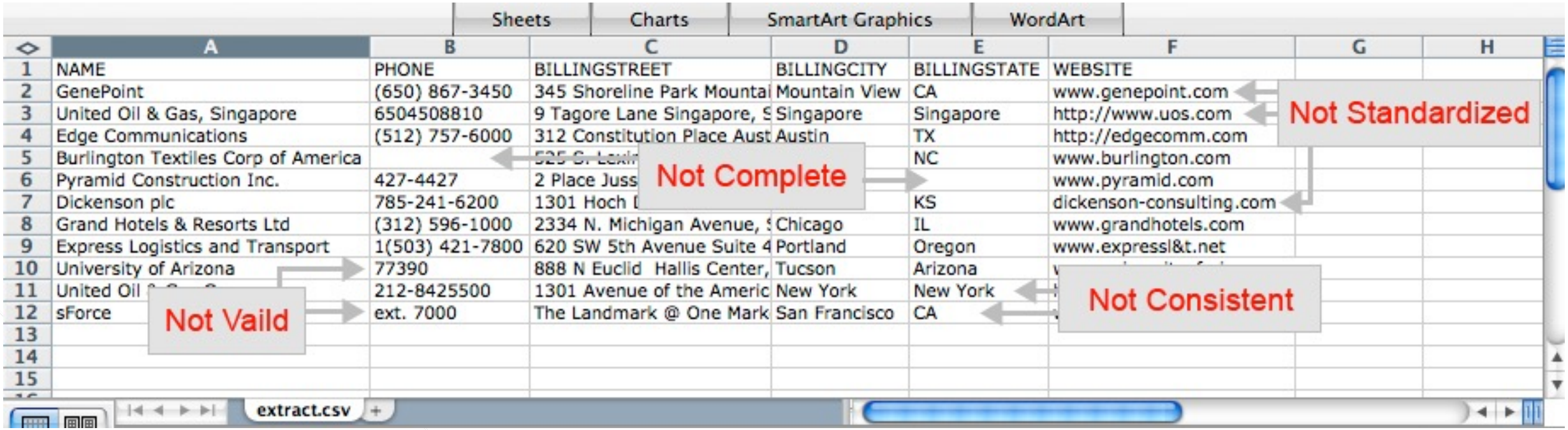


DATA QUALITY

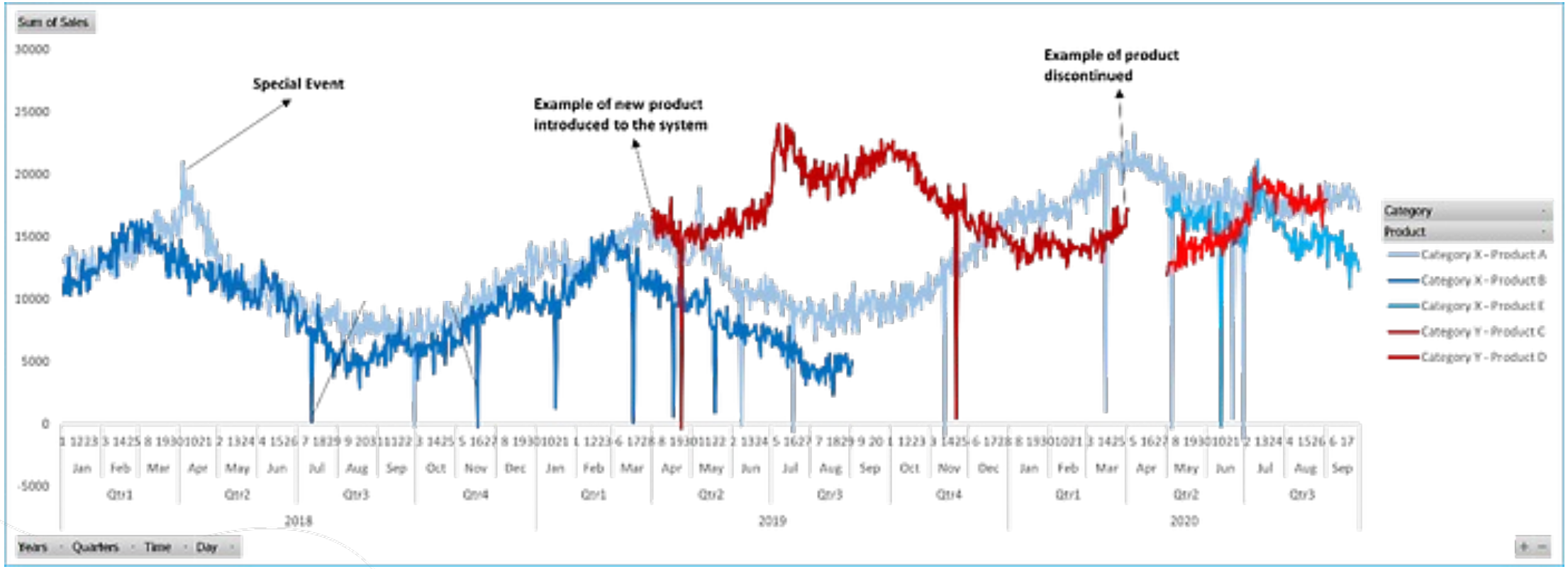
INTRODUCTION

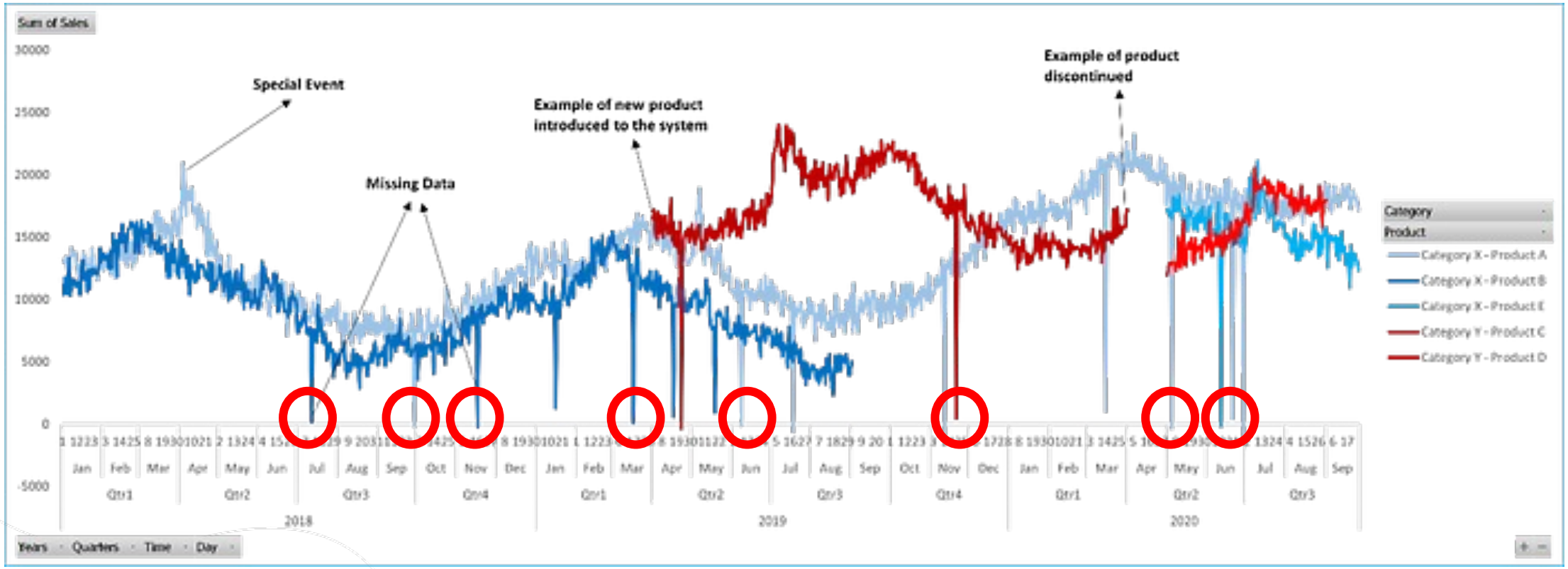
PROCESS

DQ : Example

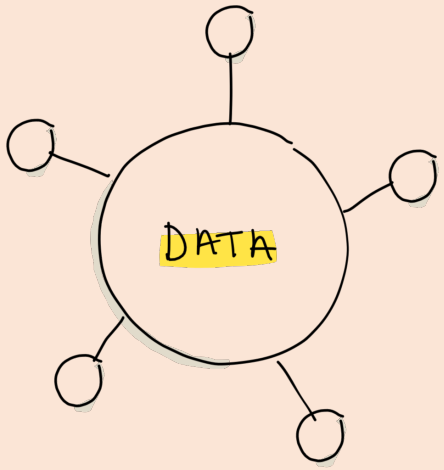


	A	B	C	D	E	F	G	H
1	NAME	PHONE	BILLINGSTREET	BILLINGCITY	BILLINGSTATE	WEBSITE		
2	GenePoint	(650) 867-3450	345 Shoreline Park Mountai	Mountain View	CA	www.genepoint.com		
3	United Oil & Gas, Singapore	6504508810	9 Tagore Lane Singapore, S	Singapore	Singapore	http://www.uos.com		
4	Edge Communications	(512)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

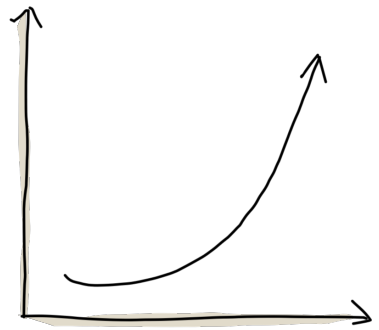

Sum of Sales
30000
25000
20000
15000
10000
5000
0
-5000
Special Event
Example of new product introduced to the system
Example of product discontinued
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Qtr1 Qtr2 Qtr3 Qtr4
2018
2019
2020
Category
Product
Category X - Product A
Category X - Product B
Category X - Product E
Category Y - Product C
Category Y - Product D
Years
Quarters
Time
Day


Sum of Sales
30000
25000
20000
15000
10000
5000
0
-5000
Special Event
Missing Data
Example of new product introduced to the system
Example of product discontinued
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep
Qtr1 Qtr2 Qtr3 Qtr4 Qtr1 Qtr2 Qtr3 Qtr4 Qtr1 Qtr2 Qtr3
2018 2019 2020
Years Quarters Time Day
Category
Product
Category X - Product A
Category X - Product B
Category X - Product E
Category Y - Product C
Category Y - Product D

Da ta Quality

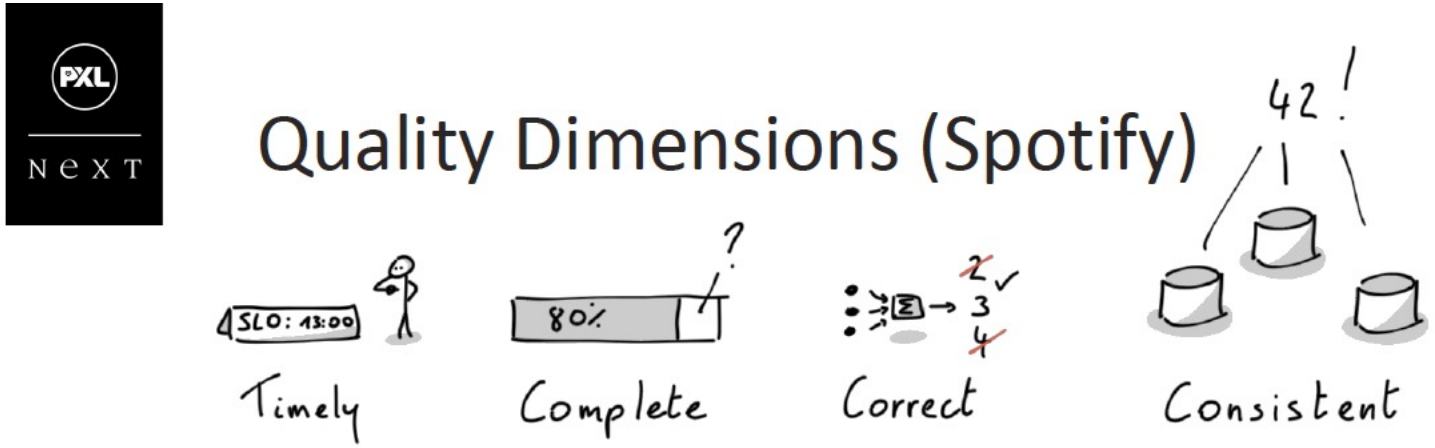


DQ DIMENSIONS



DQ PROCESS
&TOOLS

Introduced in the Previous Lesson



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

What are Data Quality Dimensions?

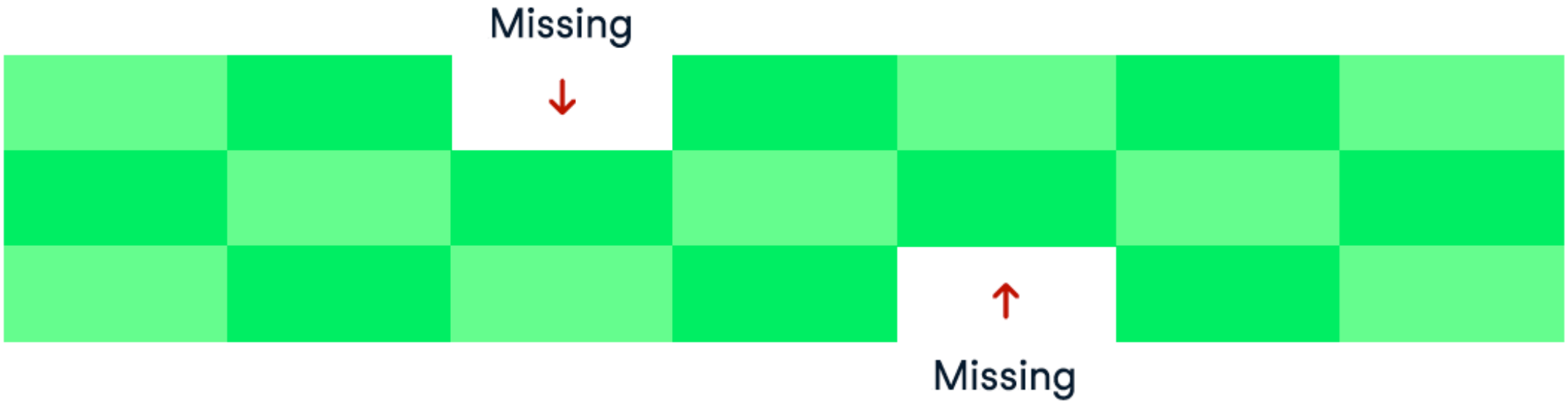
Data Quality is a measurement of the degree to which data is fit for purpose. Good data quality generates trust in data. Data Quality Dimensions are a measurement of a specific attribute of a data's quality.

Completeness – Validity – Accuracy – Uniqueness
Timeliness – Consistency

[Source: Datacamp]

> Completeness

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



[Source: Datacamp]

Completeness Example

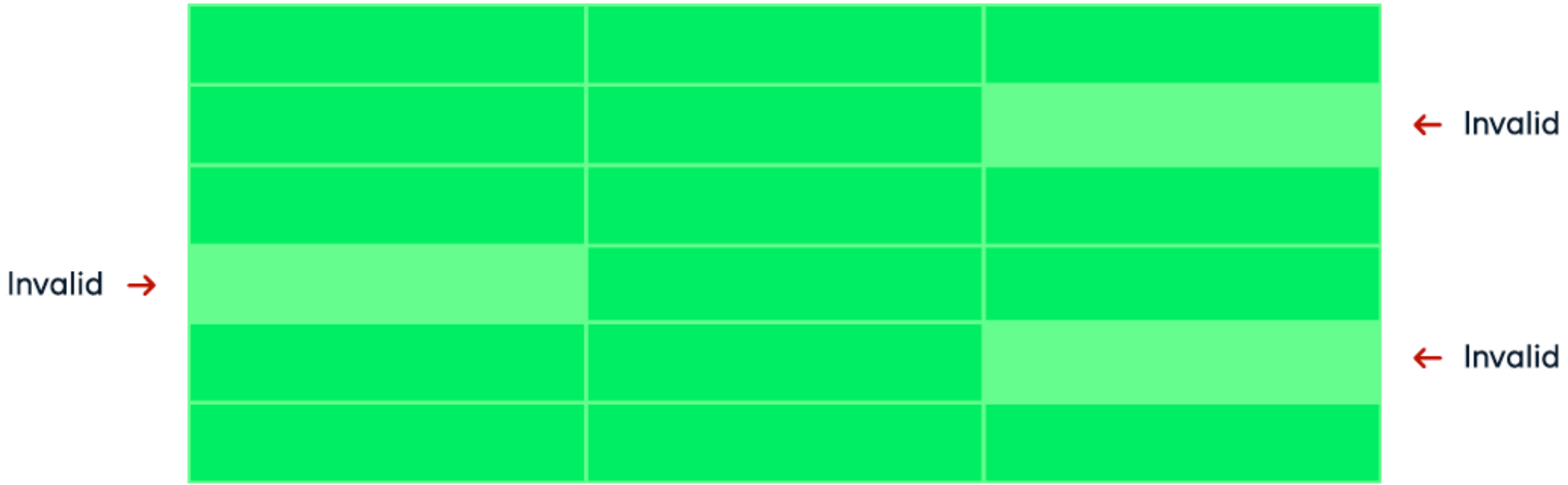
All records must have a value populated in the CustomerName field.

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

[Source: Datacamp]

Validity

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



[Source: Datacamp]

Validity Example

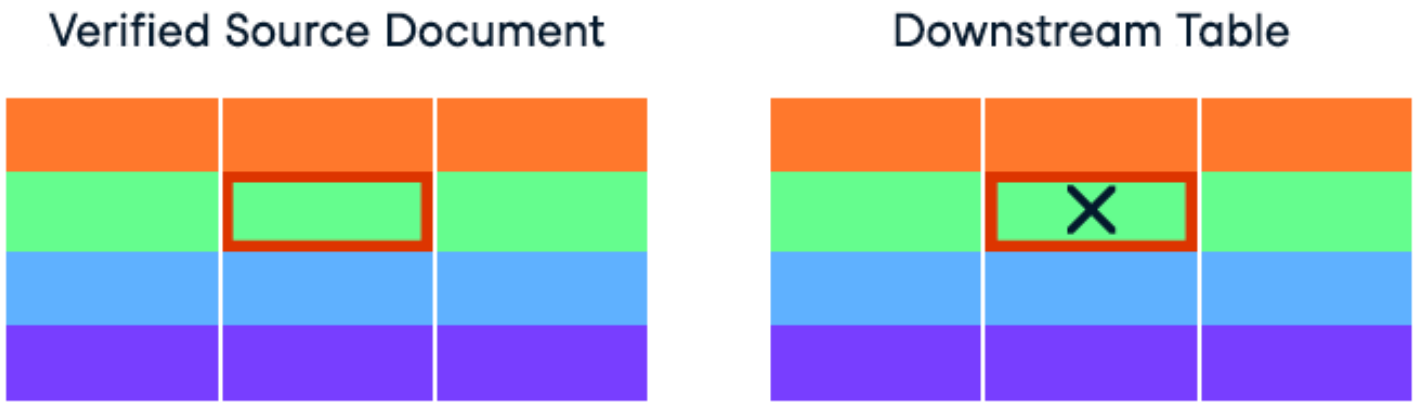
- CustomerBirthDate value must be a date in the past.
- CustomerAccountType value must be either Loan or Deposit.
- LatestAccountOpenDate value must be a date in the past.

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

[Source: Datacamp]

> Accuracy

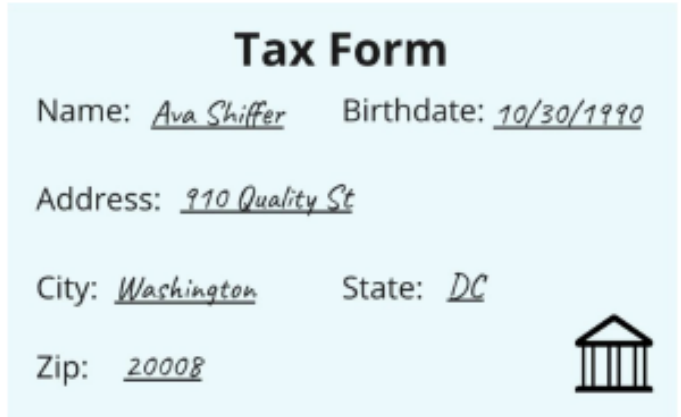
Accuracy measures the degree to which data is correct and represents the truth.



[Source: Datacamp]

Accuracy Example

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



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

[Source: Datacamp]

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

> Uniqueness

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

[Source: Datacamp]

Uniqueness Example

All records must have a unique CustomerID and CustomerName.

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

[Source: Datacamp]

> Timeliness

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

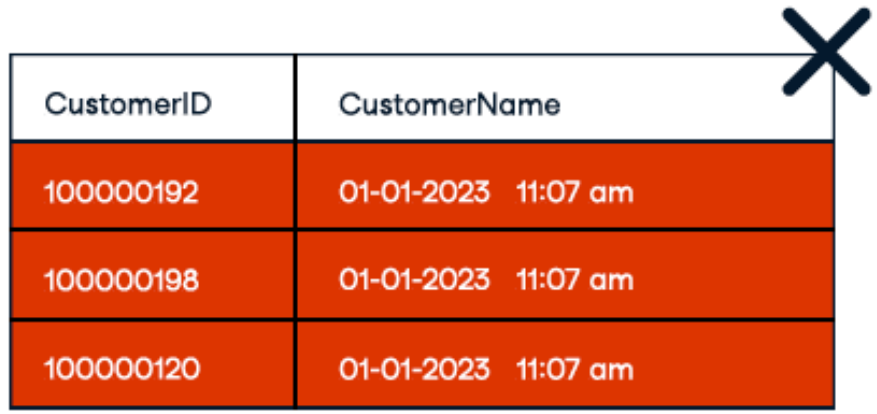
SLA	Table Load Time
08:00 am	07:59 am
10:00 am	09:59 am
11:00 am	11:01 am

← Missed the SLA

[Source: Datacamp]

Timeliness Example

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



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

[Source: Datacamp]

> Consistency

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

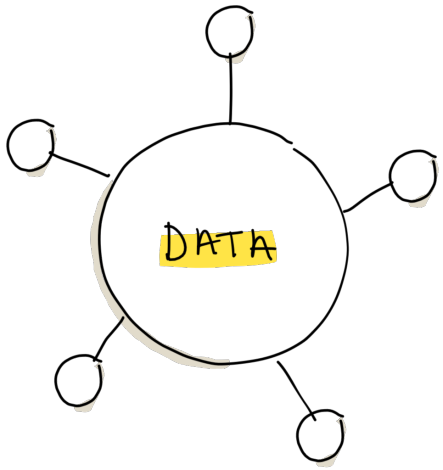
[Source: Datacamp]

Consistency Example

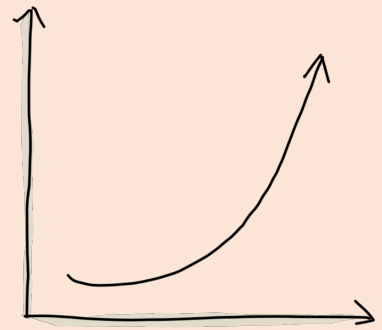
AccountTableCustomerID	CustomerTableCustomerID
108394858	108394858
192039482	192039482
203475849	NULL ✗
2930485953	NULL ✗
102832748	102832748

[Source: Datacamp]

Da ta Quality

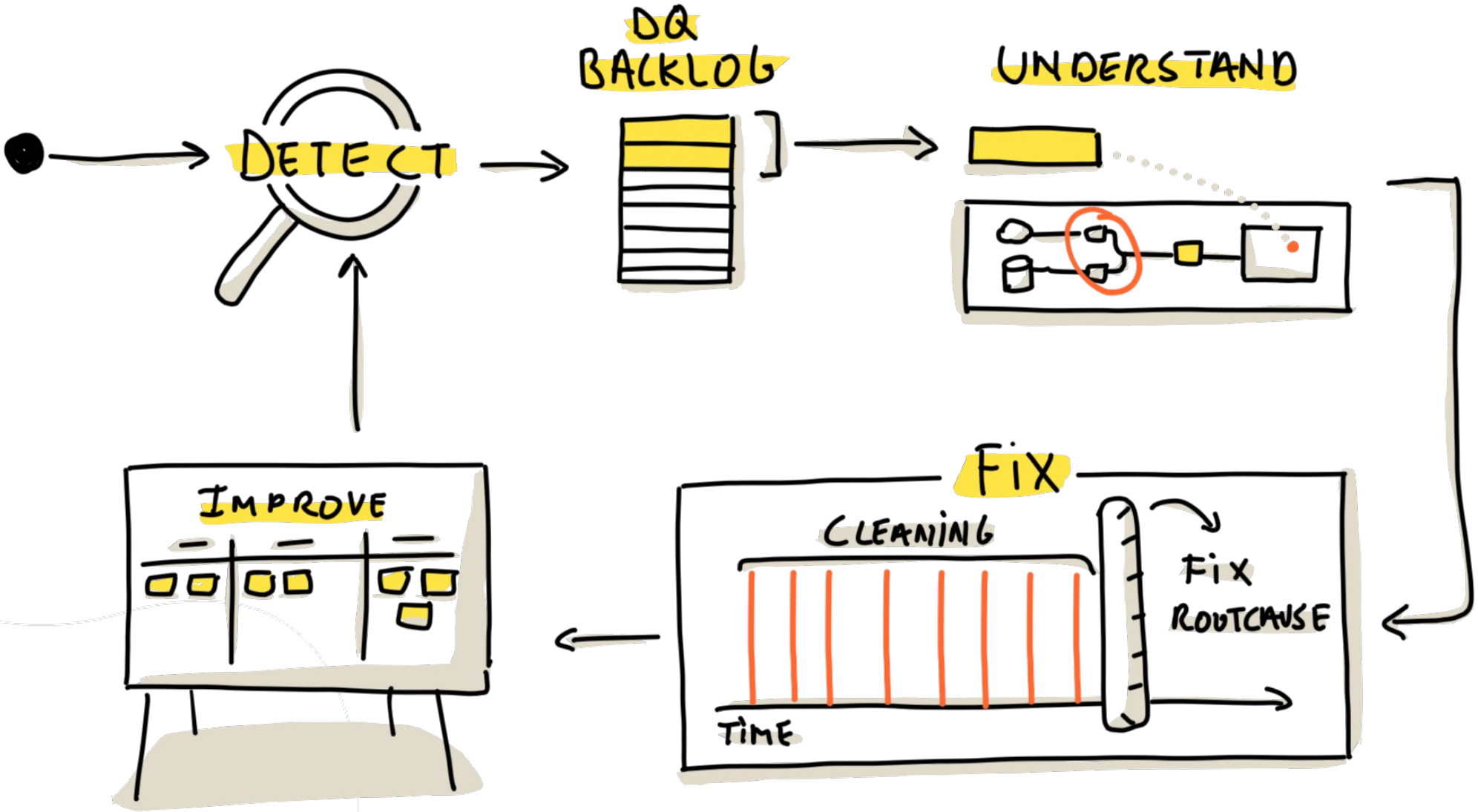


DQ DIMENSIONS

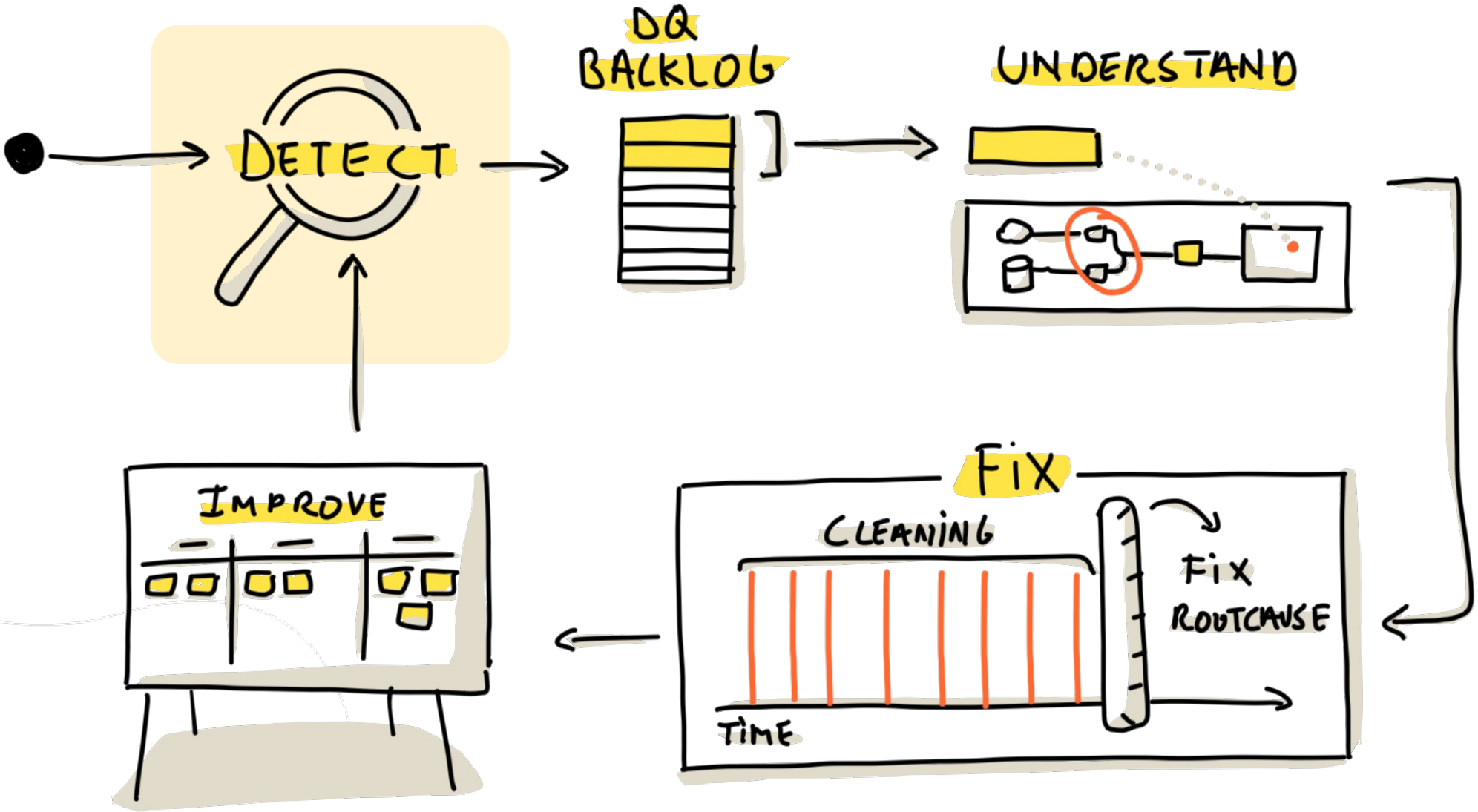


DQ PROCESS
&TOOLS

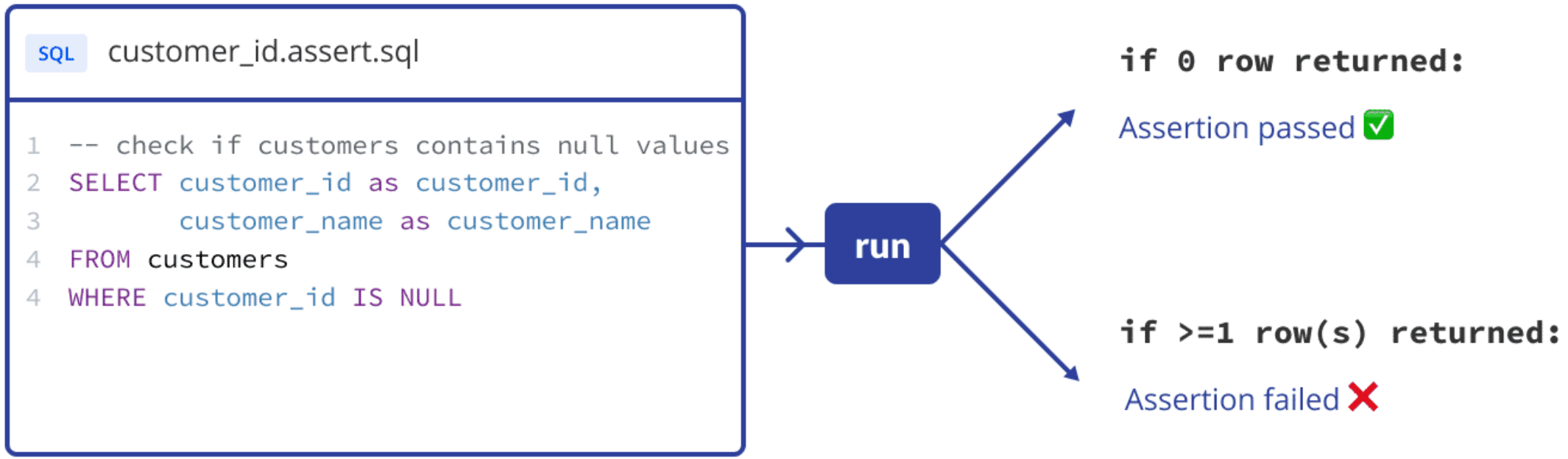
Data Quality Process



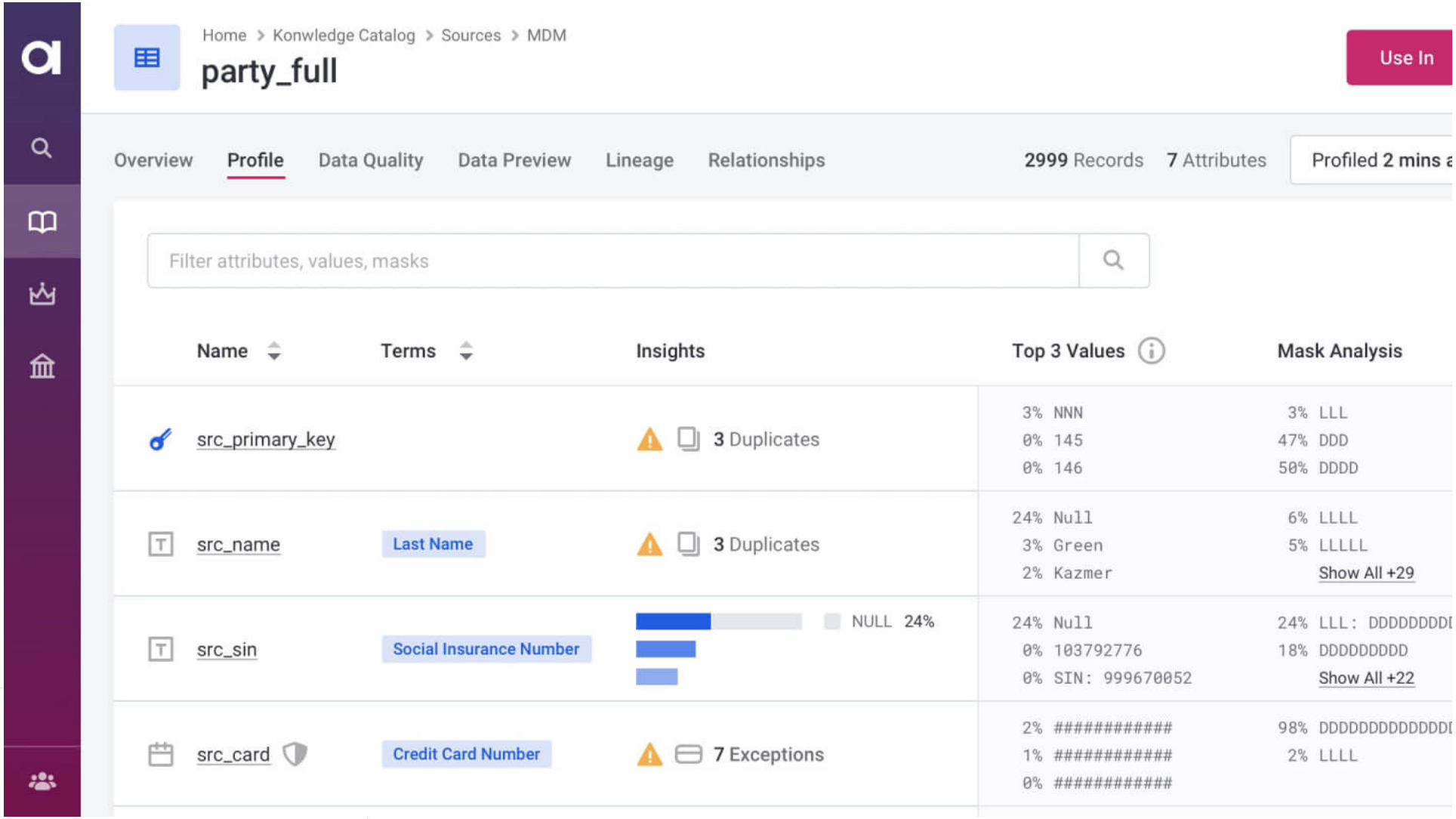
Data Quality Process



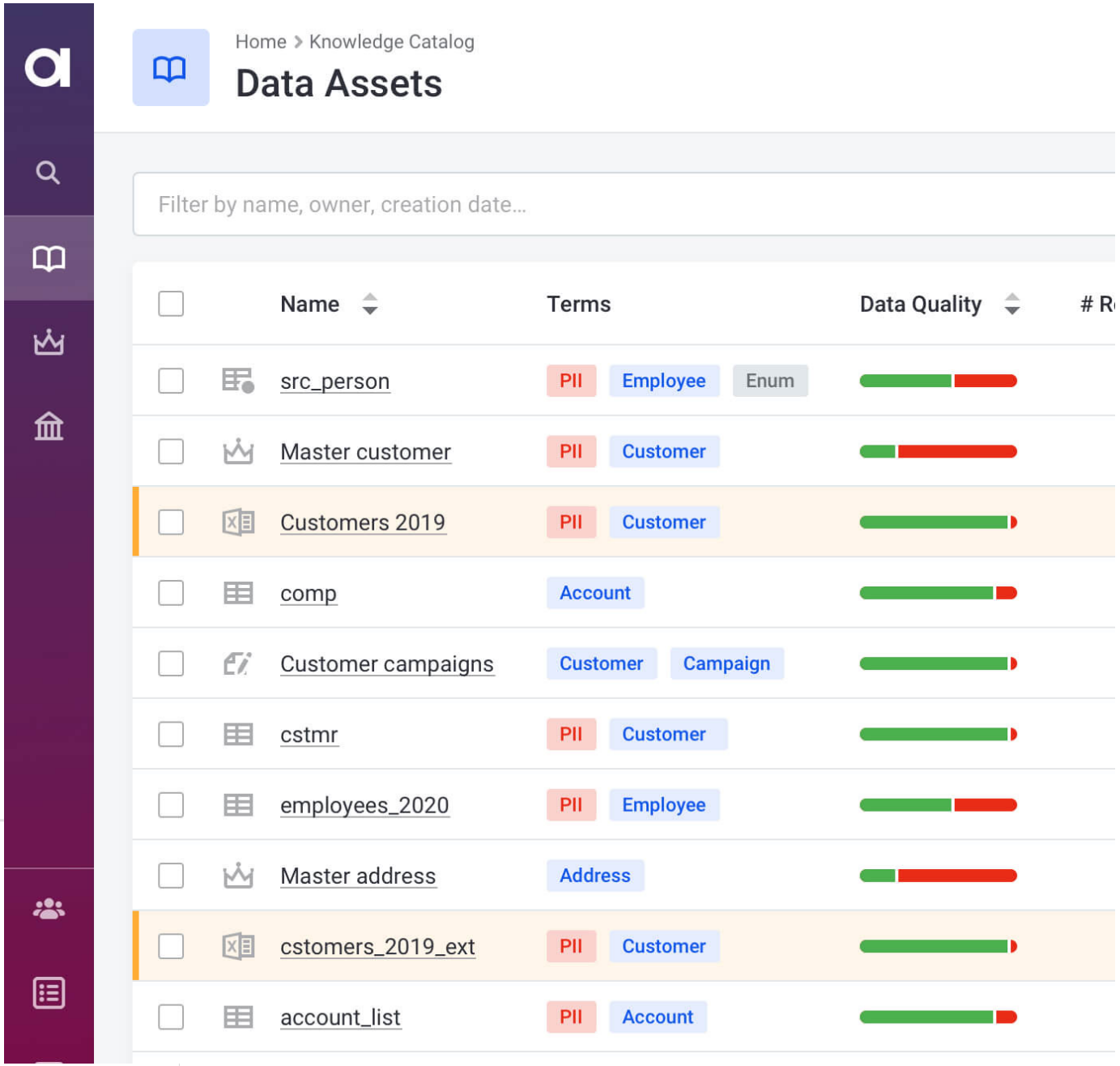
DQ Tests


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

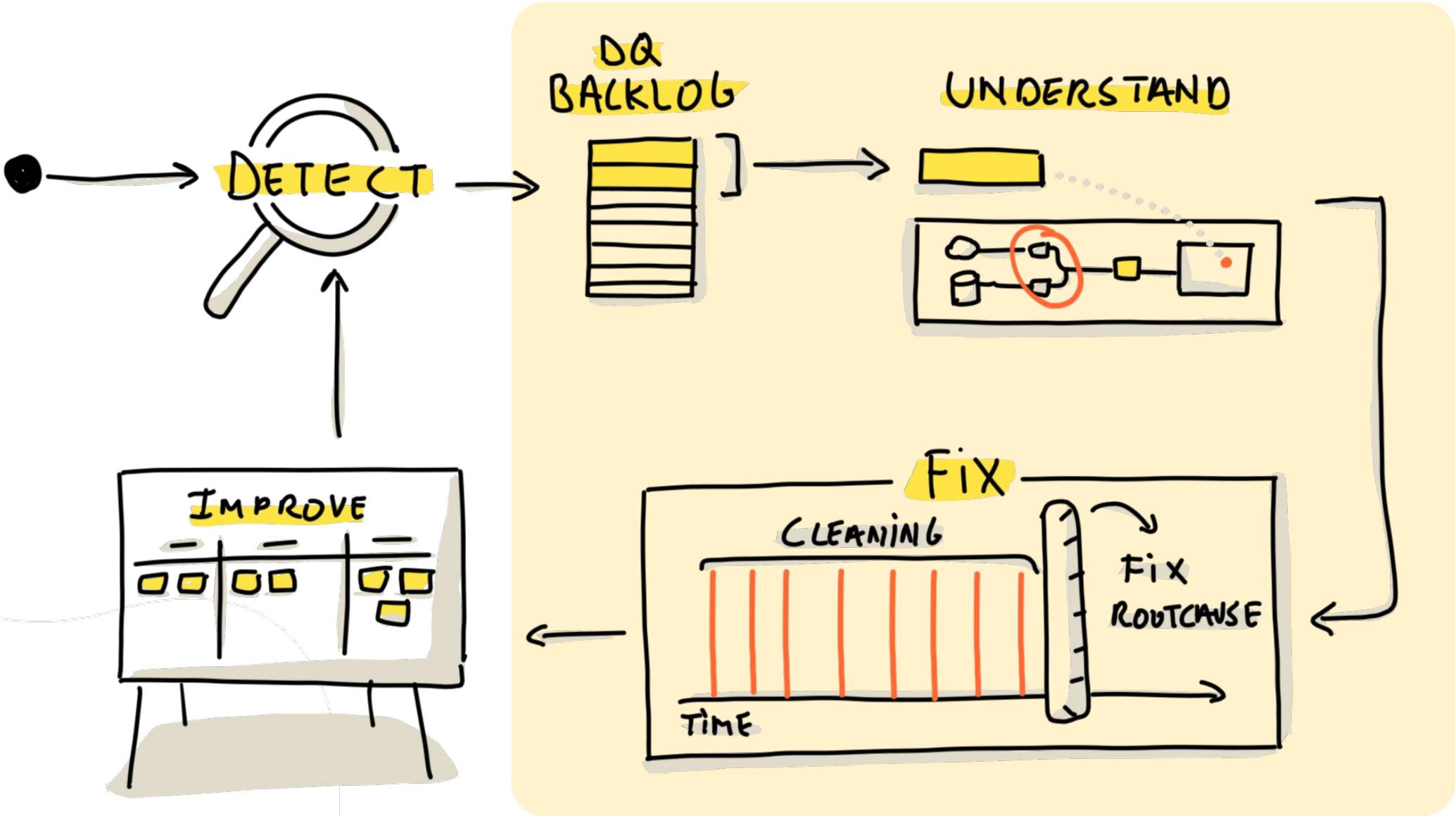


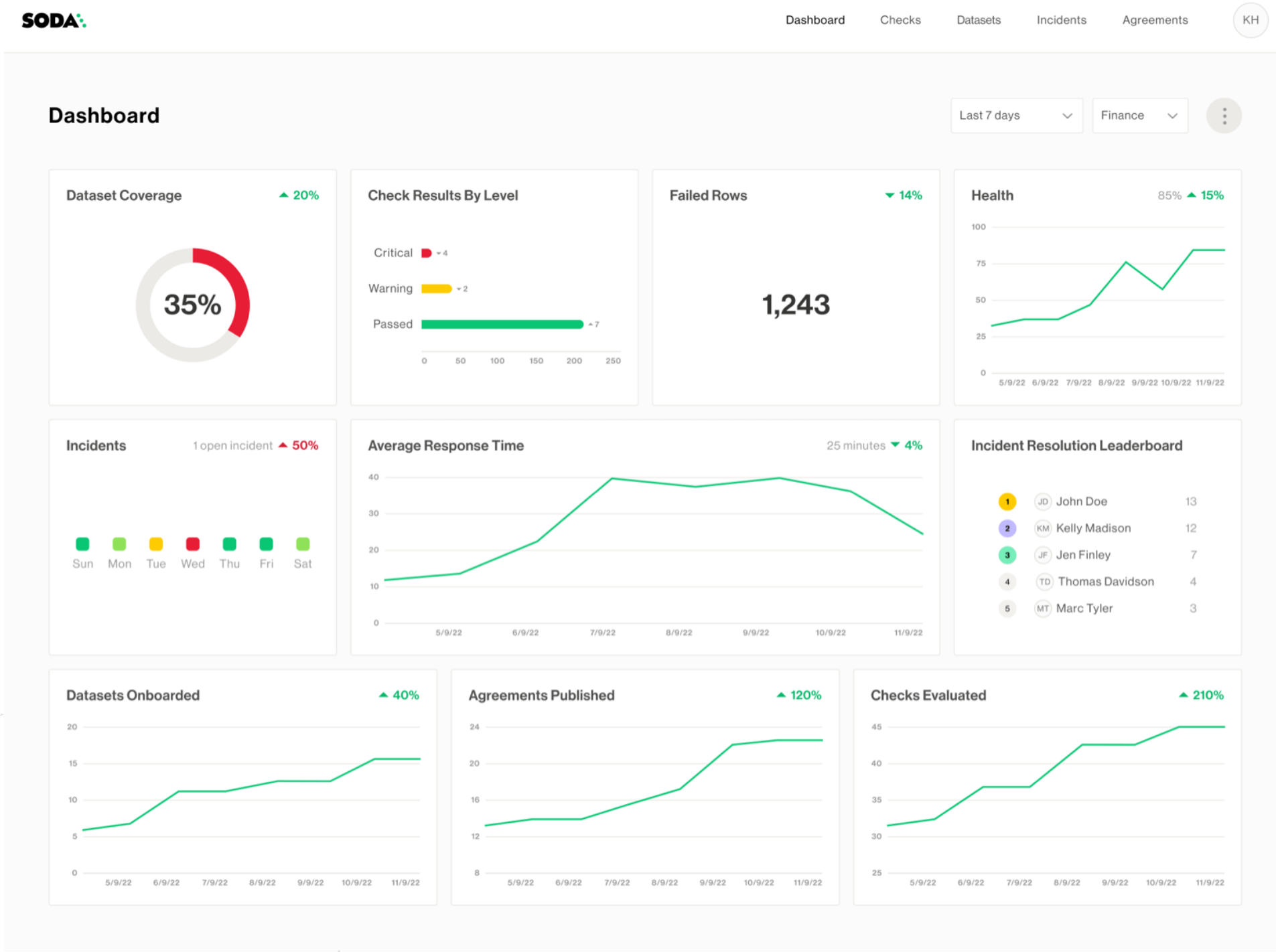
[Alation]


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

[Alation]

Data Quality Process




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

[SODA]

EXERCISE

- Which data-processes can help you to realize your use cases?
  - Data Delivery
  - Data Quality
  - Data Access Mngmnt
  - …
- How well are these processes defined today?

7.3

TECHNOLOGY

DATA TECHNOLOGY projects can become LONG RUNNING & EXPENSIVE projects

BE AWARE OF DEAD HORSES

Source: Oscar Berg

THE DEAD HORSE THEORY

ILLUSTRATED

Source: Oscar Berg

Author of:

Illustrations are part of his new book (to be announced)

The tribal wisdom of the Indians, passed on from generation to generation, says that, "When you discover that you are riding a dead horse, the best strategy is to dismount."

