EVIDEN

04 Data Gouvernance

Why data governance

Usual pitfall with data/AI projects

- Not enough Data
  - Insufficient volume to correctly learn hidden pattern inside data
  - Some technics exists to artifficially create data but it's not safe and we can miss the business objectif

- Data is biased
  - Distribution of existing data is not balanced, some population are far more present than others, difficulties to generalize and learn the global pattern

- Lack of knowledge
  - AI (technical part) is in the hands of Data Scientists while Knowledge is in the hands of domain experts

- Trust
  - Without proofs (explainability), we can't trust an AI model

- Integration with production environment
  - Lab and production are really far from each other in term of exigence and constraints

Data & ML governance

EVIDEN

Concepts

Management VS Governance

Govern

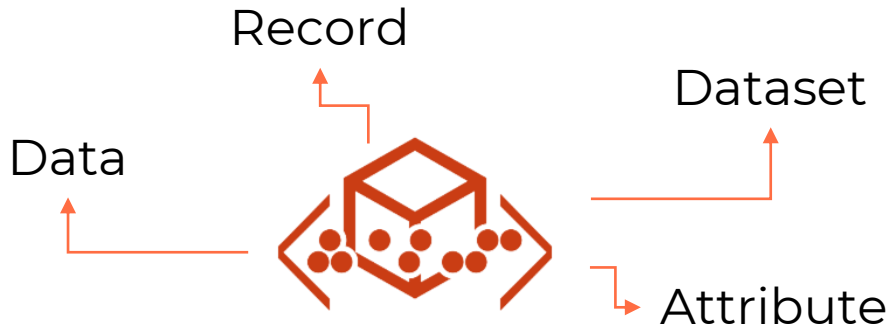
Create best practices, normalization, patterns to guide and orient peoples and technologies

Manage

Organize (people) to deliver/operate a service. Strategies are based on rules defined by the governance.

Define rules that must be applied ⇔ **Apply, Enforce rules**

Concepts
Data, a blured definition



- Data (= information = data)
  - A concept manipulated by a group of people in their daily work consisting of one or more attributes
- Attribute or Characteristic
  - Unitary element constituting a piece of data, defined by a name and a format and contains at most one value
- Record
  - Instance of a data, represented by a set of values for each attribute corresponding to the definition of the data (value that can be null)
- Dataset
  - Group of records of the same nature (linked to the same data or set of data), exhaustive or not (sample).
- Classification
  - Hierarchical organization of data and attributes to link them to business areas of the company (complementary to tags)
- Data Bundle
  - Grouping of several records of different nature (representing different data), revolving around (being attached to) a main record.

Exemplary ? **Version ?** **Transactional ?** **Exploratory ?** **Population ?**

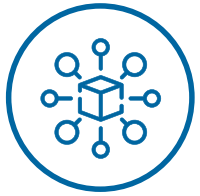
EVIDEN

4

Concepts

More definitions

REFERENCE DATA Cross-company information which structures the other information (ISO COUNTRY, etc.). These data can be external and not managed by the company,



MASTER DATA Business-critical data. Uniquely defines information specific to the company and shared within the services and therefore the IS (Third Party, Finished Product, etc.).

META DATA Additional information over the data. Groups the qualifiers/tags on each Master Data and on each attribute of these (text, integer, Date, ...), DataOwner, ...

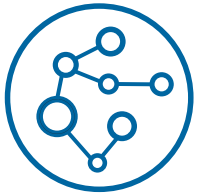
GOLDEN COPY Truth point. Currently confirmed, valid, and trusted version of a record. Consolidated vision.

DATA MODEL Set of several Master Data linked (or not) together with Relations.

DATA STEWARD Administrator, guarantor of the quality of one or more Master Data. It is the key player in Governance, a strong link between business and IT.

DATA OWNER Responsible for one or more Master Data by defining access and uses. Guarantor of the definition of the criteria of accuracy, quality, consistency, etc of these data

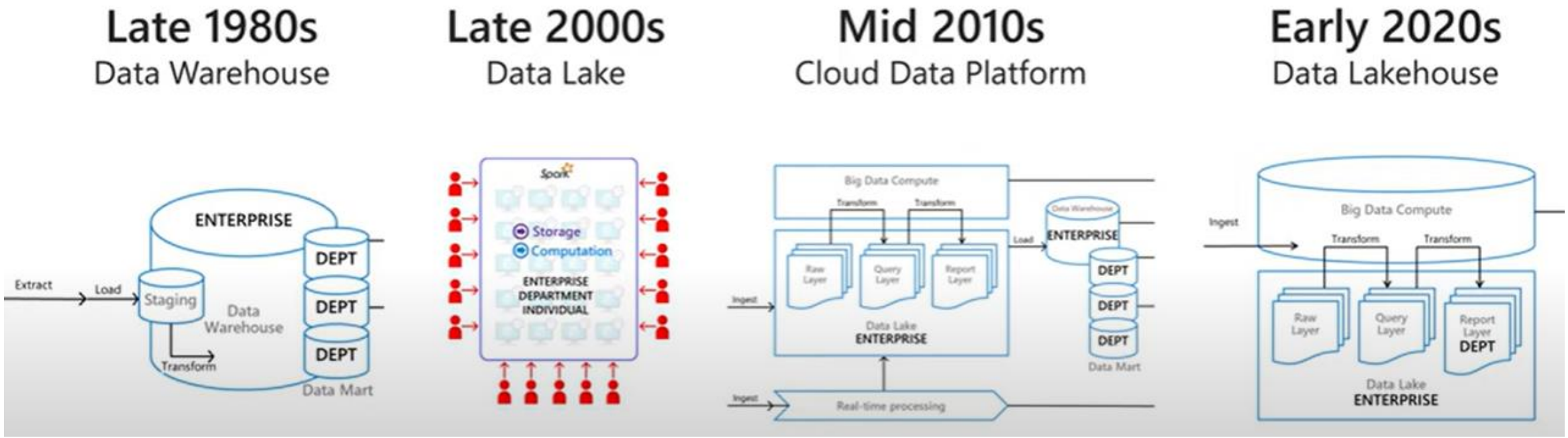
DATA LINEAGE Description of the life cycle of a data through the mapping of information flows (origin, transformation, target).



EVIDEN

From datalake to mesh

The beginning



Data Warehouse : structured storage that concentrate all the data of the enterprise, used for analytical purposes (reports, dashboards)
Issue : Hard to scaleup

Data Mart : structured storage oriented for a specific use (performance, filtering, renormalization, etc)

Data Lake : Huge amount of unstructured (and structured) storage with a scalable compute power and a centralized point for data analysis.
Issue : slow (batch oriented technologies), strong coupling between storage and compute

Data Platform : Cloud offering (easy access, agile, scalable) with a complete set of data services : from the enterprise data lake to different complementary products (streaming layer, data mart, etc) with a best of breed approach
Issue : requires strong technical skills to use or operate (especially onprem)

Data Lakehouse : scalable structured processing on top of heterogenous data in a lake
Issue : still a centralized approach with potential bottleneck on central data engineering team

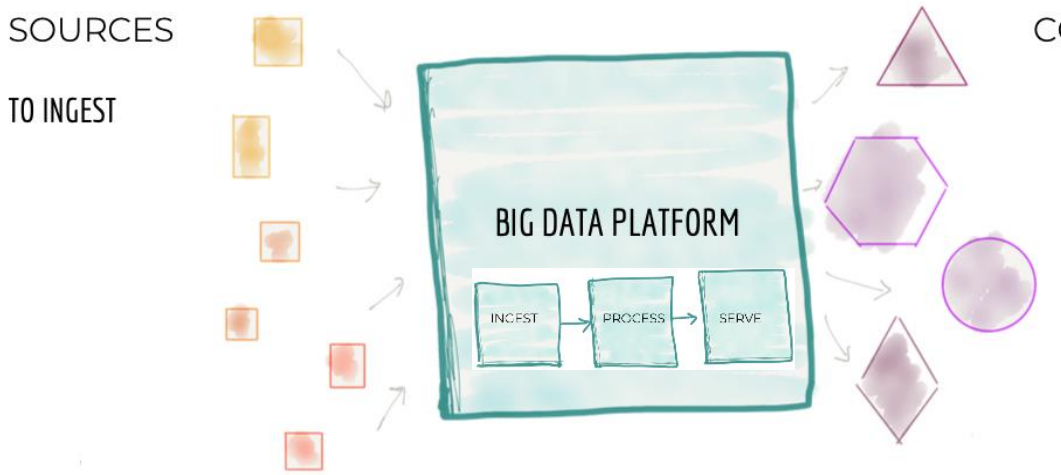
Actual Data Architecture

Monolith datalake, the limits

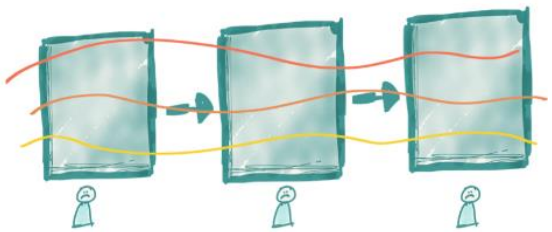
Objectives
- Ingest data from several sources
- Cleanse and enrich data
- Expose to various heterogeneous consumers

Fail 1 : Monlith and centrilized approach
Pressure points
- Data and sources proliferation
- Innovation with short expérimentations

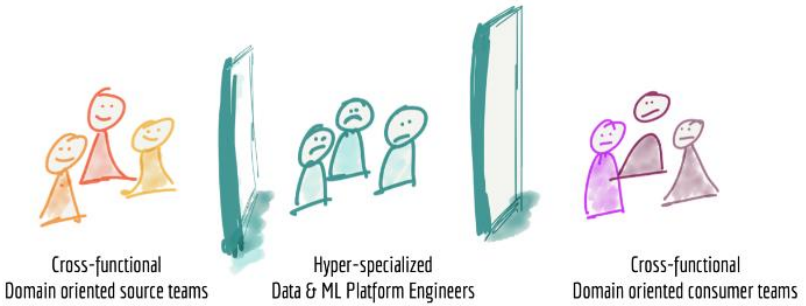
Central team is bottleneck and hard to scale efficiently



Fail 2 : Highly coupled pipeline
Highly coupled data pipelines
Delivering a new feature requires
an end2end new pipeline



Fail 3 : Siloed teams and data ownership
Central team manipulate data they don't master
Domain teams fight for priority on backlog