Source: Oscar Berg

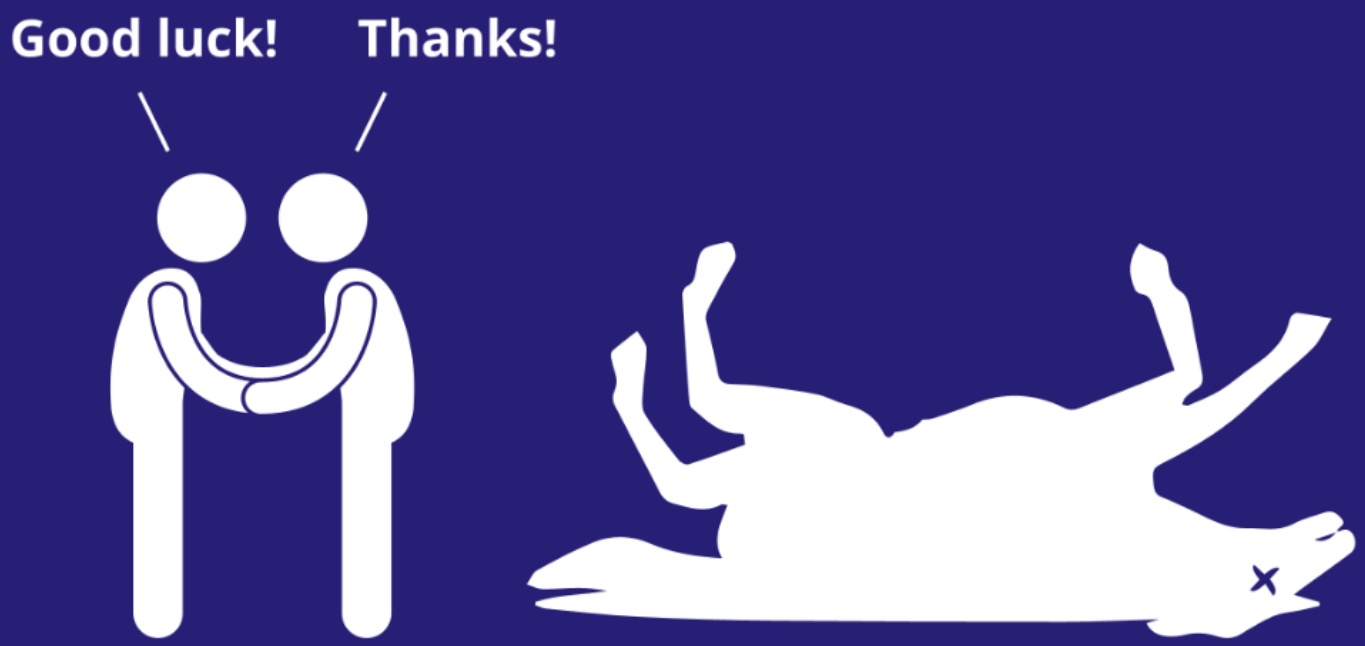
The Dead Horse Theory goes on to say that in modern business, education and government, far more advanced strategies are often employed, such as:

Source: Oscar Berg

1. Buying a stronger whip.

Source: Oscar Berg

2. Changing riders.

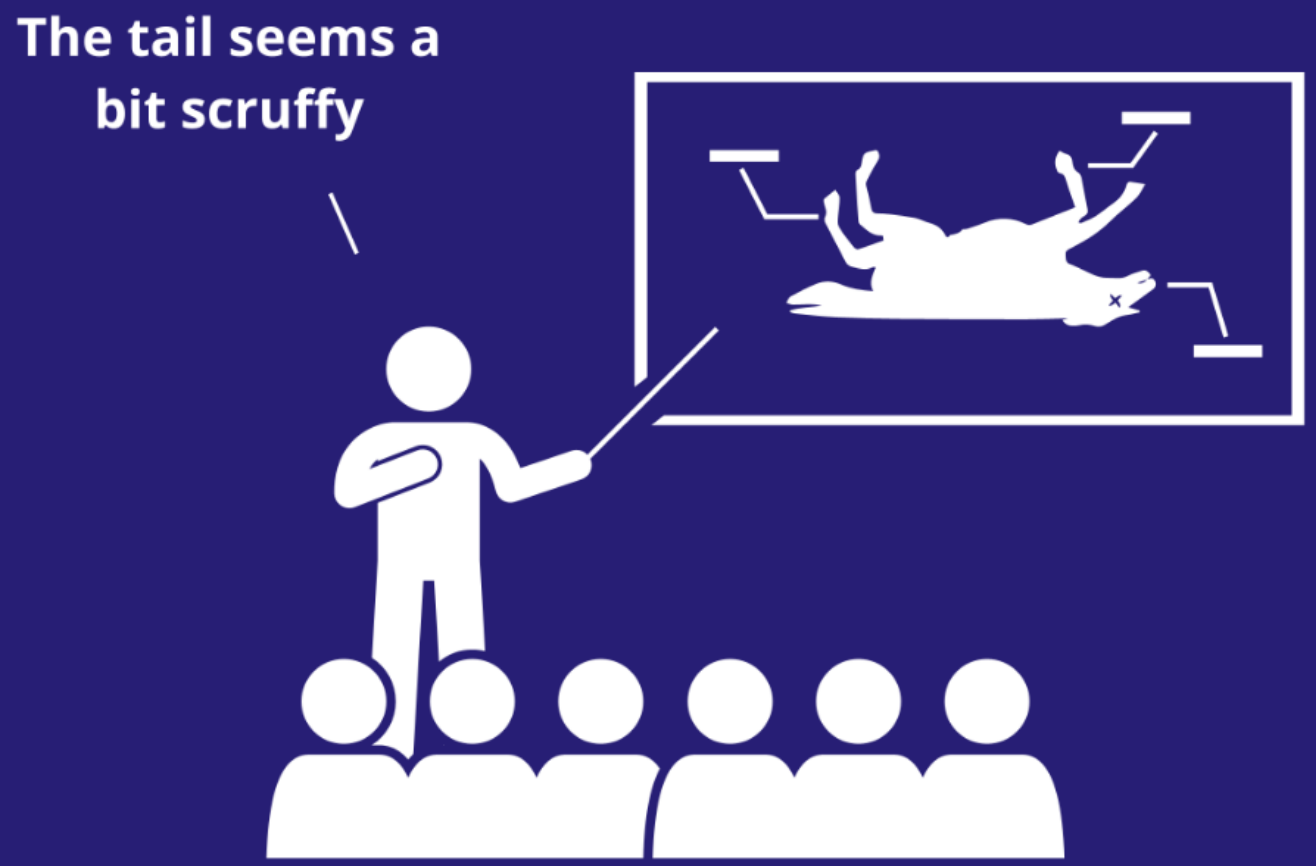


Source: Oscar Berg

3. Threatening the horse with termination.

Source: Oscar Berg

4. Appointing a committee to study the horse.



Source: Oscar Berg

5. Arranging to visit other countries to see how others ride dead horses.

Source: Oscar Berg

6. Lowering the standards so that dead horses can be included.

Source: Oscar Berg

7. Re-classifying the dead horse as ‘living-impaired’.

Source: Oscar Berg

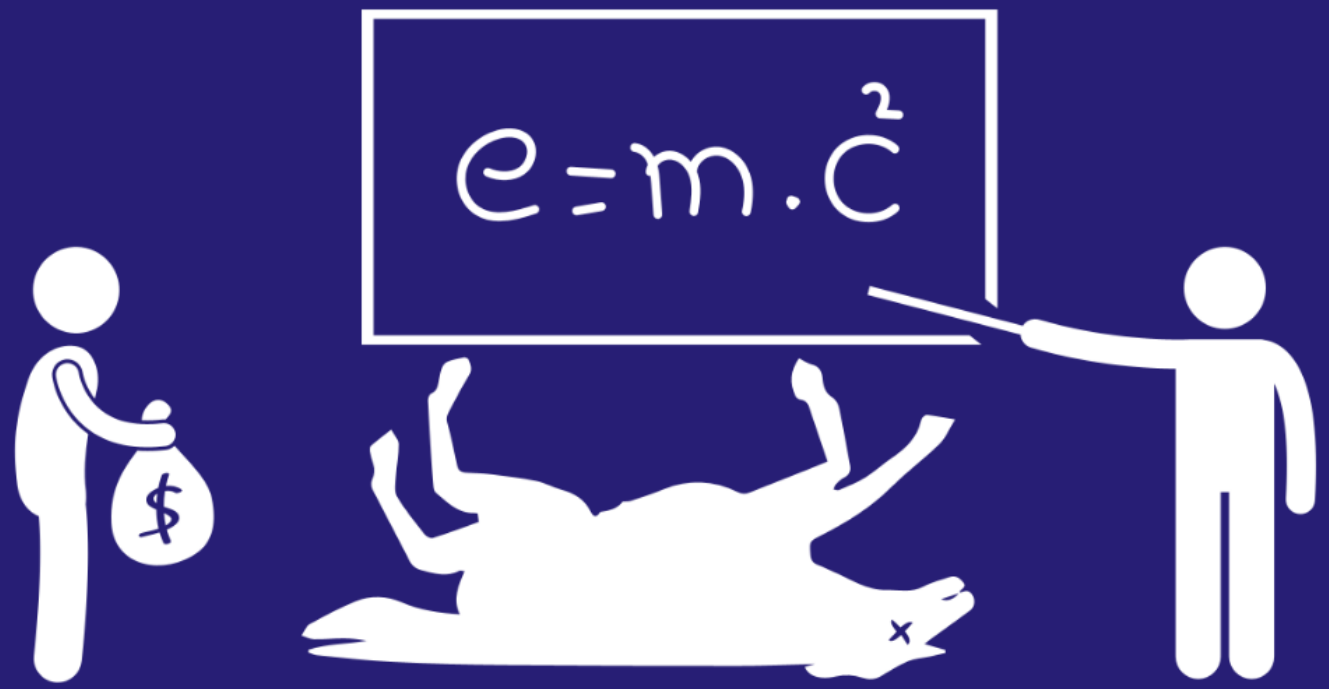
8. Hiring outside contractors to ride the dead horse.

Source: Oscar Berg

9. Harnessing several dead horses together to increase the speed.

Source: Oscar Berg

10. Providing additional funding and/or training to increase the dead horse's performance.



Source: Oscar Berg

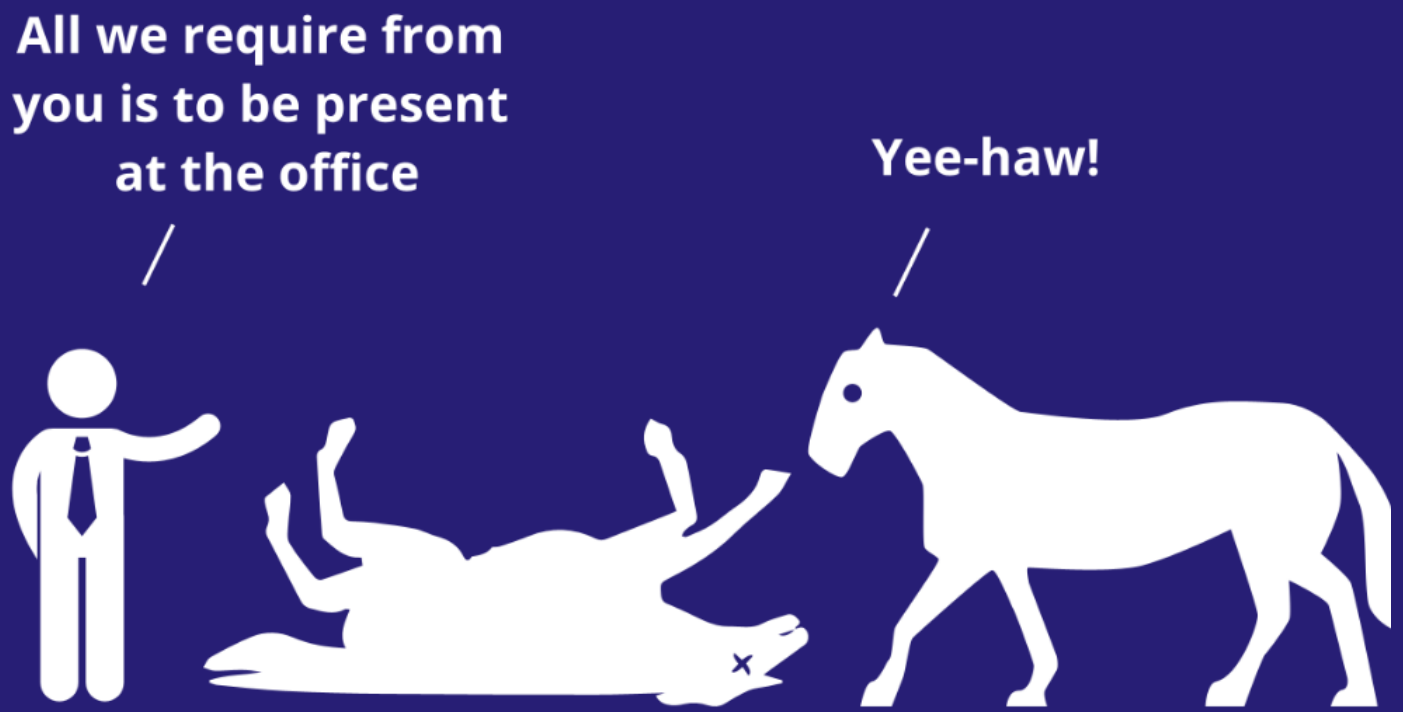
11. Doing a productivity study to see if lighter riders would improve the dead horse's performance.

Source: Oscar Berg

12. Declaring that as the dead horse does not have to be fed, it is less costly, carries lower overhead and, therefore, contributes substantially more to the bottom line of the economy than do some other horses.

Source: Oscar Berg

13. Re-writing the expected performance requirements for all horses.

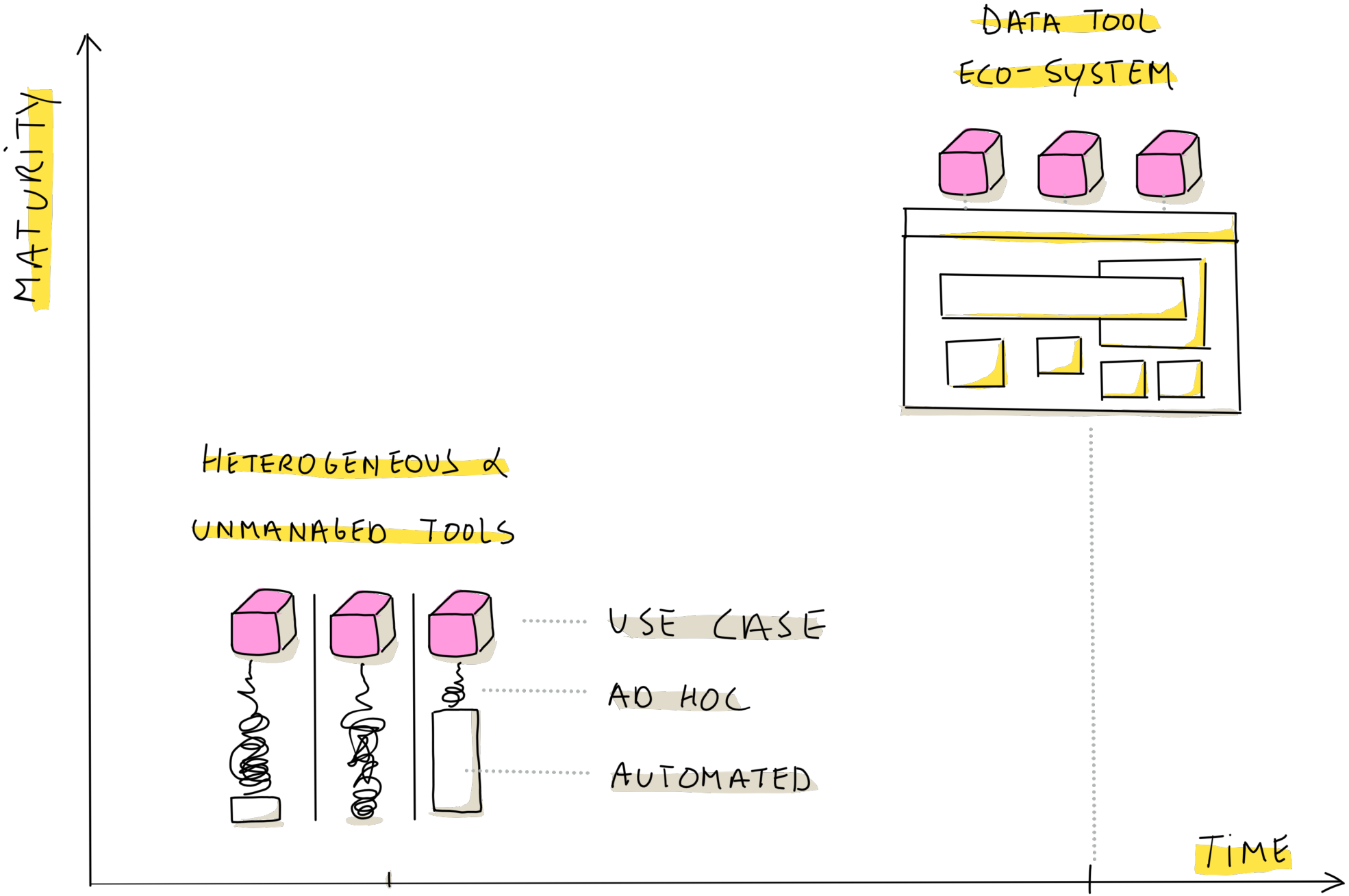


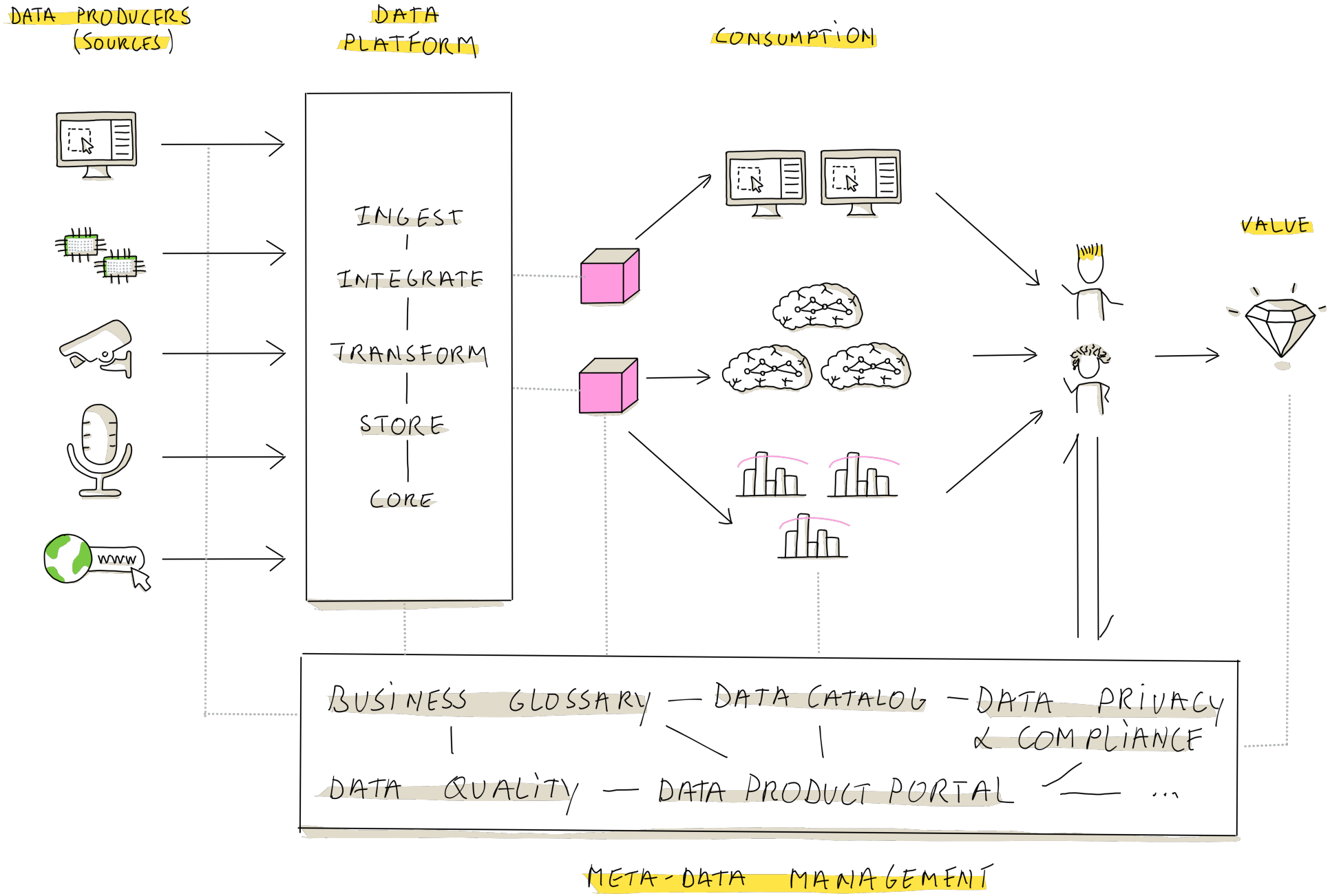
Source: Oscar Berg

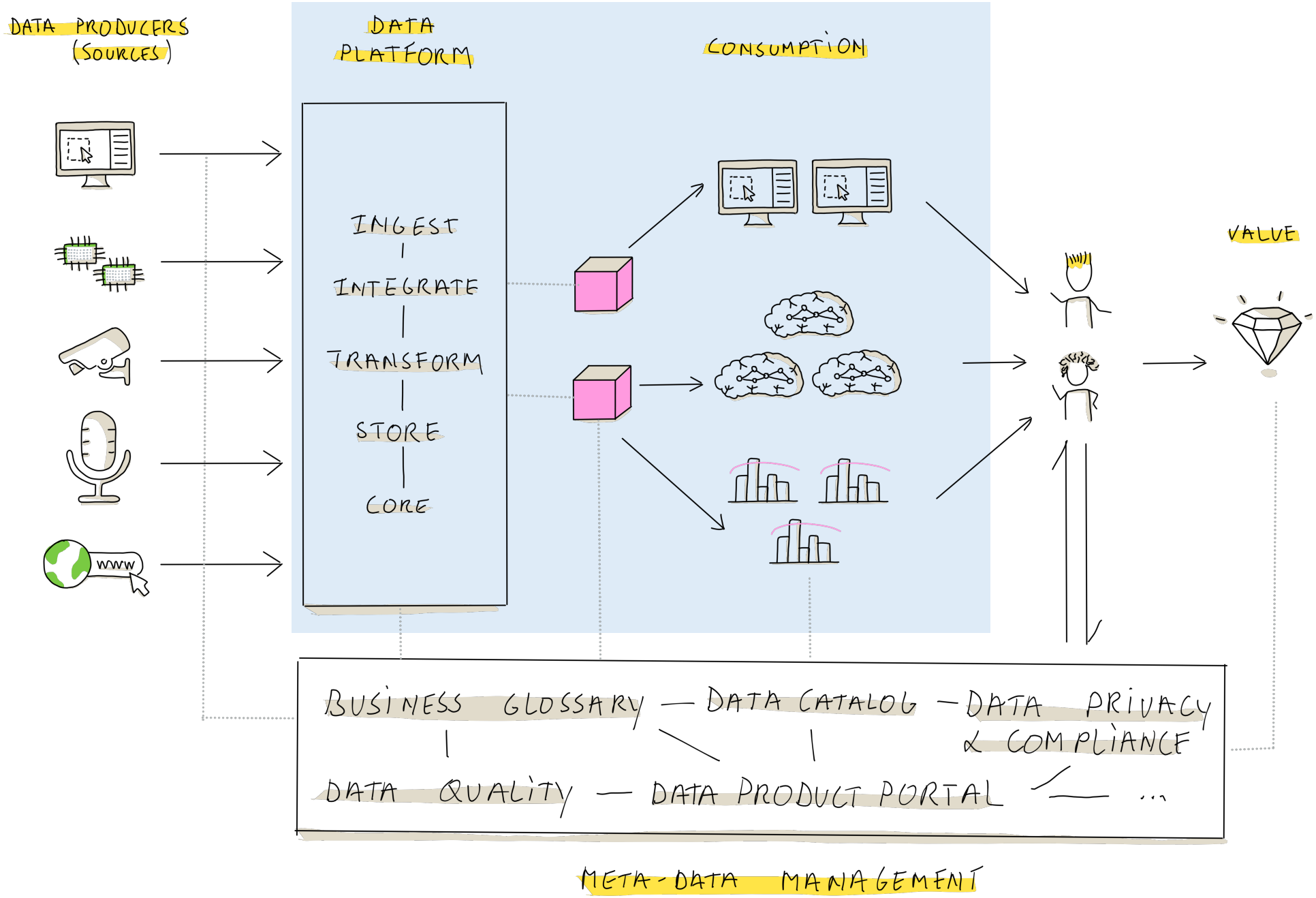
Did you ever worked with dead horses?

7.3

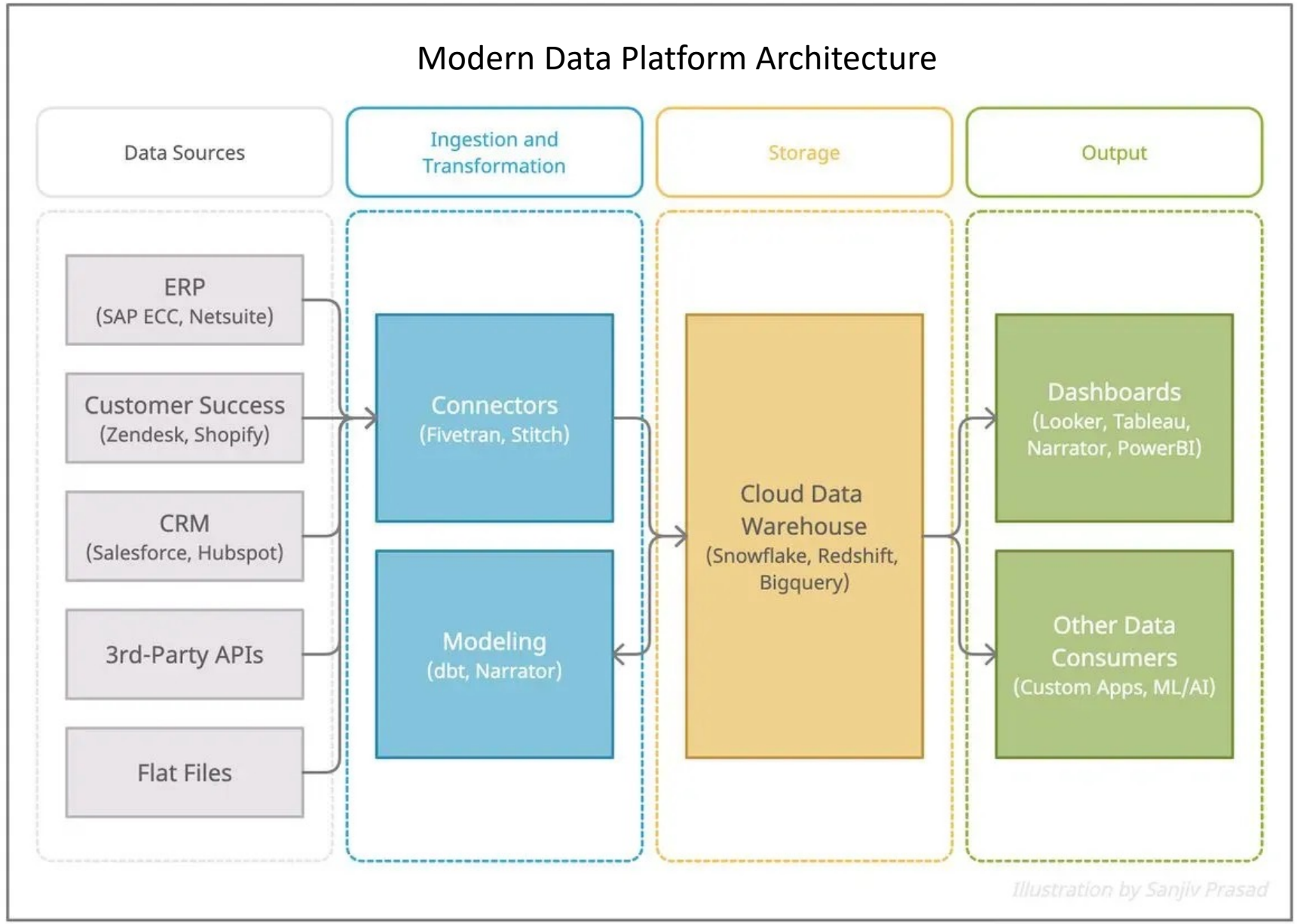
TECHNOLOGY

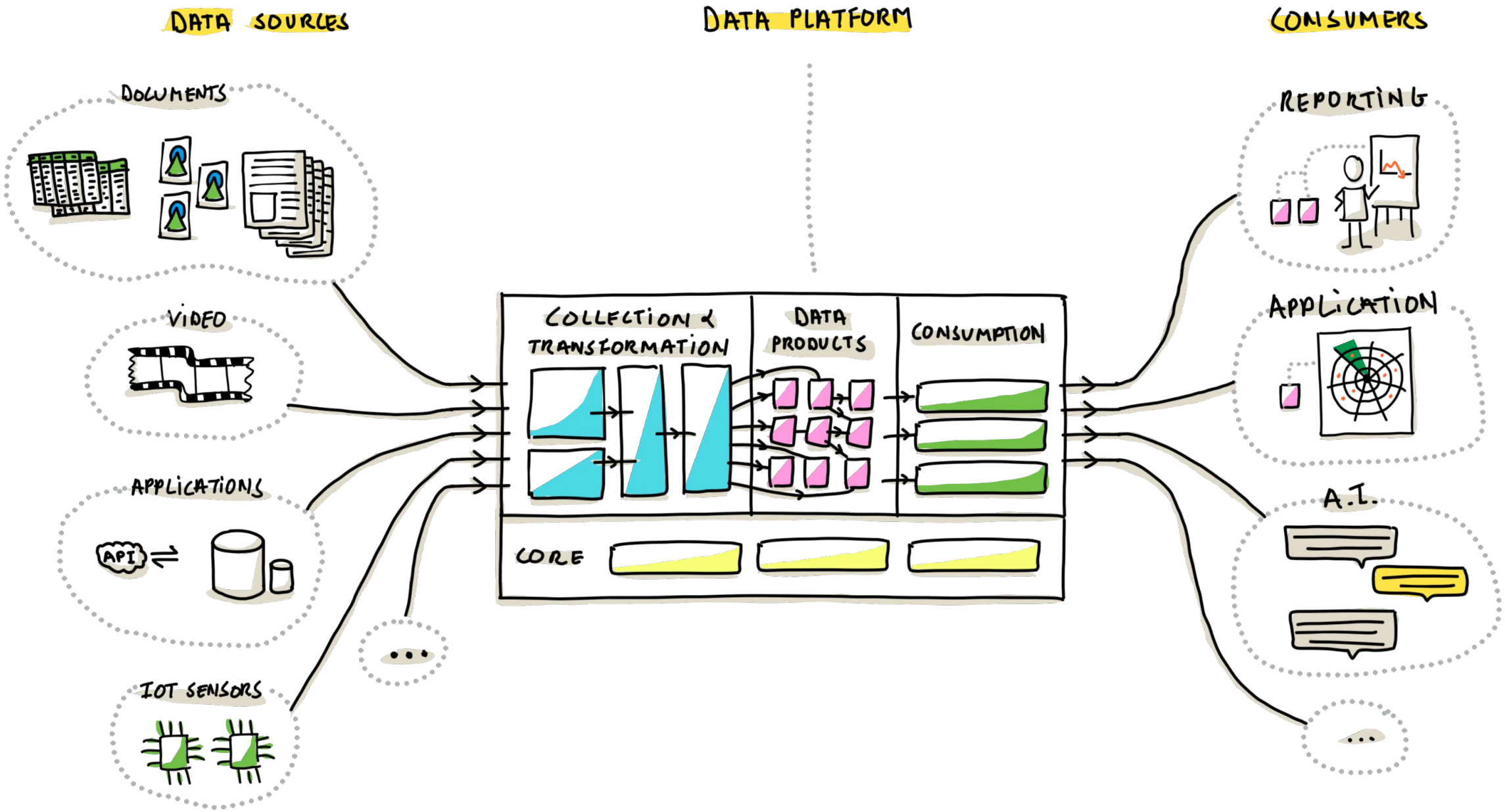

MATURITY
DATA TOOL
ECO-SYSTEM
HETEROGENEOUS &
UNMANAGED TOOLS
USE CASE
AD HOC
AUTOMATED
TIME

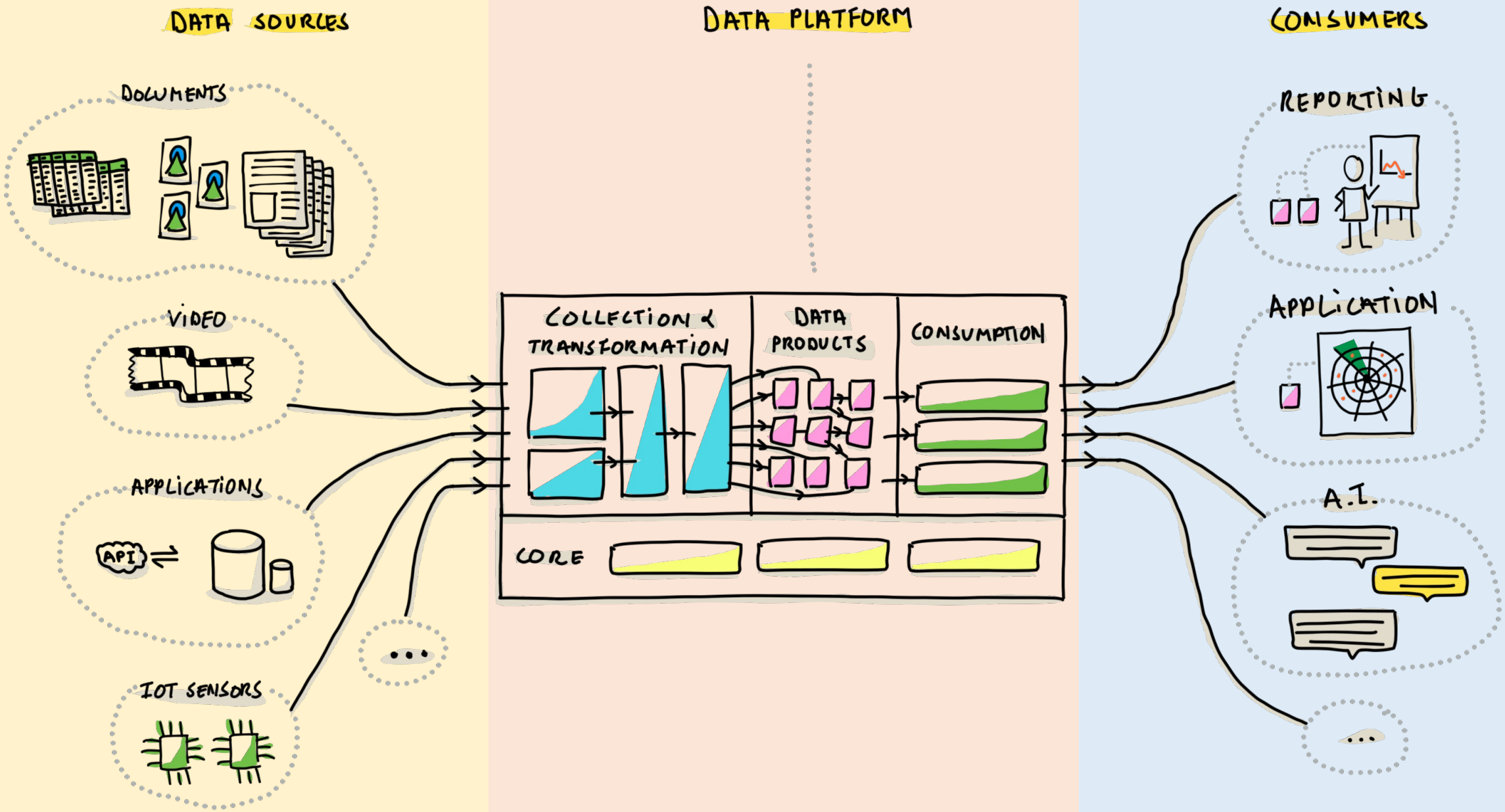

DATA PRODUCERS (SOURCES)
DATA PLATFORM
CONSUMPTION
VALUE
INGEST
INTEGRATE
TRANSFORM
STORE
CORE
www
BUSINESS GLOSSARY
DATA CATALOG
DATA PRIVACY & COMPLIANCE
DATA QUALITY
DATA PRODUCT PORTAL
...
META-DATA MANAGEMENT


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

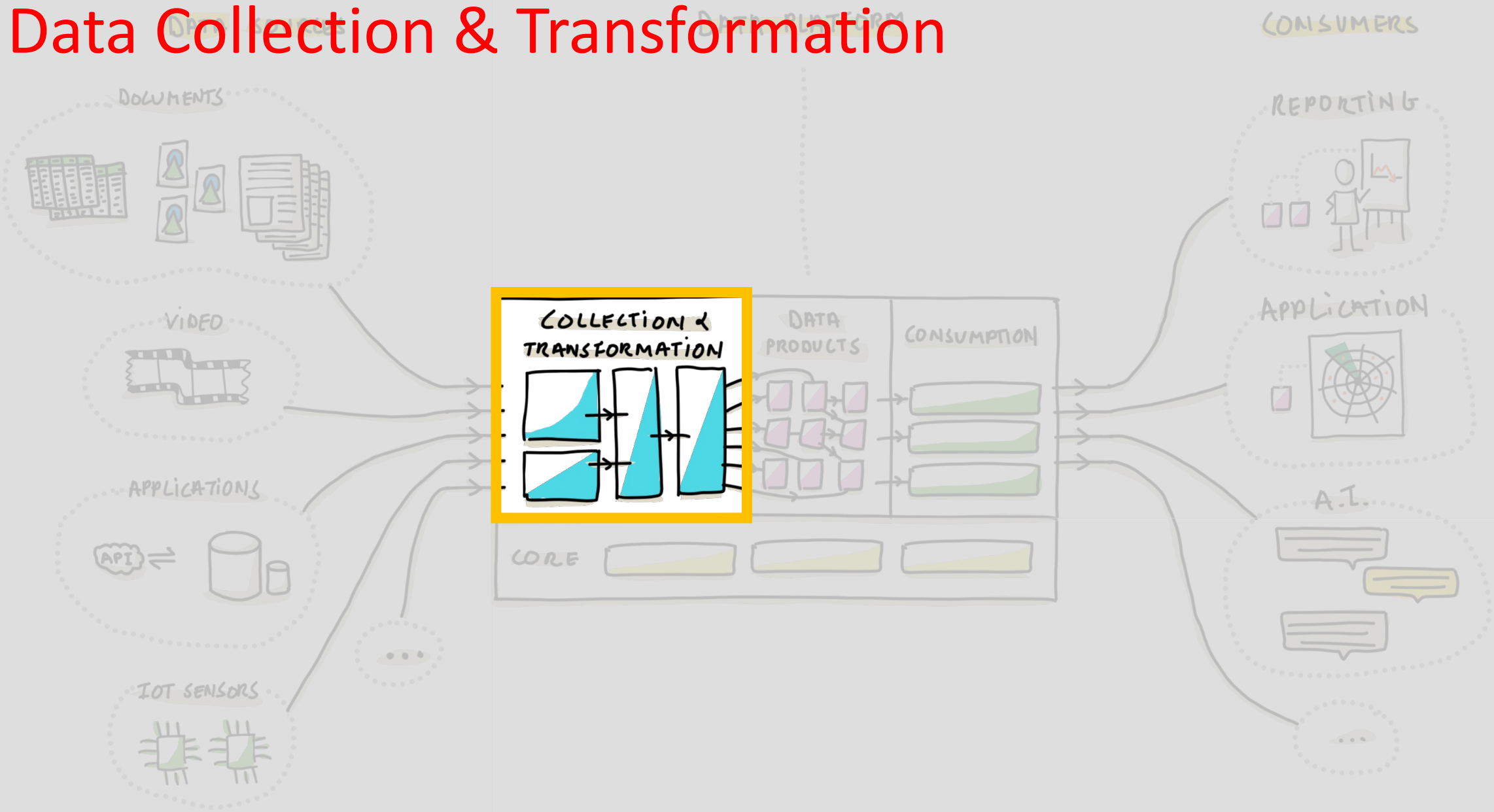
Modern Data Platform Architecture




DATA SOURCES
DATA PLATFORM
CONSUMERS
DOCUMENTS
VIDEO
APPLICATIONS
API
IOT SENSORS
COLLECTION & TRANSFORMATION
DATA PRODUCTS
CONSUMPTION
CORE
REPORTING
APPLICATION
A.I.


DATA SOURCES
DOCUMENTS
VIDEO
APPLICATIONS
API
IOT SENSORS
DATA PLATFORM
COLLECTION & TRANSFORMATION
DATA PRODUCTS
CONSUMPTION
CORE
CONSUMERS
REPORTING
APPLICATION
A.I.
"Operational Plane"
"Analytical Plane"
"Operational / Analytical Plane"

1. Data Collection & Transformation



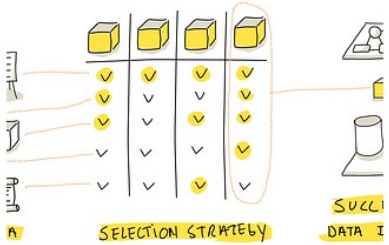
"Operational Plane" "Analytical Plane" "Operational / Analytical Plane"

See also my Medium Blog

janmeskens in The Modern Scientist

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...

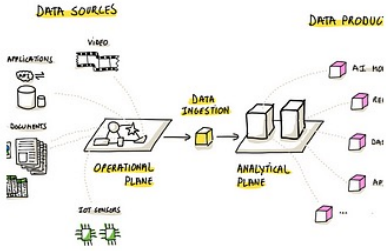


Jan 17 715 5

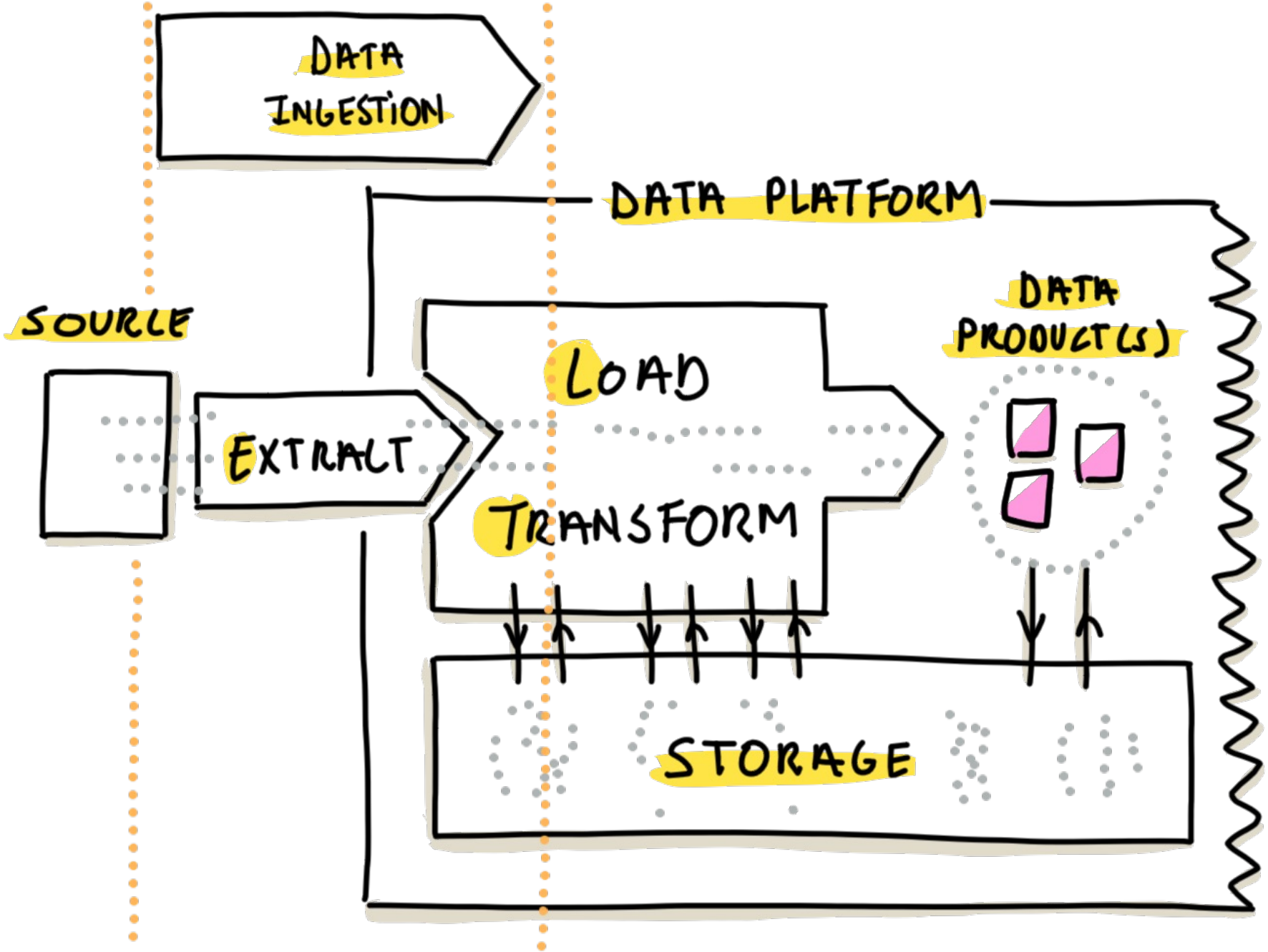
janmeskens in The Modern Scientist

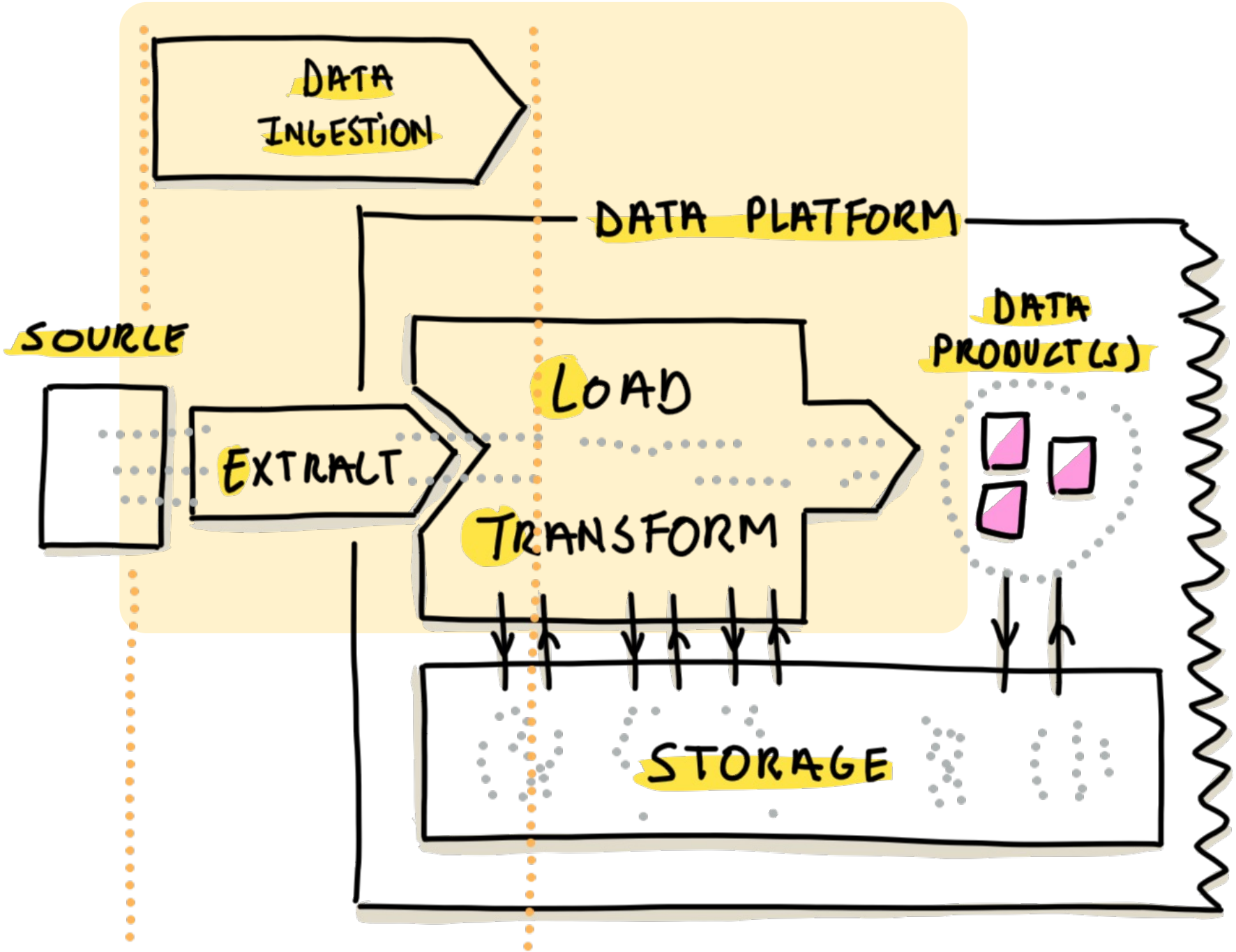
Data Ingestion — Part 1: Architectural Patterns

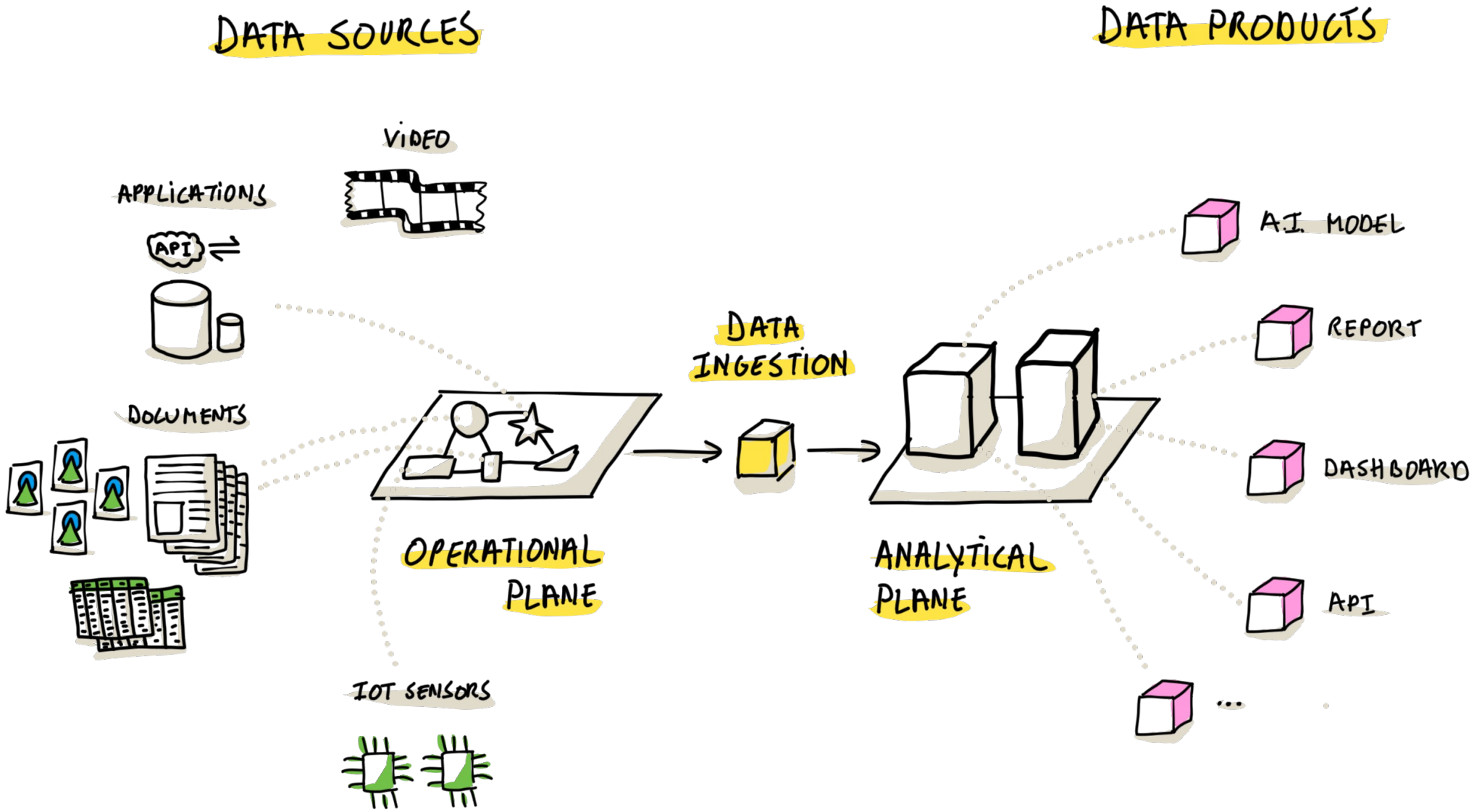
Over the course of two articles, I will thoroughly explore data ingestion, a fundamental process that bridges the operational...