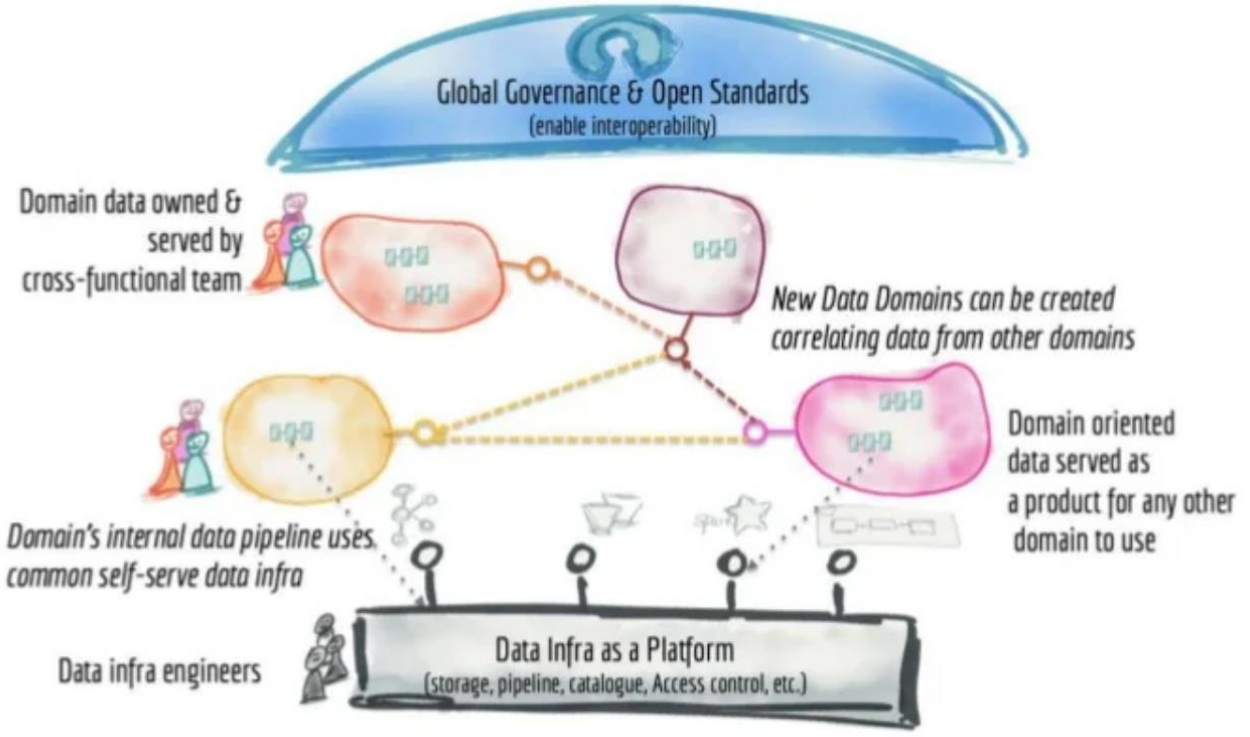
EVIDEN

Data Mesh Architecture

Concepts

Domain Driven
Business domain responsibilities should **extend to the data**

- Source oriented domains focus on **curating data**
- Consumer oriented focus on modeling and **extracting value**



Self serve platform
Every domain needs data and big data **capabilities** (storage, compute, automatization, ...)
Duplicating domain data platform will increase cost, so we should use shared infra that provide **Big Data as a Service**

Pifall
Don't build a fragmented siloed enterprise with **inaccessible data**

Data Products
Domains owner should consider their Data assets as any other valuable products (APIs, Applications, etc). **Data consumer experience** should be their main focus.

EVIDEN

Data Mesh governance

New roles and responsibilities

Data Product Owner

He's responsible for the **roadmap** and the **vision** behind every data products of his domain. He decides what data and on which form is exposed to the company. He's the domain local relay of governance best practices and norms.

He's focussed on **consumers satisfaction** and **product quality**. His main objective is to provide more and better data, well defined and documented, easy to use and with rich value. He defines KPIs to measure QoS and ensure that every engagements are respected (SLO, SLA).

This new role is created to ensure that **operational** and **data ambitions** (analytics, ML, etc) are well managed, integrated and included into the core requirements of every data product's lifecycle

Rest of the business domain team

Data Engineers

Usually DE are a centralized ressource, but now DE need to be in every domain team to develop domain data pipelines (data cleansing, processing, exposition pipelines)

Software Developers

They develop, support and integrate domain applications (business interfaces, customer websites, midlleware engines, etc)

Data Scientists

They explore, analyse, cross data to discover patterns and implement data pipelines to achieve business objectives

Infra Engineers

They instanciate and maintain a data platform for their busines domain, they bring support on industrialization of every data product during its lifecycle

Data Analysts

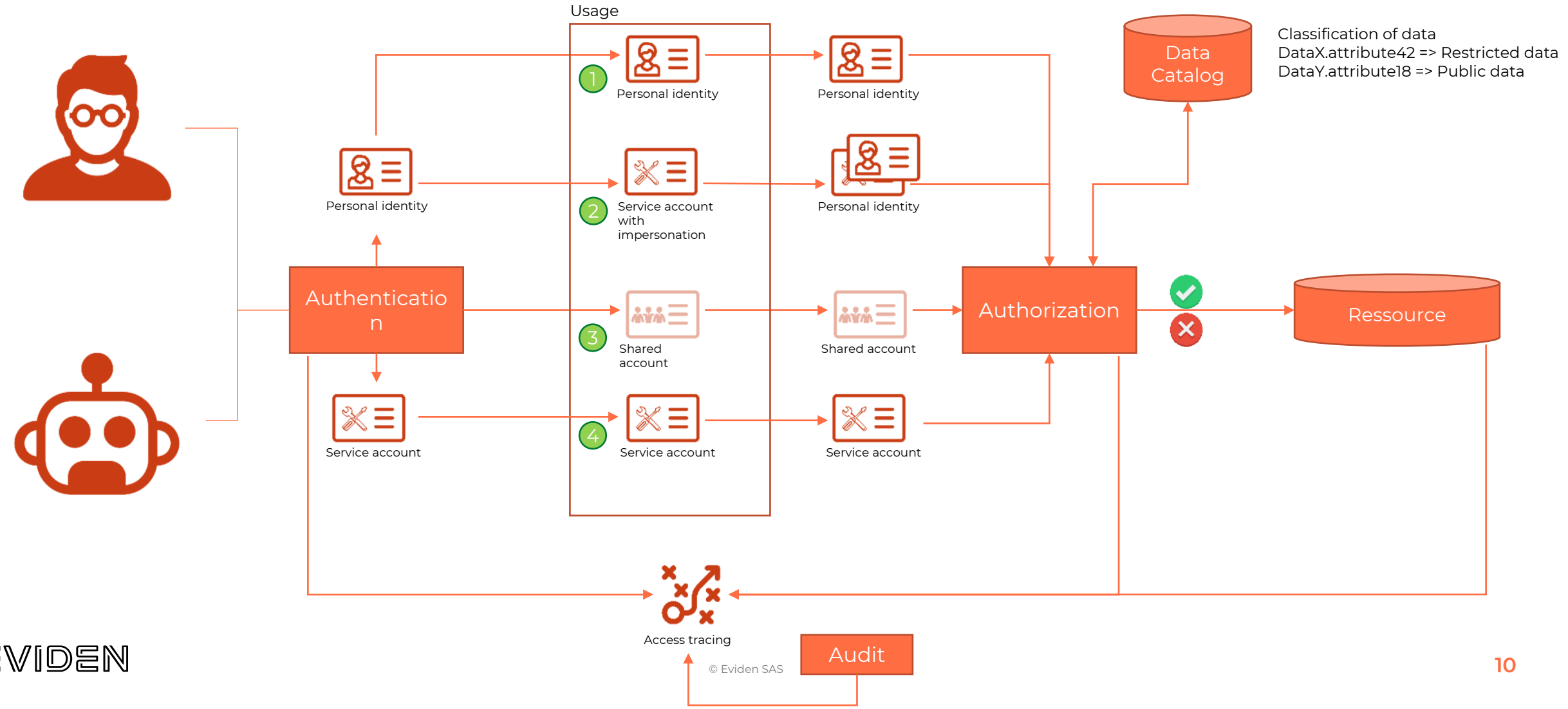
They build KPIs and produce dashboard to follow business indicators

EVIDEN

Data security

Authentication, Authorization, Audit

1. The user accesses the resources in his name (**recommended**)
2. The user goes through a service account while still transmitting his personal identity (used a lot in the Hadoop ecosystem)
3. The user does not use his identity and goes through a shared account, not recommended
4. Programs use their service account to access resources (**recommended**)



Data Organization

Data lifecycle inside dataplatform

Also known as “Medalion architecture”

Bronze **Silver**

Ingest

Staging zone used as technical interface between two systems

Majestic_entries

123654851|2018-05-24|book|Tallinn, Estonia
123736540|2018-05-24|book|Doha, Qatar
123793204|2018-05-24|audio|Nashville, USA
123835264|2018-05-24|audio|Panama City, Panama
123862351|2018-05-24|book|Nabgkok, Tailand
123965841|2018-05-24|book|Bishkek, Kyrgyzstan

CSV

Standard

Cleansed and standardized data

Majestic_clean

orderNb	date	item	itemNb	site	siteNb
123654851	2018-05-24	book	42	Tallinn, Estonia	412
123736540	2018-05-24	book	42	Doha, Qatar	155
123793204	2018-05-24	audio	15	Nashville, USA	632
123835264	2018-05-24	audio	15	Panama City, Panama	540
123862351	2018-05-24	book	42	Nabgkok, Tailand	325
123965841	2018-05-24	book	42	Bishkek, Kyrgyzstan	611

Parquet OU

Refined

Data structured in Business Objects (BOM) ready to be used in any use case

Company_Item_sales

orderNb	date	item	itemNb	site	siteNb	Price	Margin %Discount
123654851	2018-05-24	book	42	Tallinn, Estonia	412	42	30 15
123736540	2018-05-24	book	42	Doha, Qatar	155	24	28 0
123793204	2018-05-24	audio	15	Nashville, USA	632	10	9 5
123835264	2018-05-24	audio	15	Panama City, Panama	540	8	7 5
123862351	2018-05-24	book	42	Nabgkok, Tailand	325	31	22 15
123965841	2018-05-24	book	42	Bishkek, Kyrgyzstan	611	48	24 15

Expose

Output zone to transfer augmented data to other systems

Books

date	item	itemNb	Price	Margin %Discount
2018-05-24	book	42	42	30 15
2018-05-24	book	42	24	28 0
2018-05-24	book	42	31	22 15
2018-05-24	book	42	48	24 15

Audio

date	item	itemNb	Price	Margin %Discount
2018-05-24	audio	15	10	9 5
2018-05-24	audio	15	8	7 5