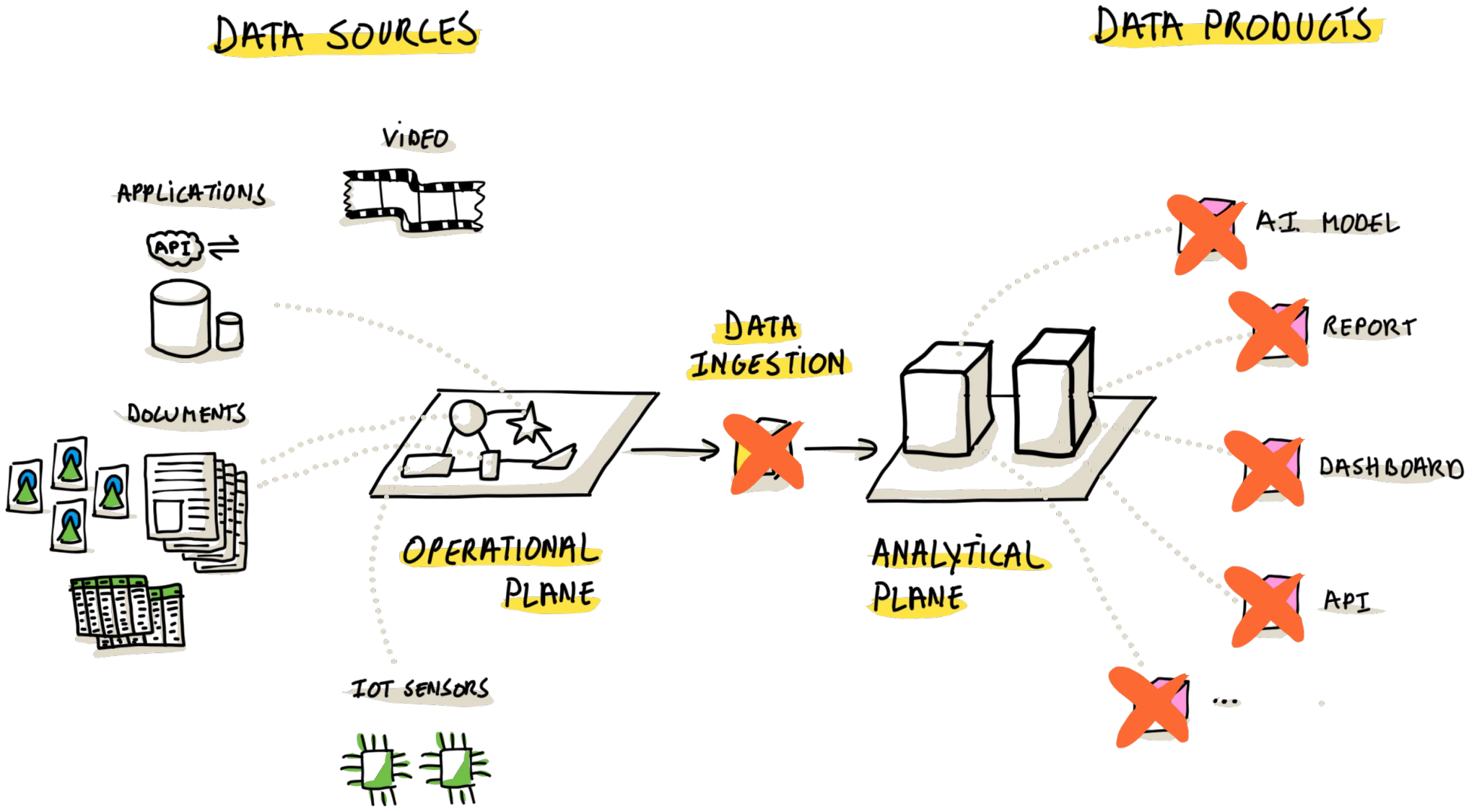


Nov 27, 2023 2.7K 26


DATA
INGESTION
DATA PLATFORM
SOURCE
EXTRACT
LOAD
TRANSFORM
DATA
PRODUCT(S)
STORAGE


DATA
INGESTION
DATA PLATFORM
SOURCE
EXTRACT
LOAD
TRANSFORM
DATA
PRODUCT(S)
STORAGE


DATA SOURCES
DATA PRODUCTS
VIDEO
APPLICATIONS
API
DOCUMENTS
IOT SENSORS
DATA INGESTION
OPERATIONAL PLANE
ANALYTICAL PLANE
A.I. MODEL
REPORT
DASHBOARD
API
...


DATA SOURCES
DATA PRODUCTS
VIDEO
APPLICATIONS
API
DOCUMENTS
IOT SENSORS
DATA INGESTION
OPERATIONAL PLANE
ANALYTICAL PLANE
A.I. MODEL
REPORT
DASHBOARD
API
...

Definition

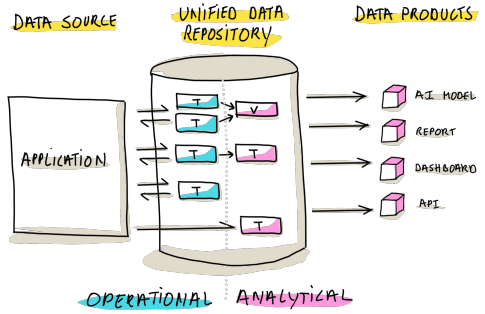
Data ingestion refers to the tools & processes used to **collect data from various sources and move it to a target site**, either in **batches or in real-time**. The data ingestion layer is **critical to your downstream** data science, BI, and analytics systems which depend on **timely, complete, and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

[Qlik – Data Ingestion]

Data Ingestion Patterns

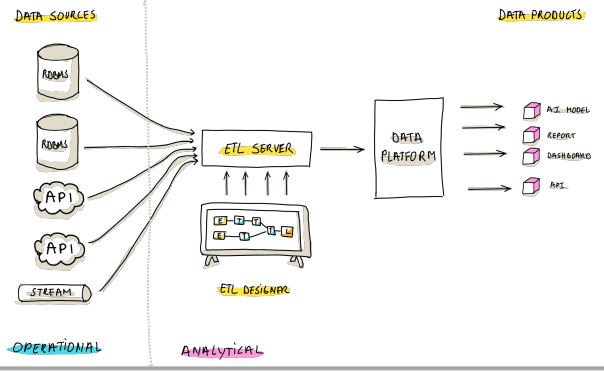
Pattern #1

UNIFIED DATA REPOSITORY



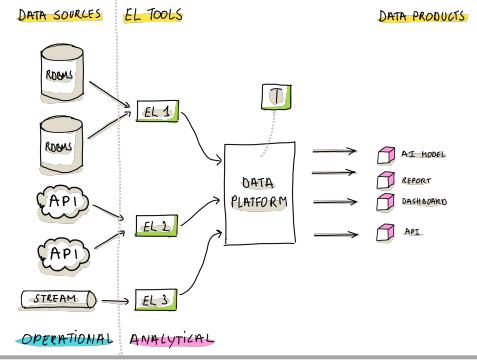
Pattern #2

ETL – Extract Transform Load



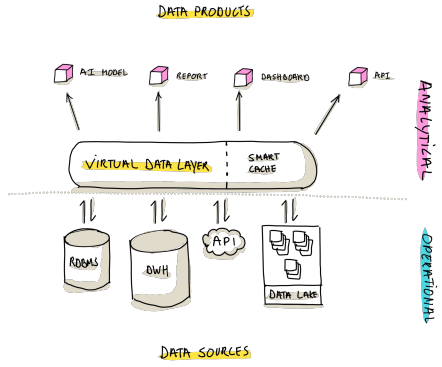
Pattern #3

ELT – Extract Transform Load



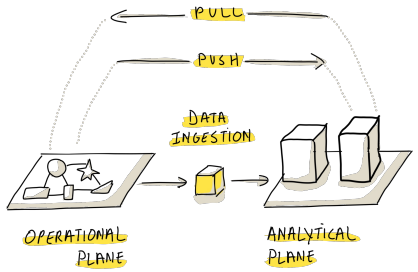
Pattern #4

Data Virtualization



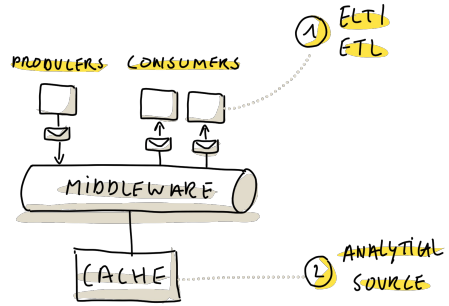
Pattern #5

Push vs Pull



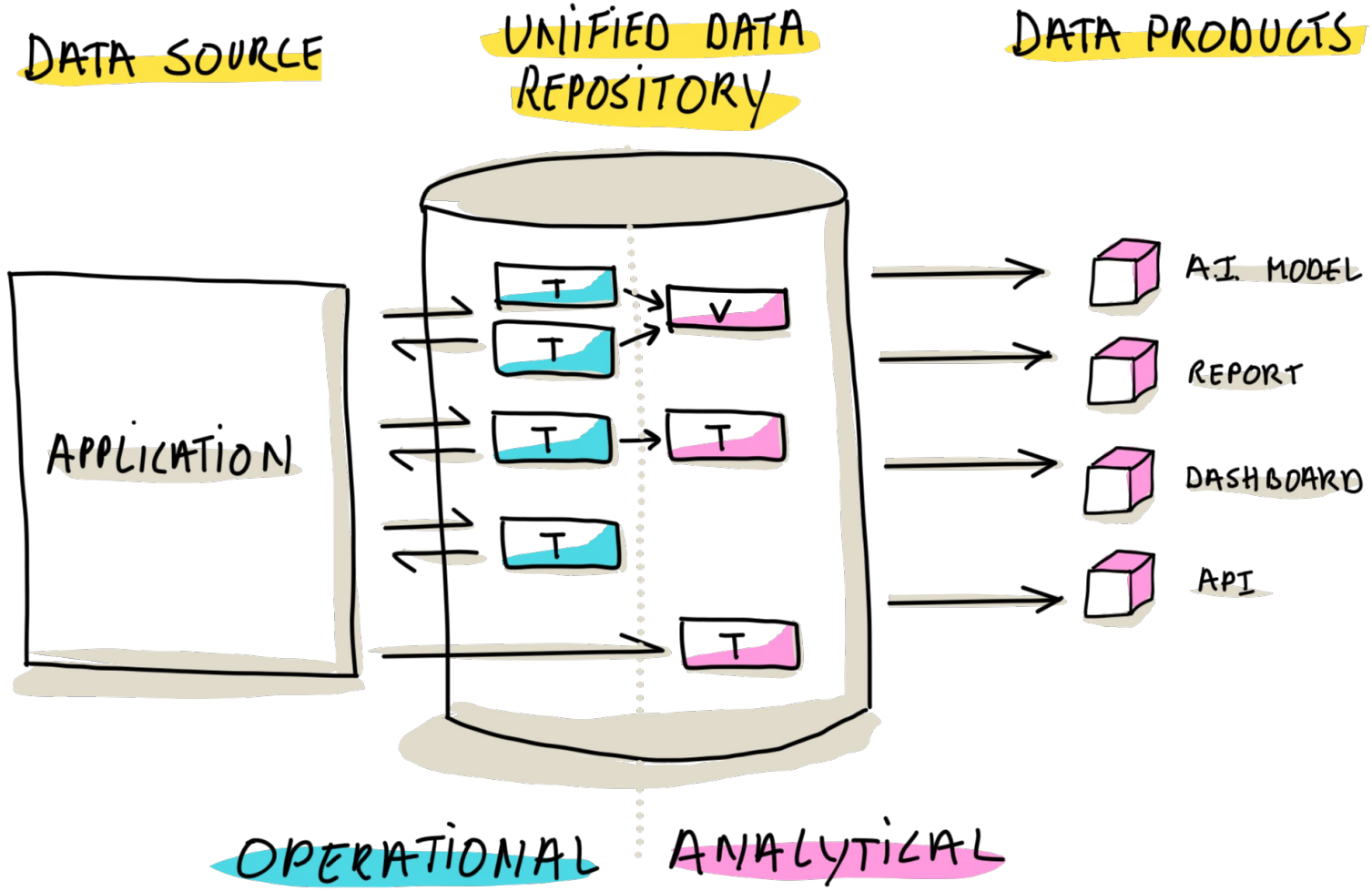
Pattern #6

Streaming Data



Pattern #1

UNIFIED DATA REPOSITORY



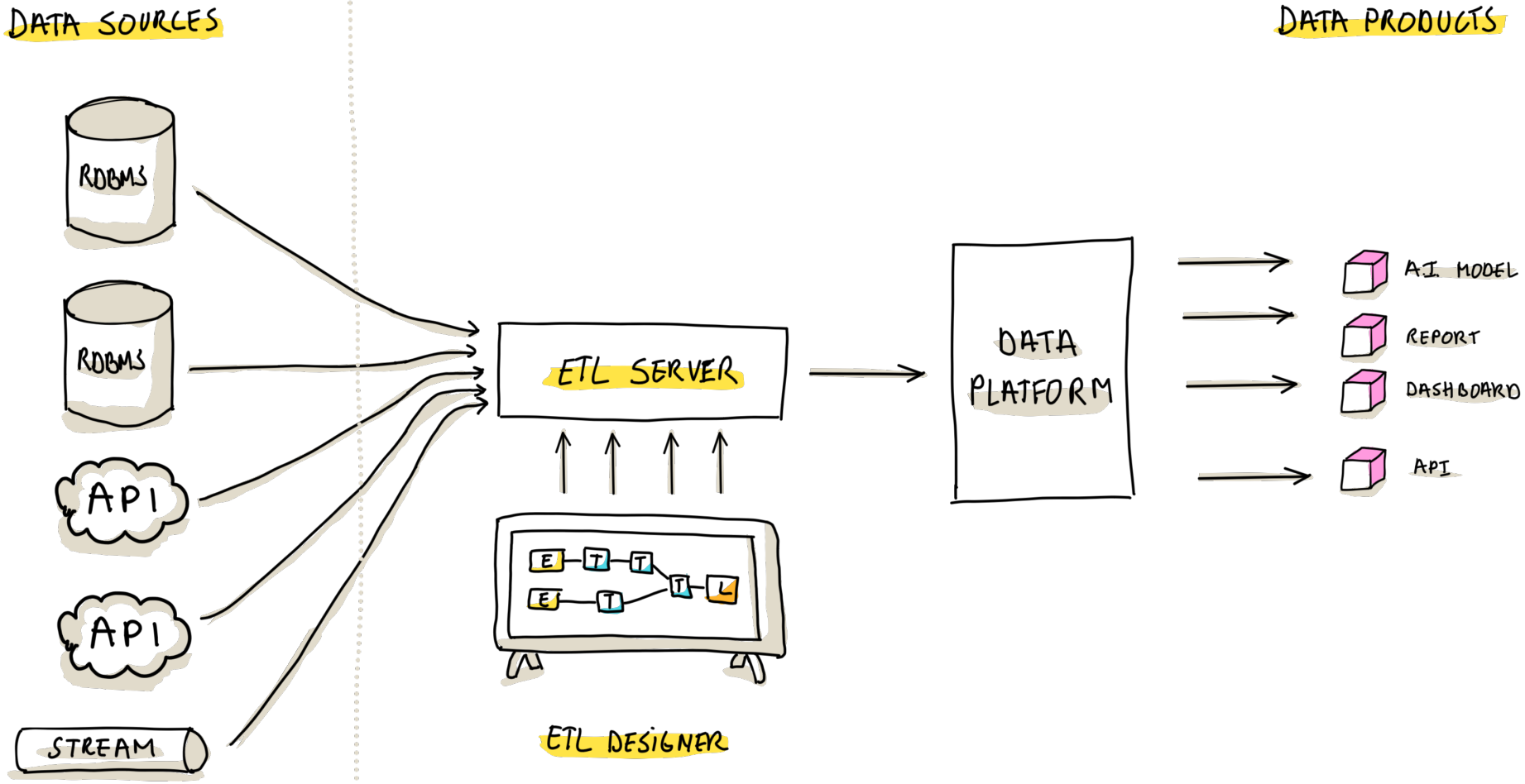
Unified Data Repository = A single storage system caters to both the operational application needs and analytical processing

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*

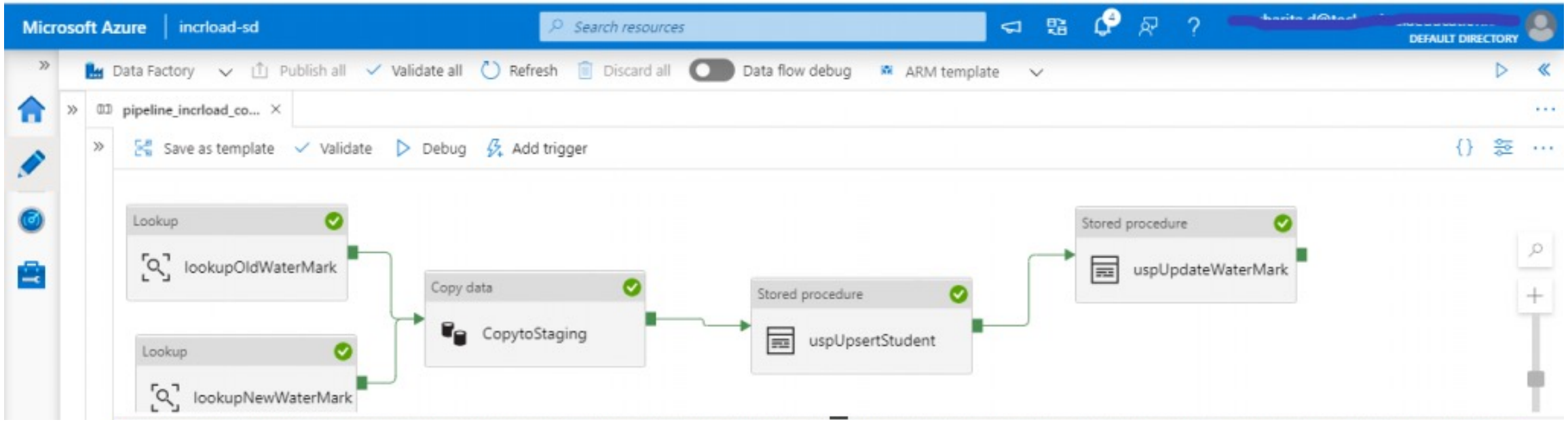
3 steps:

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 **T**ransform **L**oad

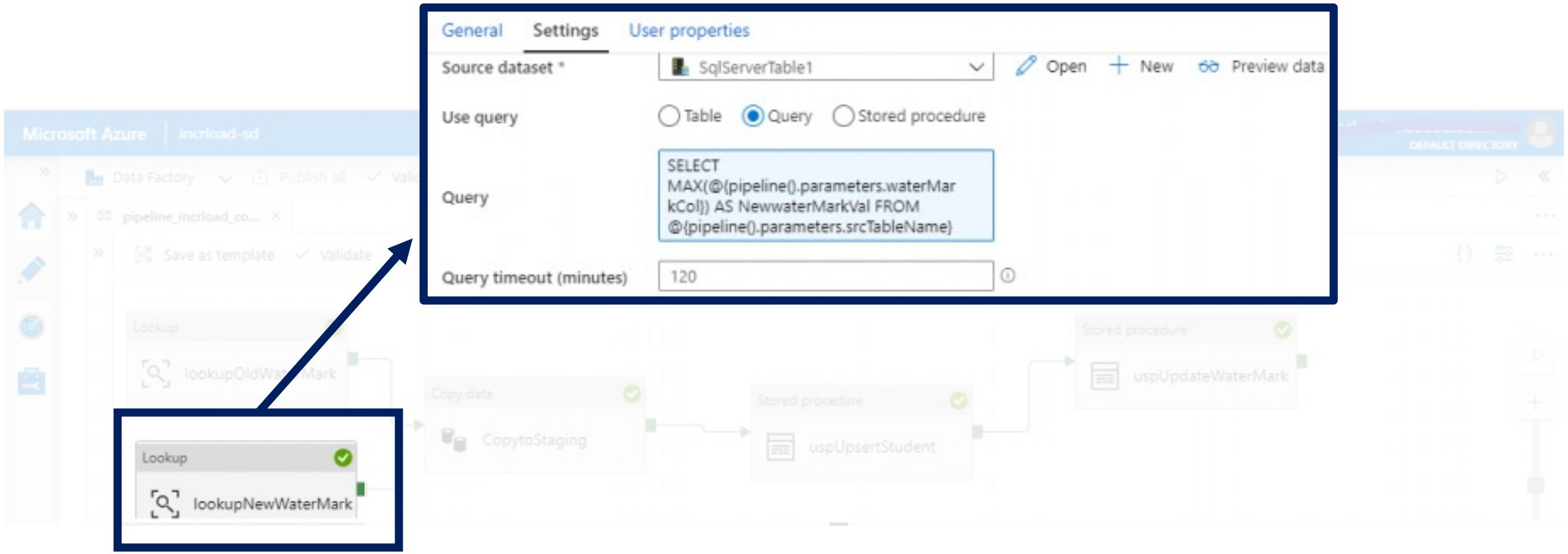


Extract Transform **Load**

[Data Loading using Azure Data Factory]

Pattern #2

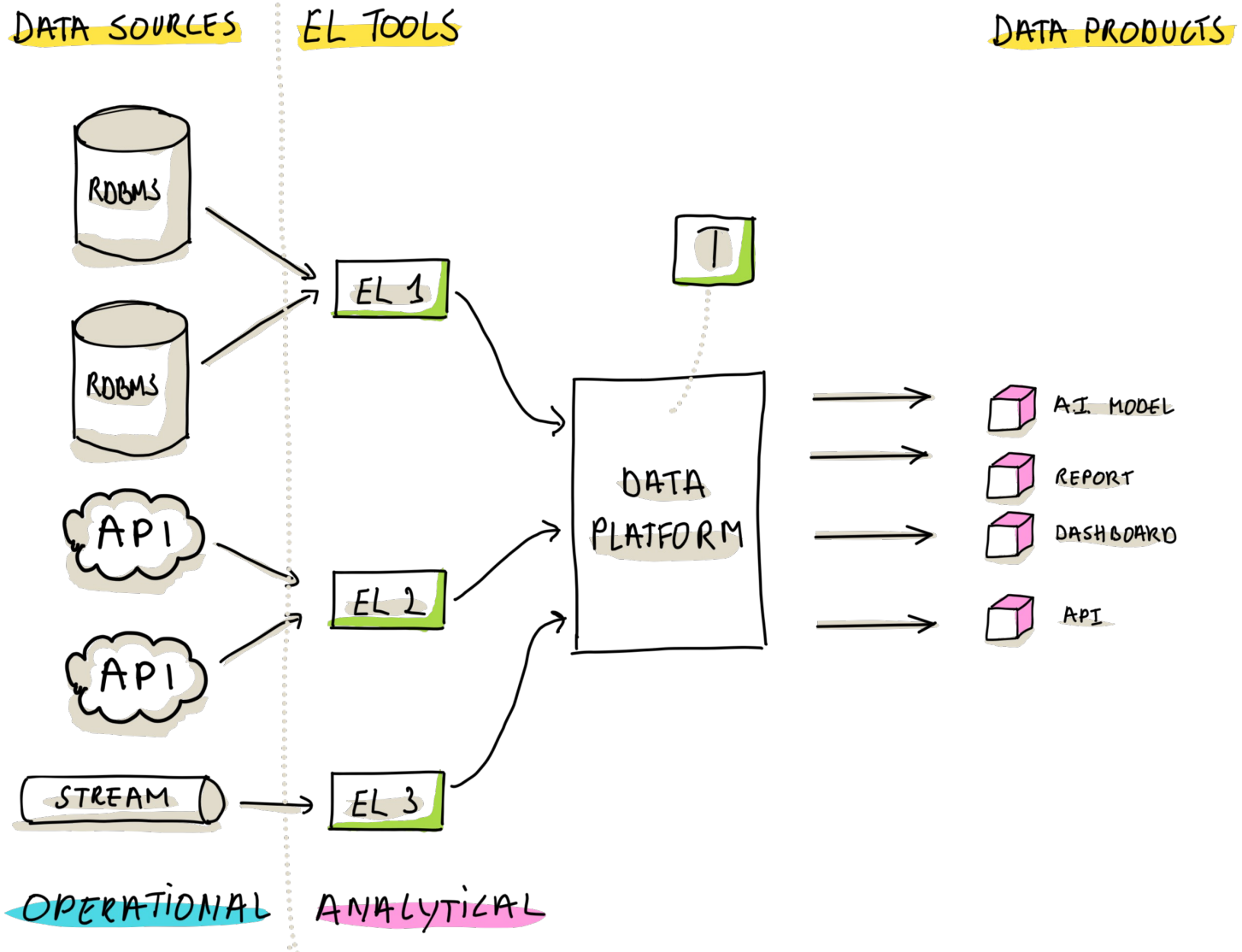
ETL – Extract Transform Load



[Data Loading using Azure Data Factory]

Pattern #3

ELT – Extract **L**oad **T**ransform



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

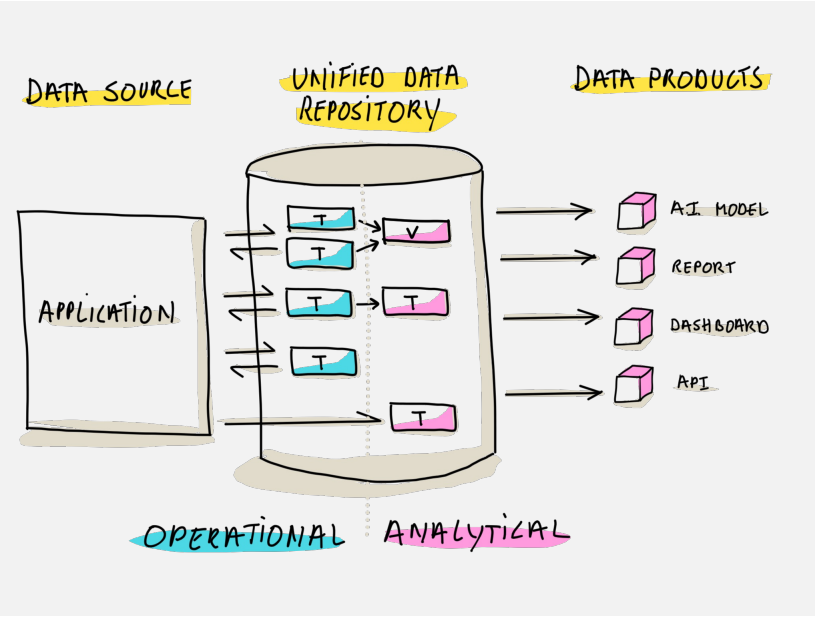
In ELT:

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 : STRENGTS & WEAKNESSES

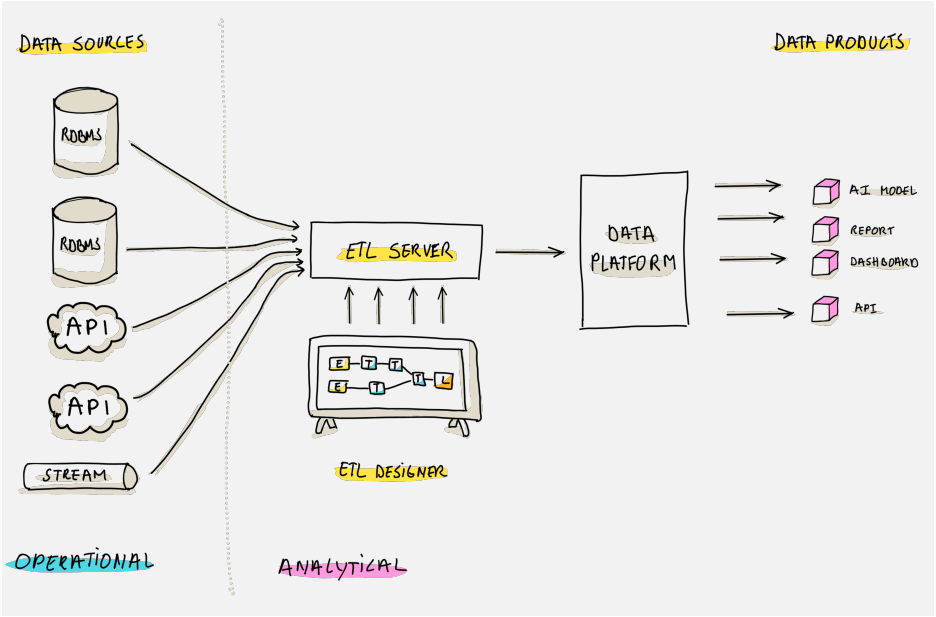
#1

Unified Data Repository



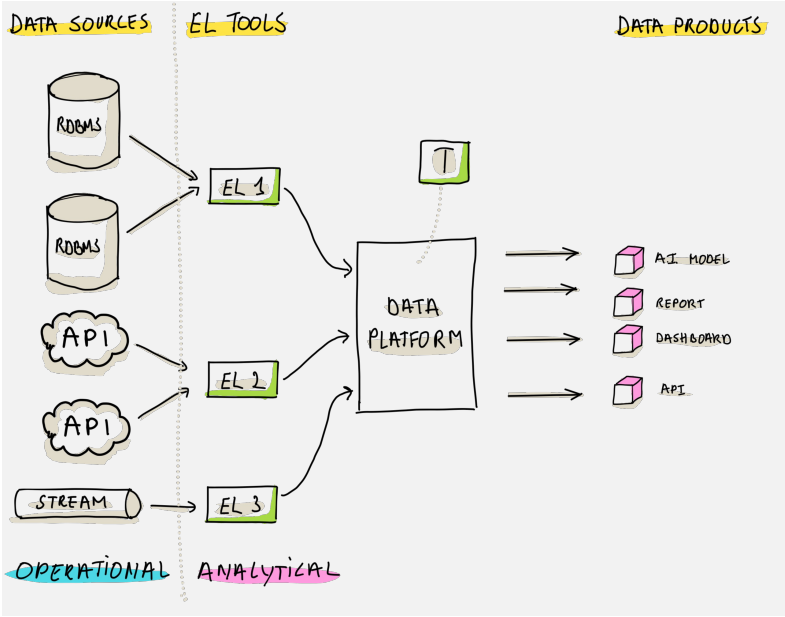
#2

ETL



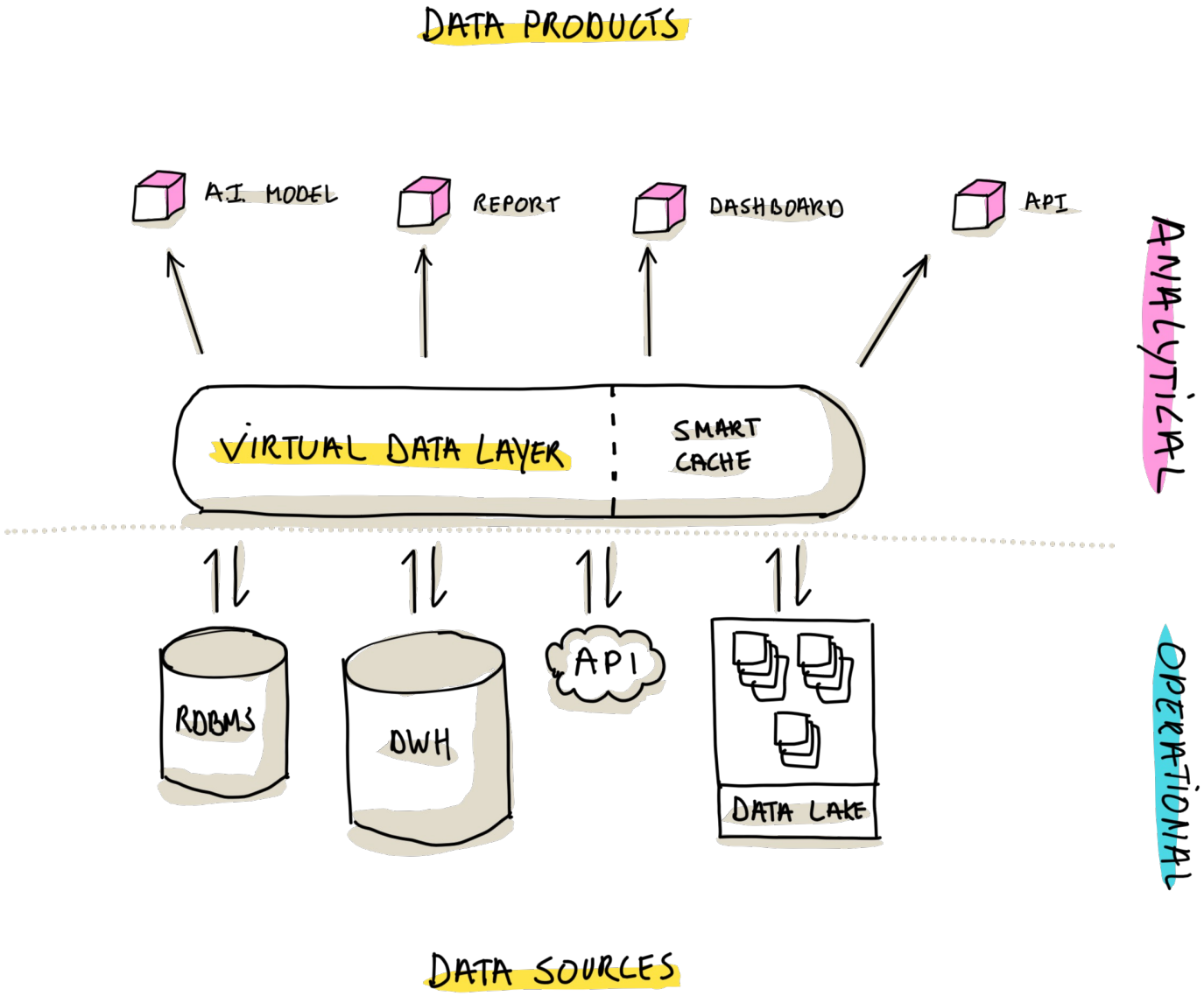
#3

ELT



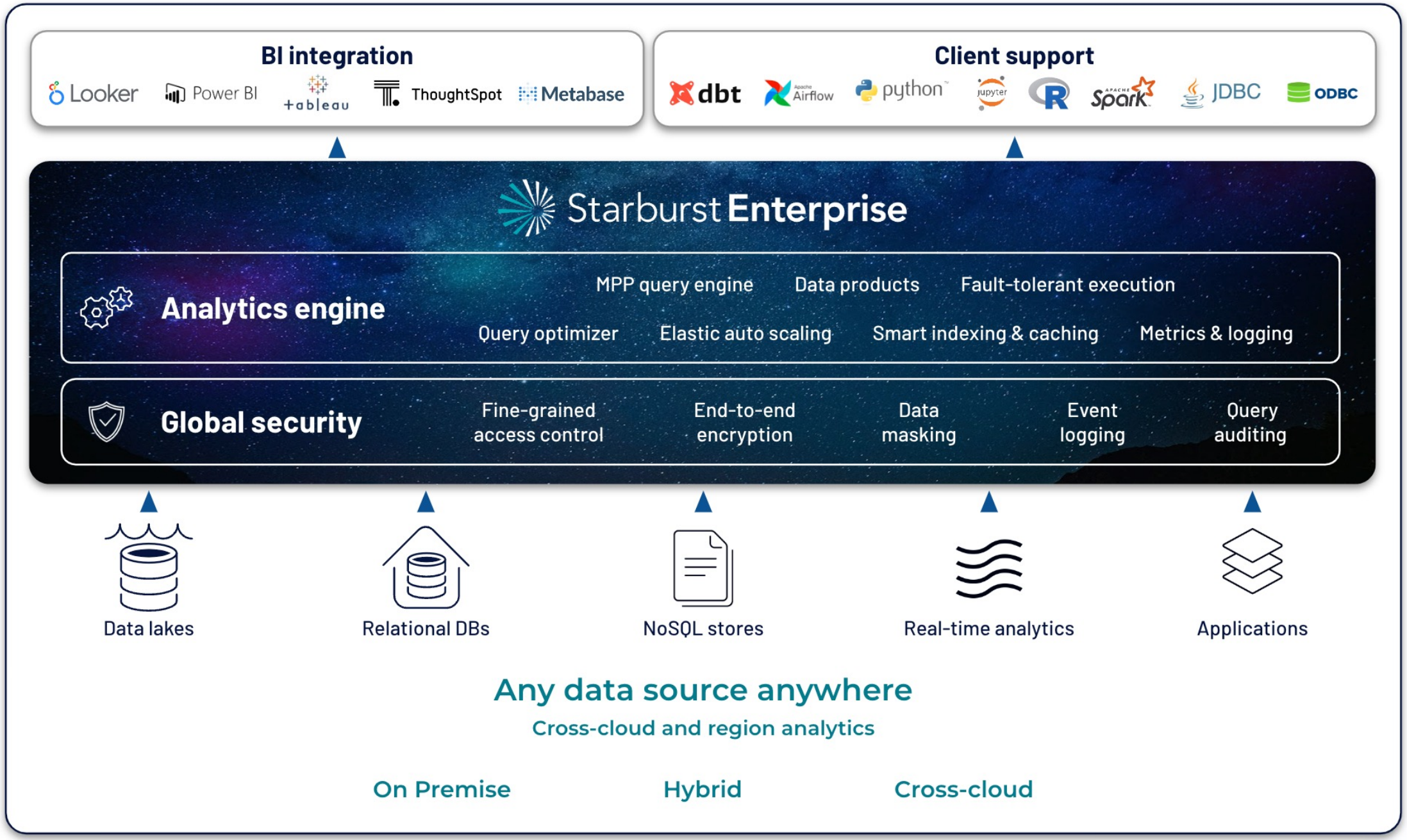
Pattern #4

Data Virtualization



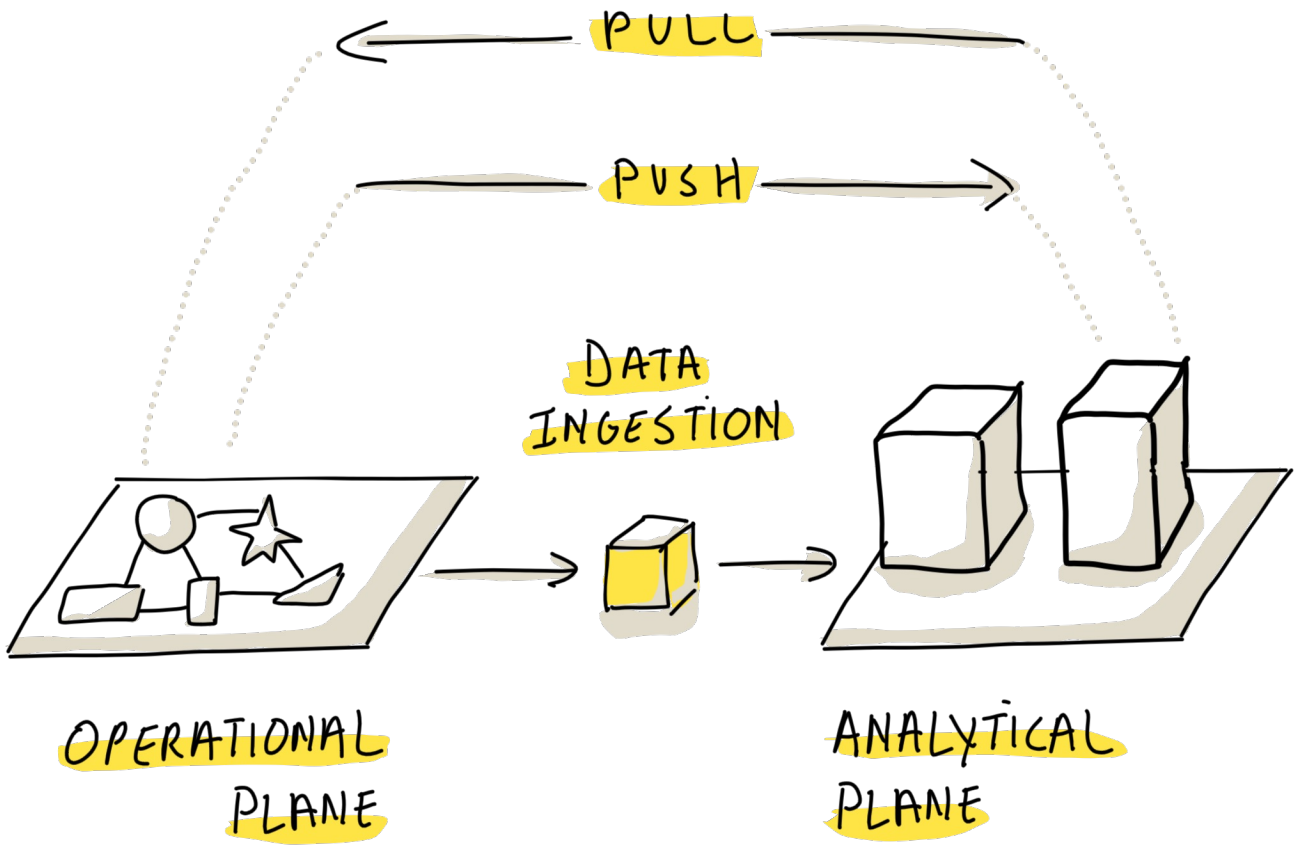
Data Virtualization = specialized software to establish a virtualized data layer over multiple underlying data sources. This intermediary layer allows for the execution of queries that are partially processed by the original data sources, integrating the results into a cohesive dataset for analysis.