OU **API**

Archive

Used for legal and forensic analysis

CSV

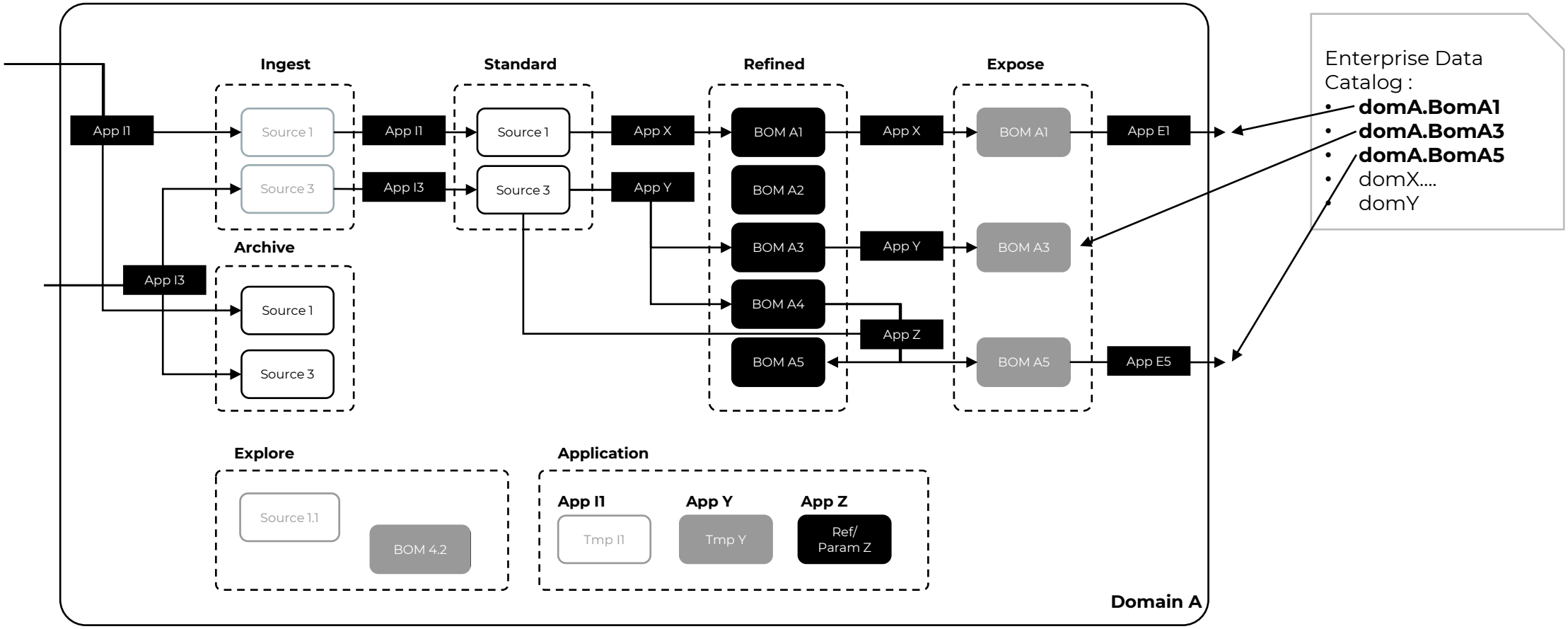
EVIDEN

Data lifecycle

Focus on a domain

Synonyms

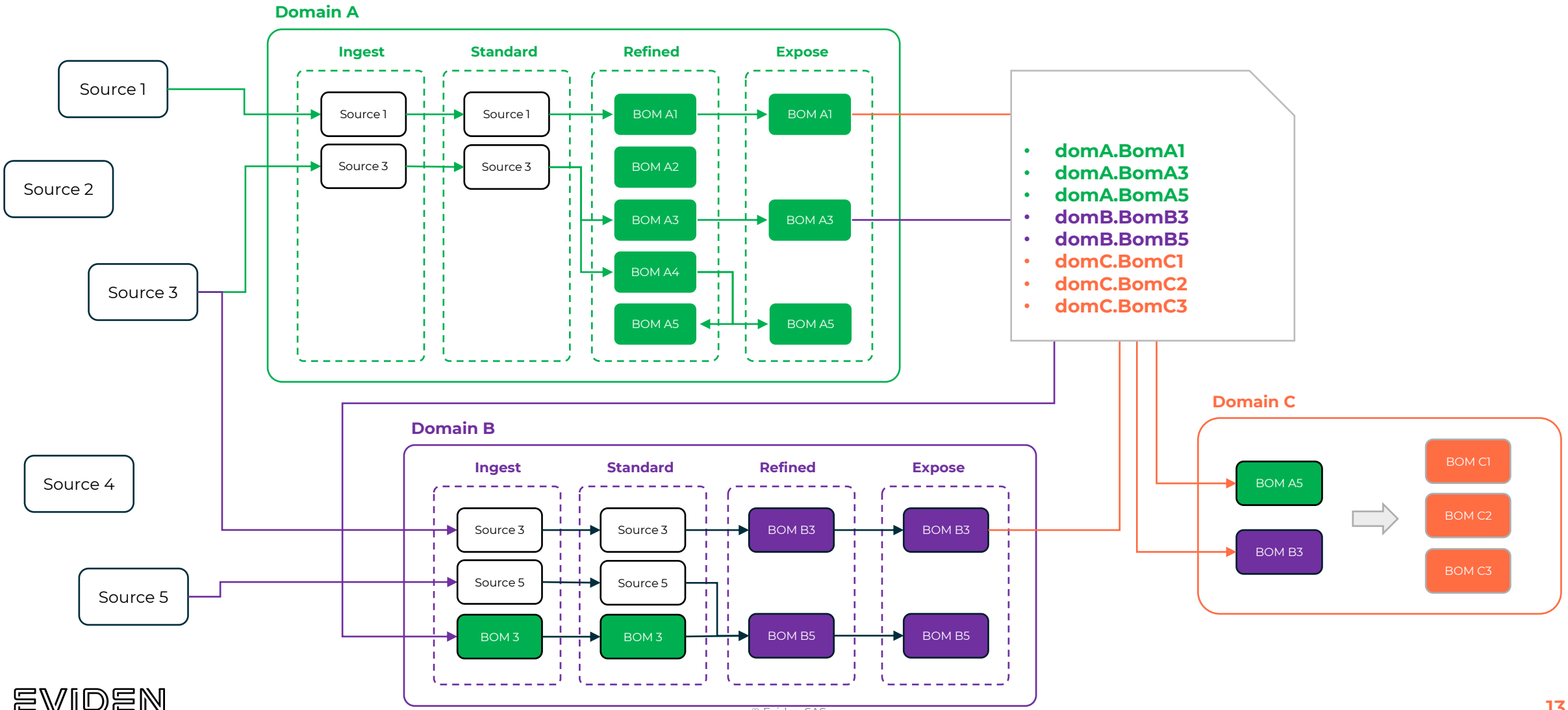
Ingest : Consume, Staging, In, Input, Raw, Bronze
Standard : Normalized, Silver
Refined : Business, Golden
Expose : Serve
Archive : Raw, Bronze, Cold, Backup
Explore : Lab
Application : Processing, Project, Work



Ingest zone should keep data only during ingestion process, for technical purpose only, once standardized, input data could be erased
Explore zone should keep data only during exploration project, not over long term
Application zone could have temporary data needed for application and long-term parameters or referentials
Expose zone should have pointers over refined data rather than duplicating data

Data lifecycle

Multi domains



Data Governance

Sumup

Concepts

Best practices for data and workflow organization/lifecycle.
Normalization rules, architecture patterns, security classifications, ...