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


BI integration
Looker
Power BI
tableau
ThoughtSpot
Metabase
Client support
dbt
Airflow
python
jupyter
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"

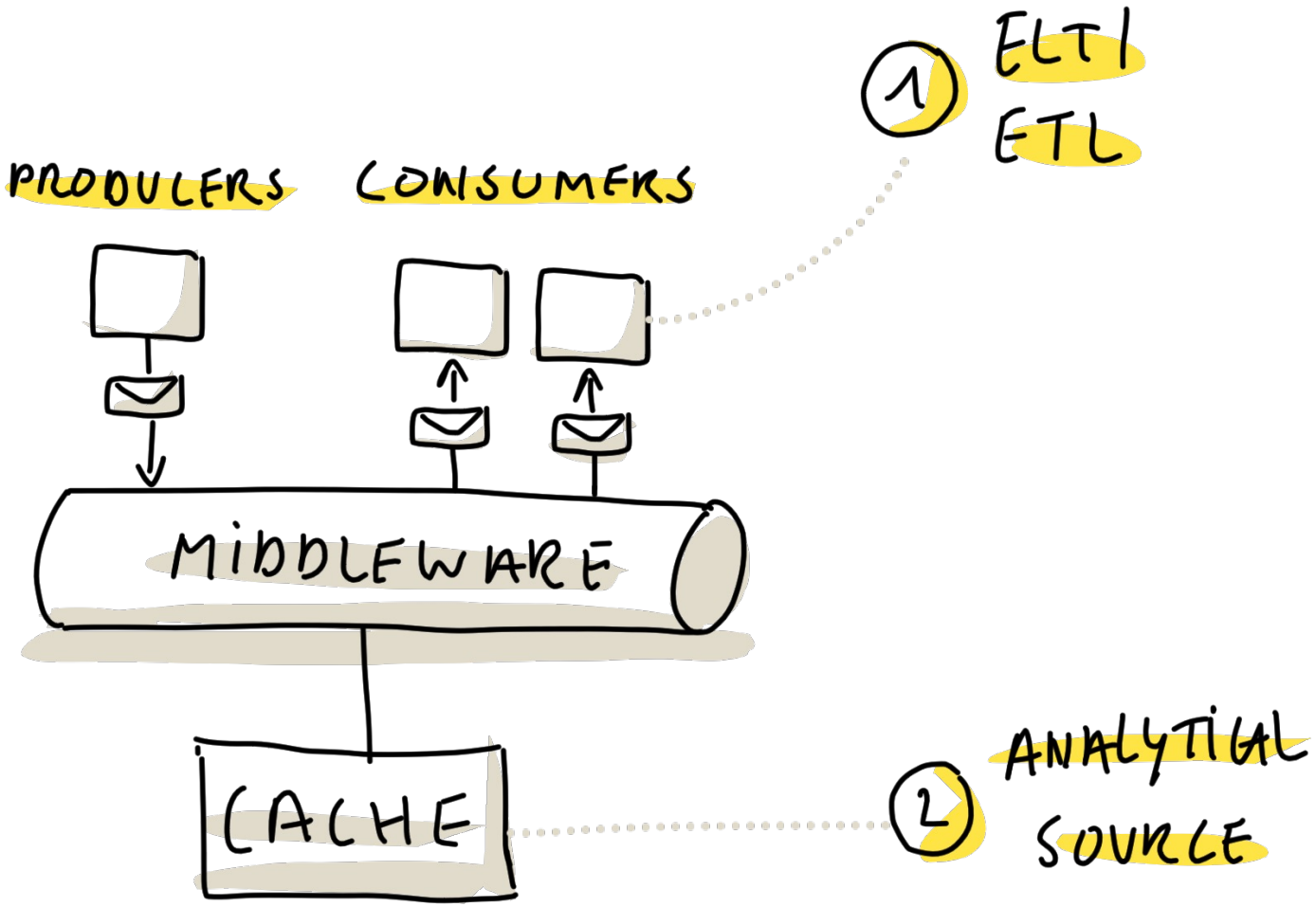
Tables
represent state

"The state of the board"

[Streams and Tables in Apache Kafka – A Primer]

Pattern #6

Streaming Data



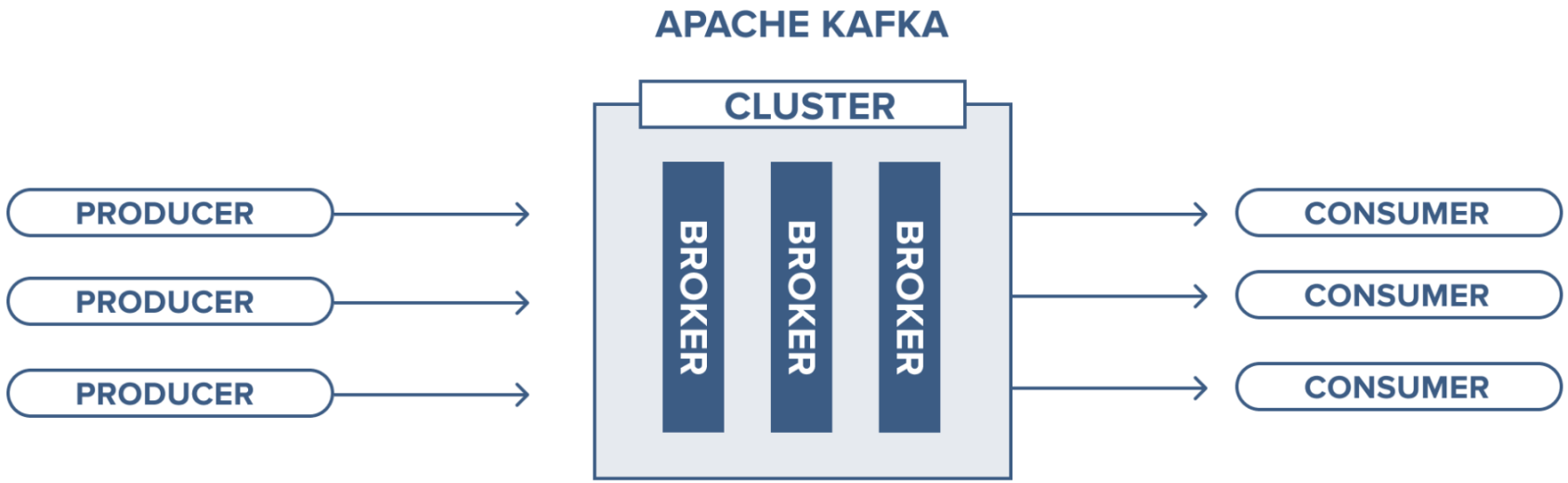
Stream processing = the continuous flow of data as it's generated, enabling real-time processing and analysis for immediate insights.

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

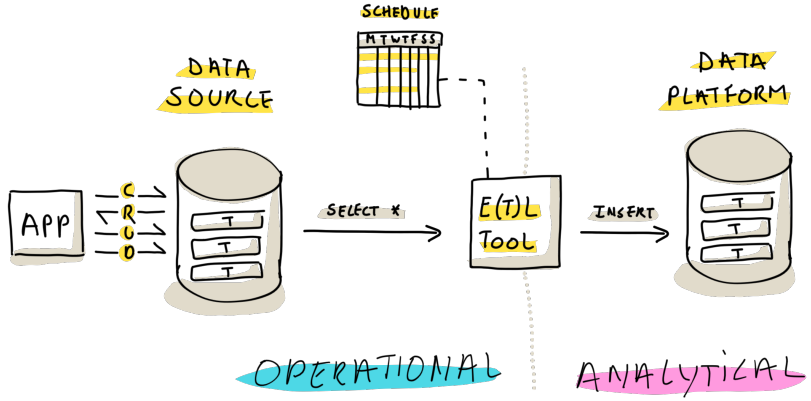
Example: Kafka



Data Ingestion Tool Flavors

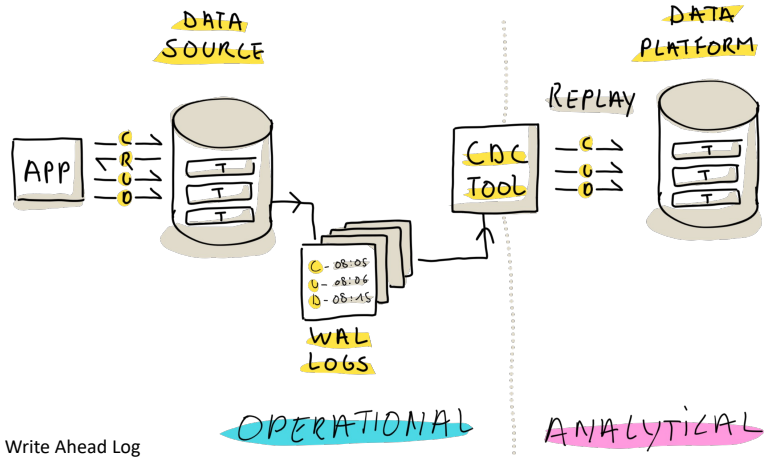
Flavor #1

Batch Loading



Flavor #2

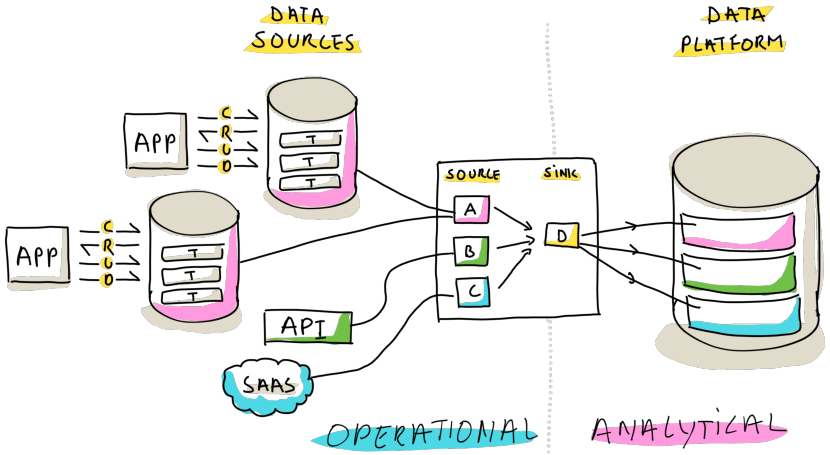
CDC – Change Data Capture



WAL = Write Ahead Log

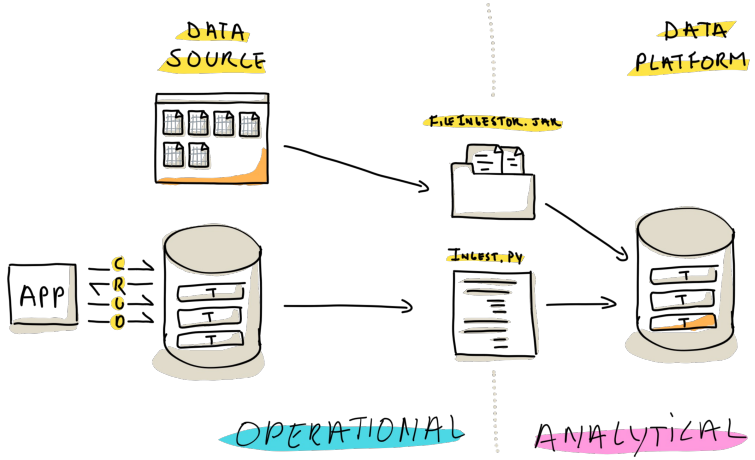
Flavor #3

Connector Based



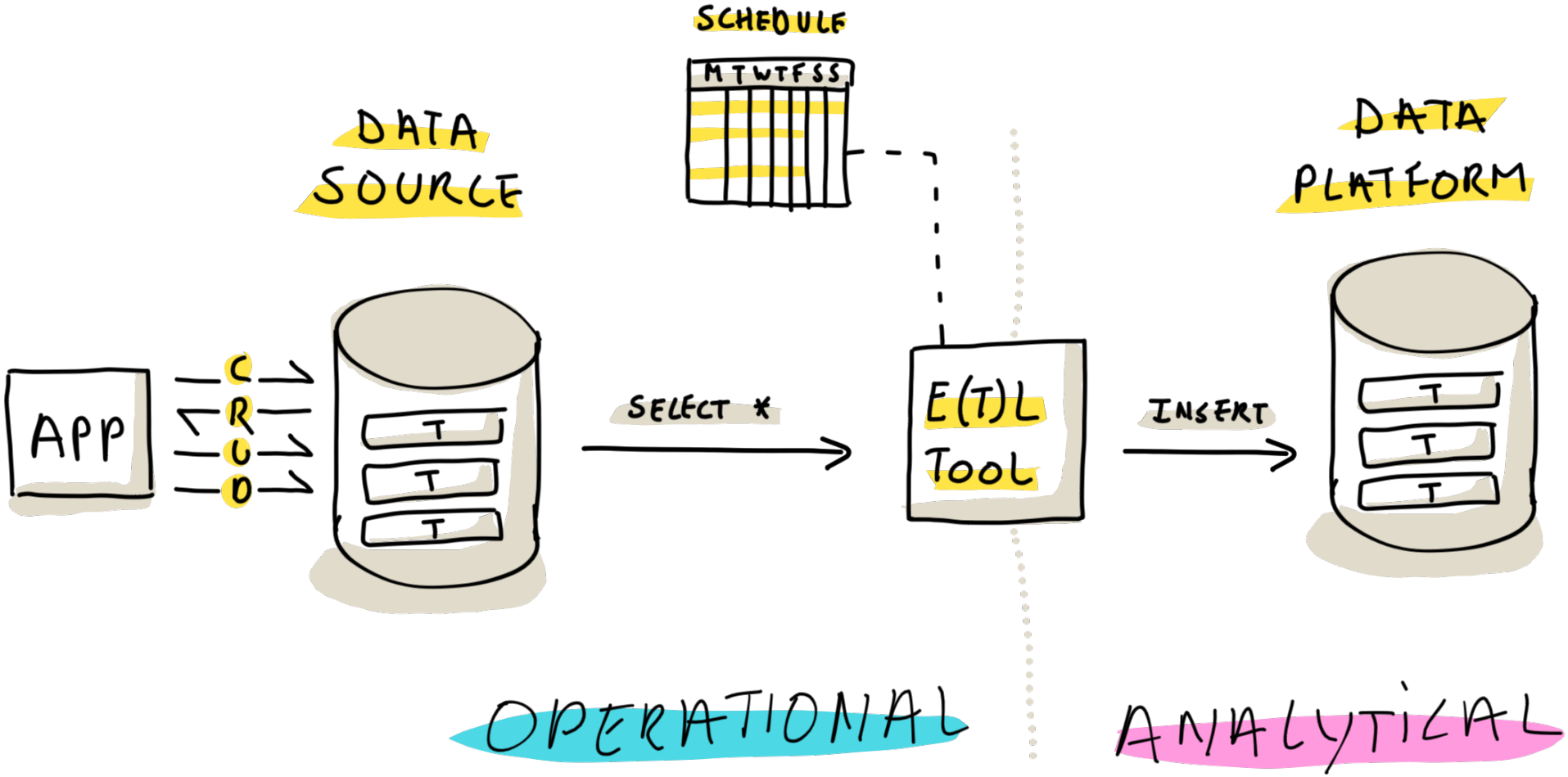
Flavor #4

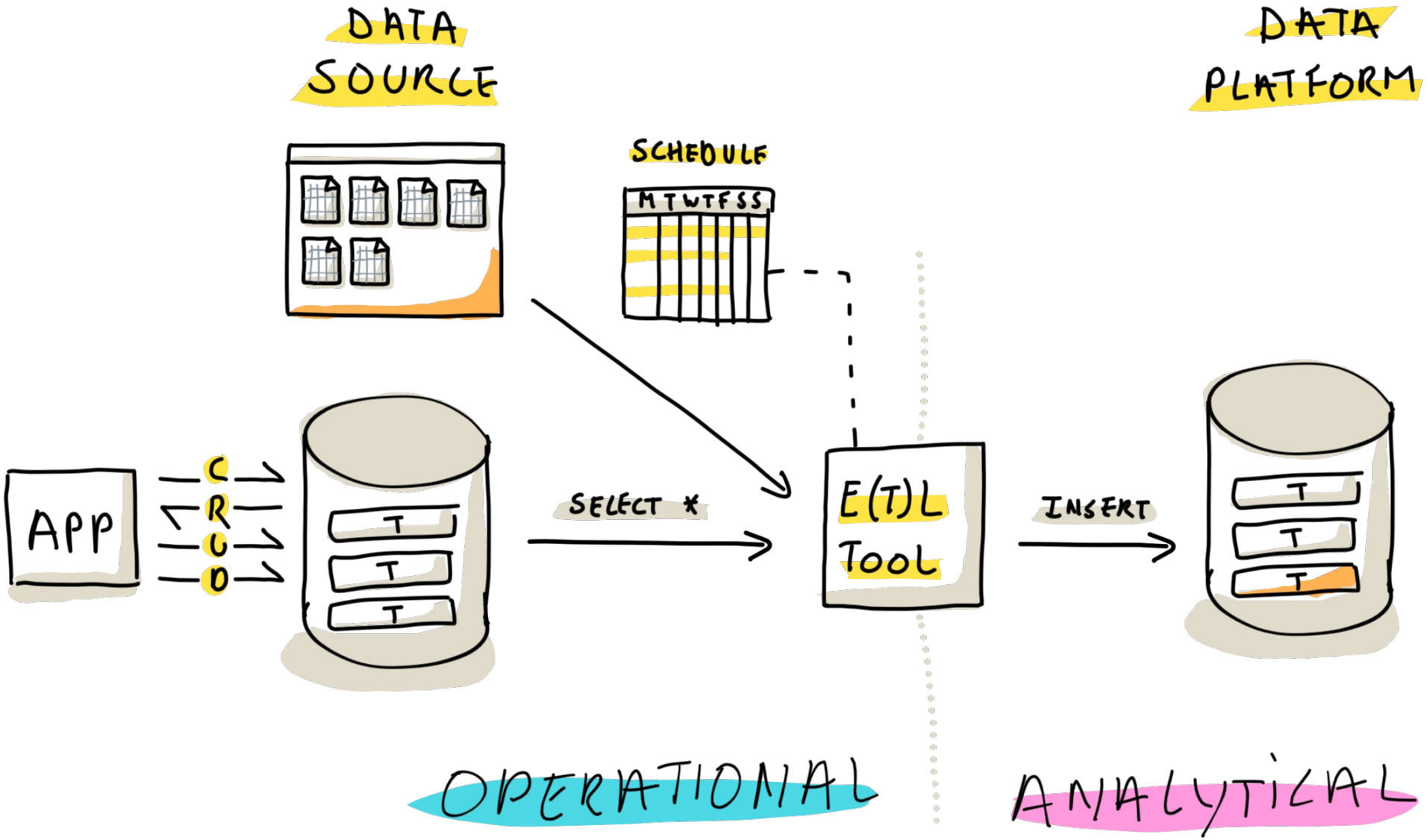
Custom Builds



Flavor #1

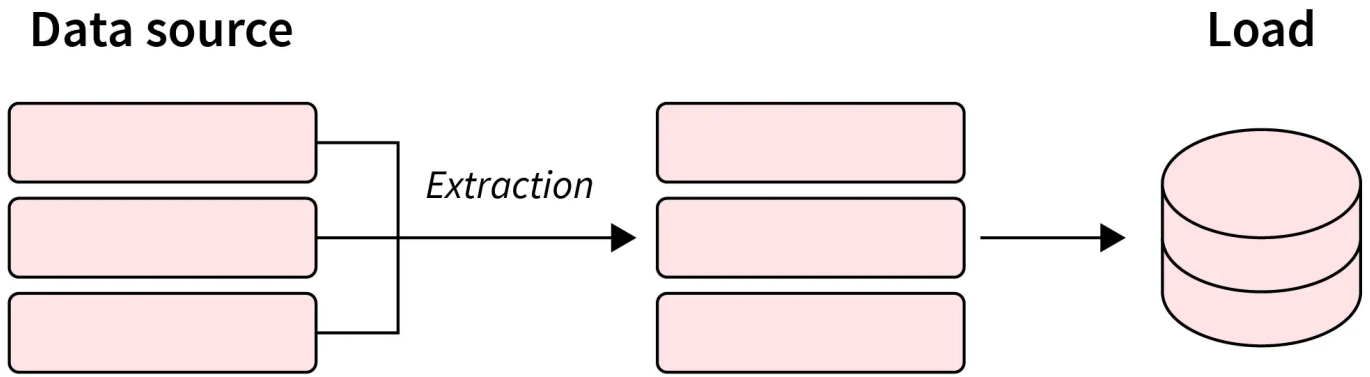
Batch Loading




DATA SOURCE
DATA PLATFORM
SCHEDULE
M T W T F S S
App
C
R
U
D
T
T
T
SELECT *
E(T)L Tool
INSERT
T
T
T
OPERATIONAL
ANALYTICAL

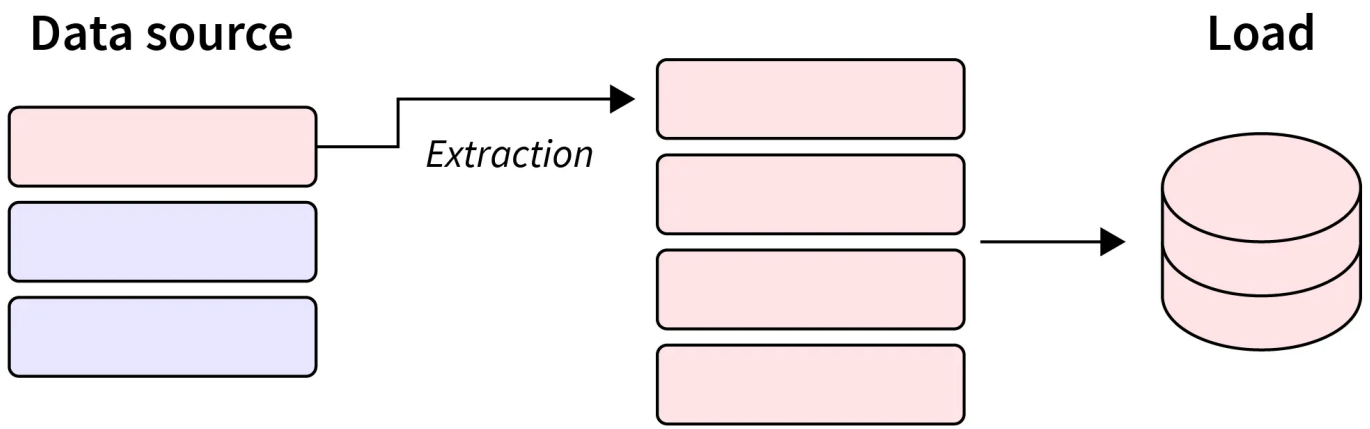
Full load

All available data is extracted at the same time



Incremental load

The occuring changes in the source data is incrementally extracted and loaded



[What is Data Extraction?]

EXERCISE: DESCRIBE INCREMENTAL LOADING

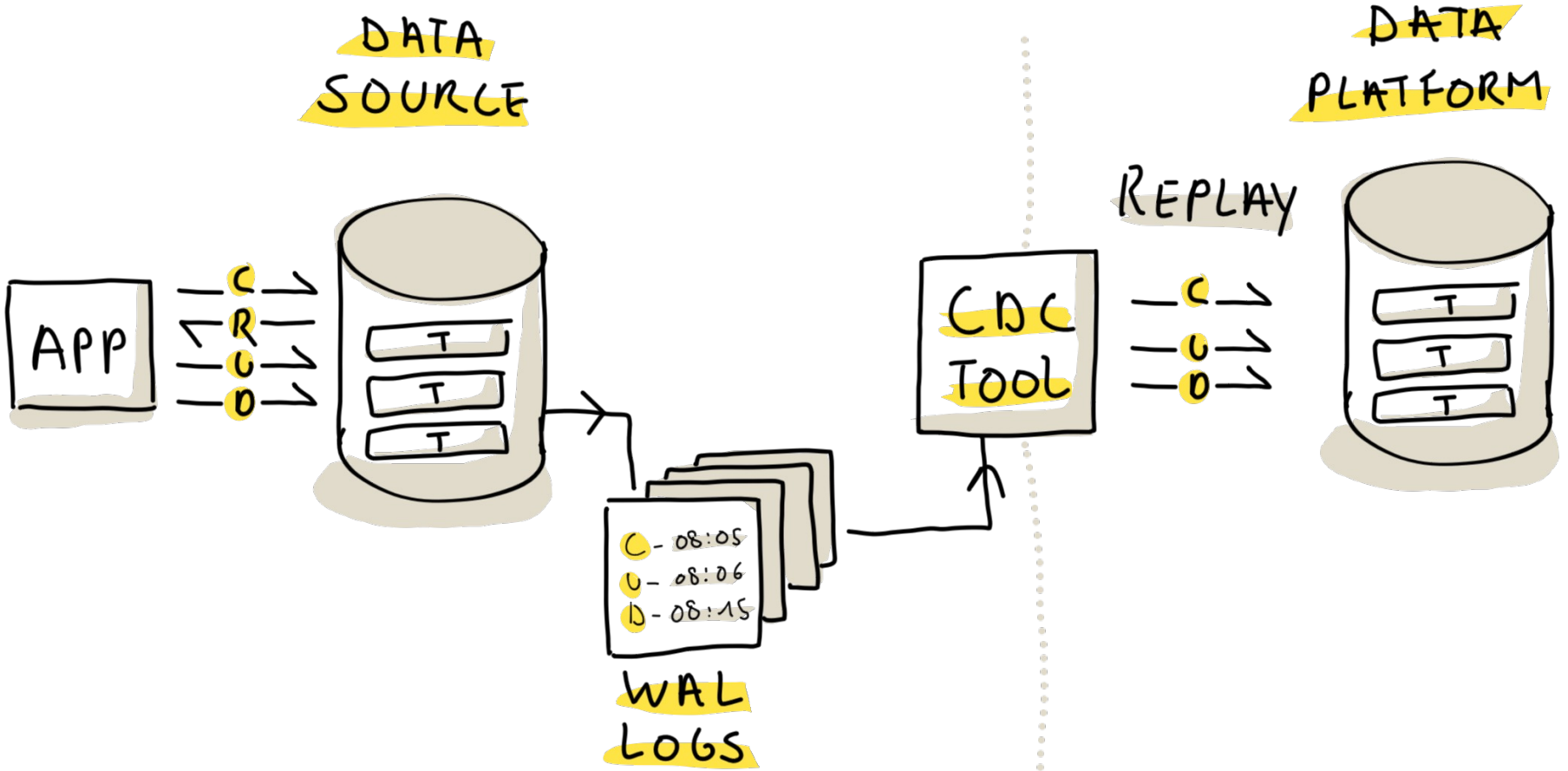
m_transaction

OFF

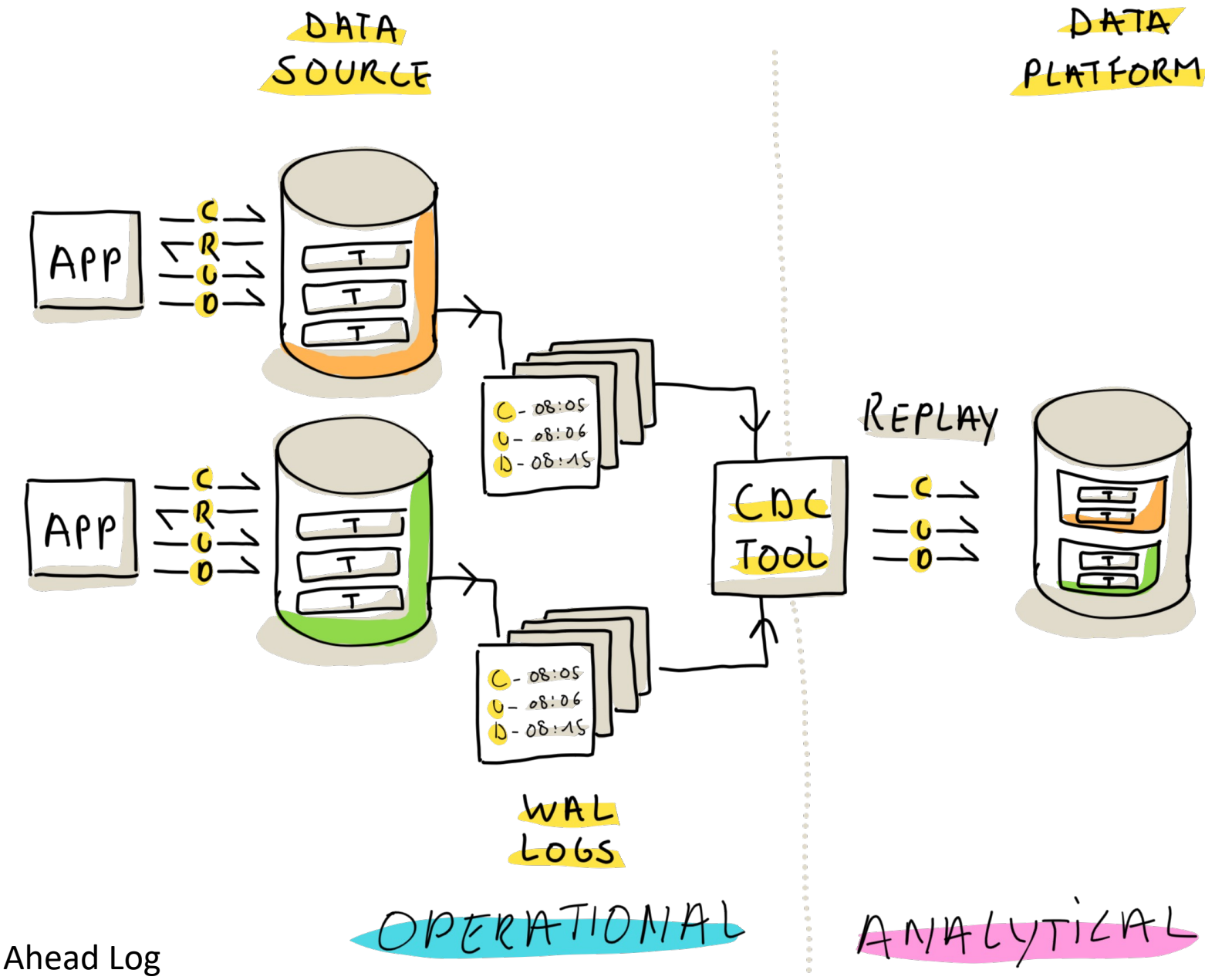
	m_transaction_id	ad_client_id	ad_org_id	isactive	created	createdby	updated
1	1002856	1000010	1000061	Y	2020-03-18 14:40:01.253	100	2020-03-18 14:40:01.253
2	1002857	1000010	1000061	Y	2020-03-18 14:53:40.809	100	2020-03-18 14:53:40.809
3	1002858	1000010	1000061	Y	2020-03-18 15:20:00.275	100	2020-03-18 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 **D**ata Capture

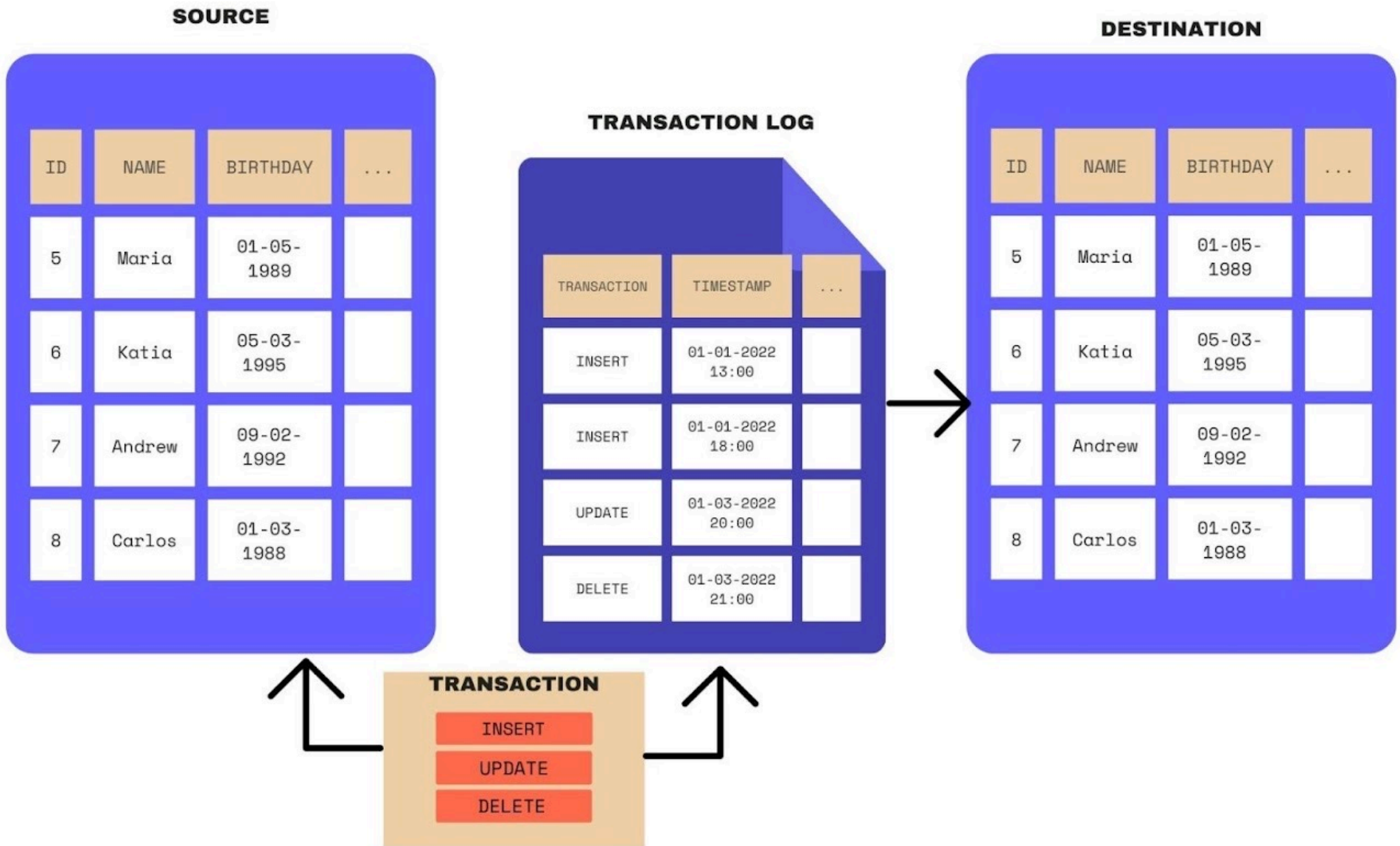


WAL = Write Ahead Log

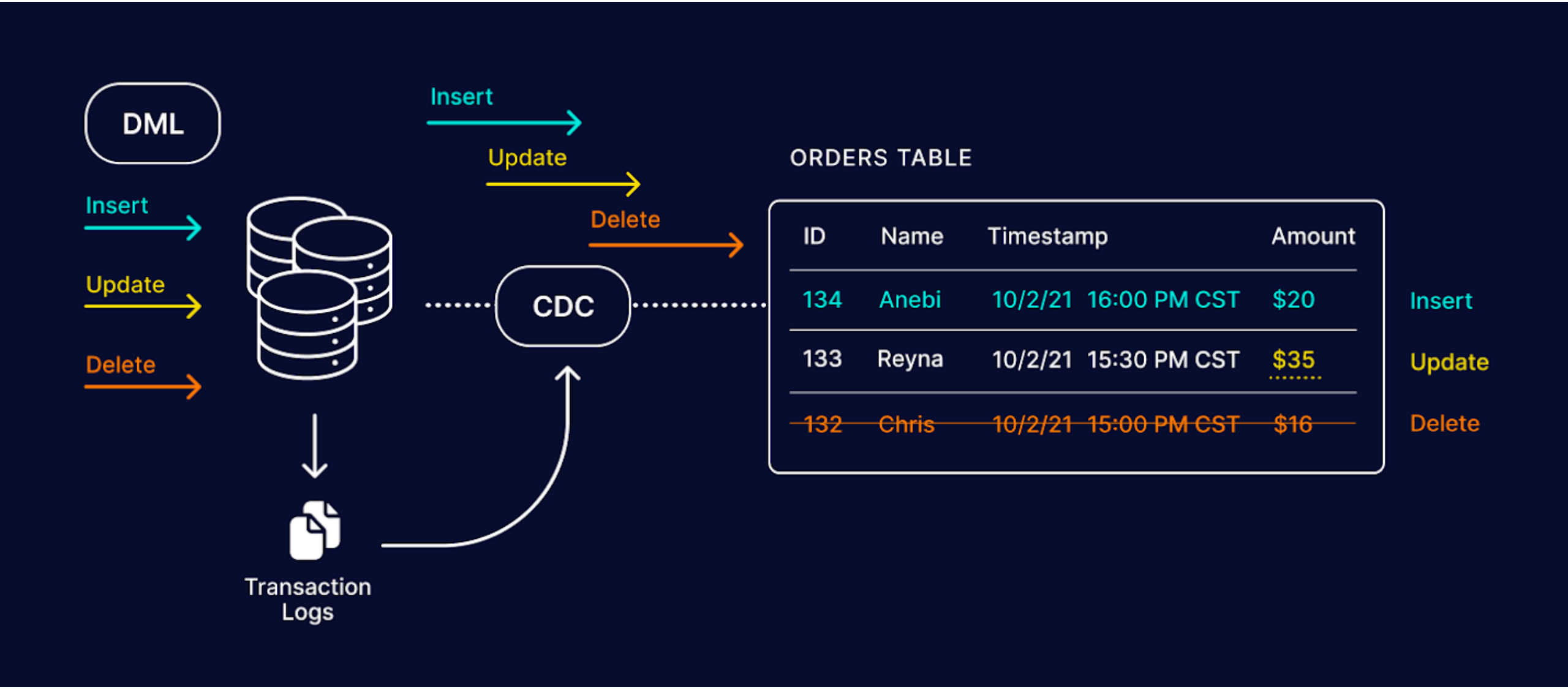


WAL = Write Ahead Log

Log-based CDC technique



[Understanding Change Data Capturing]



[How to Implement CDC with Apache Kafka]

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.

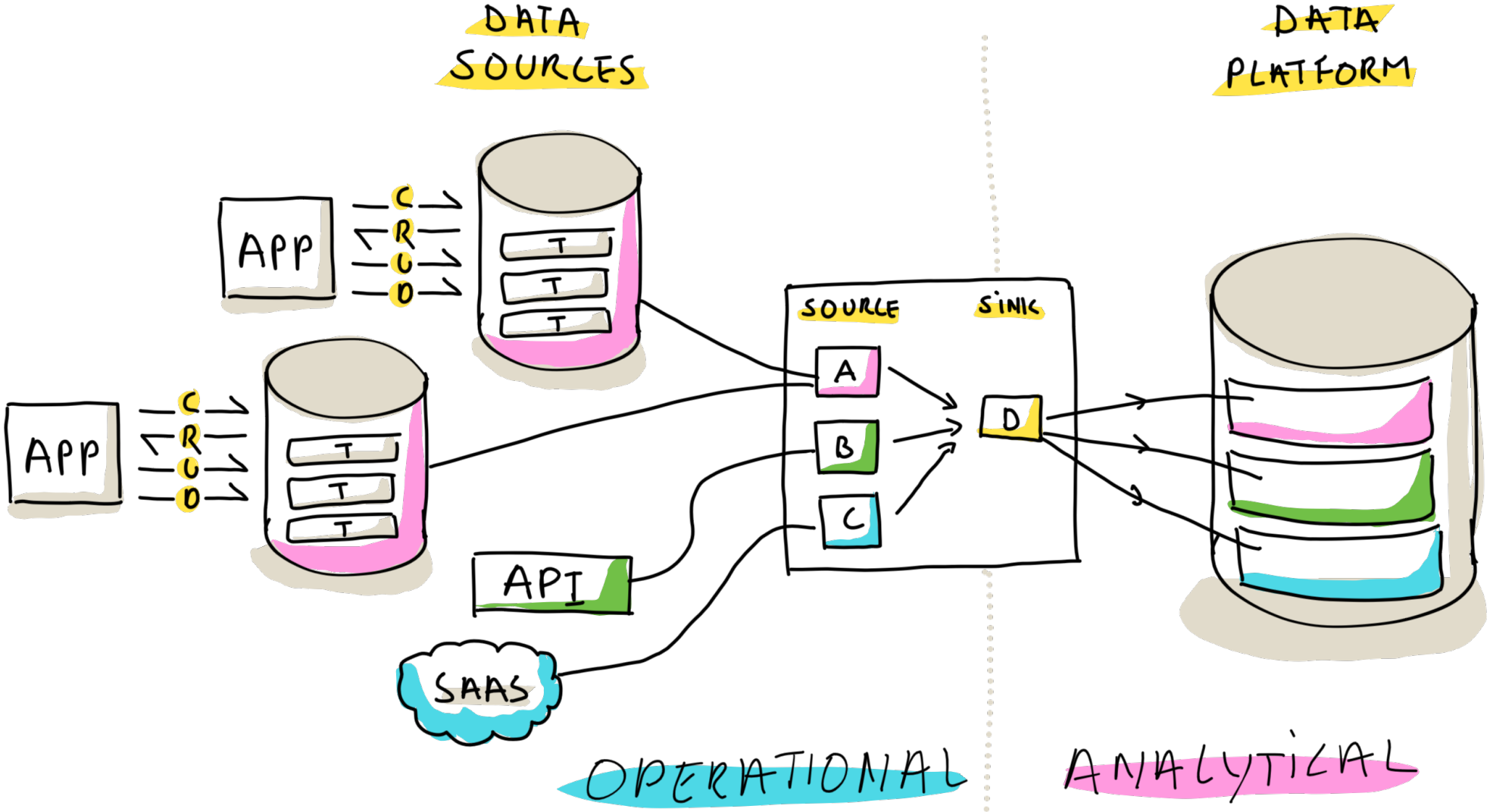
Various techniques exist: meta-data based, trigger based, **log based**

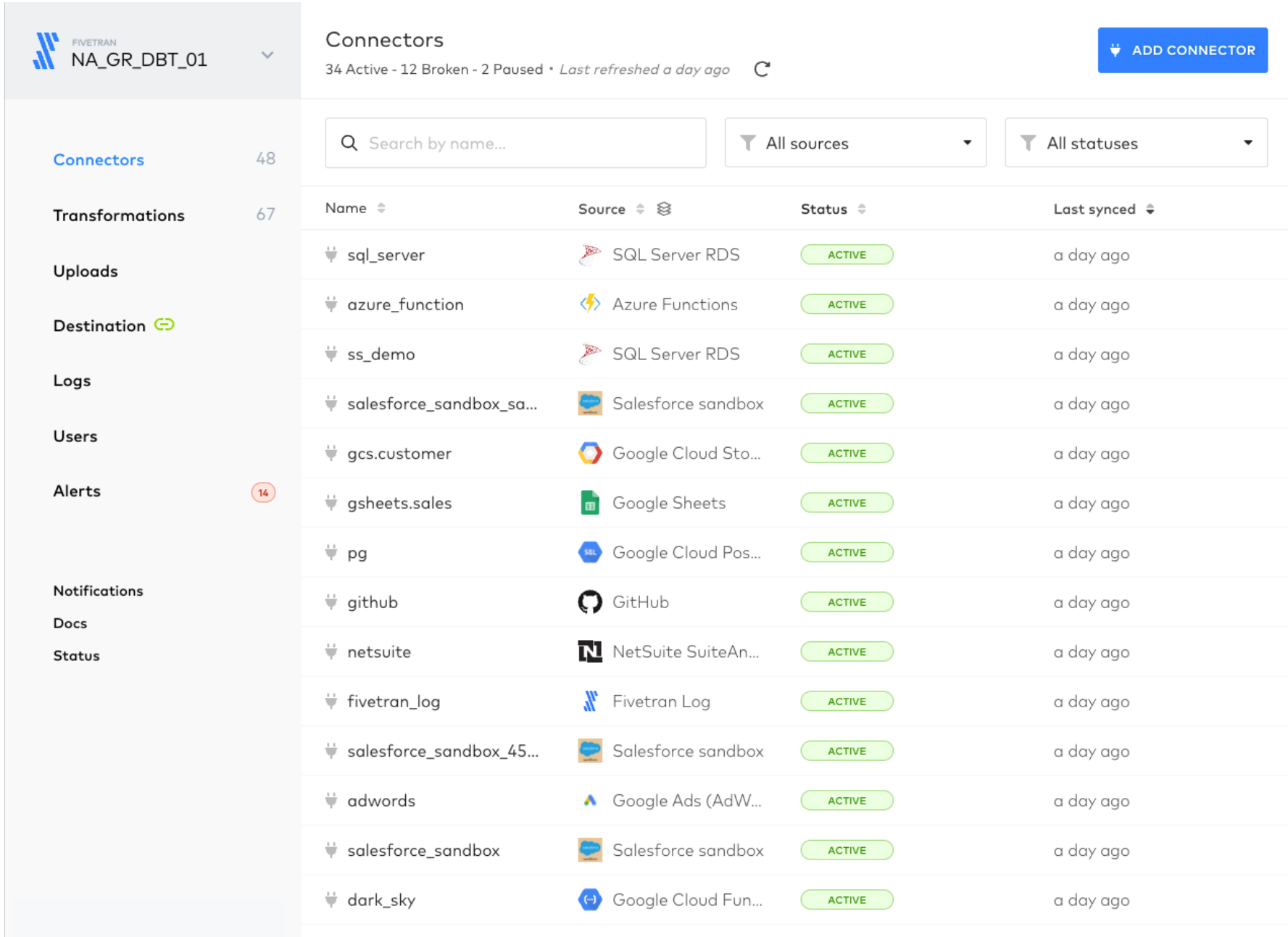
Log based CDC systems read data directly from the database Change Data Capture logs to identify changes in a database (not from the actual database)

[Understanding Change Data Capturing]

Flavor #3

Connector Based




FIVETRAN
NA_GR_DBT_01
Connectors
48
Transformations
67
Uploads
Destination
Logs
Users
Alerts
14
Notifications
Docs
Status
Connectors
34 Active - 12 Broken - 2 Paused • Last refreshed a day ago
ADD CONNECTOR
Search by name...
All sources
All statuses
Name
Source
Status
Last synced
sql_server
SQL Server RDS
ACTIVE
a day ago
azure_function
Azure Functions
ACTIVE
a day ago
ss_demo
SQL Server RDS
ACTIVE
a day ago
salesforce_sandbox_sa...
Salesforce sandbox
ACTIVE
a day ago
gcs.customer
Google Cloud Sto...
ACTIVE
a day ago
gsheets.sales
Google Sheets
ACTIVE
a day ago
pg
Google Cloud Pos...
ACTIVE
a day ago
github
GitHub
ACTIVE
a day ago
netsuite
NetSuite SuiteAn...
ACTIVE
a day ago
fivetran_log
Fivetran Log
ACTIVE
a day ago
salesforce_sandbox_45...
Salesforce sandbox
ACTIVE
a day ago
adwords
Google Ads (AdW...
ACTIVE
a day ago
salesforce_sandbox
Salesforce sandbox
ACTIVE
a day ago
dark_sky
Google Cloud Fun...
ACTIVE
a day ago

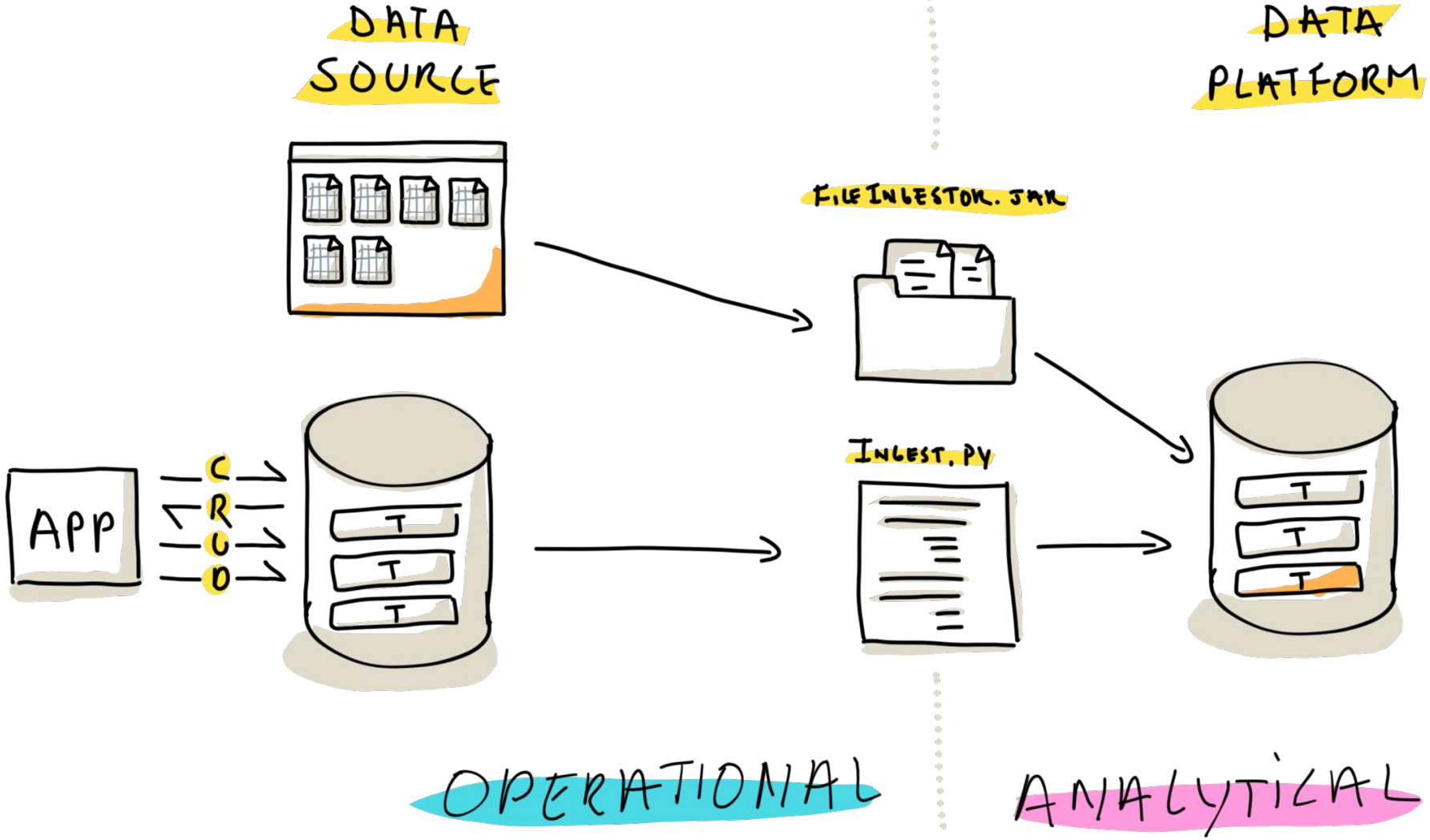
[Fivetran]

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.
- Higly **flexible** but **No control** over the individual connectors

Flavor #4

Custom Builds



Disadvantages:

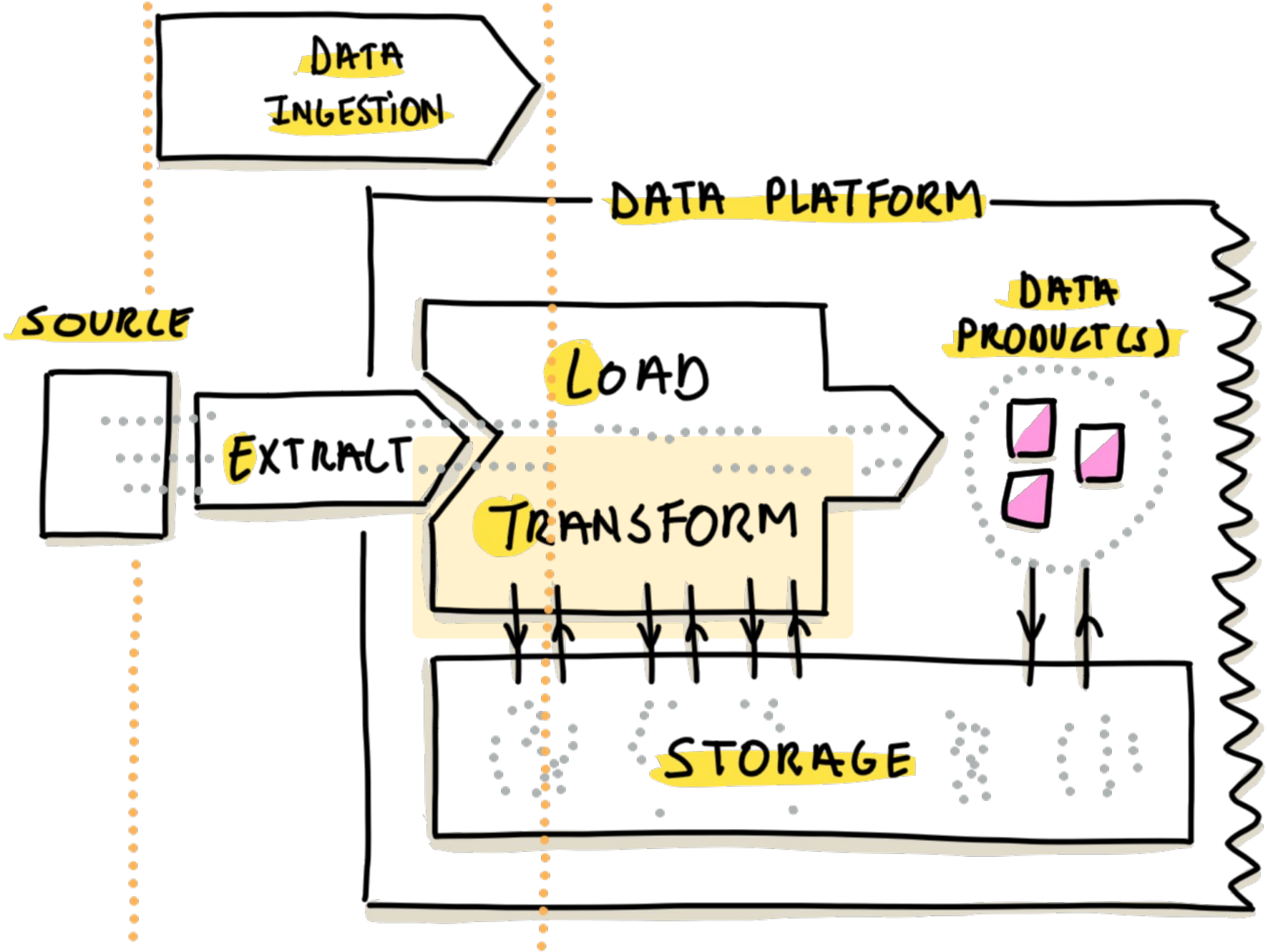
- Building ingestion pipelines is usually more expensive than expected
- Specific programming knowledge & team needed
- High maintenance cost

Advantages:

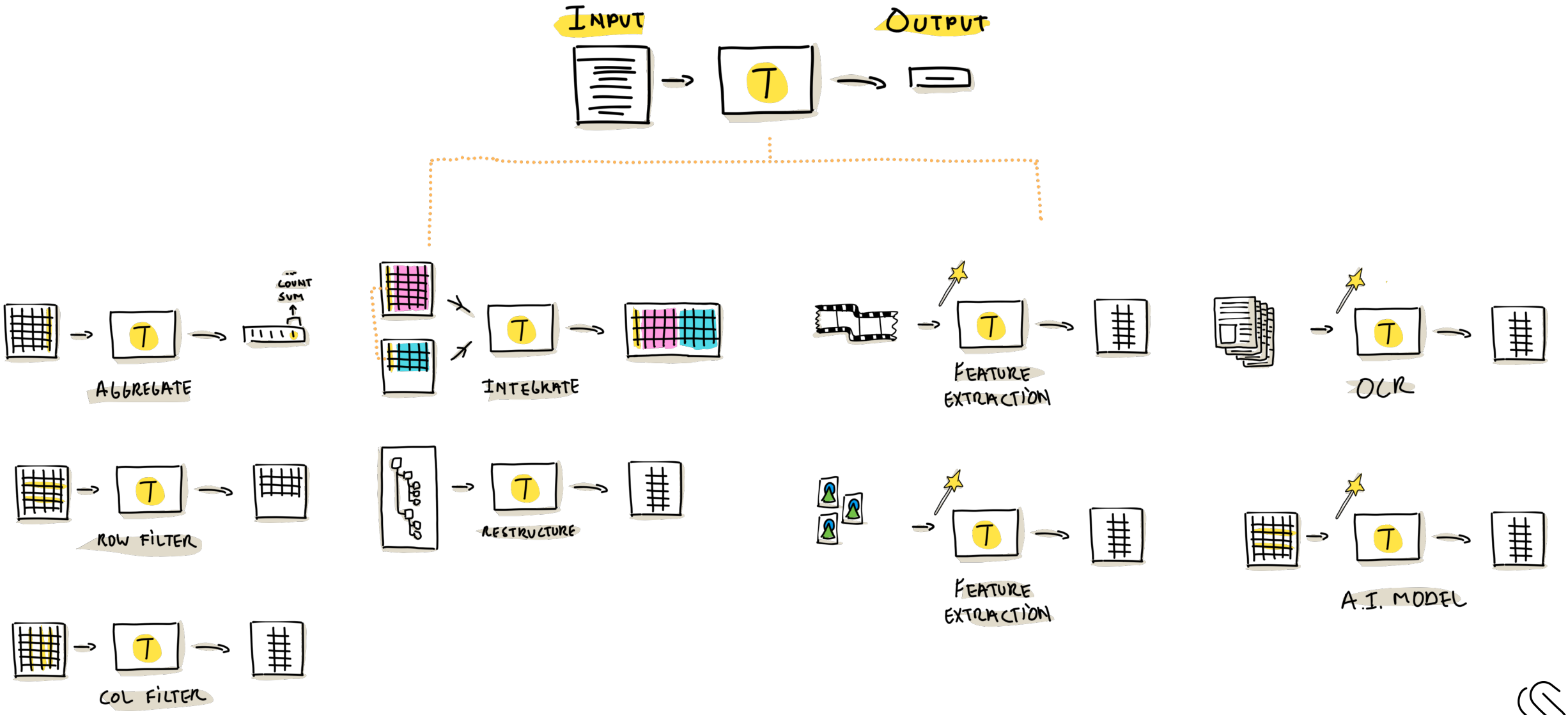
- Full control
- Allows to ingest very specific / unique / exotic data sources

Building	VS	Buying
Customization and scale		Lower upfront cost
Greater control No license fees		Rapid deployment
Competitive edge		Updates and maintenance
Easy to modify		Has active userbase

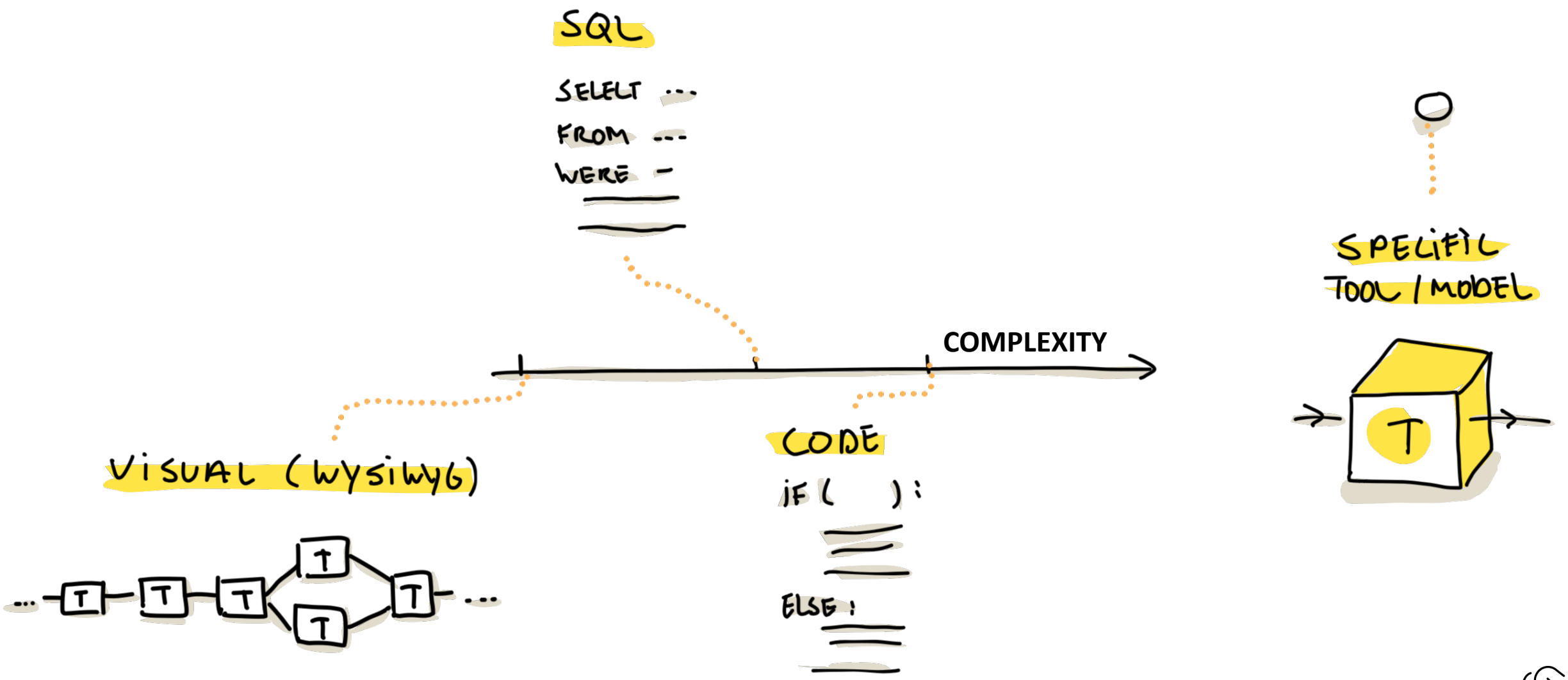
[Build vs Buy: The Pros and Cons]


DATA
INGESTION
DATA PLATFORM
SOURCE
EXTRACT
LOAD
TRANSFORM
DATA
PRODUCT(S)
STORAGE

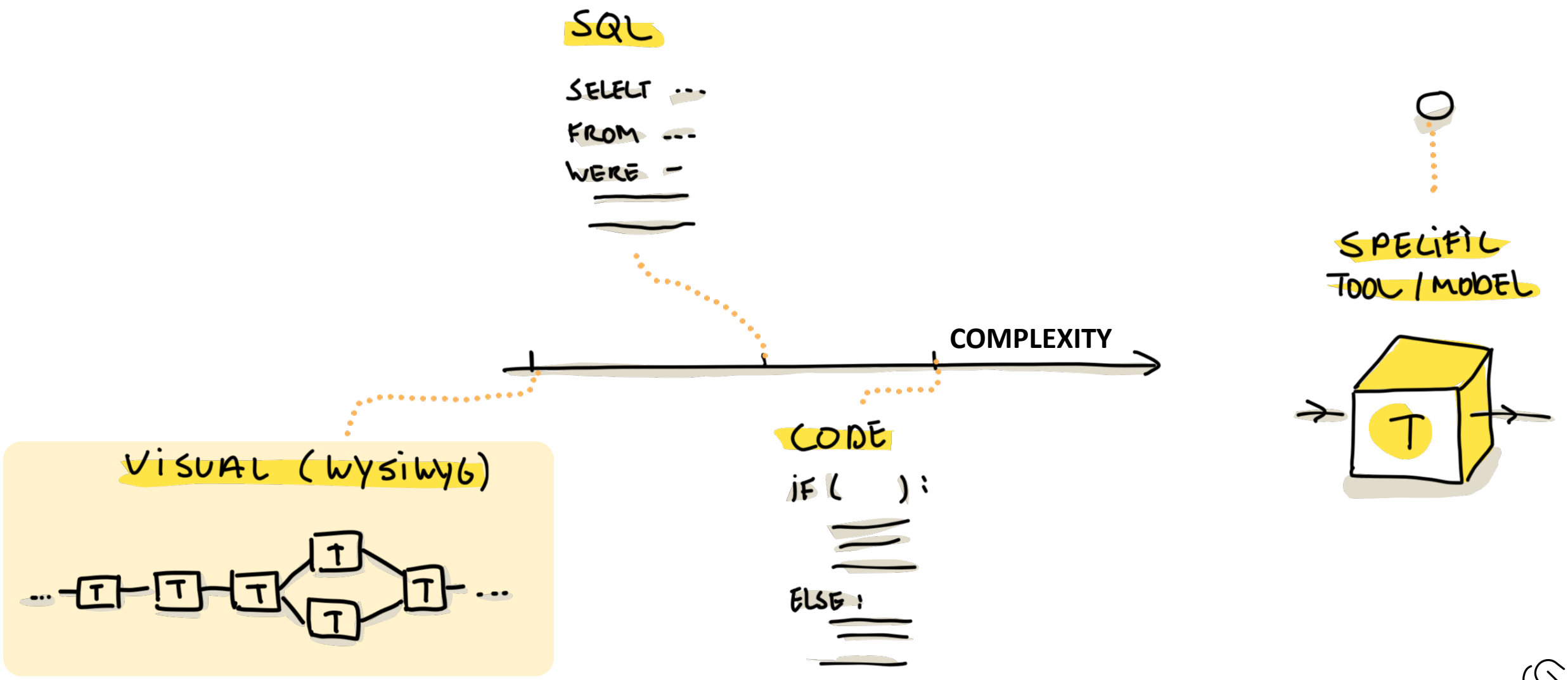
Data Transformations



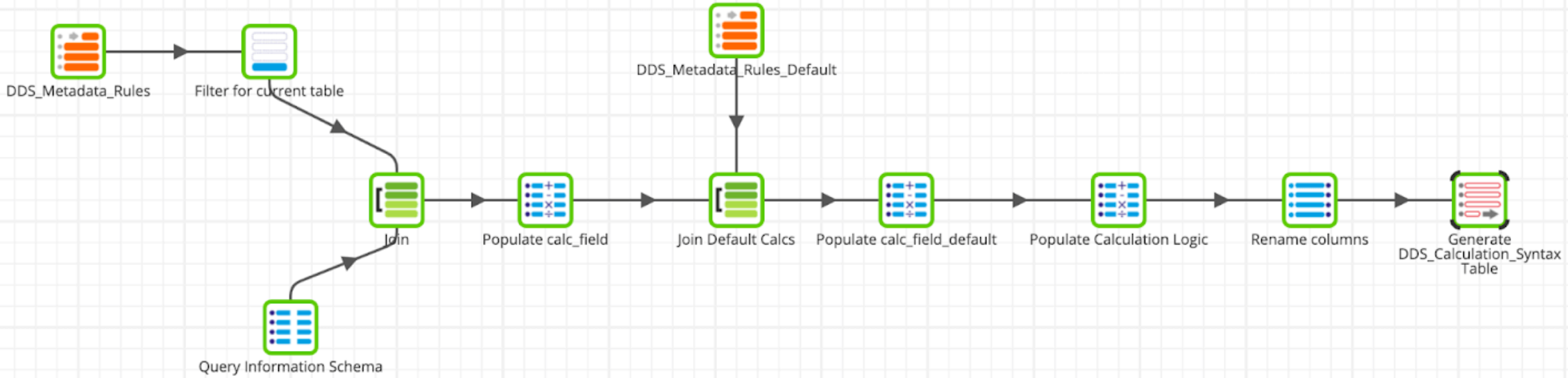
Data Transformation Tool Flavors



Data Transformation Tool Flavors



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'.



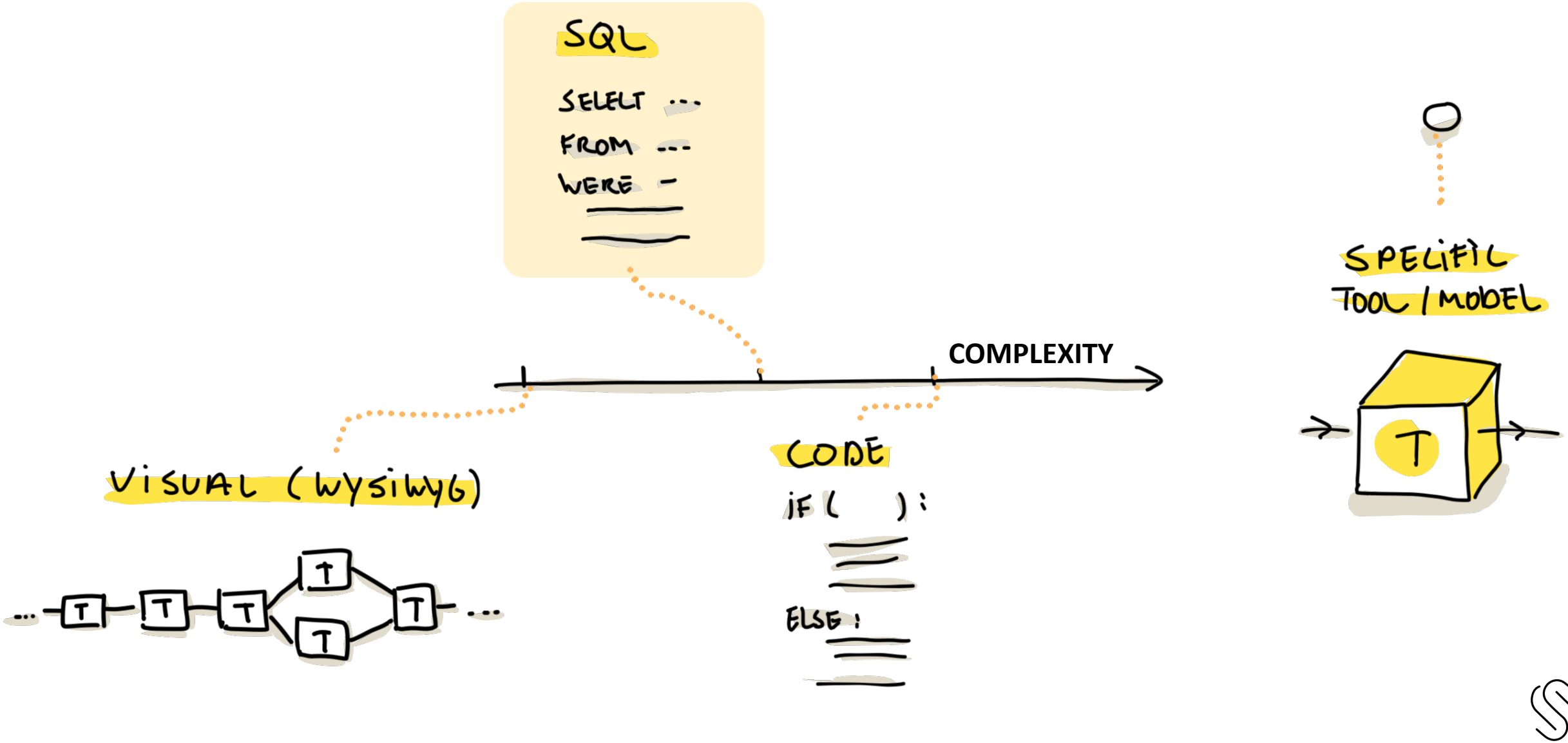
Properties | Export | Sample | Metadata | SQL | Plan | Help

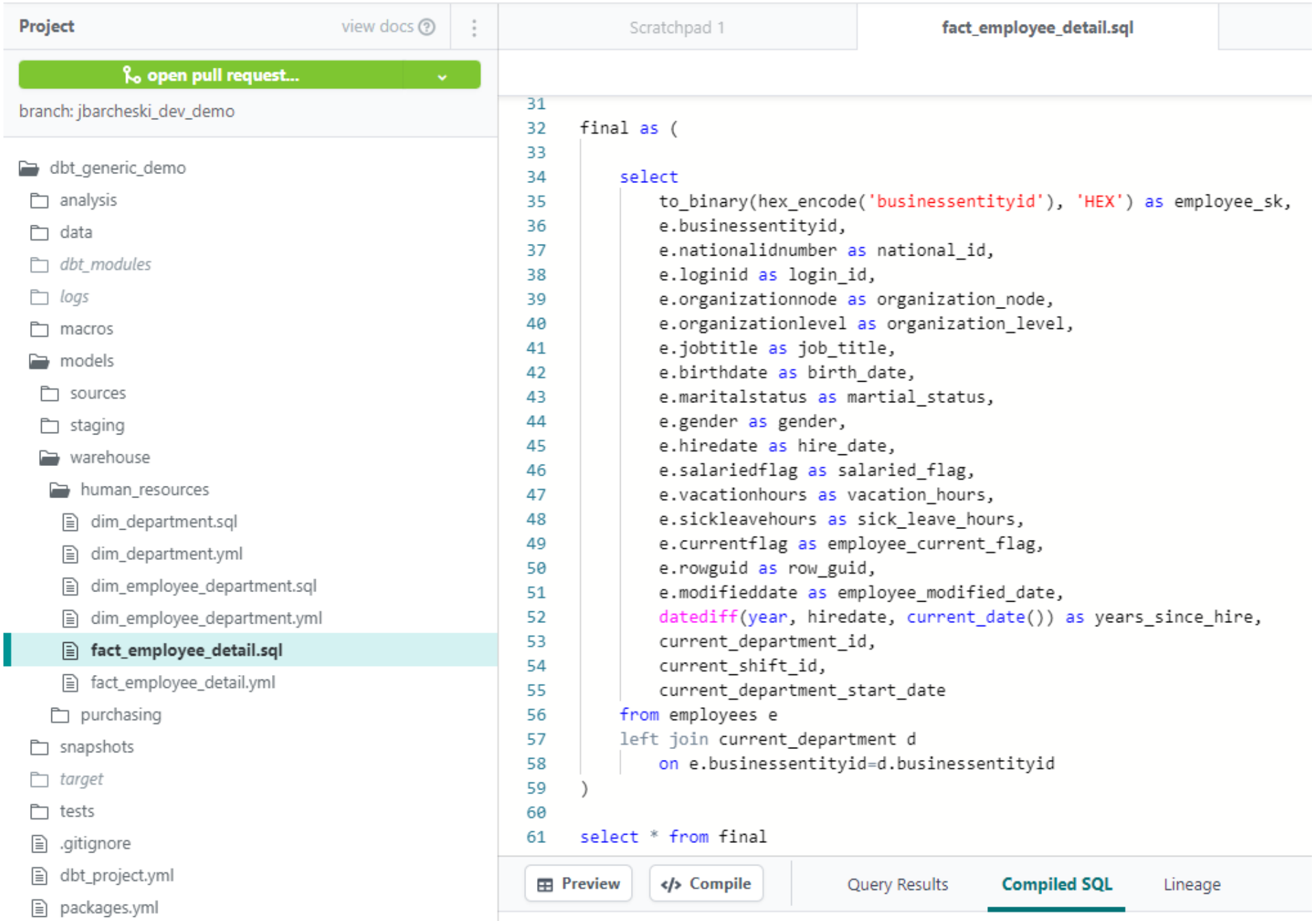
Data | Row Count | Filter Not Set | Export Sa

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"),' '))

[Matillion]

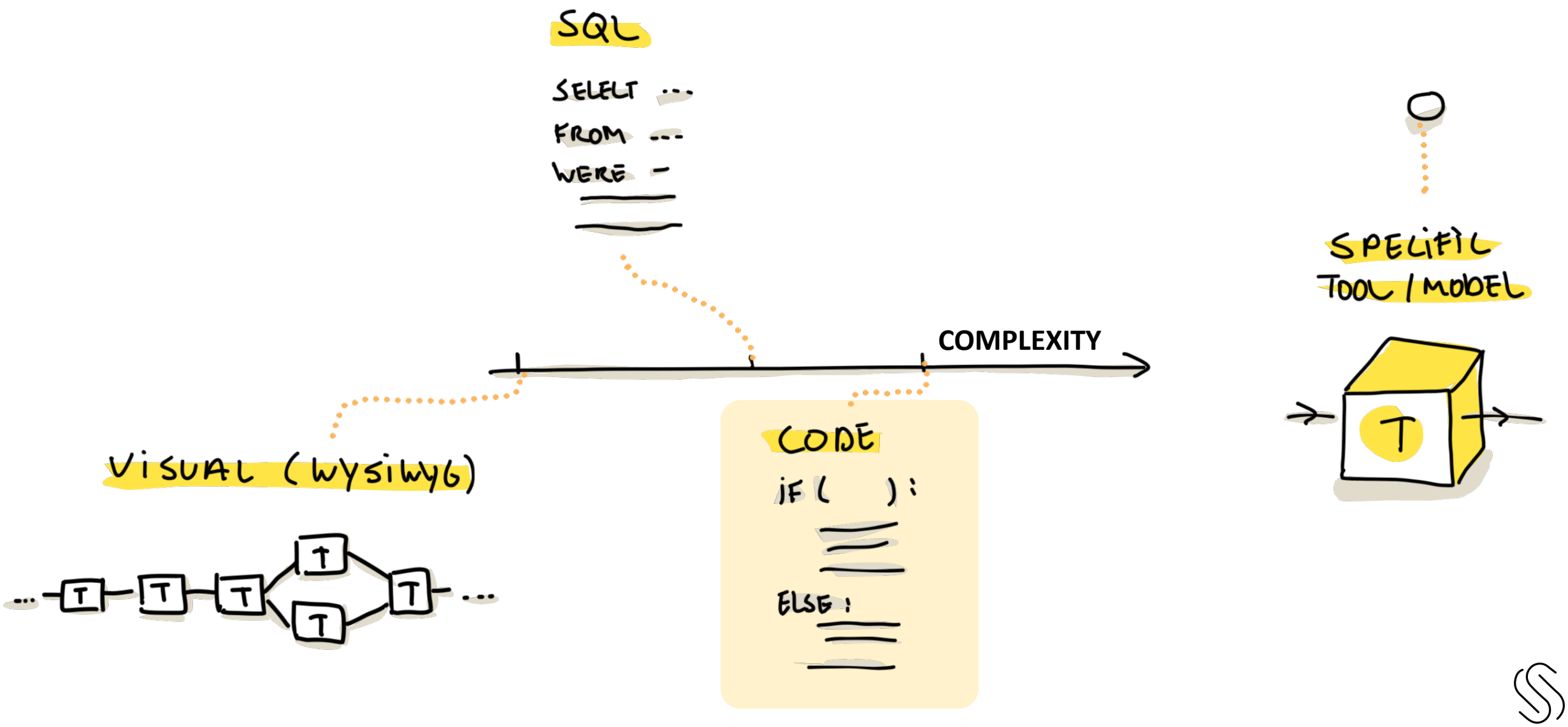
Data Transformation Tool Flavors




Project
view docs
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
Scratchpad 1
fact_employee_detail.sql
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

[DBT]

Data Transformation Tool Flavors

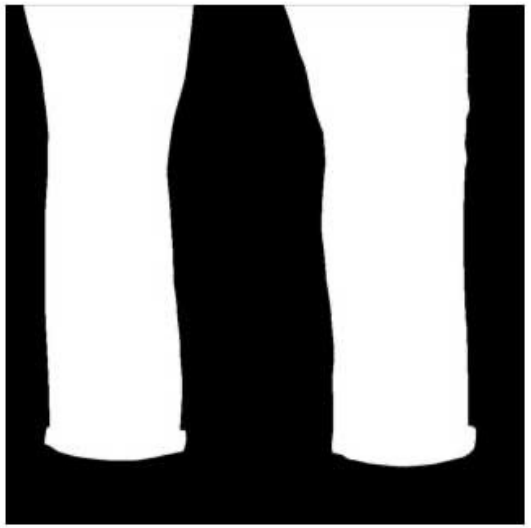


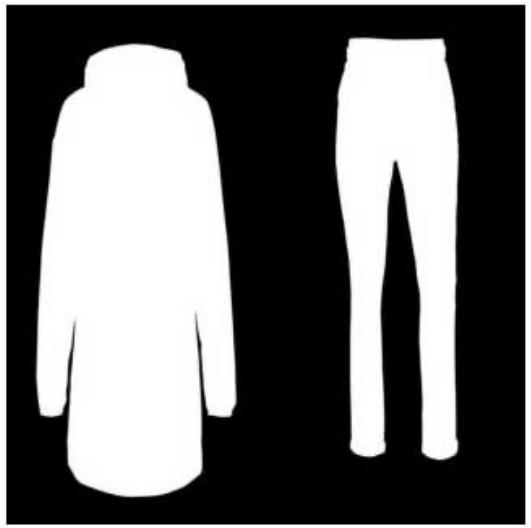
Background-color change

Background-color change

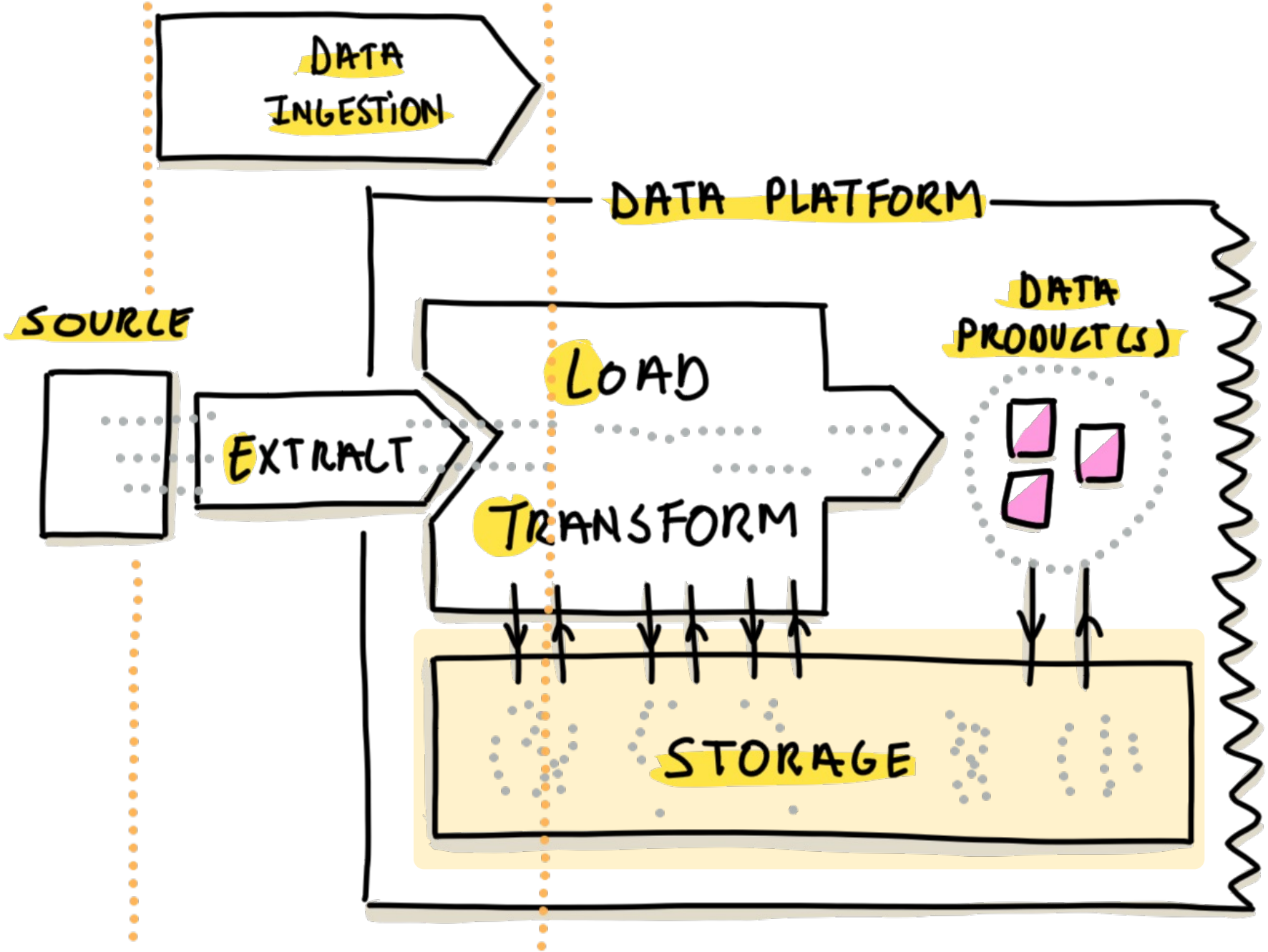
Cropping

Combined images

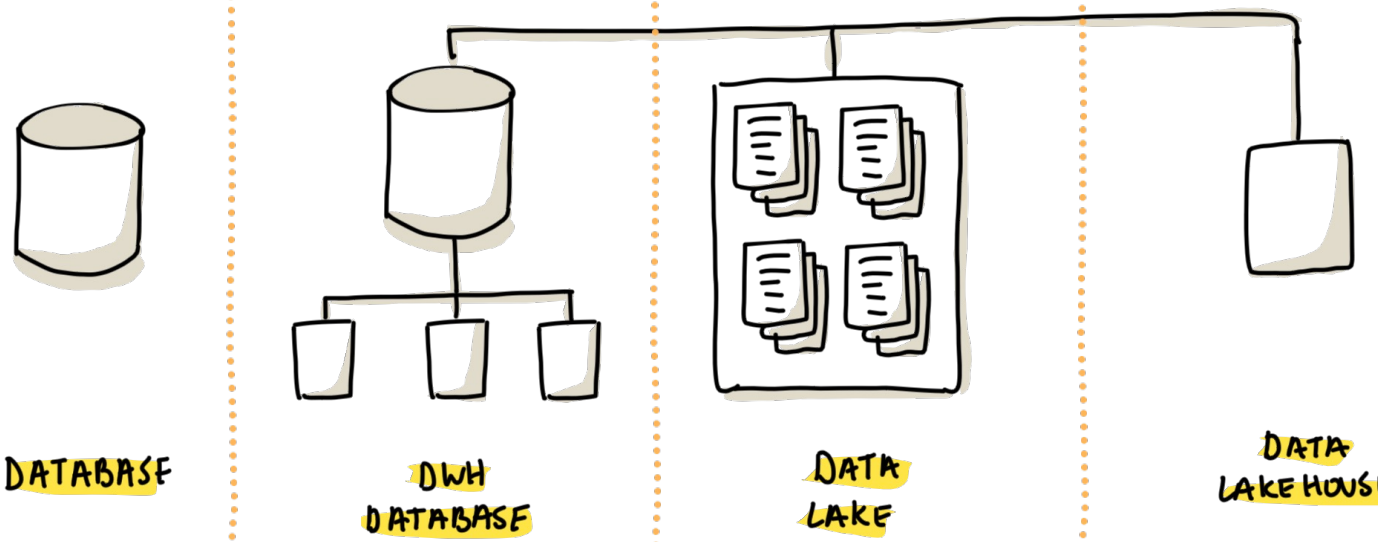




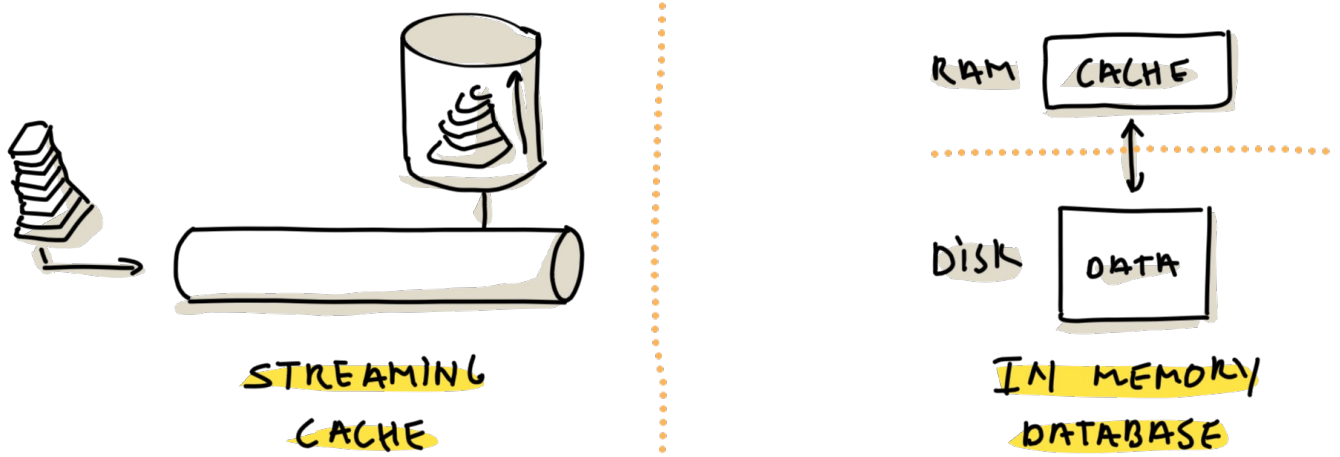
[Automated Background Removal in E-commerce Fashion Image Processing Using PyTorch on Databricks]


DATA
INGESTION
DATA PLATFORM
SOURCE
EXTRACT
LOAD
TRANSFORM
DATA
PRODUCT(S)
STORAGE

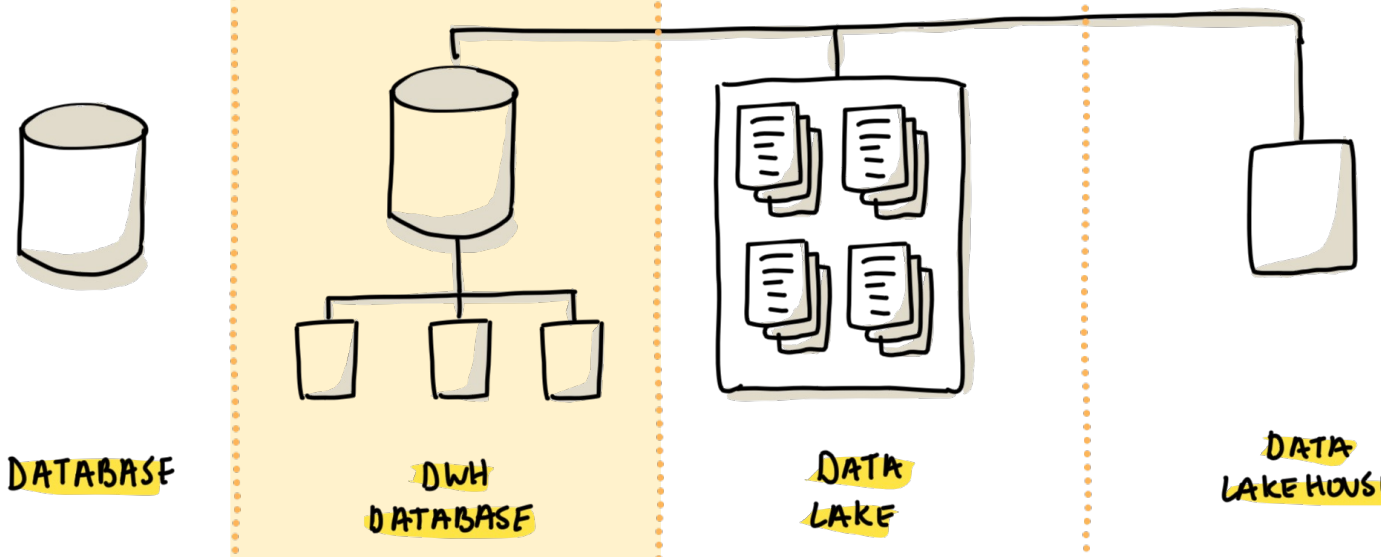
COMMON



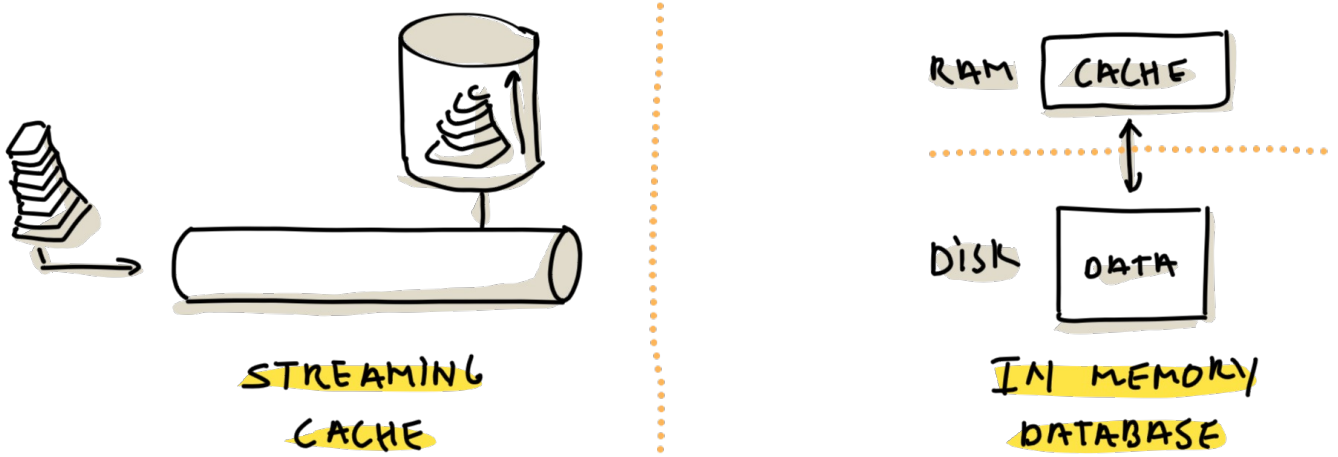
SPECIFIC

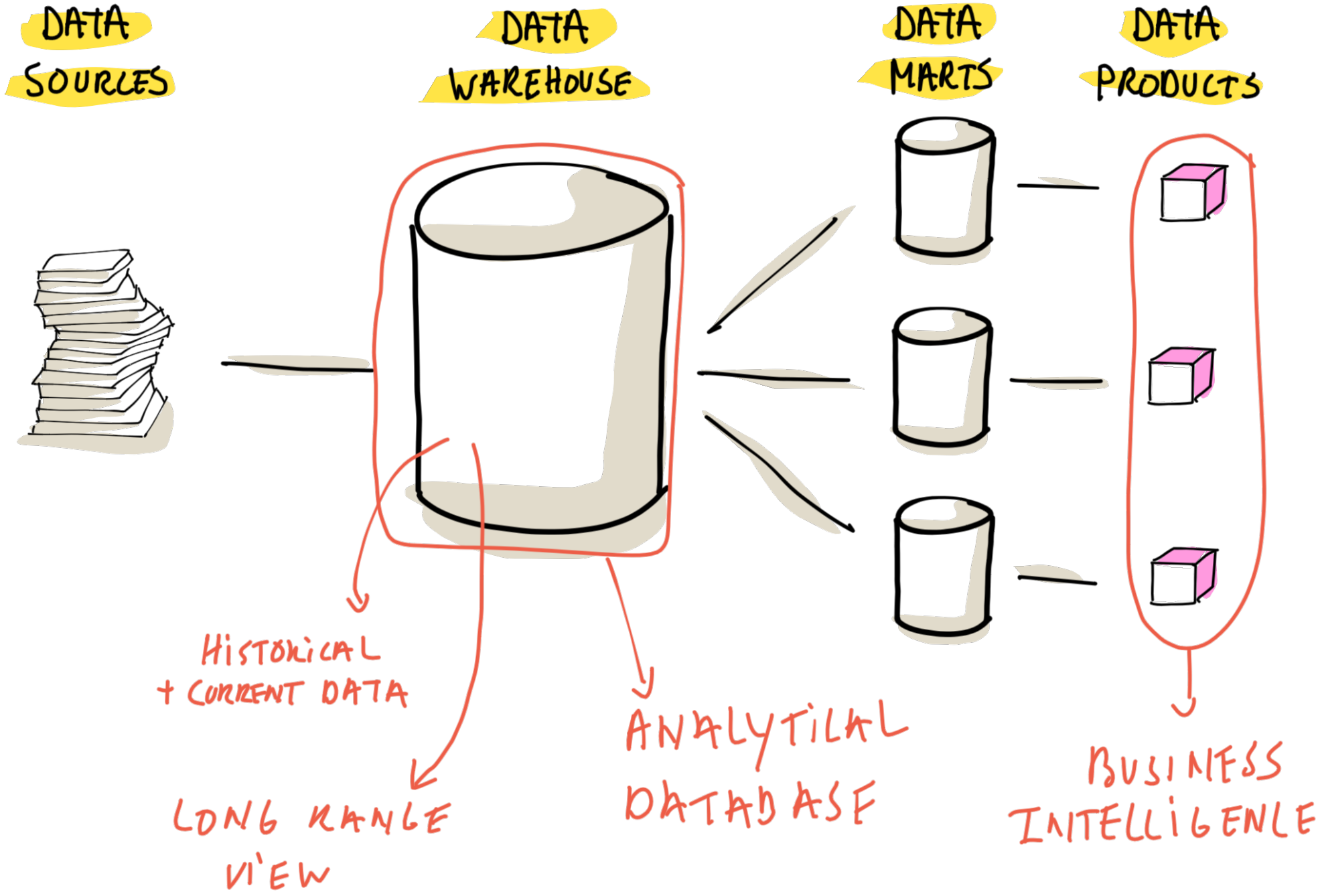


COMMON



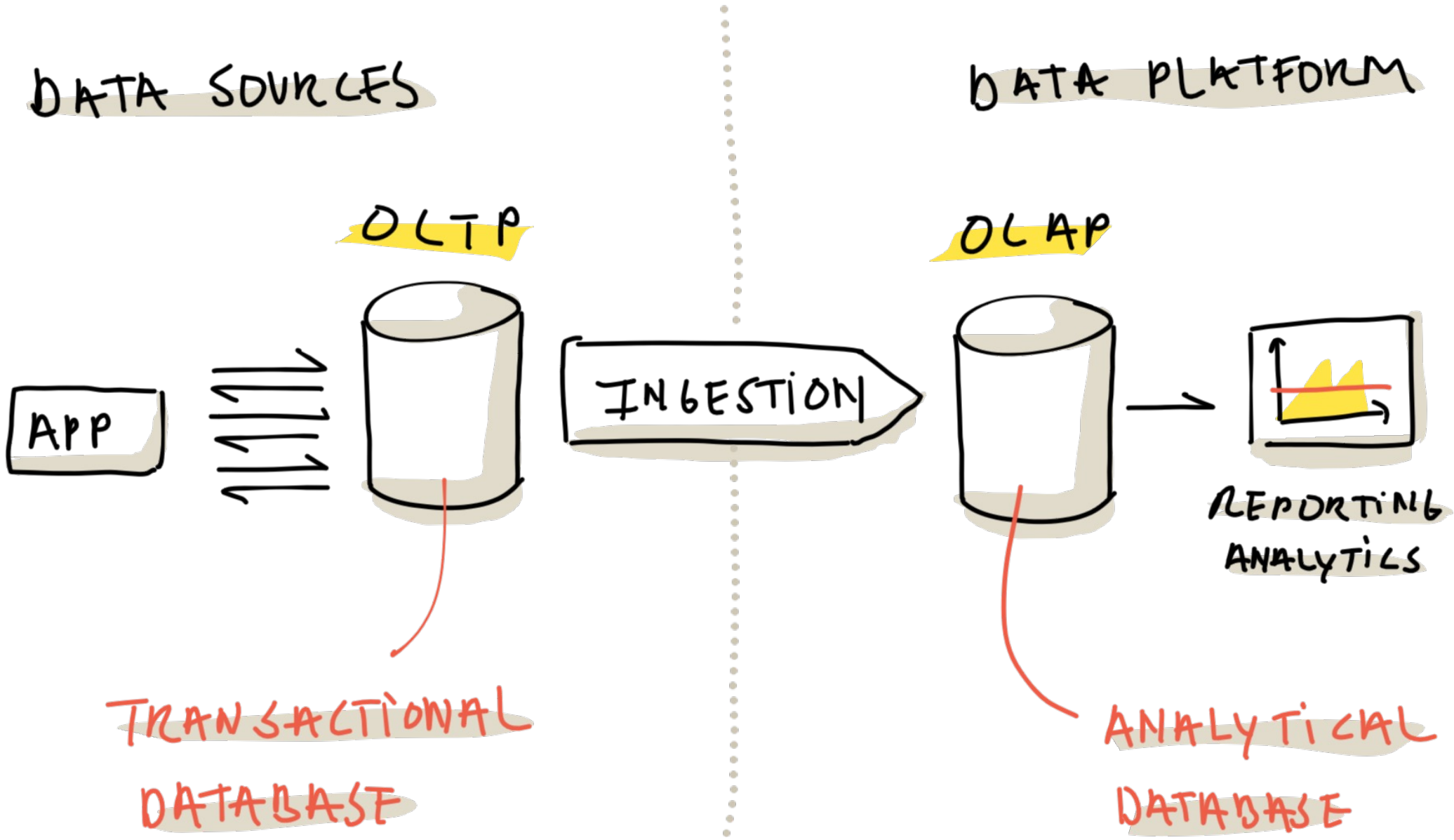
SPECIFIC




DATA SOURCES
DATA WAREHOUSE
DATA MARTS
DATA PRODUCTS
HISTORICAL + CURRENT DATA
LONG RANGE VIEW
ANALYTICAL DATABASE
BUSINESS INTELLIGENCE

Data Warehouse

- = An enterprise data platform used for the analysis and reporting of **mainly structured and semi-structured data** from multiple data sources
- Also called an Enterprise Data Warehouse (EDW)
- A data warehouse:
  - Can store both current and historical data in one place
  - Designed to give a long-range view of data over time
  - A primary component of business intelligence.
- Data Warehouses are built on top of an Analytical Database


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

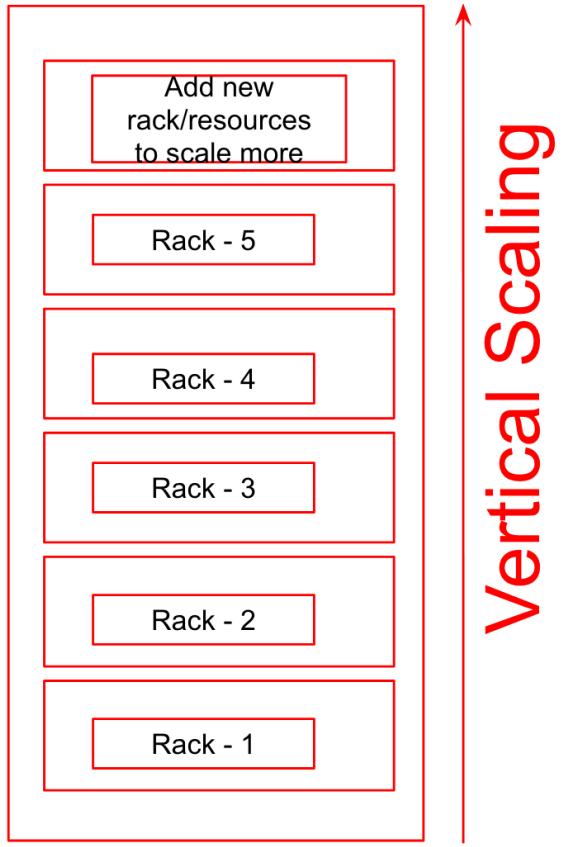
OLTP vs OLAP

- Online Transactional Processing (OLTP):
  - Designed to handle large volumes of transactional data involving multiple users.
  - Rapidly update, insert, or delete small amounts of data in real-time
  - Databases that serve business applications (ERPs, CRM, …)
- Online Analytical Processing (OLAP):
  - Designed to process large amounts of data quickly
  - Enabling in-depth data analysis
  - The core of Analytical Databases, suited for DWHs

Data Mart

- A specialized view of data that caters to a particular group of users within an organization.
- For example, a ‘Finance Mart’
  - **Goal**: Supporting the finance department to perform predictive analysis for some of its customers.
  - **Solution**: A specific subset of consumer data from the data warehouse to create a data mart that best serves the predictive analysis.
- Advantages:
  - **Faster Data Access**: Quick and focused access, faster time-to-market compared to crawling the complete DWH
  - **Faster Decision Making:** Focused on supporting decision making of a specific department.

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

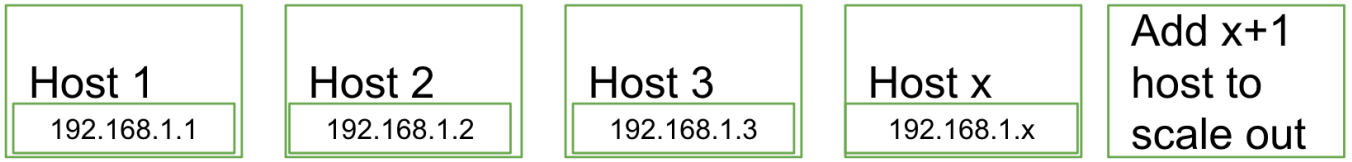


Vertical Scaling

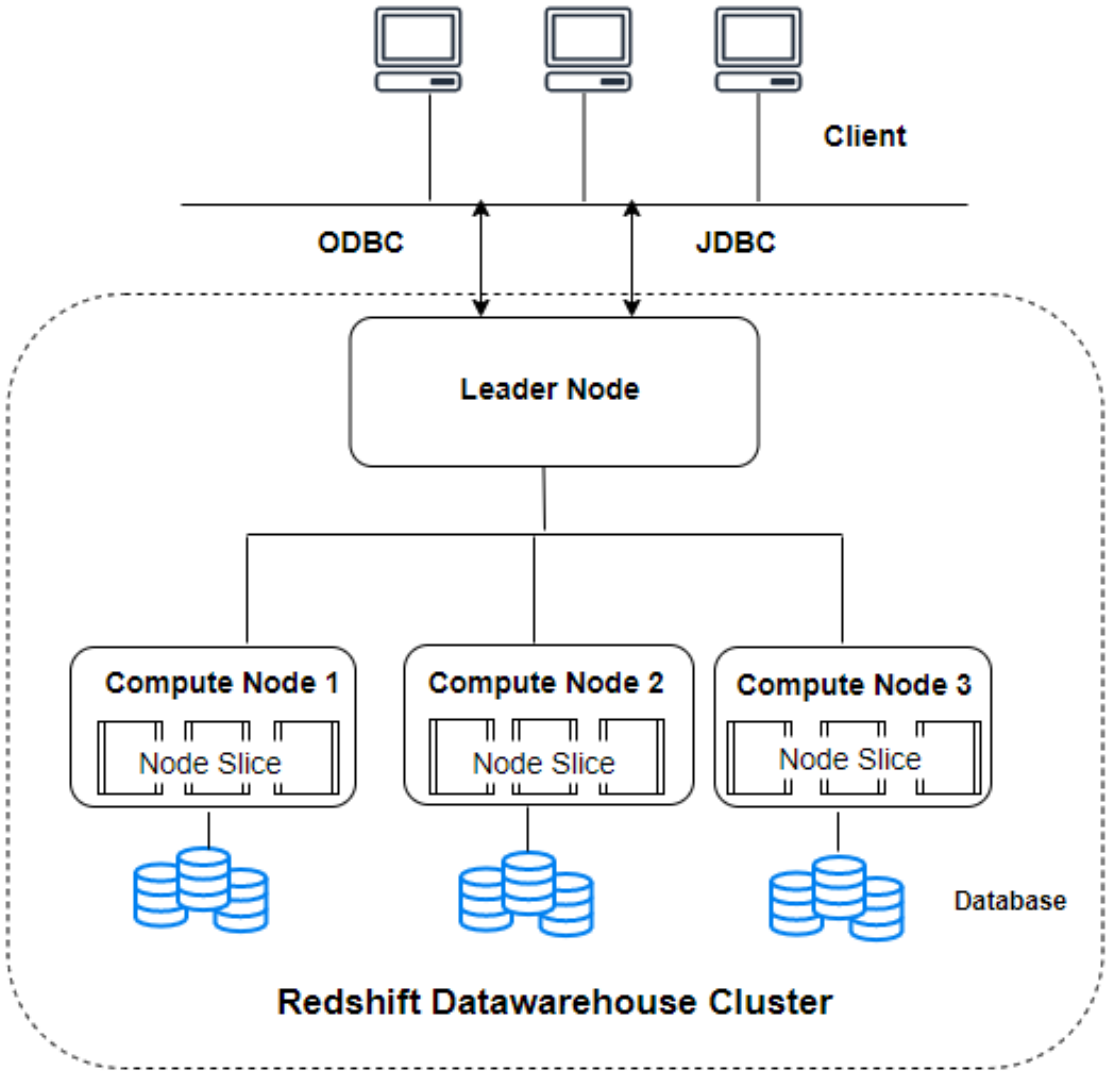
To scale more, Add more RAM, CPU, Memory to the **one existing machine**

Horizontal Scaling

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

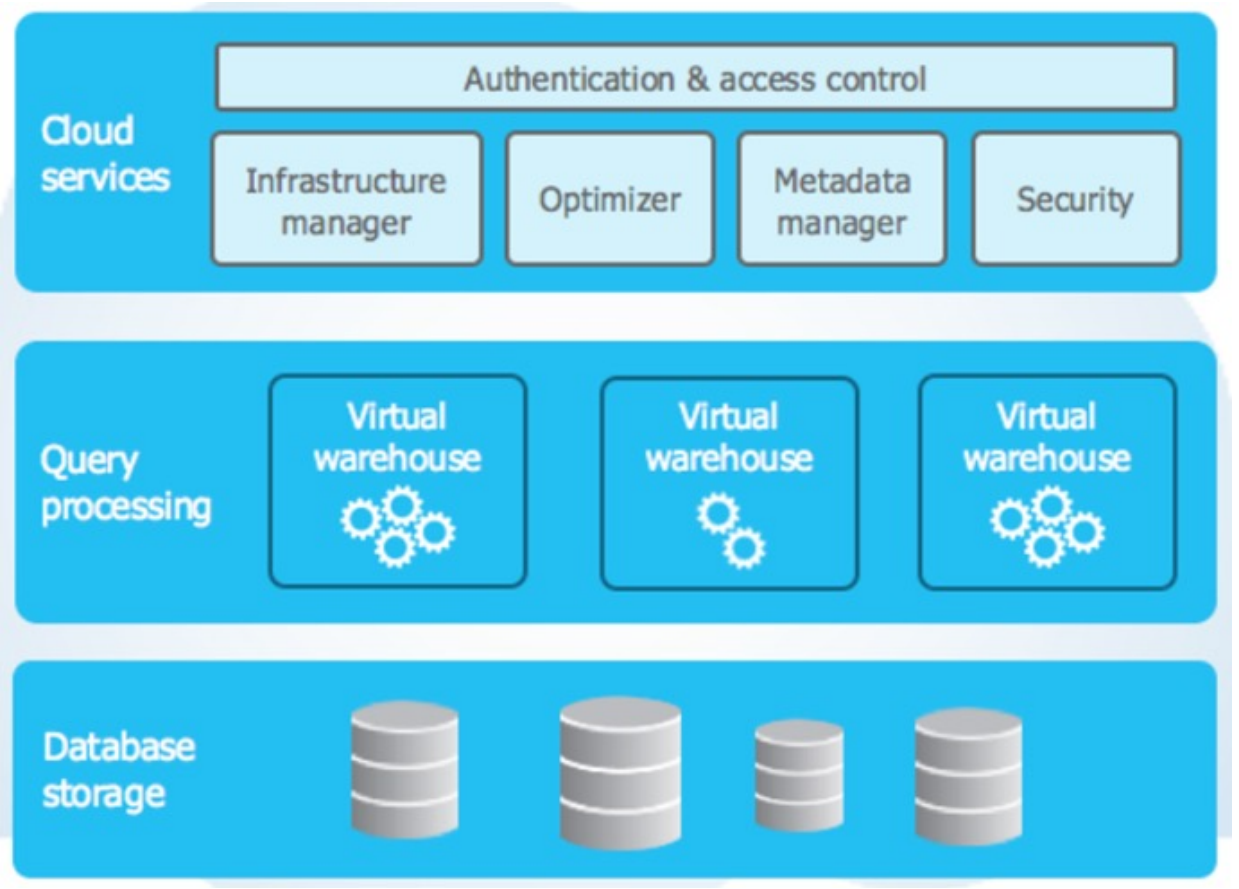


Example: Amazon Redshift



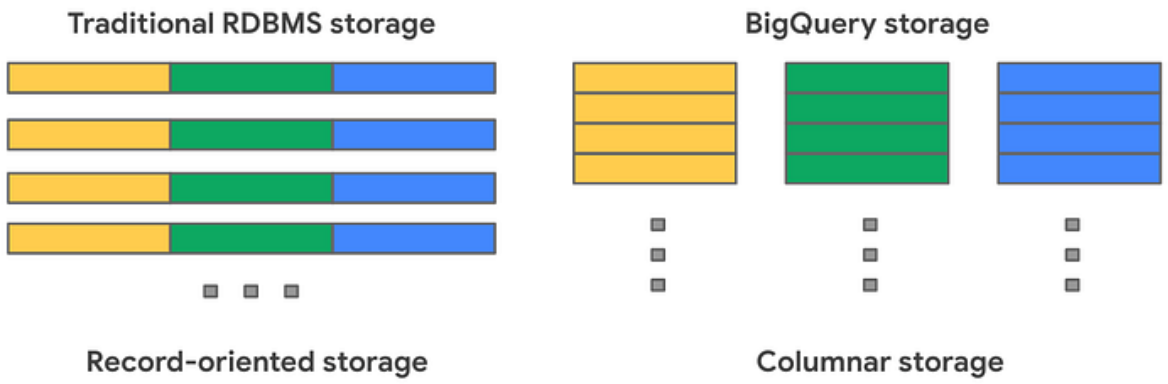
Scaling = Refconfiguring total databasenv

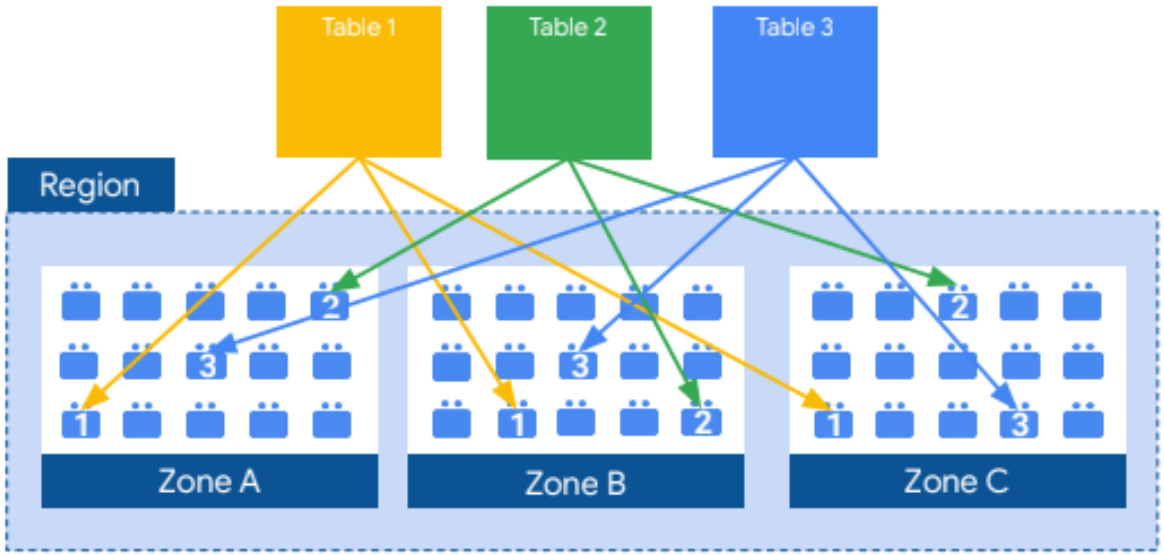
Example: Snowflake



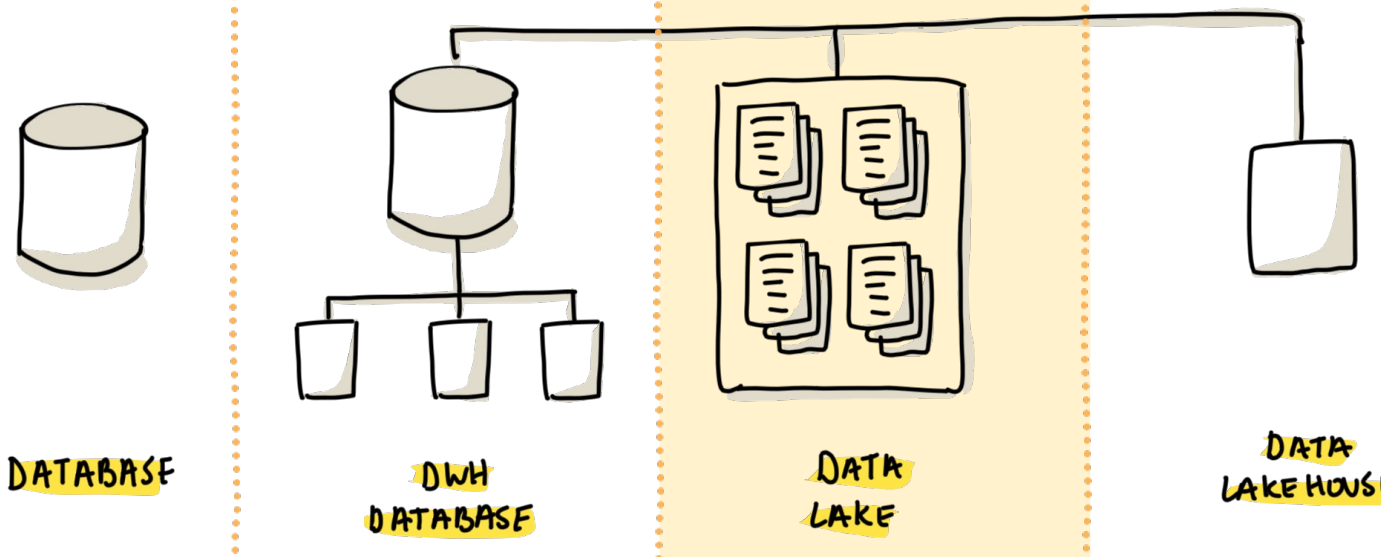
Independent Scaling

Example: Google BigQuery

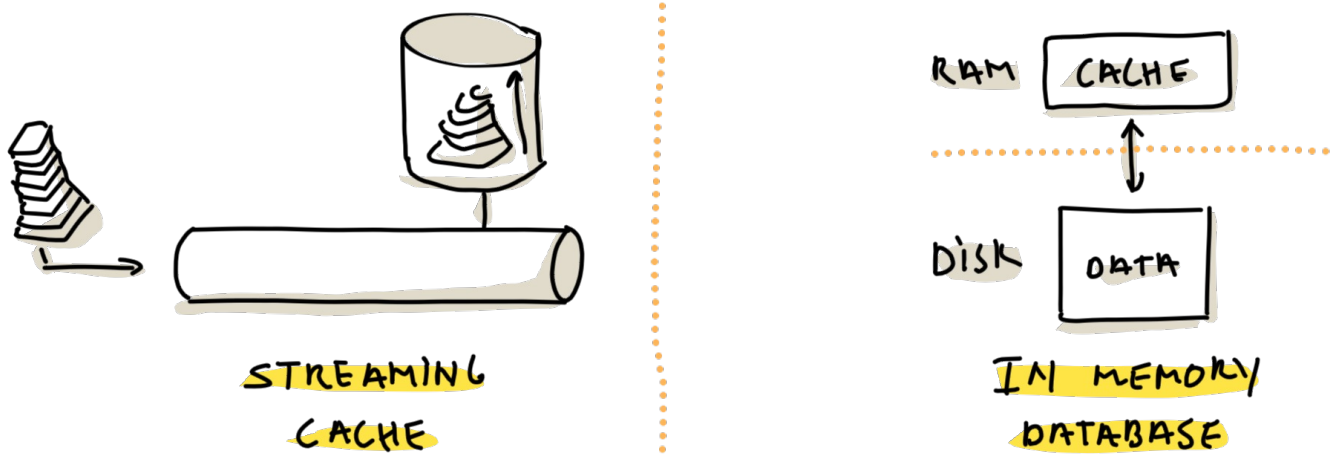




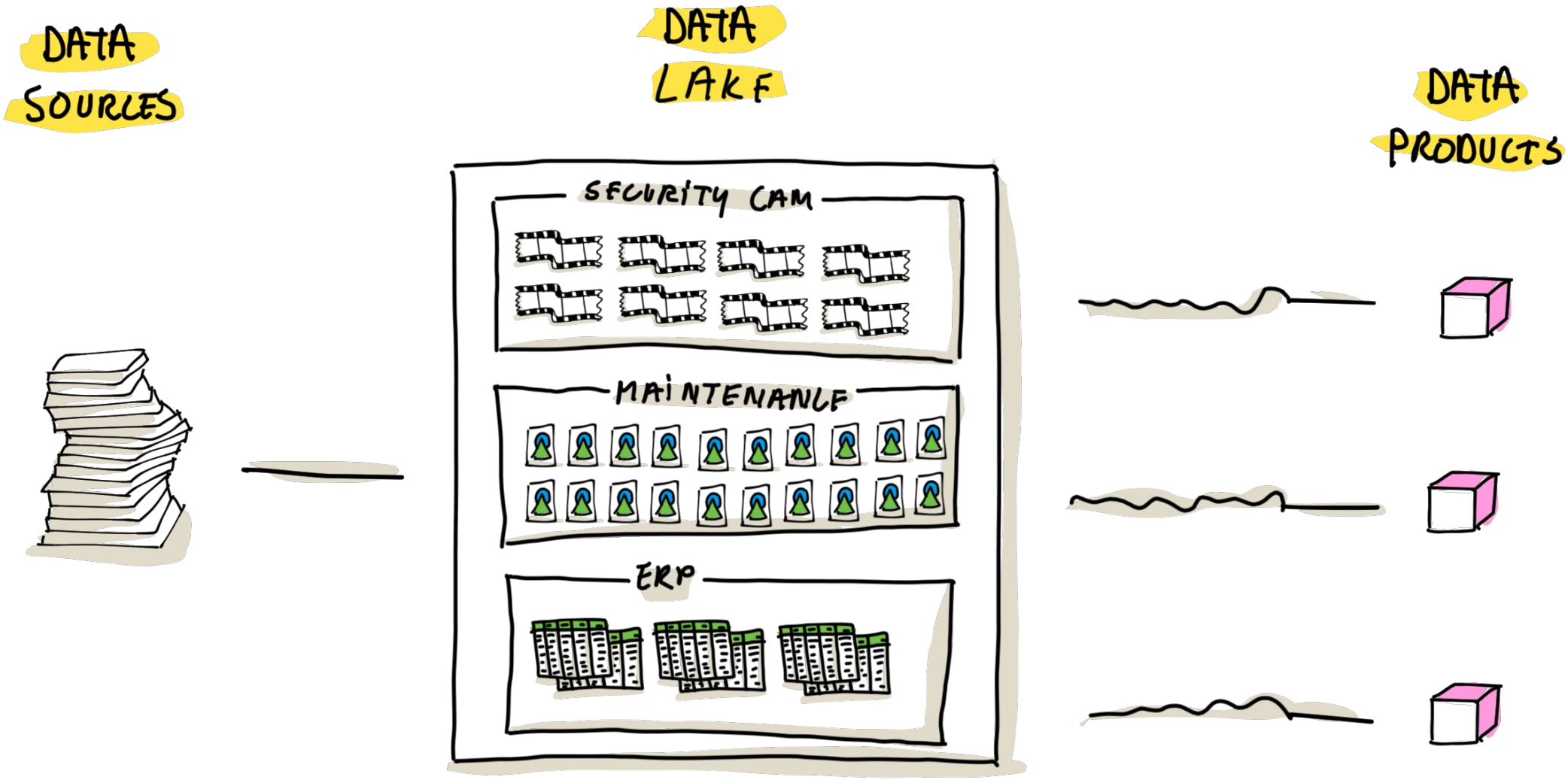
COMMON

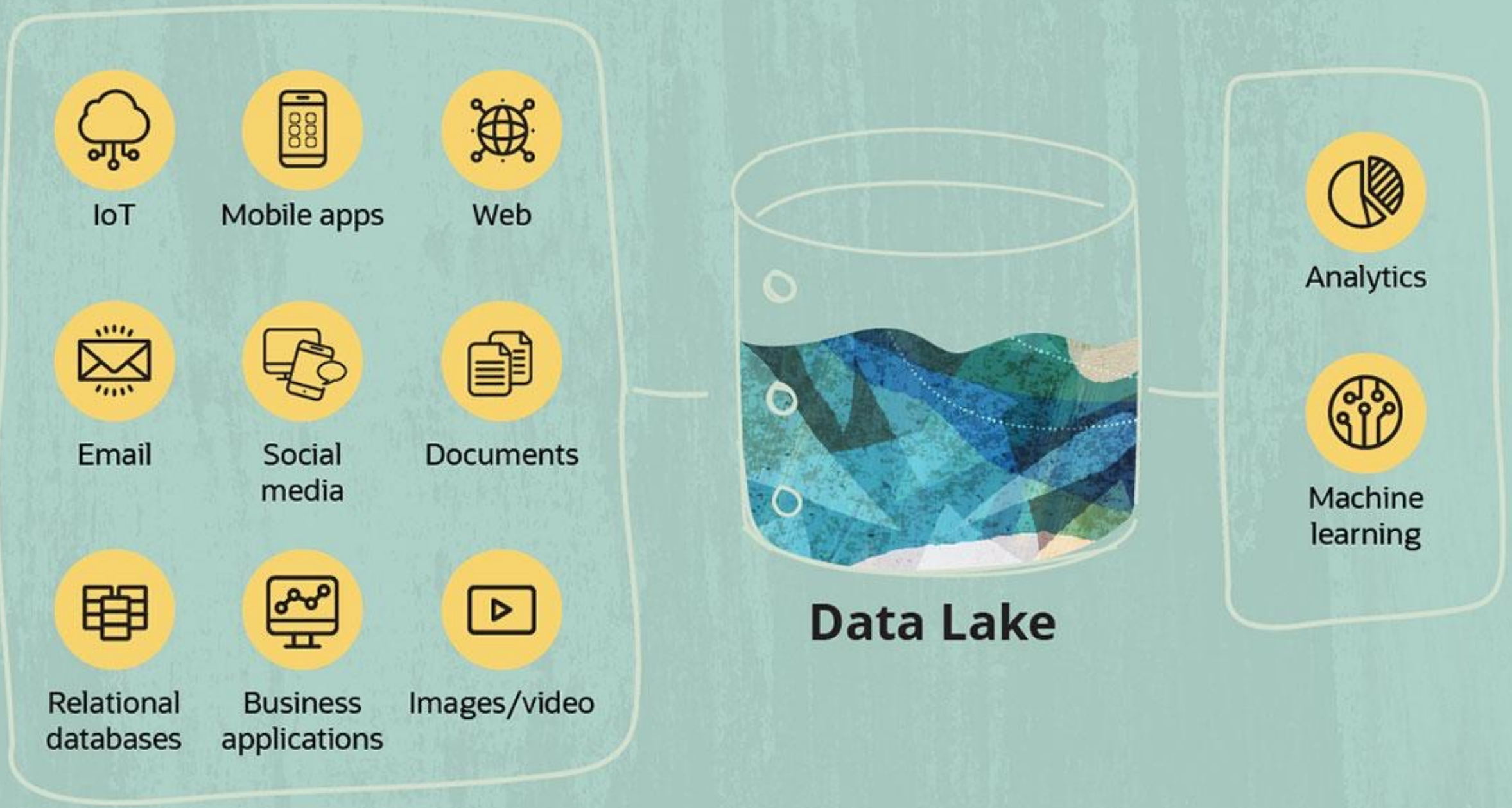


SPECIFIC



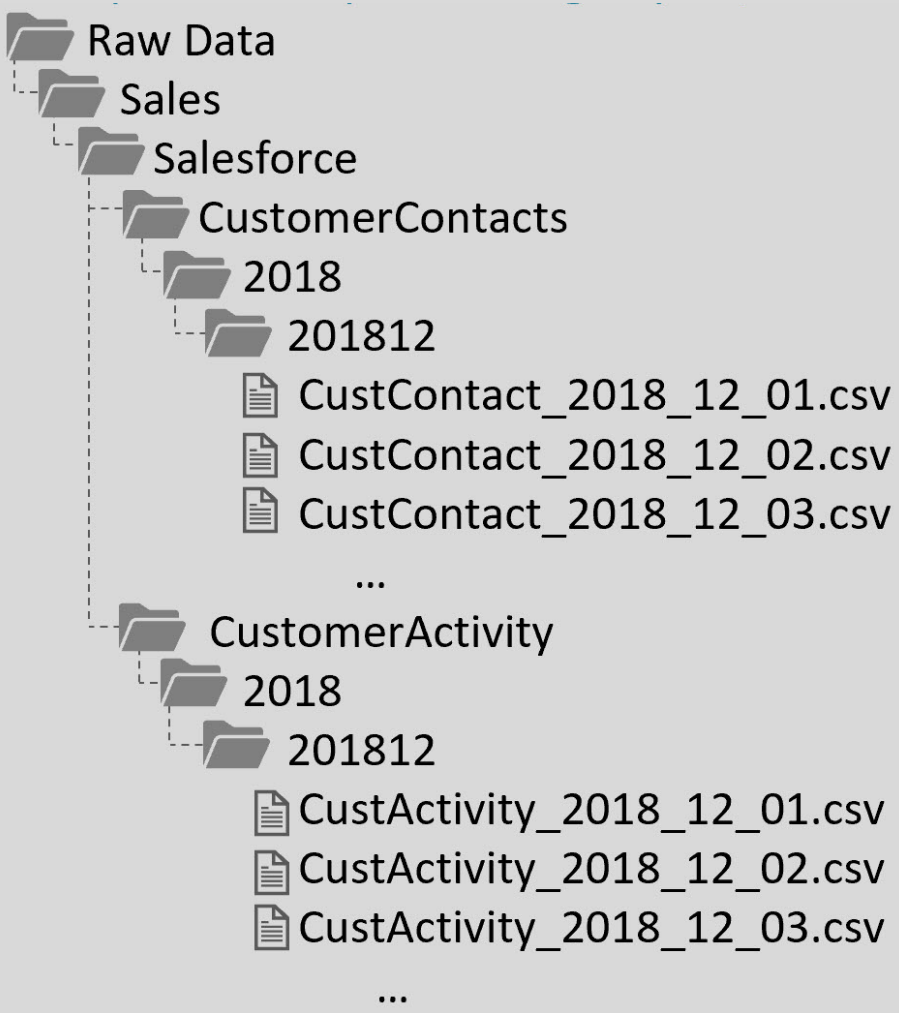
Data Lake


DATA SOURCES
DATA LAKE
DATA PRODUCTS
SECURITY CAM
MAINTENANCE
ERP

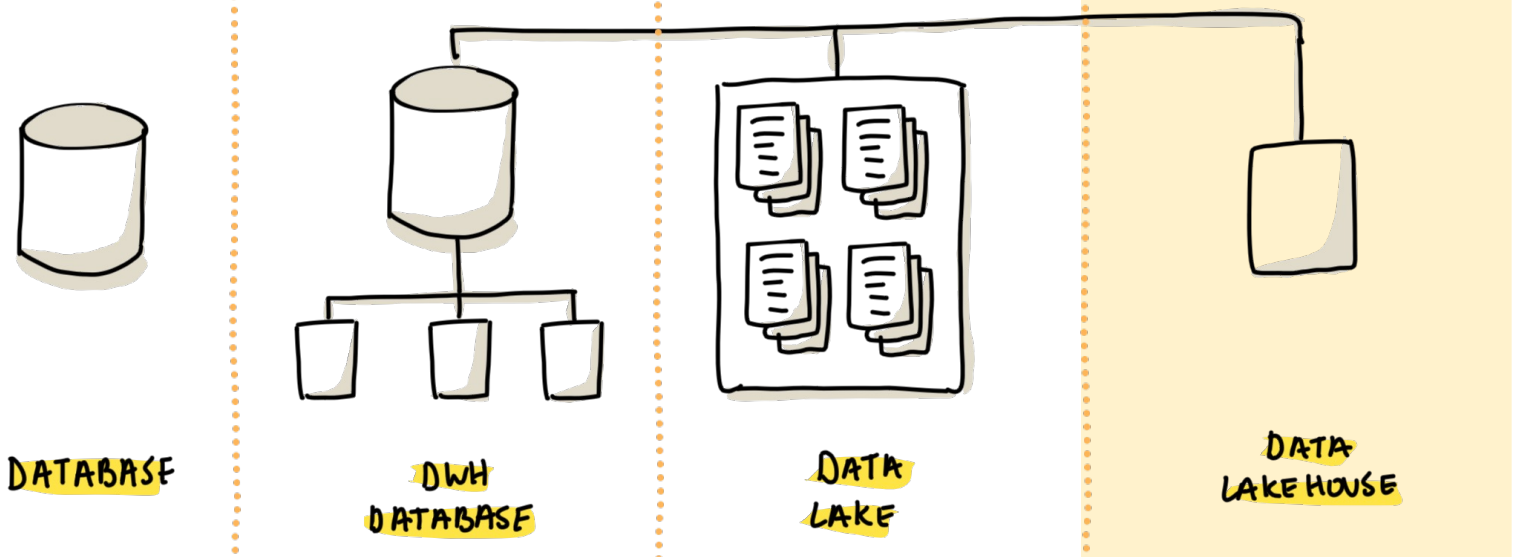


[What is a Data Lake? Elements and Best Practices]

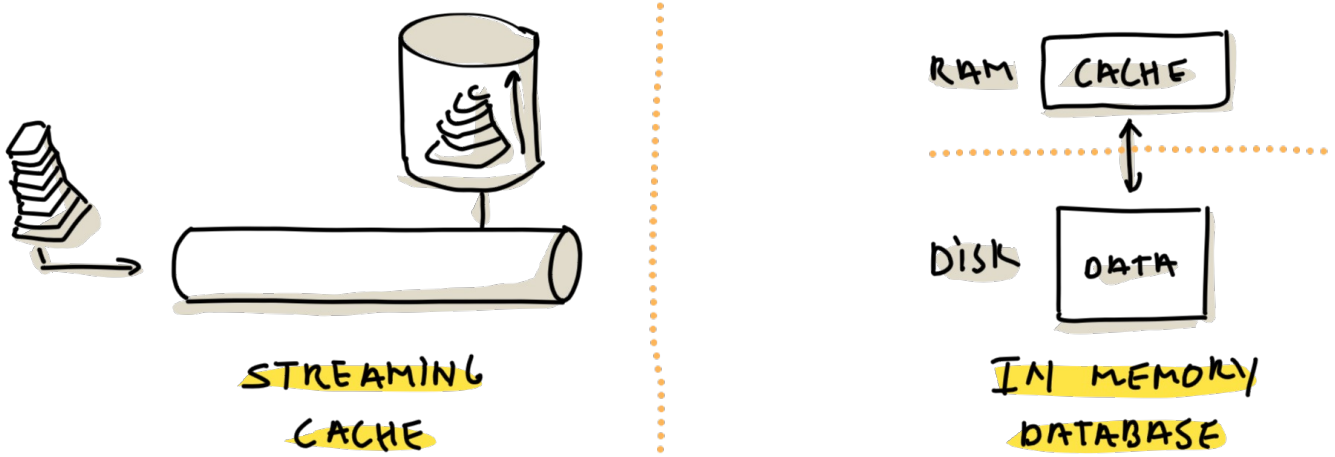
Data Lake: A **Structured** File Repository



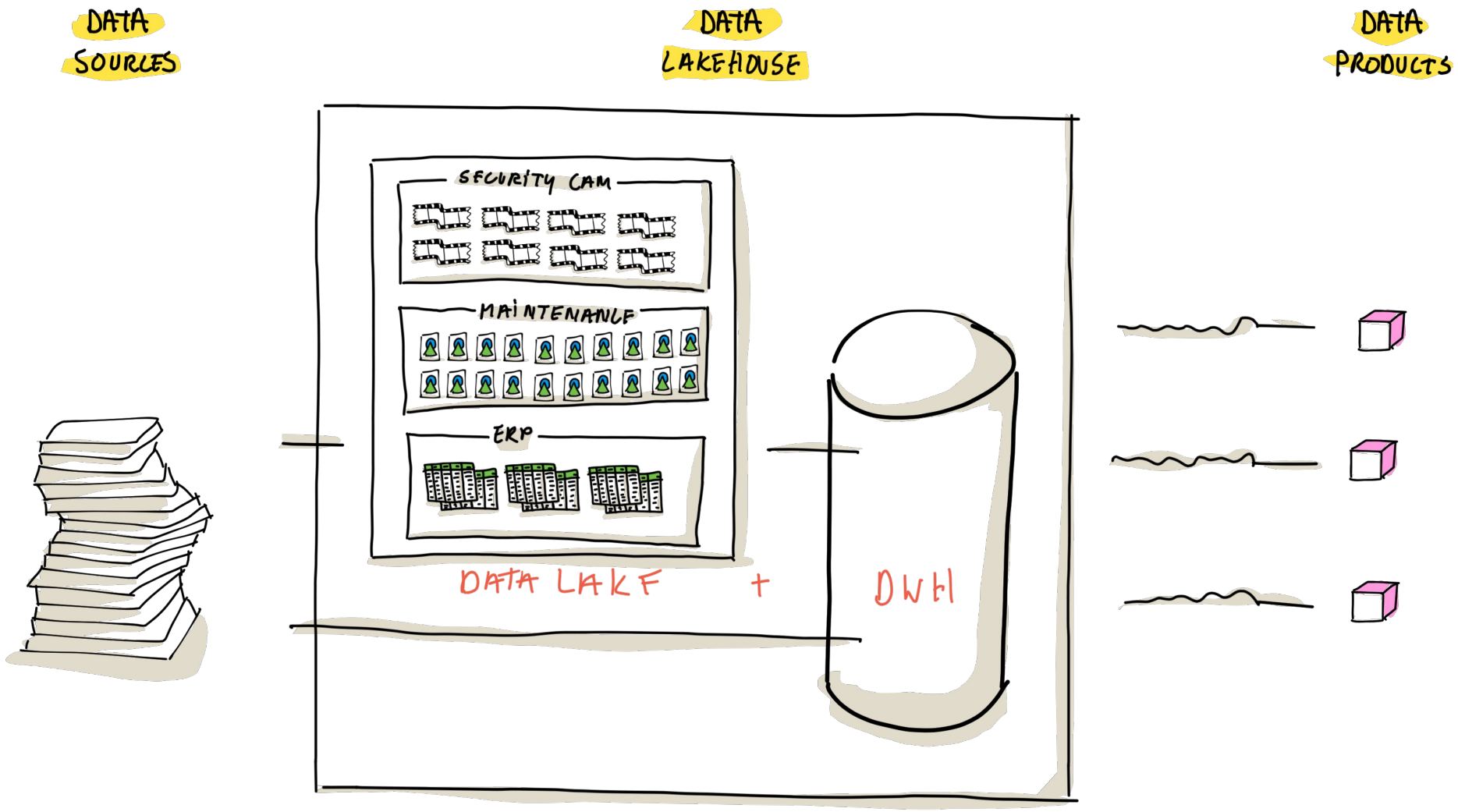
COMMON



SPECIFIC

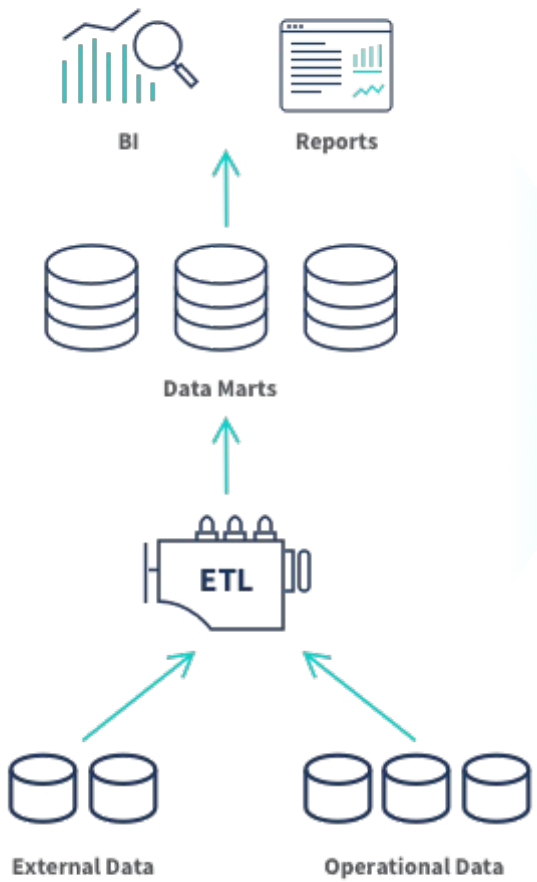


Data Lakehouse



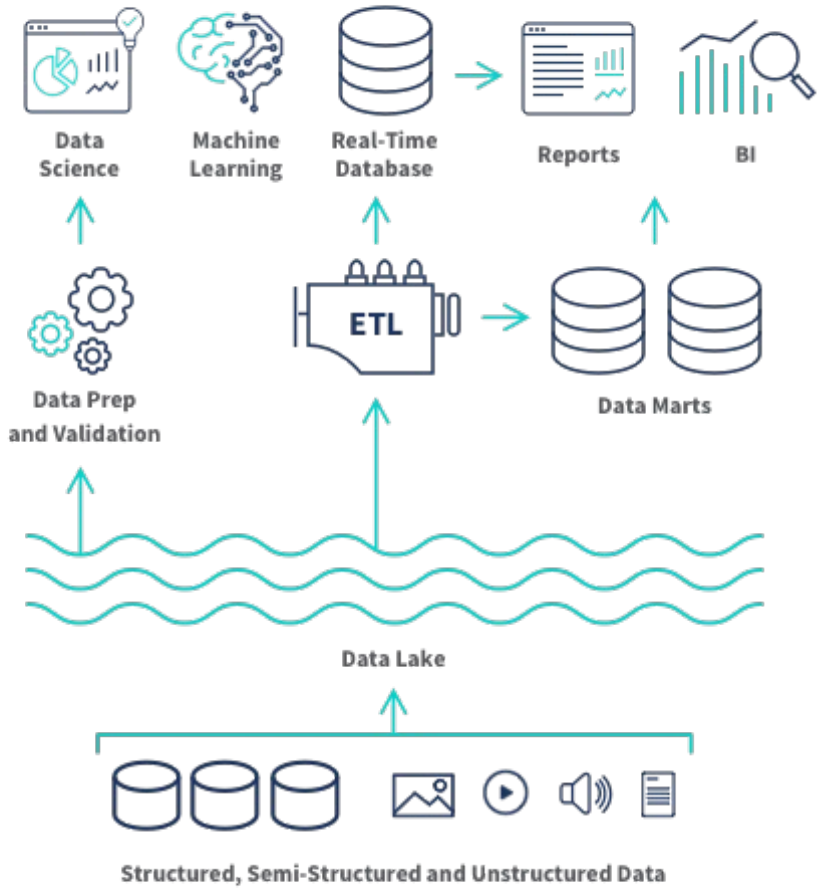
LATE 1980'S

Data Warehouse



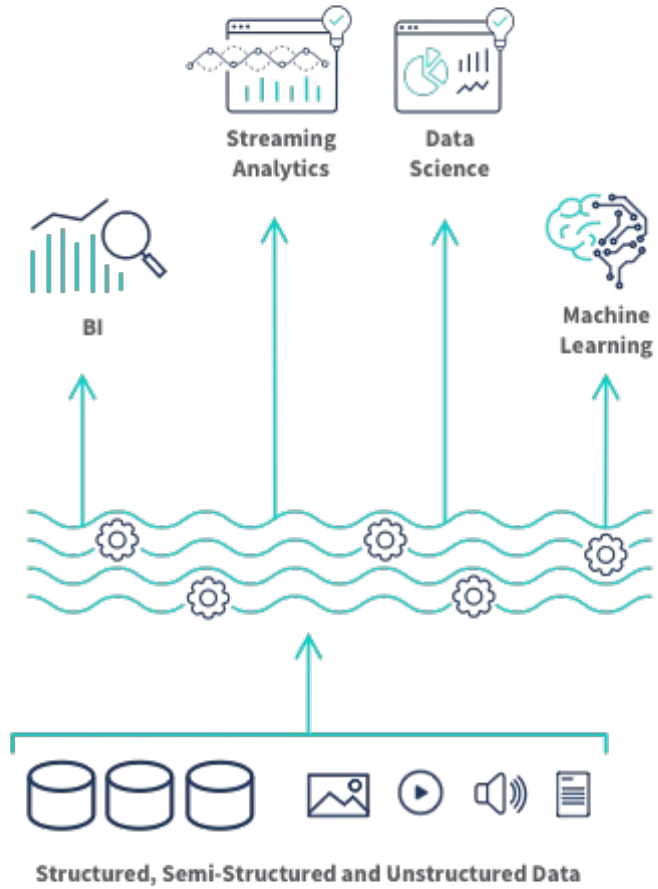
2011

Data Lake



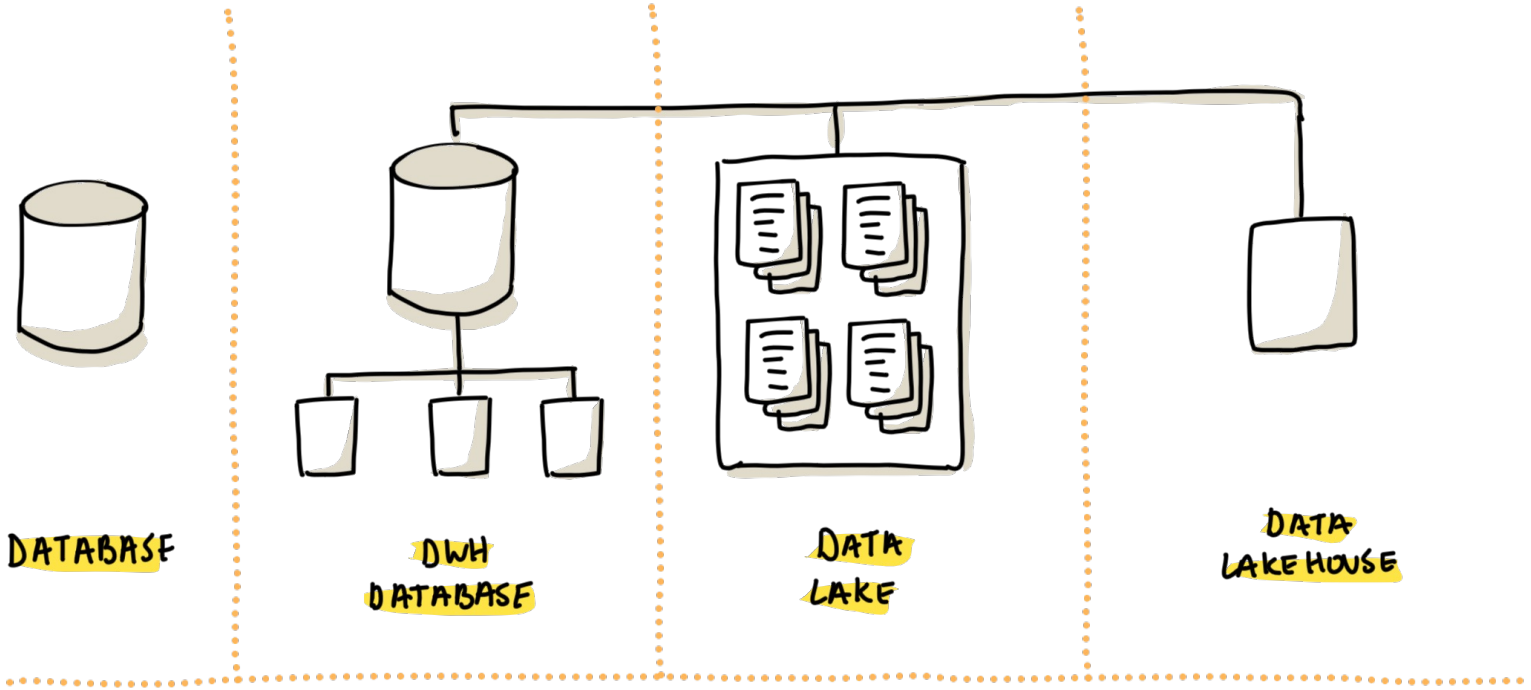
2020

Lakehouse

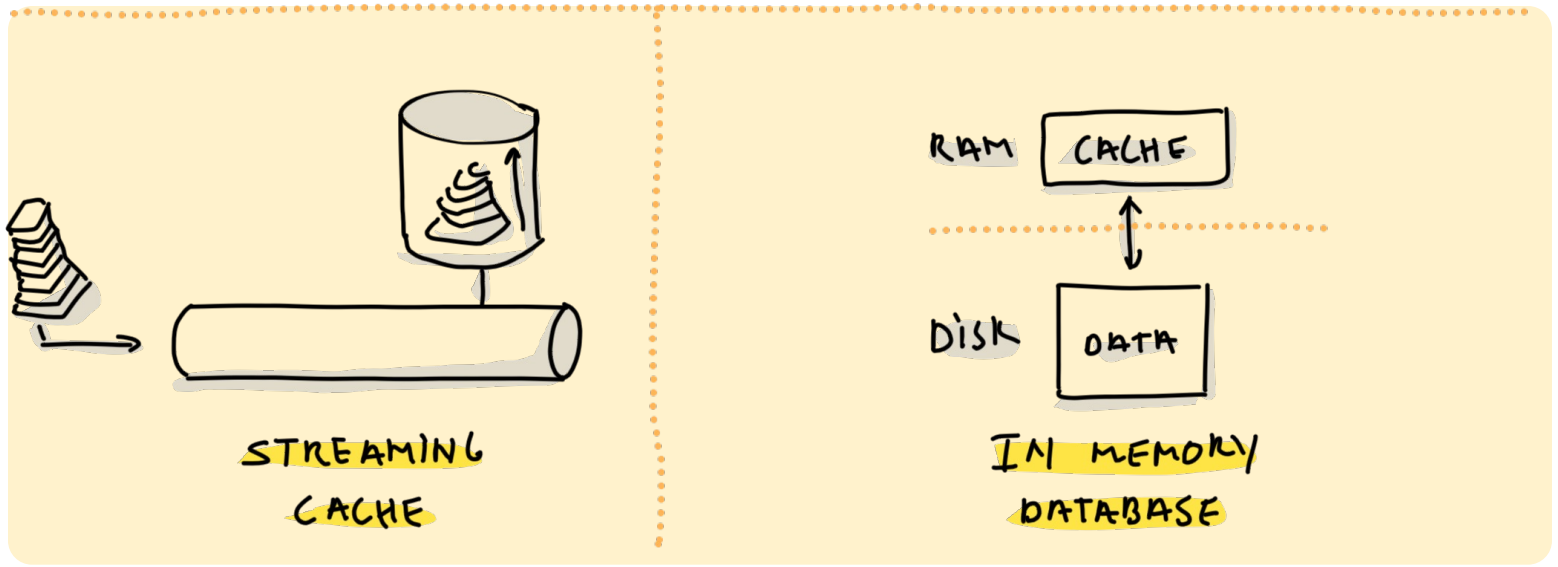


	Database	DWH Database	Data Lake	Data Lakehouse
Cost	+++	+++	+	++
Scaling	Vertical (expensive)	Horizontal (cheaper)		Horizontal (cheaper)
Volume	++	+++	+++++	+++++
Type of Data	Structured	Structured & Semi-Structured	Structured, Semi-Structured & Unstructured	Structured, Semi-Structured & Unstructured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