Peoples

Data product owner that is responsible of the roadmap of its data products

Tools

Mesh/Data catalog that index every valuable data of the company

EVIDEN

EVIDEN

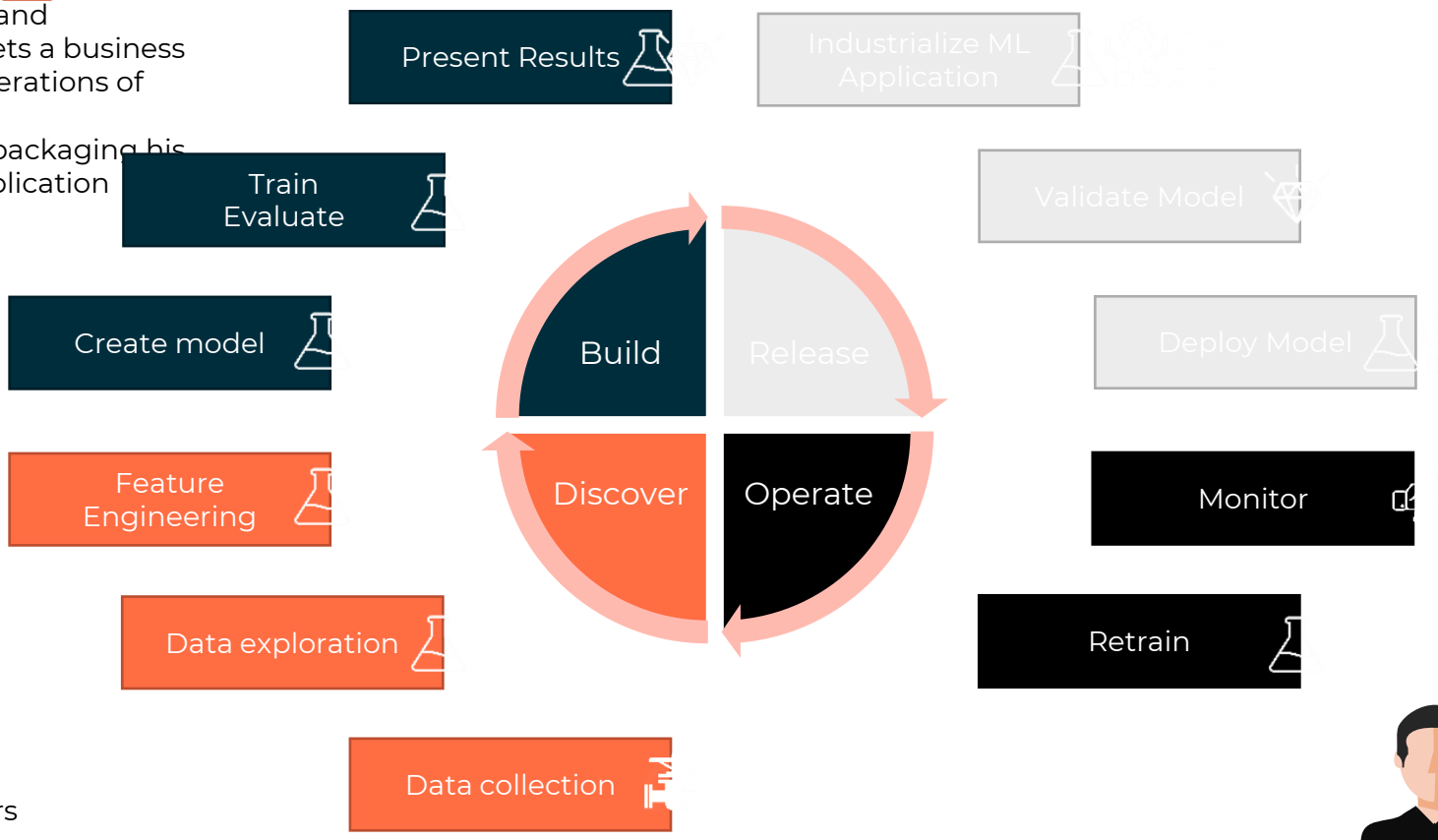
05 ML Governance

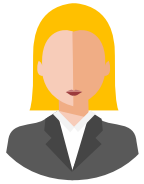
ML lifecycle and personae

Data Scientist
Focus on algorithms and mathematics, he meets a business need through agile iterations of R&D.
He is responsible for packaging his model into an ML application

ML Engineer
He helps Data Scientists with technical support and he is responsible for the industrialization of the ML application

Business owner
Based on the technical and business performance metrics of the model, he validates or not the new model.



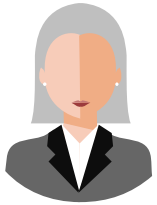


Data Engineer
He builds data pipelines to ingest the data in the platform and others pipelines to expose refined data generated by ML applications

Operationals
He leads the deployment of the new ML app and he is responsible for following technical metrics in production

ML lifecycle and personae

Support profiles



Data Product owner

He’s responsible for the roadmap and the vision behind every data products of his domain. He gives business requirements and direction to the engineers responsible to develop his product

Scrum master

Delivery profil, he’s responsible to organise and follow all the evolutions made by the agile team. He support the team gathering through the company all necessary information so that the backlog can move forward.

DPO

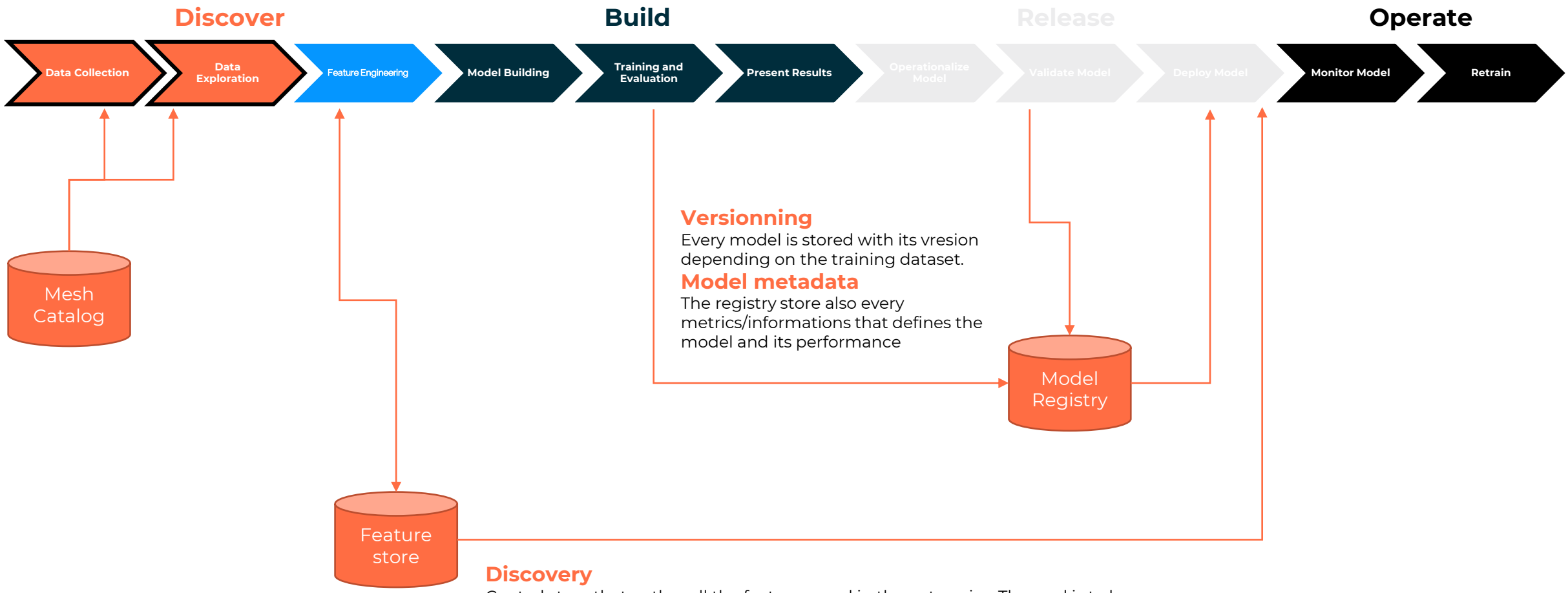
In collaboration with the project teams, he ensures that the ML application complies with the safety and ethical rules of the company and the country.

Data Architect

It ensures the consistency of data processing, services and applications within the company's specific context. It can also support the company's data strategy.

ML Governance

Governance tooling



Versionning
Every model is stored with its vresion depending on the training dataset.

Model metadata
The registry store also every metrics/informations that defines the model and its performance

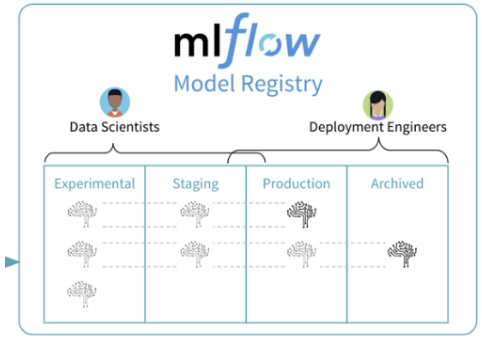
Discovery
Central store that gather all the features used in the enterprise. The goal is to have an exact same way to define features over multiple projects

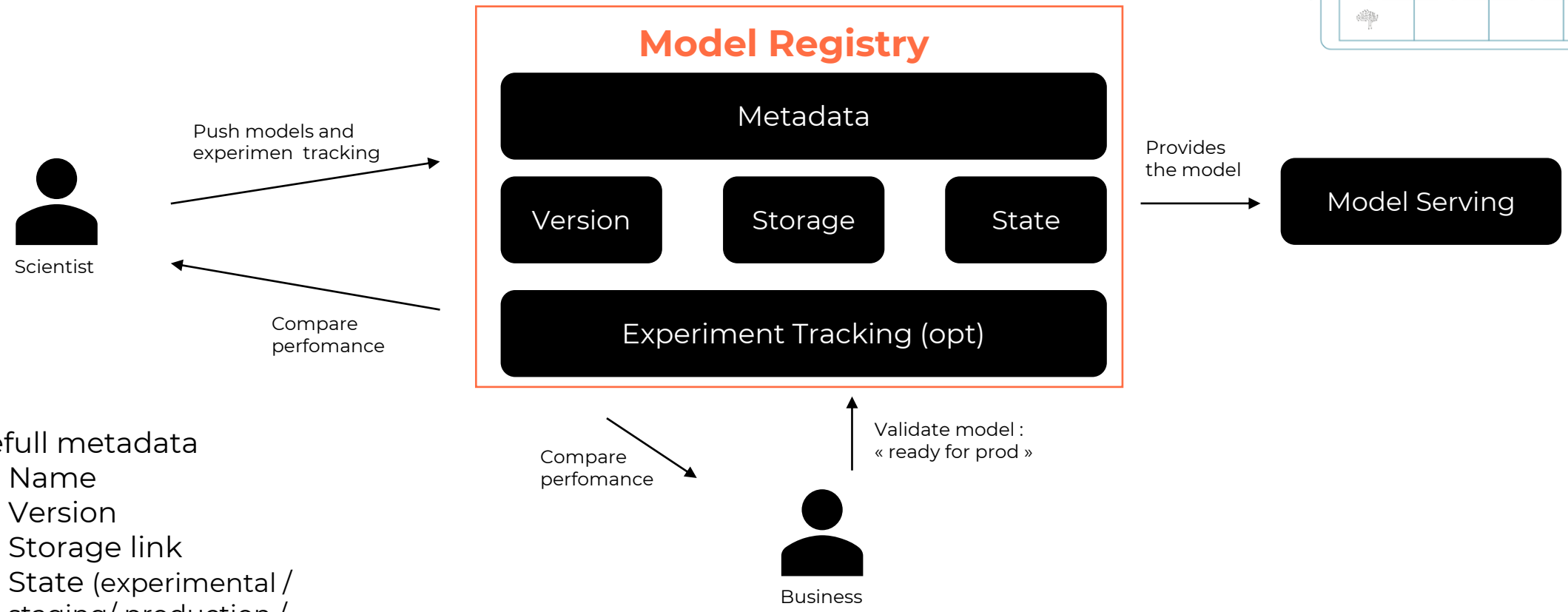
Versionning
Every time input dataset has new records, features should be recalculated and versionned