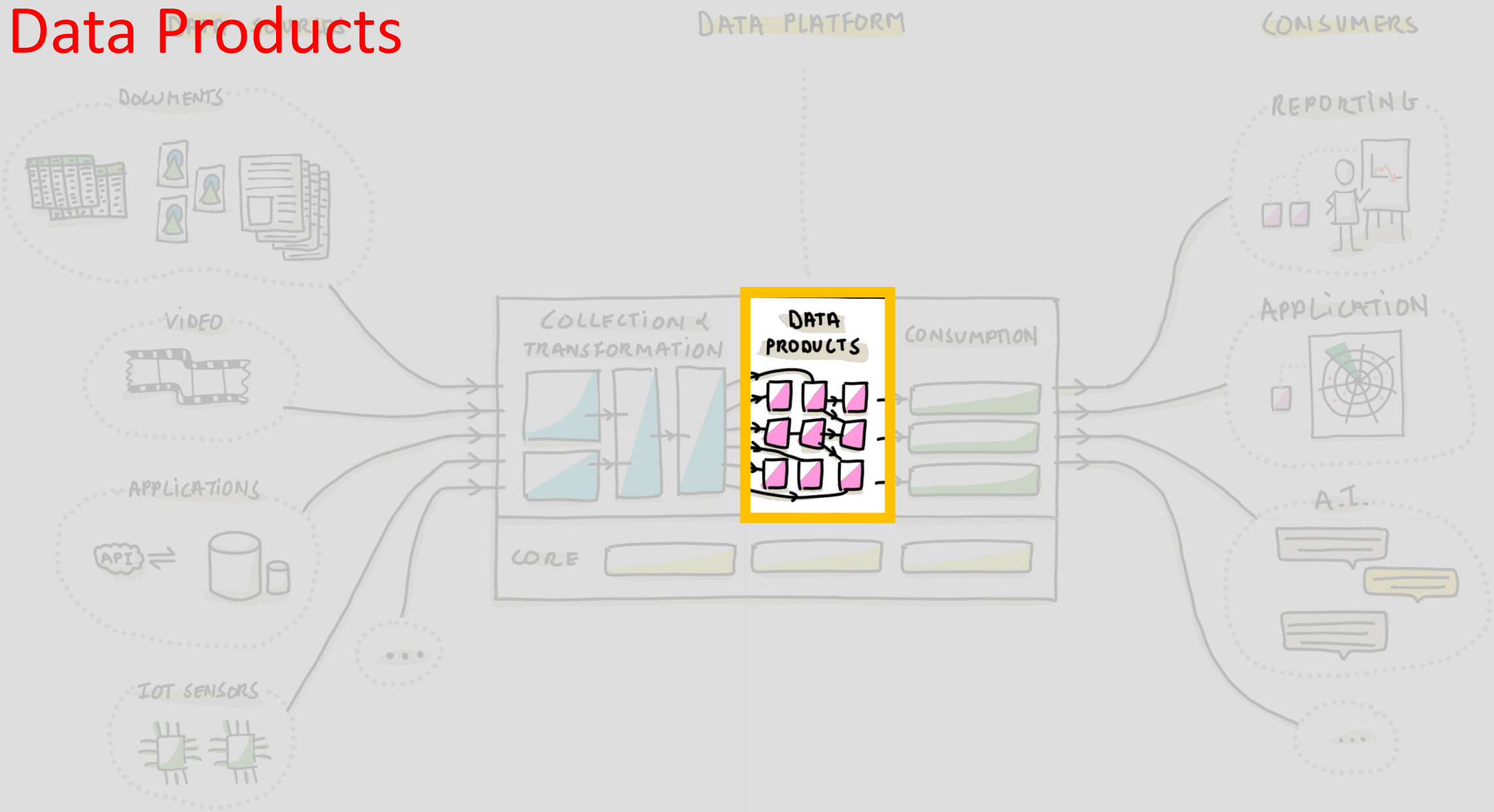
COMMON



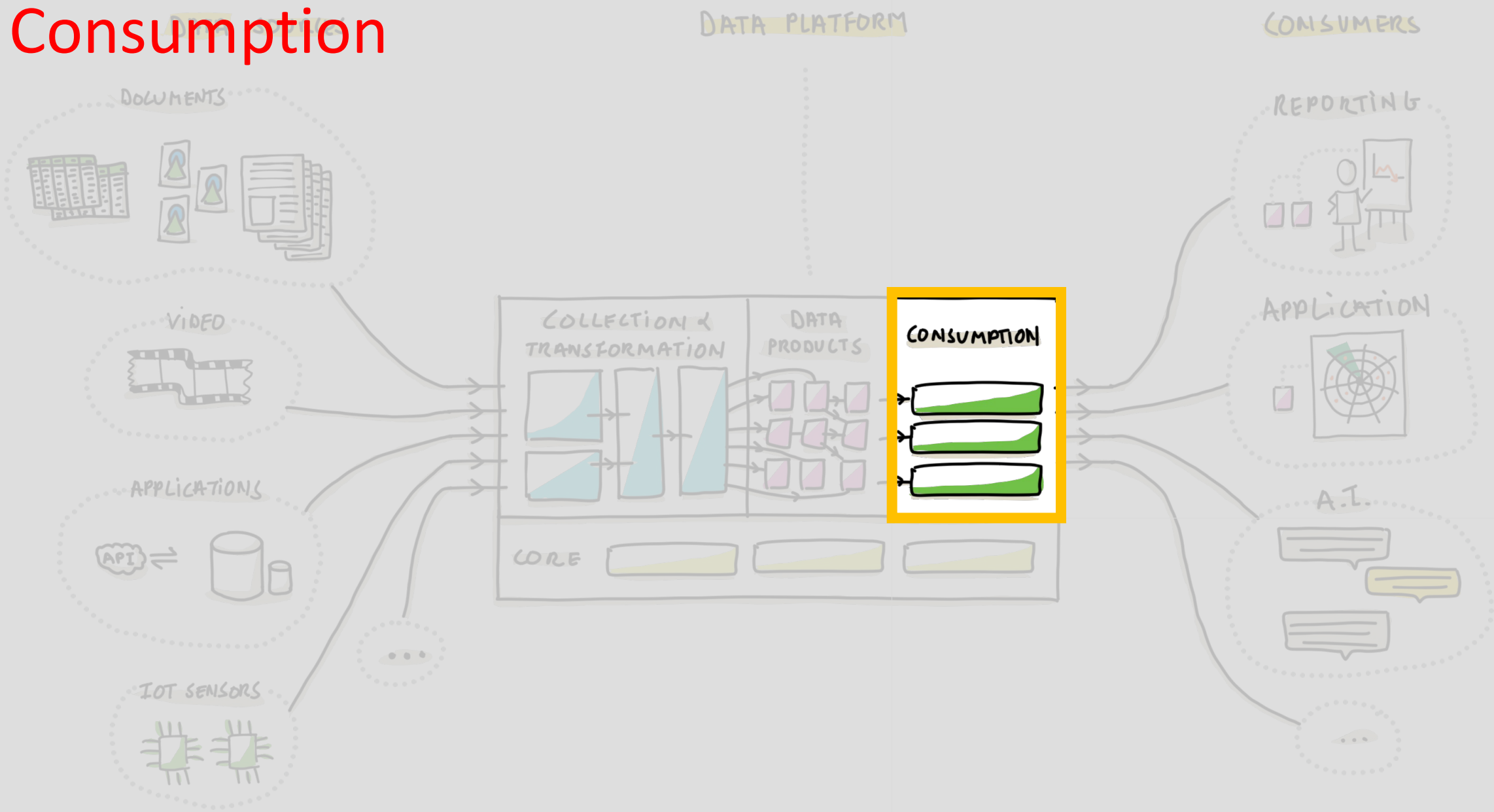
SPECIFIC



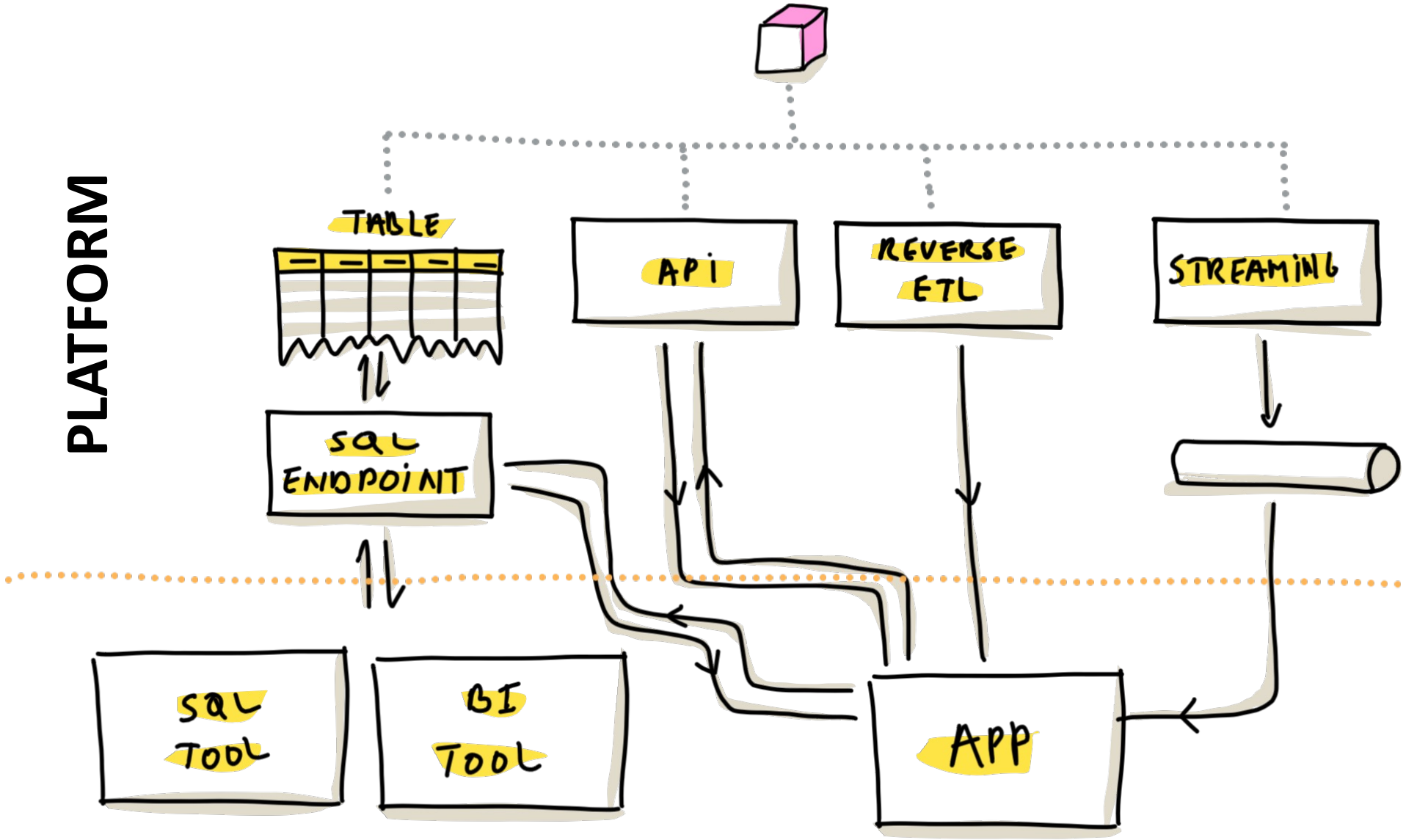
2. Data Products



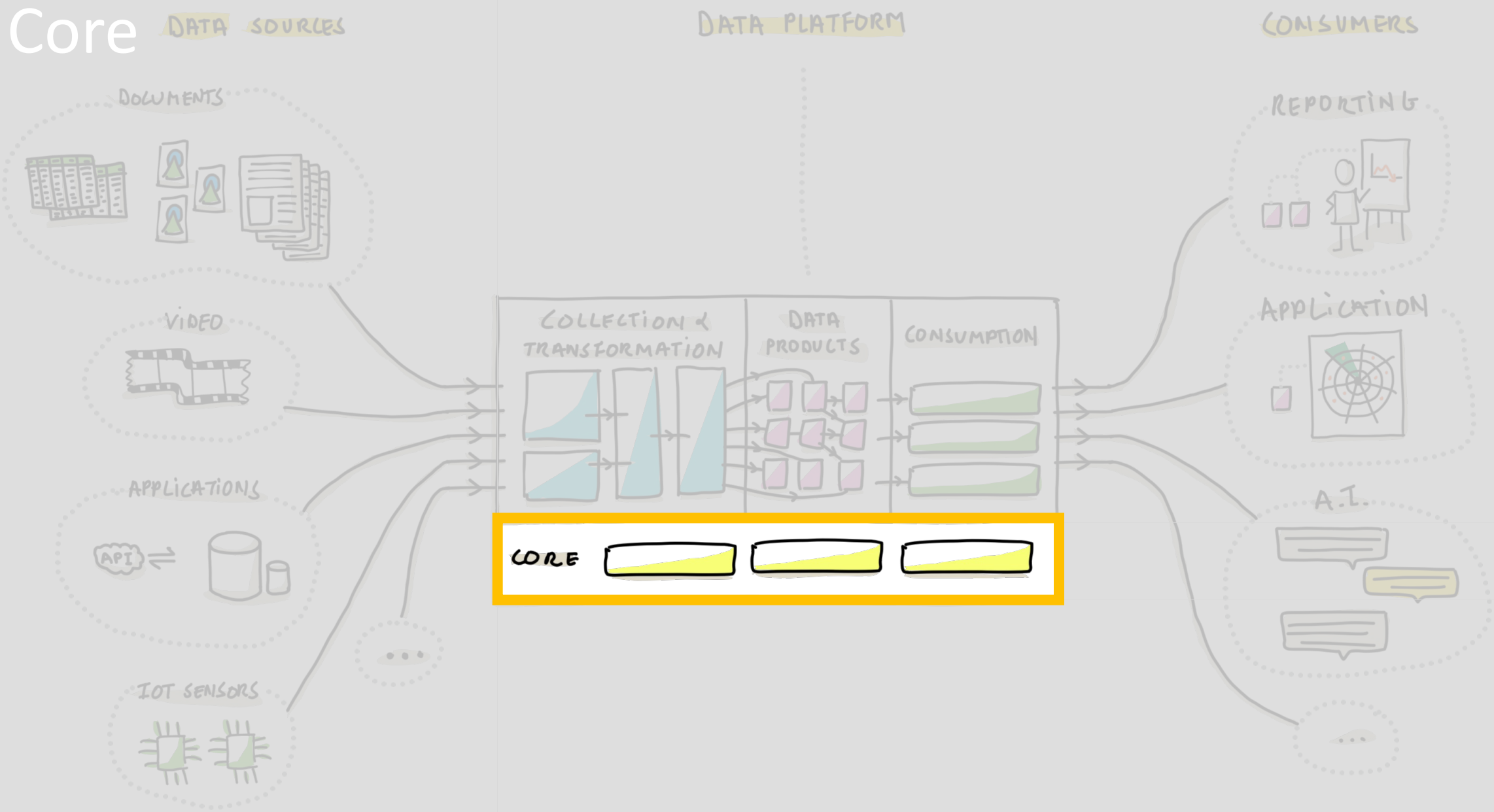
3. Consumption



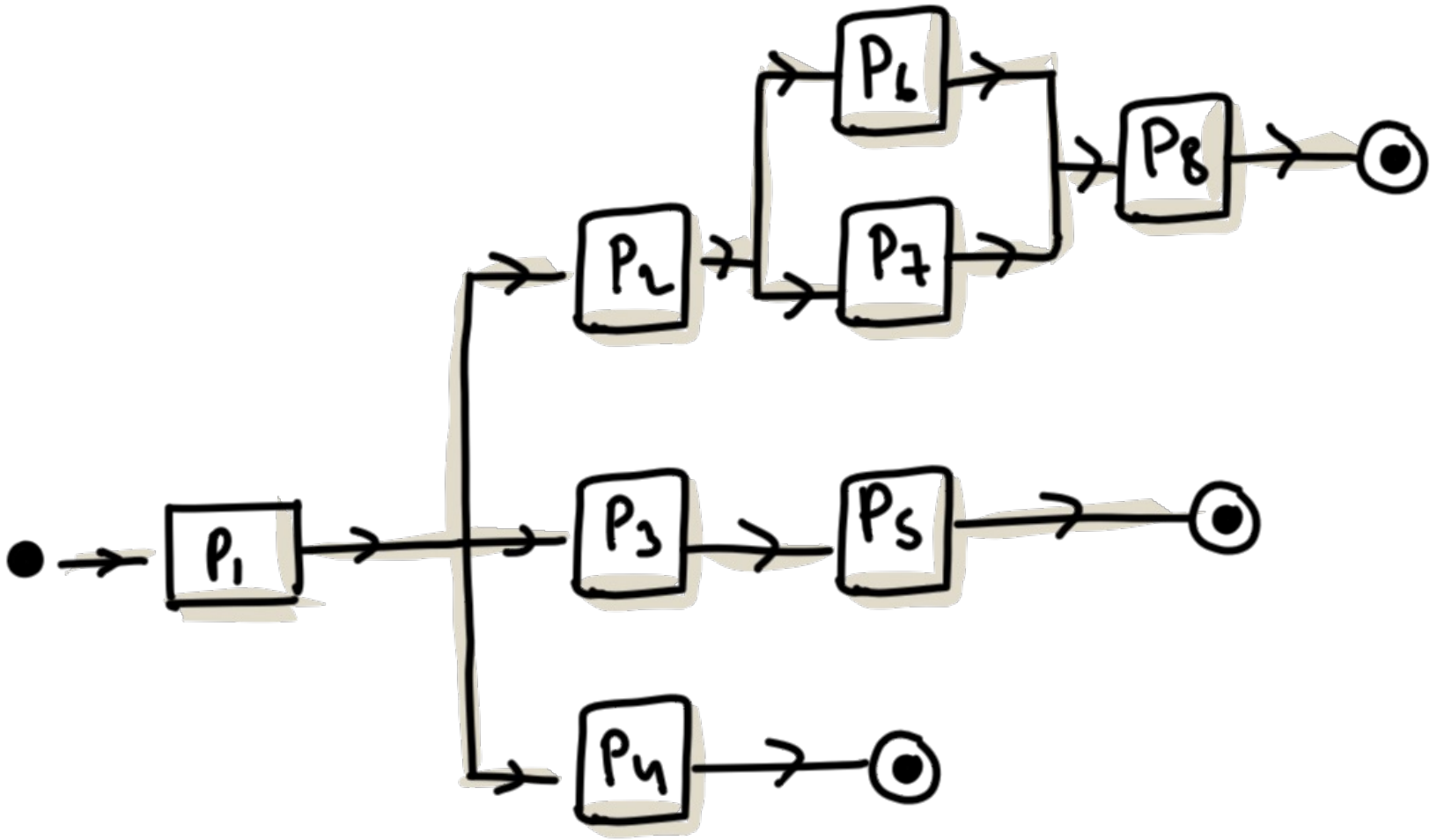
Common Consumption Patterns



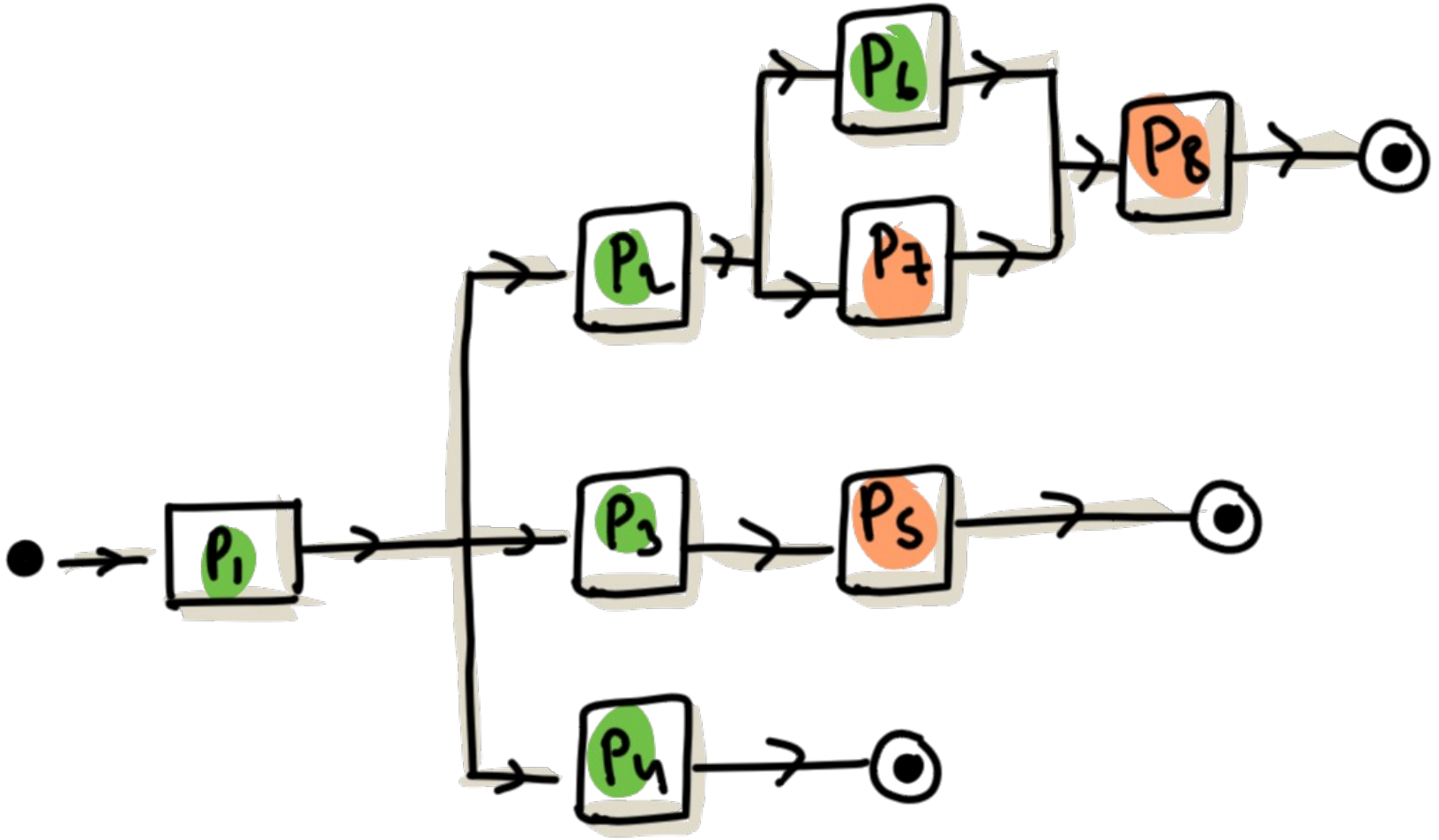
4. Core


DATA SOURCES
DATA PLATFORM
CONSUMERS
DOCUMENTS
VIDEO
APPLICATIONS
API
IOT SENSORS
COLLECTION & TRANSFORMATION
DATA PRODUCTS
CONSUMPTION
CORE
REPORTING
APPLICATION
A.I.
"Operational Plane"
"Analytical Plane"
"Operational / Analytical Plane"

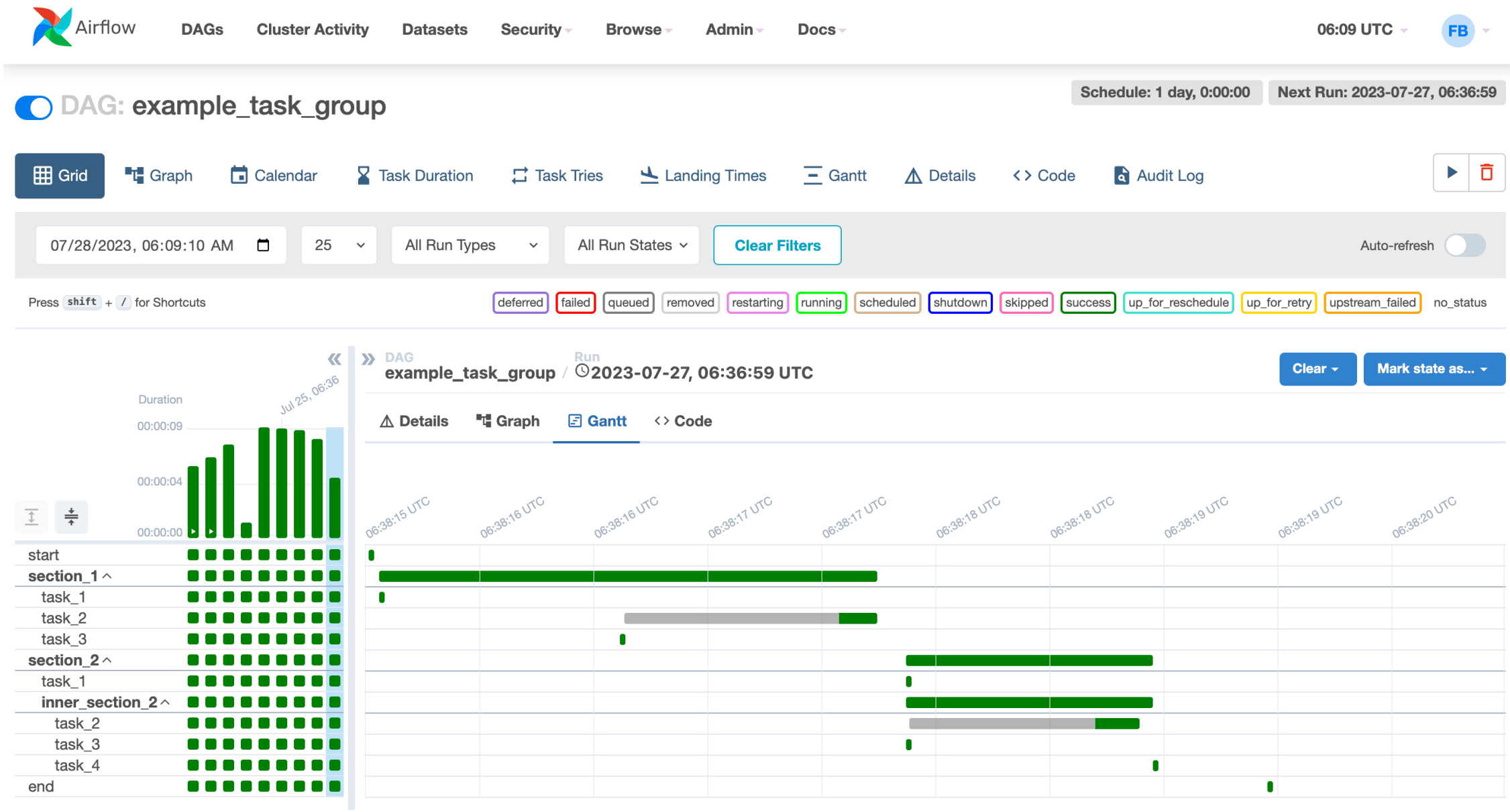
Orchestration



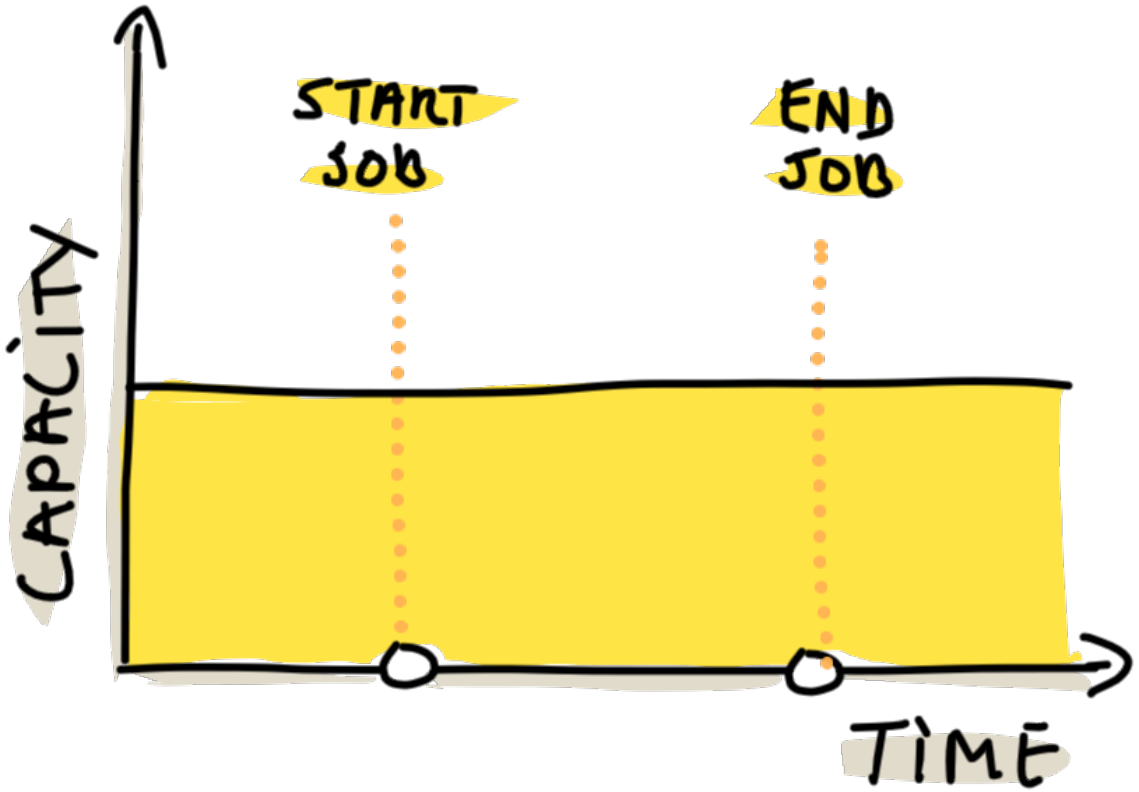
Orchestration



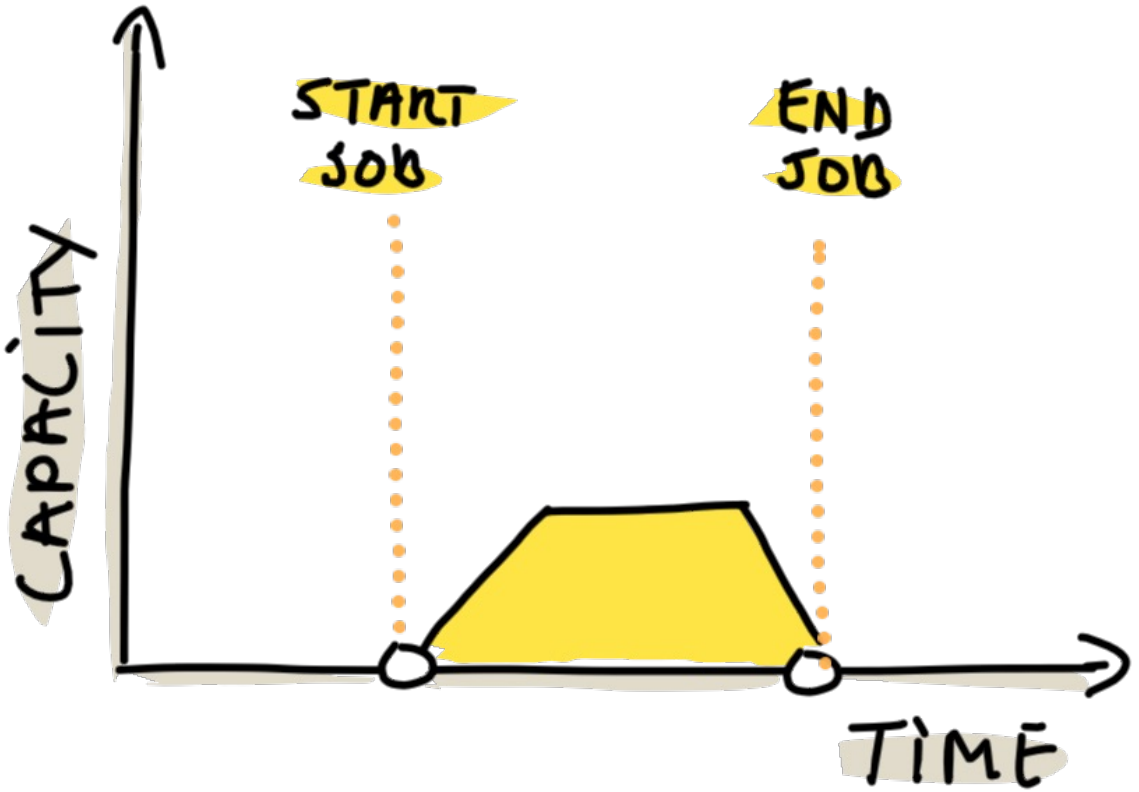
Example: Apache Airflow



Infrastructure Management (Compute)

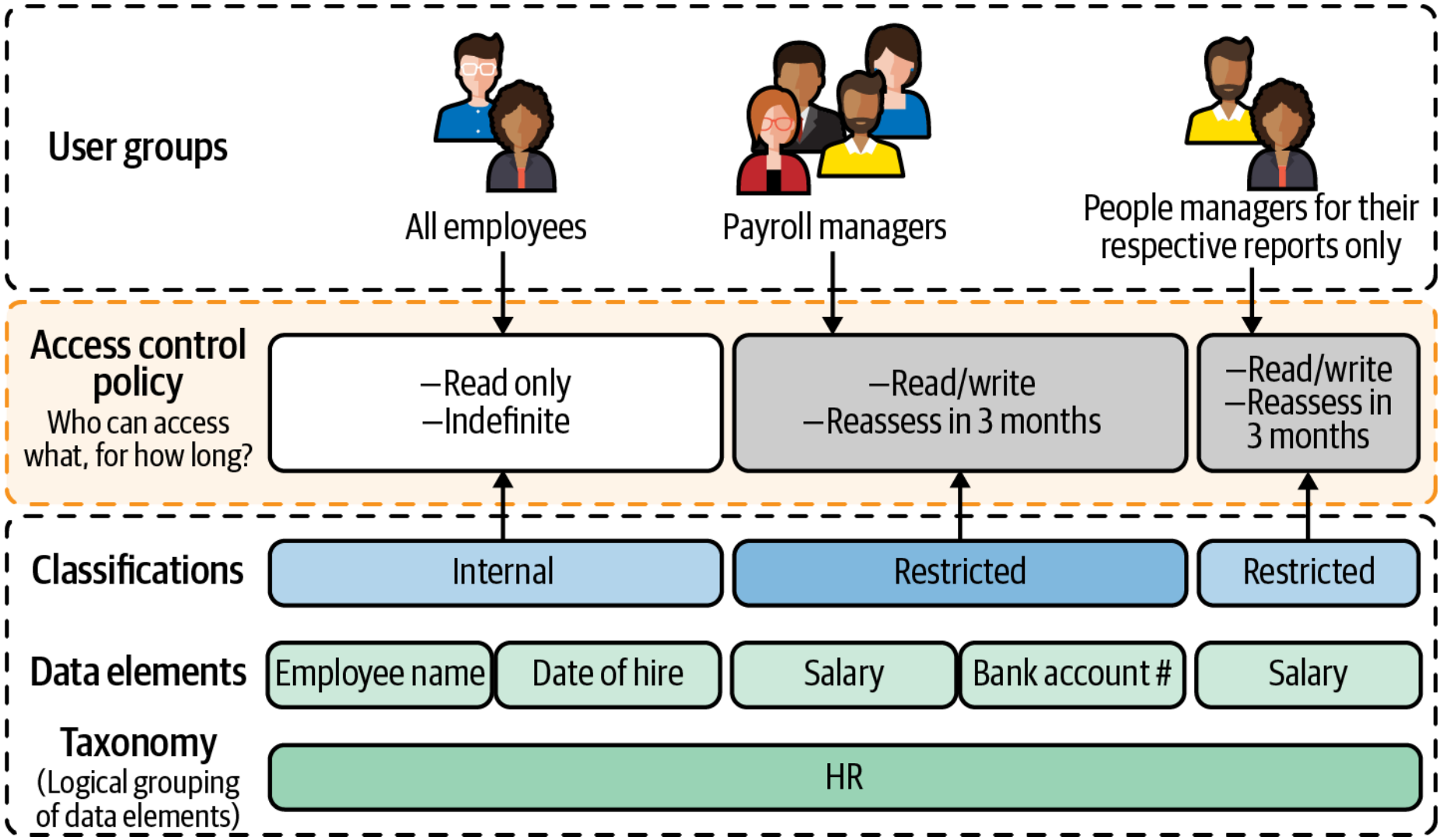


Static



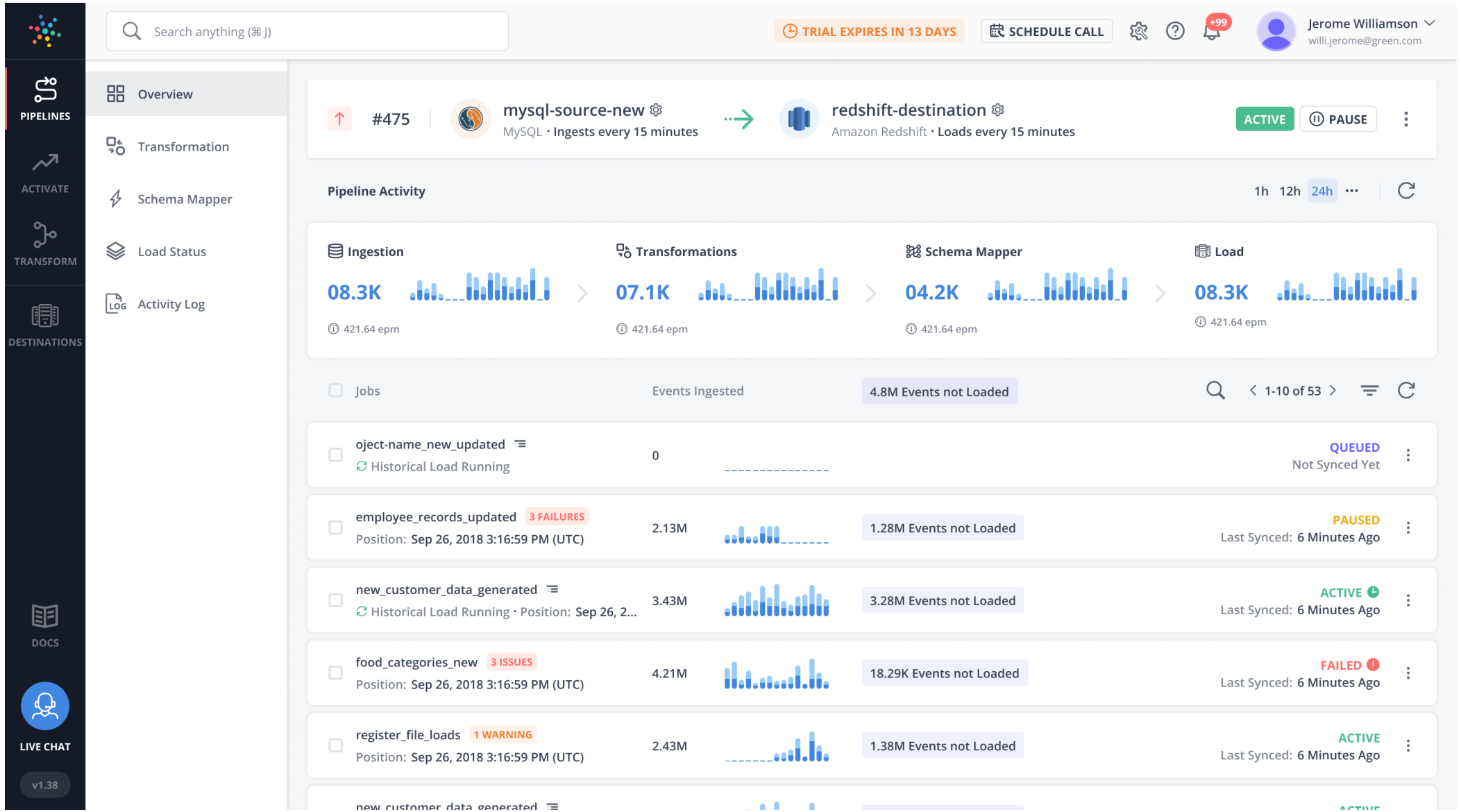
Dynamic

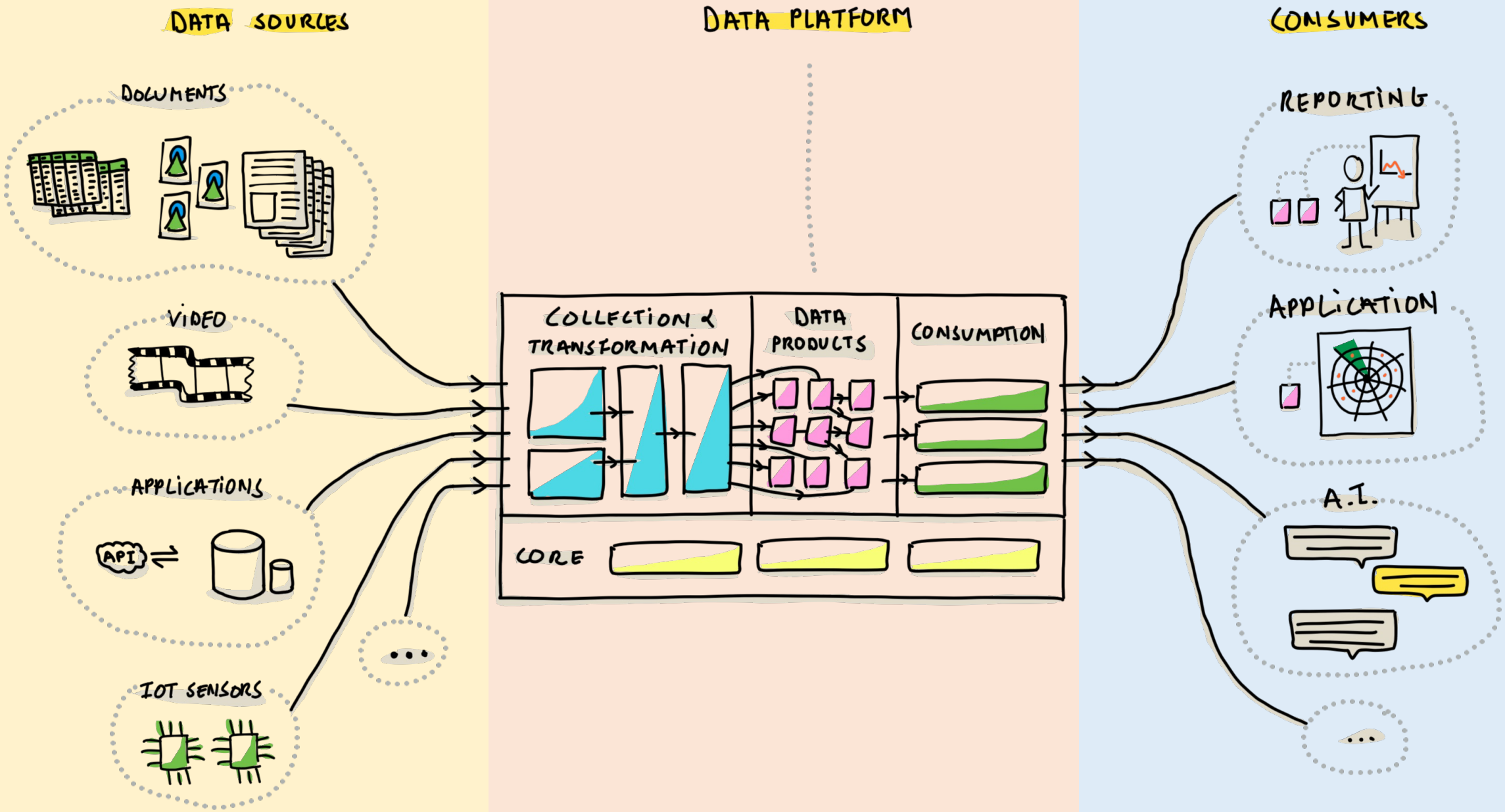
Security & Access Control



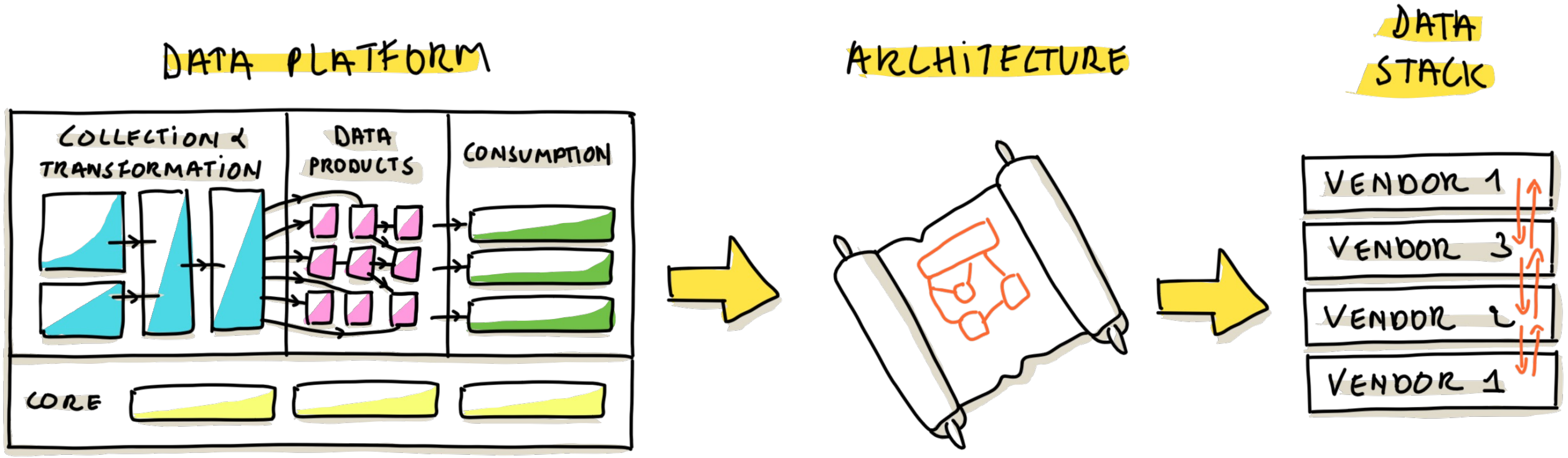
[Source]

Monitoring


Search anything (⌘ J)
TRIAL EXPIRES IN 13 DAYS
SCHEDULE CALL
+99
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
oject-name_new_updated
Historical Load Running
0
QUEUED
Not Synced Yet
employee_records_updated
3 FAILURES
Position: Sep 26, 2018 3:16:59 PM (UTC)
2.13M
1.28M Events not Loaded
PAUSED
Last Synced: 6 Minutes Ago
new_customer_data_generated
Historical Load Running · Position: Sep 26, 2...
3.43M
3.28M Events not Loaded
ACTIVE
Last Synced: 6 Minutes Ago
food_categories_new
3 ISSUES
Position: Sep 26, 2018 3:16:59 PM (UTC)
4.21M
18.29K Events not Loaded
FAILED
Last Synced: 6 Minutes Ago
register_file_loads
1 WARNING
Position: Sep 26, 2018 3:16:59 PM (UTC)
2.43M
1.38M Events not Loaded
ACTIVE
Last Synced: 6 Minutes Ago

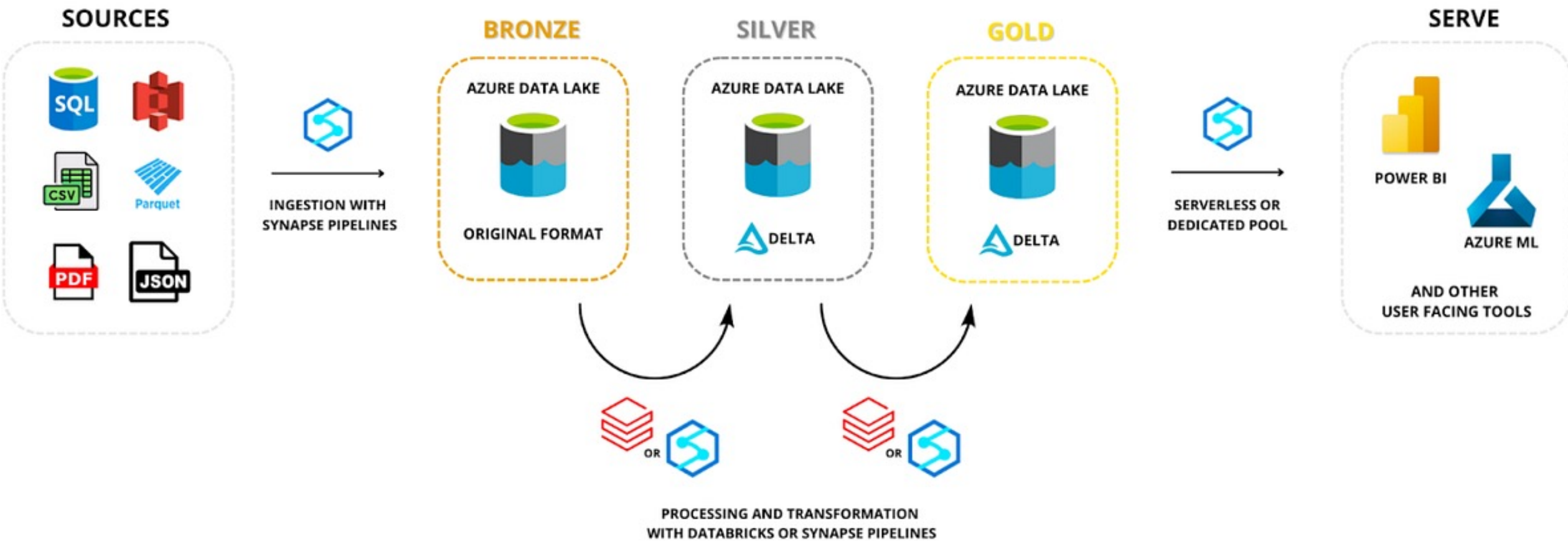

DATA SOURCES
DOCUMENTS
VIDEO
APPLICATIONS
API
IOT SENSORS
DATA PLATFORM
COLLECTION & TRANSFORMATION
DATA PRODUCTS
CONSUMPTION
CORE
CONSUMERS
REPORTING
APPLICATION
A.I.
"Operational Plane"
"Analytical Plane"
"Operational / Analytical Plane"

From Platform to Implementation



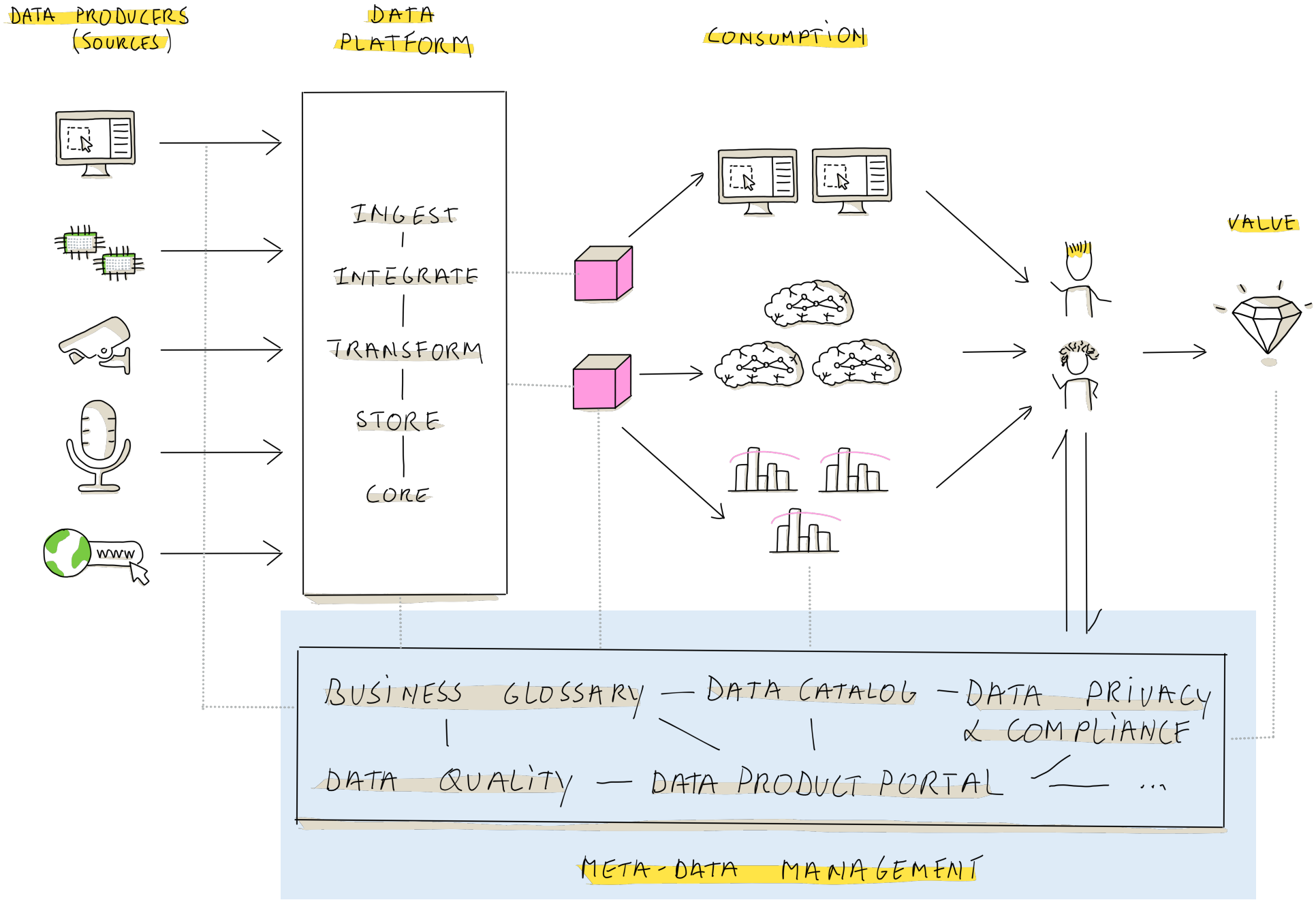
Medaillon Reference Architecture

LAKEHOUSE STORAGE LAYERS

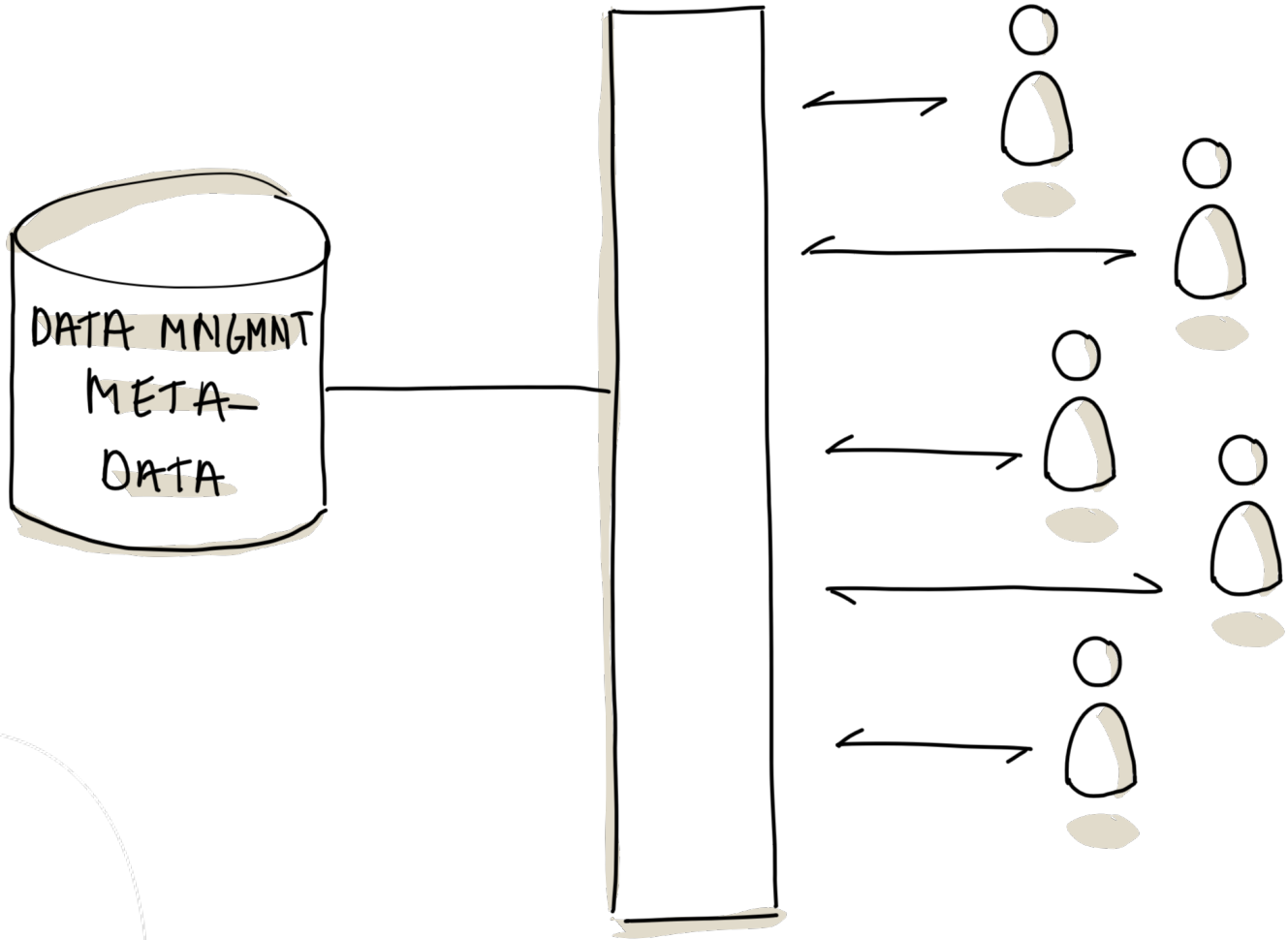


EXERCISE

- Which data-stack do you have today (very high level)?
- Which capabilities are missing for your UCs?
  - Ingesting data from certain systems?
  - Integrating data across systems?
  - Transforming data in the right shape?
  - Storing the needed amount of data?
  - …


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

META-DATA PORTAL



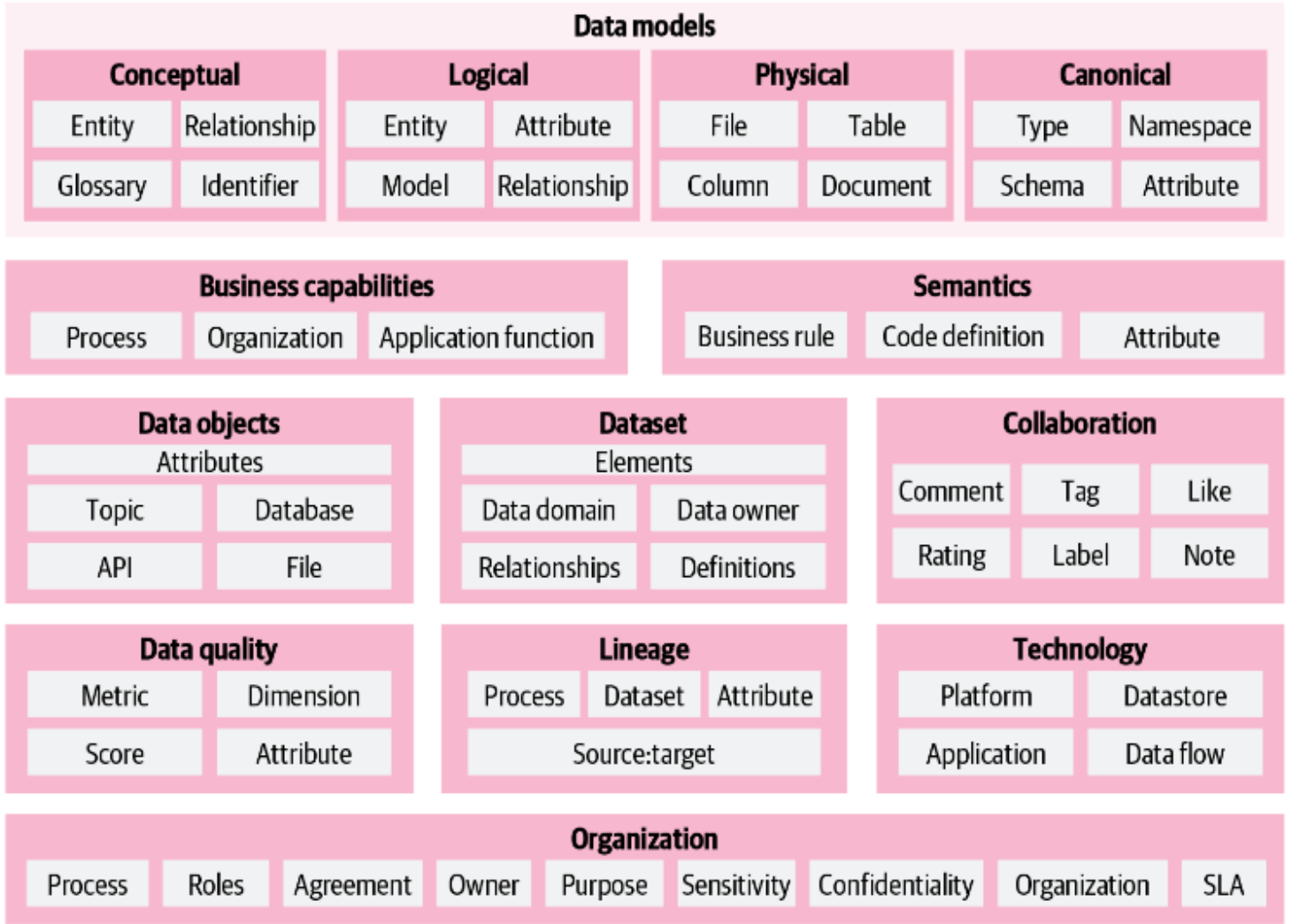


Figure 10-2. An overview showing all the major metadata management subject areas for the enterprise. These subject areas are nonexhaustive. Each organization uses the areas differently.

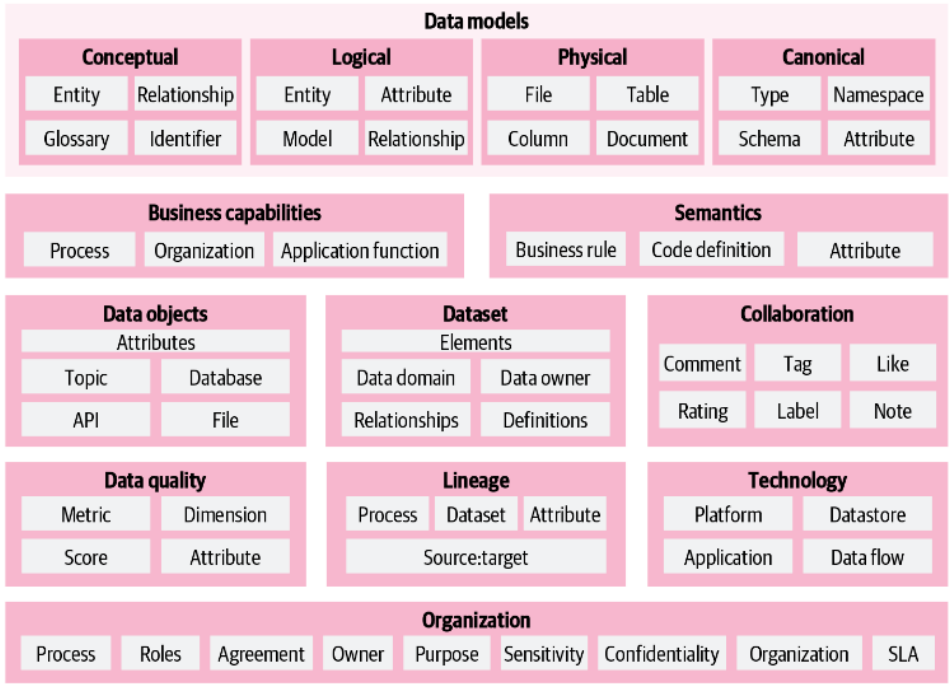
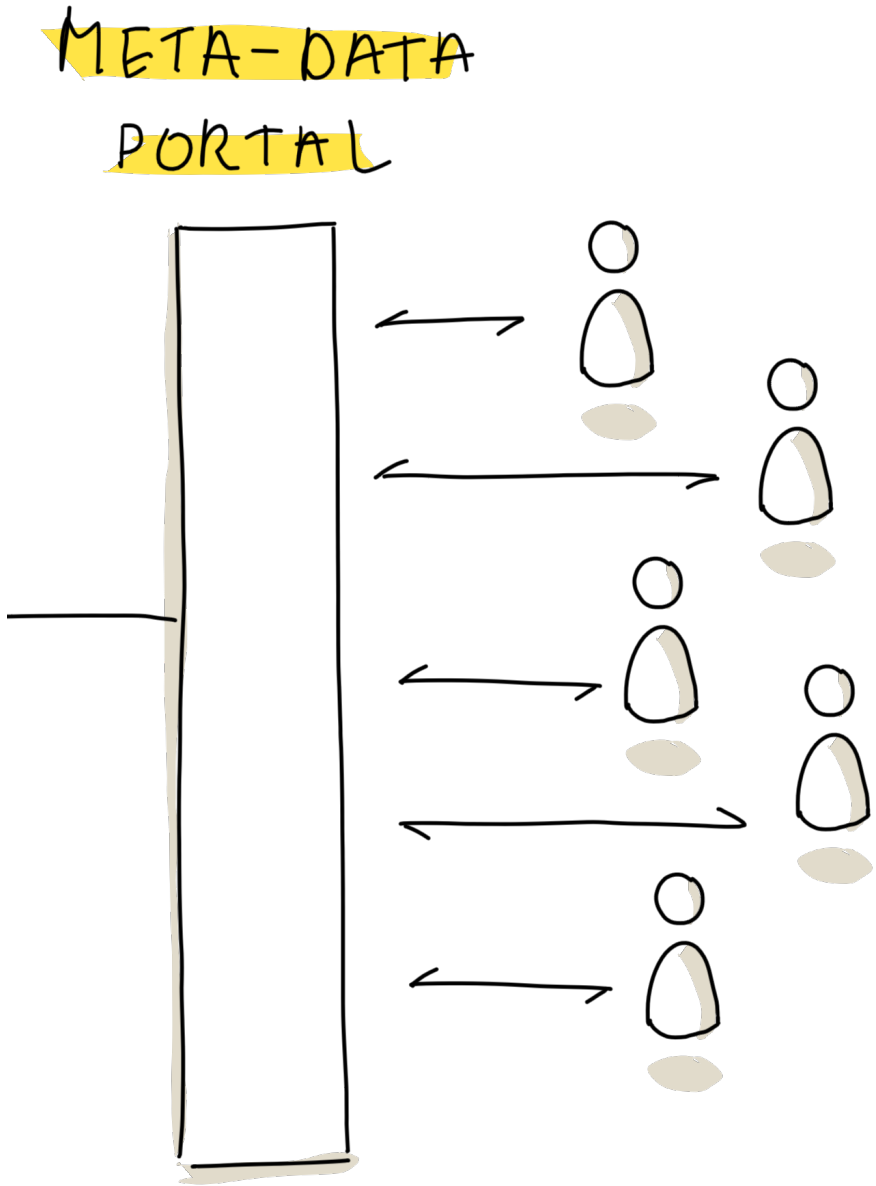
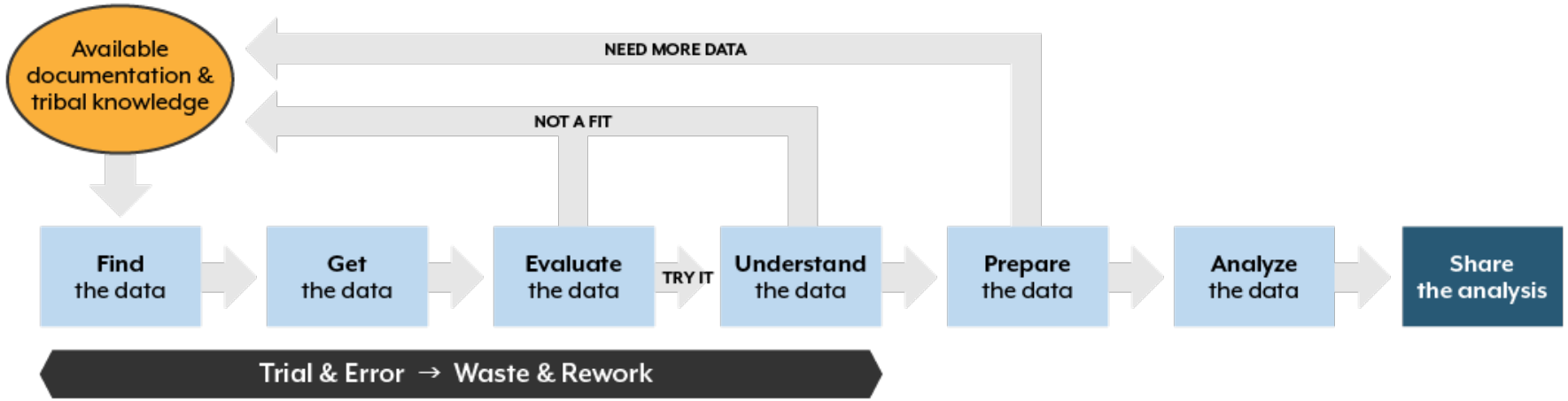


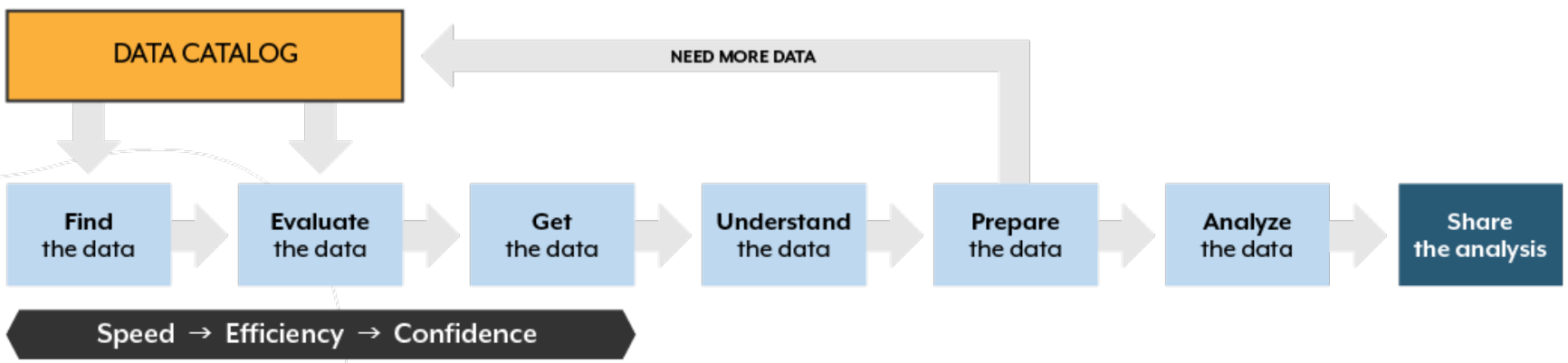
Figure 10-2. An overview showing all the major metadata management subject areas for the enterprise. These subject areas are nonexhaustive. Each organization uses the areas differently.



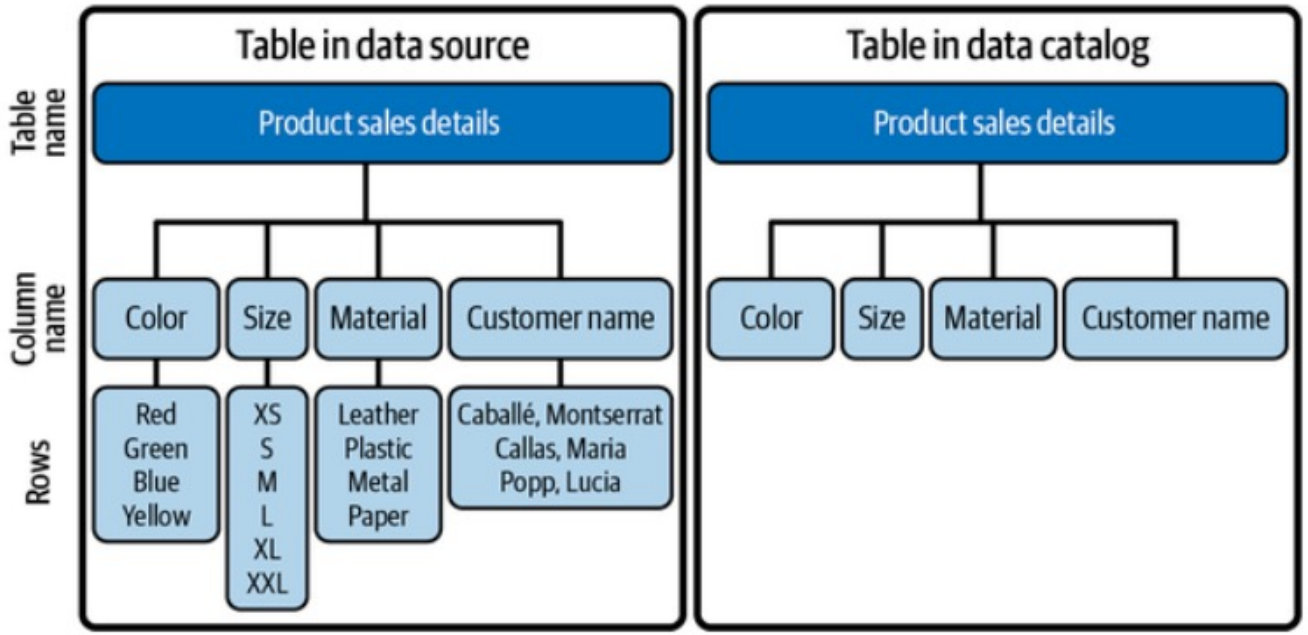
Without Data Catalog

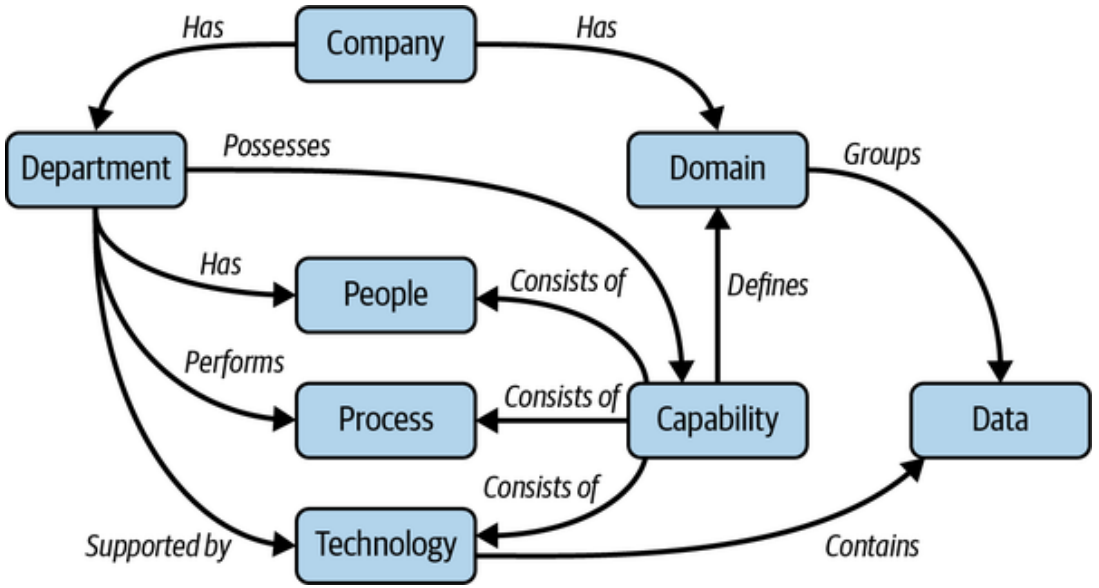


With Data Catalog

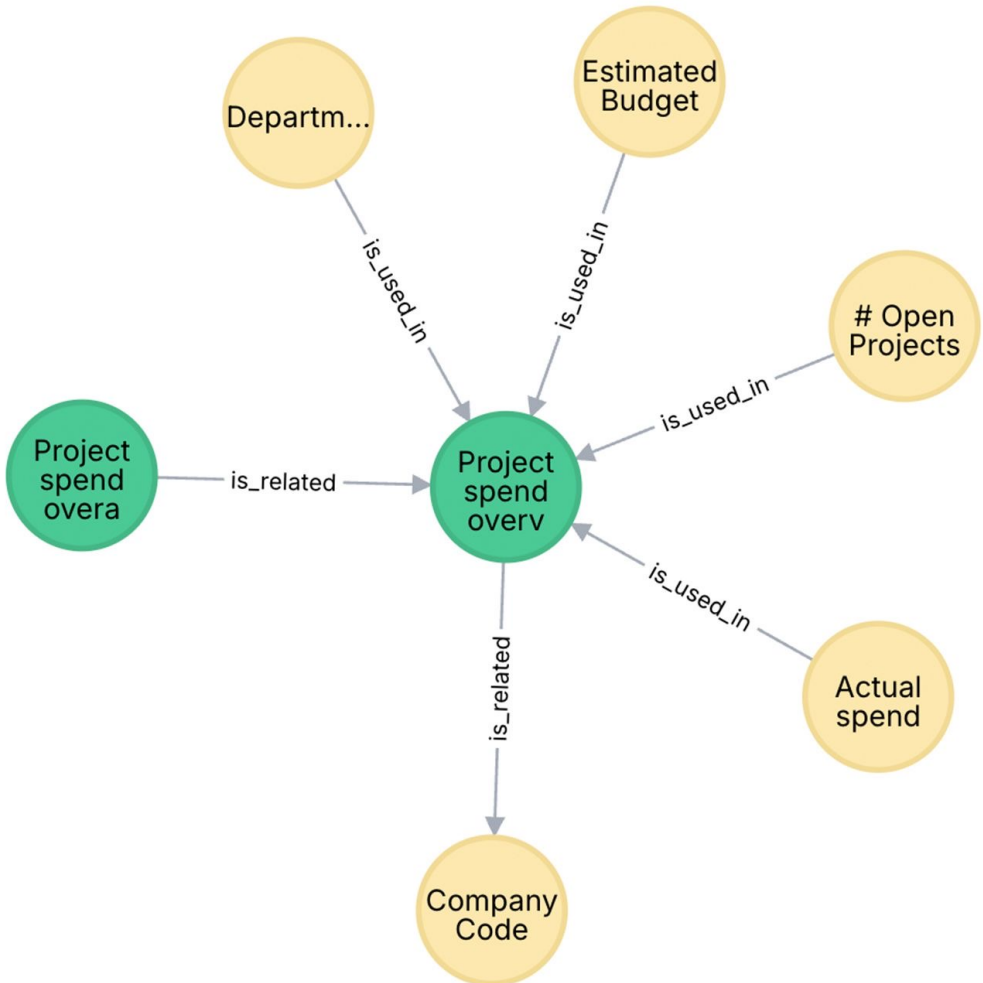


[What is a Data Catalog]





[The Enterprise Data Catalog, Ole Olesen-Bagneux]



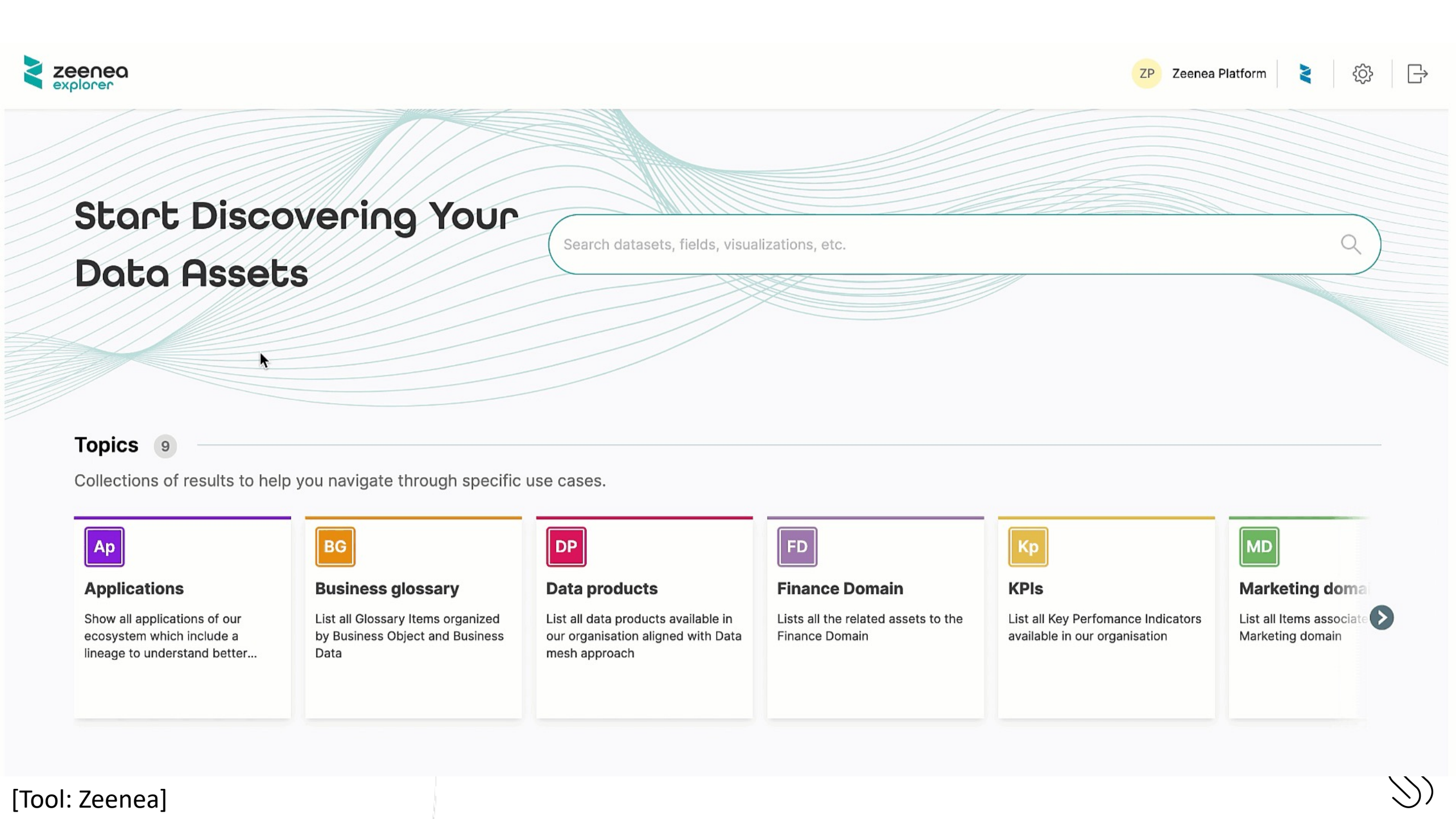
Asset Types

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

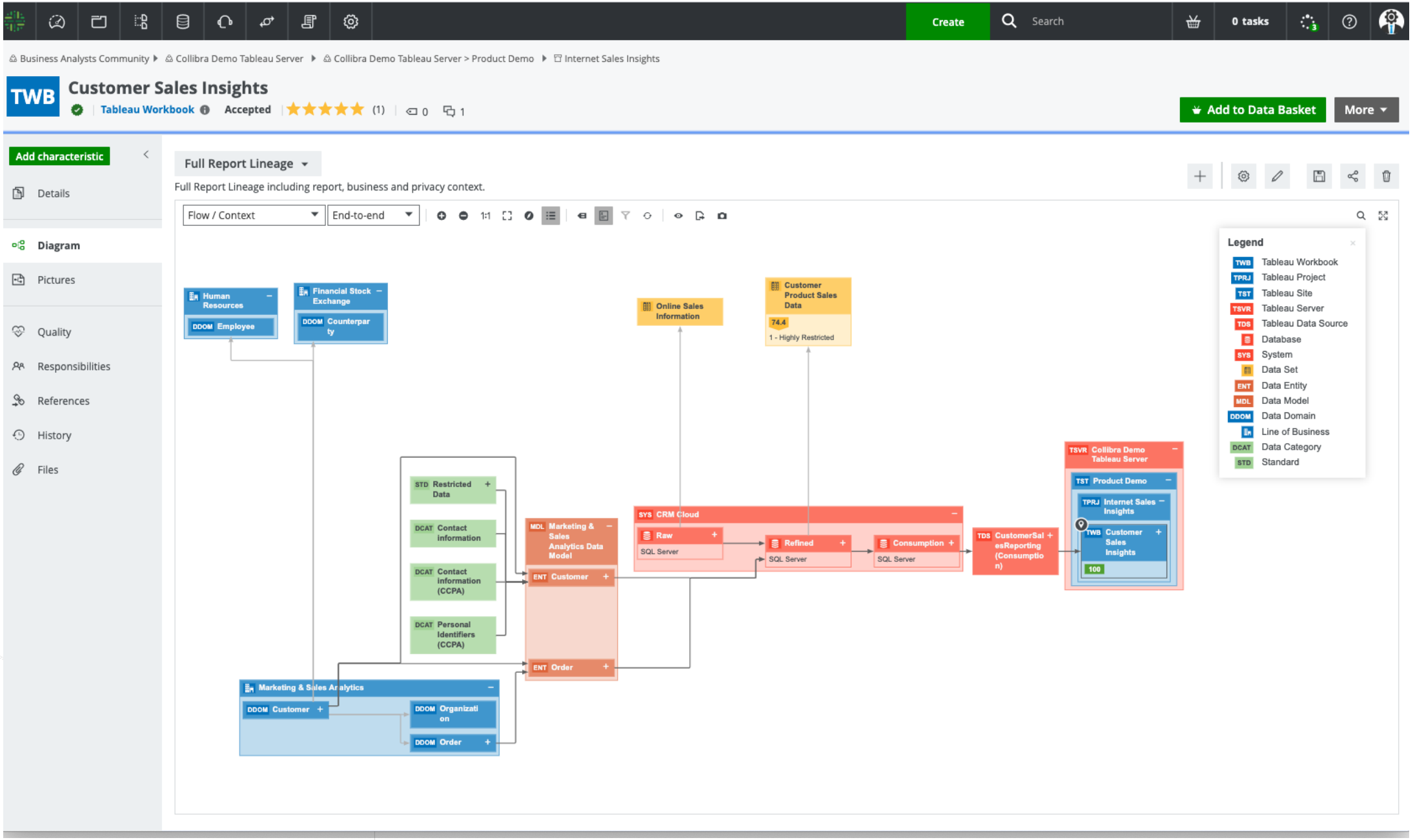
Relation Type

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

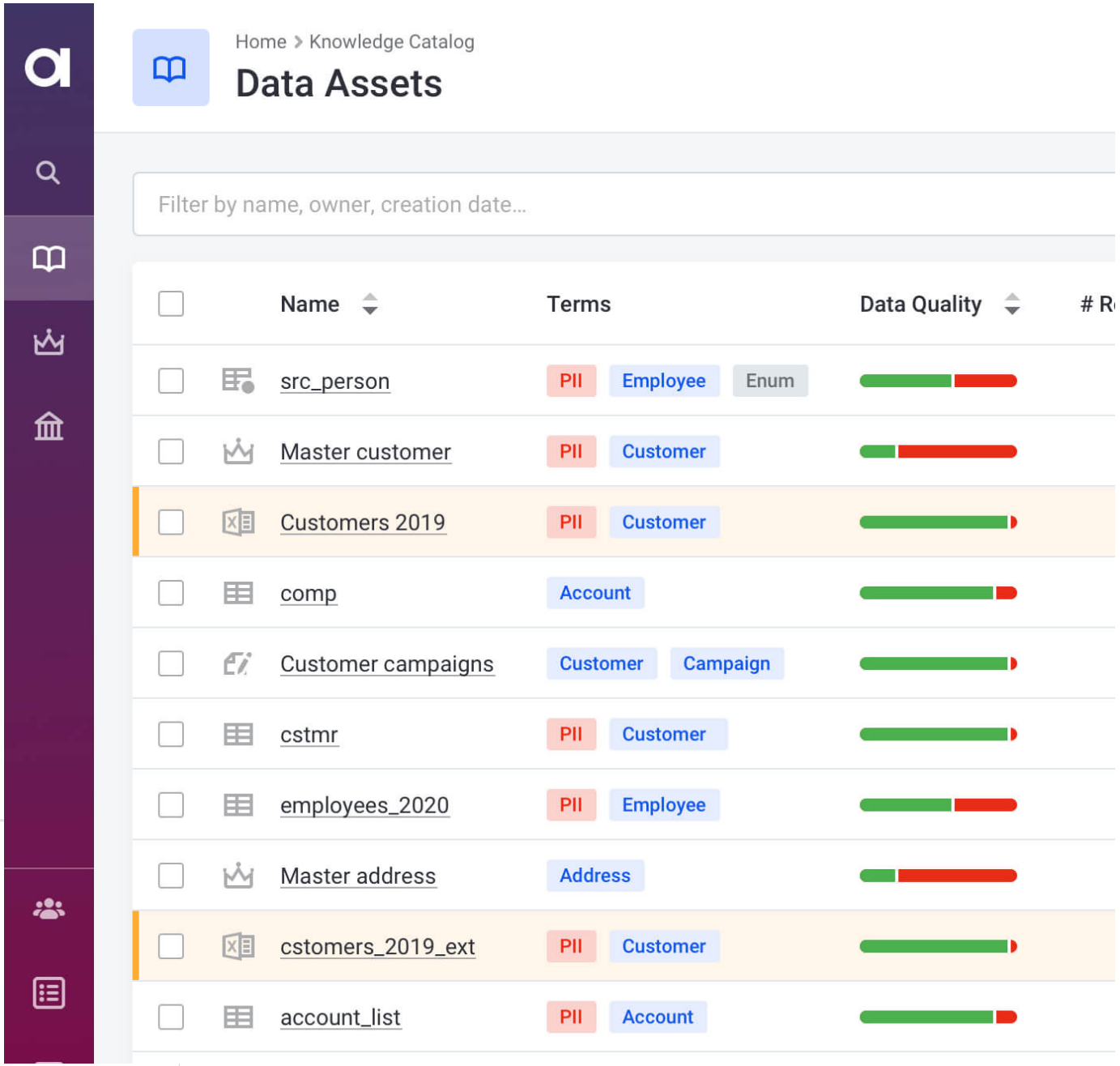
[Tool: Dscribe]


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

[Tool: Zeenea]

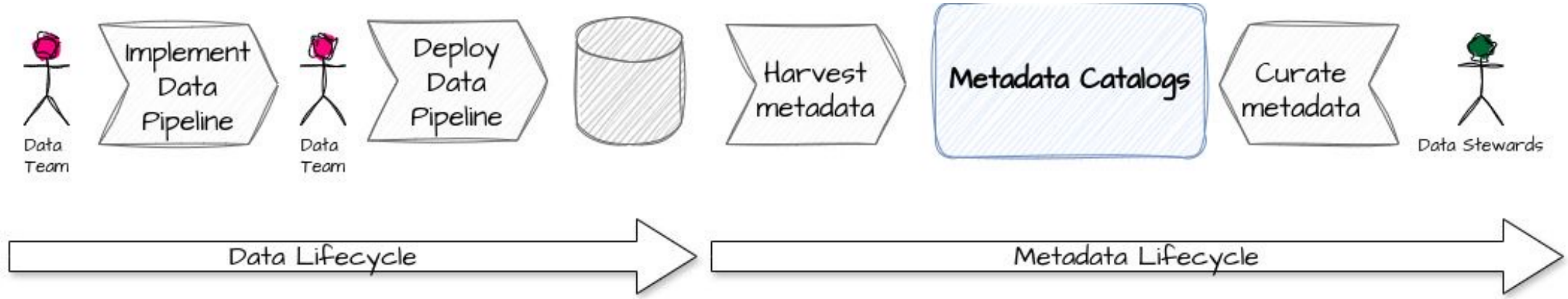


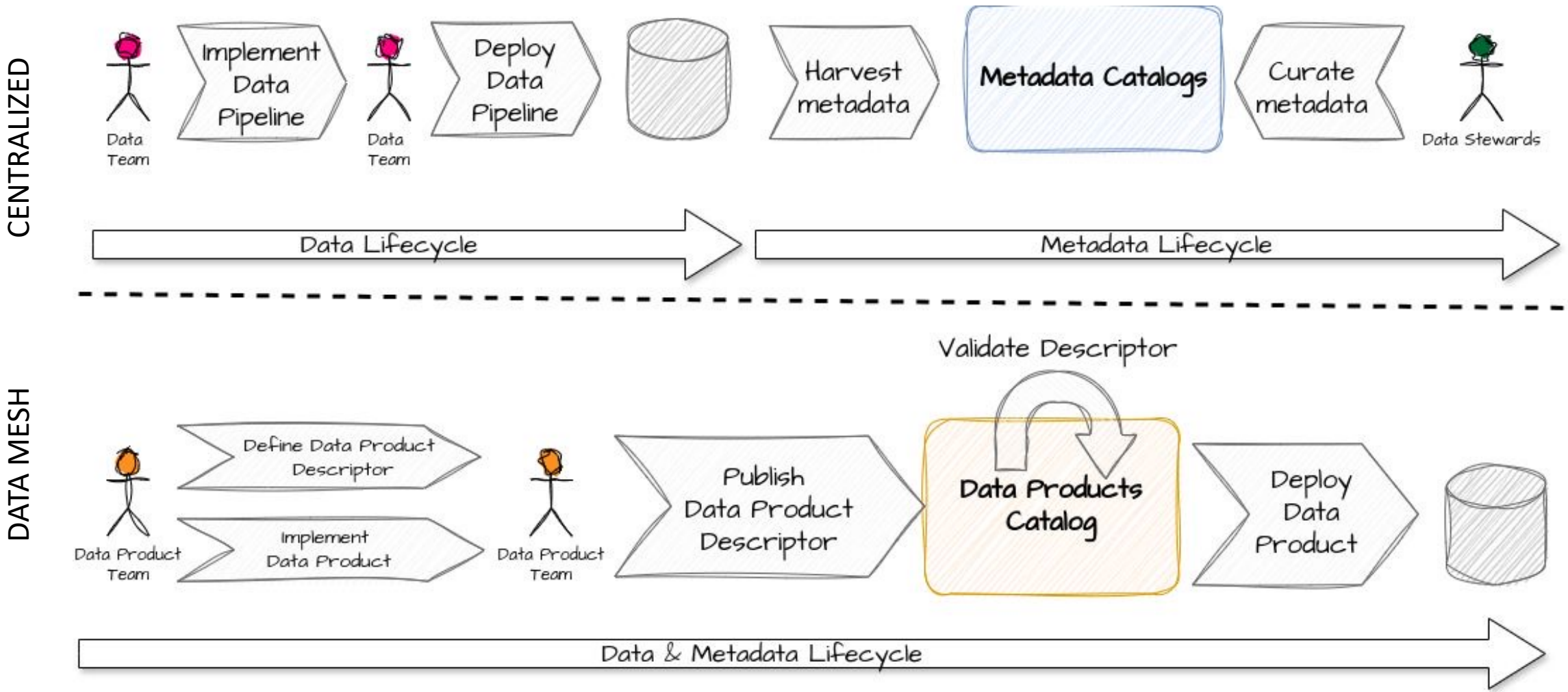
[Tool: Collibra]


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

[Alation]

CENTRALIZED





[Copyright: Andrea Gioia]

Meta-data tools : Summary

- Select a tool that connects with your data platform setup (pipelines, databases, …)
- Installation/deployment of the tool is just step 1
- It takes resources/time to:
  - Discuss definitions
  - Agree upon these definitions
  - Integrate these in the catalog / business glossary
  - …
- You can already start to create those definitions, even before the final tool is begin selected

EXERCISE

- Do you create meta-data (models, definitions , …) about the use cases today?
- Is this meta-data being stored following a certain tooling?
- Can a meta-data tool help to realize your UCs?