Serving
During inference, model should use the central feature definition before prediction

ML Governance

Model registry





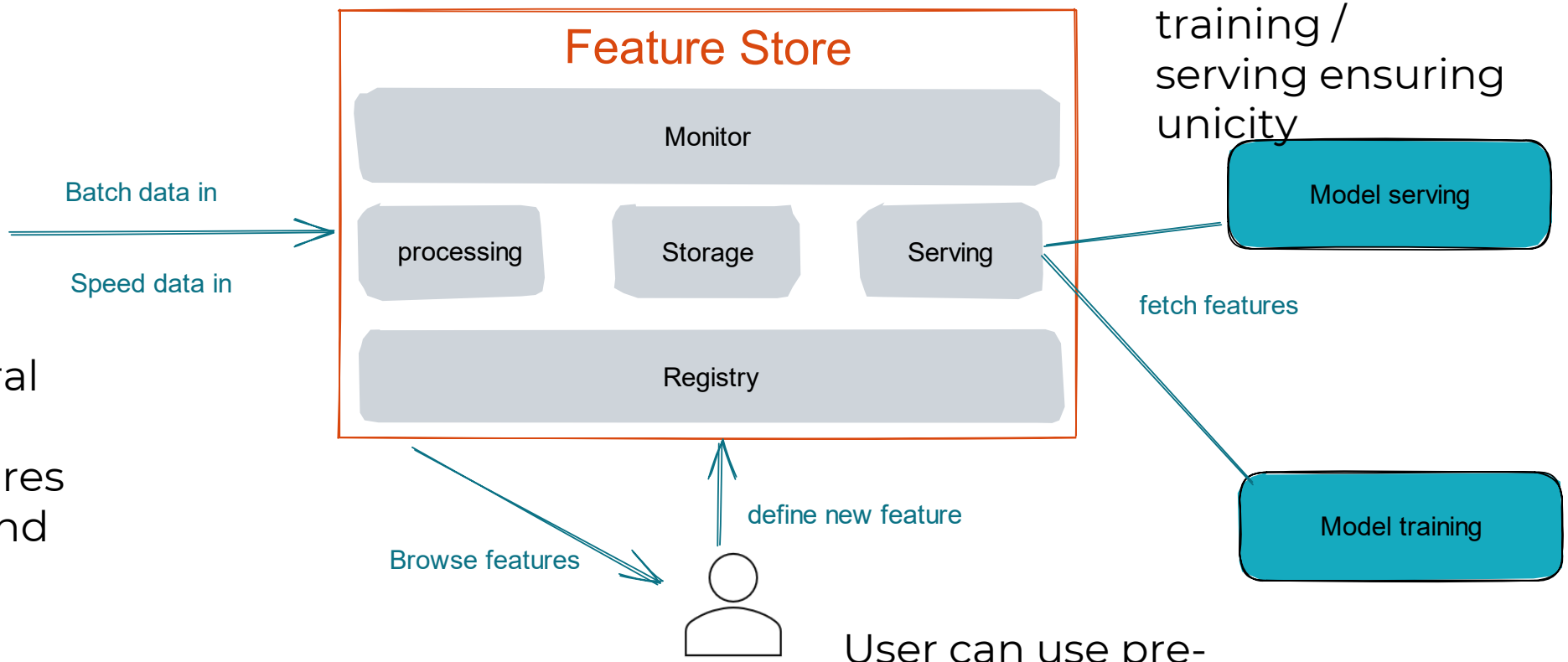
Usefull metadata

- Name
- Version
- Storage link
- State (experimental / staging/ production / archived)

ML Governance

Feature store

Expose via API features for training / serving ensuring unicity



Ingest several sources to create features from data and store it

User can use pre-builded features, add new or update some features

AI Project management

Traditional approach

Framing workshop

- Precise identification of the data that will be used
- Removal of technical and functional gray areas
- GDPR analysis (CADP, AIDP)
- Creation of specifications (backlog) for the first sprints to go "flying start"

Quick data analysis

- Analysis of data availability and accessibility
- Quality analysis (quantity, distribution, accuracy, etc.)

Proof Of Concept/Technology/Value

Agile iterative cycles of sprints to explore the problem and try to converge towards a solution

At each end of the cycle, an inventory is made of the value delivered to record whether or not the experience is continued.

Pilote

Industrialization of the results of the POC: taking into account the industrial constraints of the IS, automation, security, scalability, performance, etc.

Phase 1 – Framing

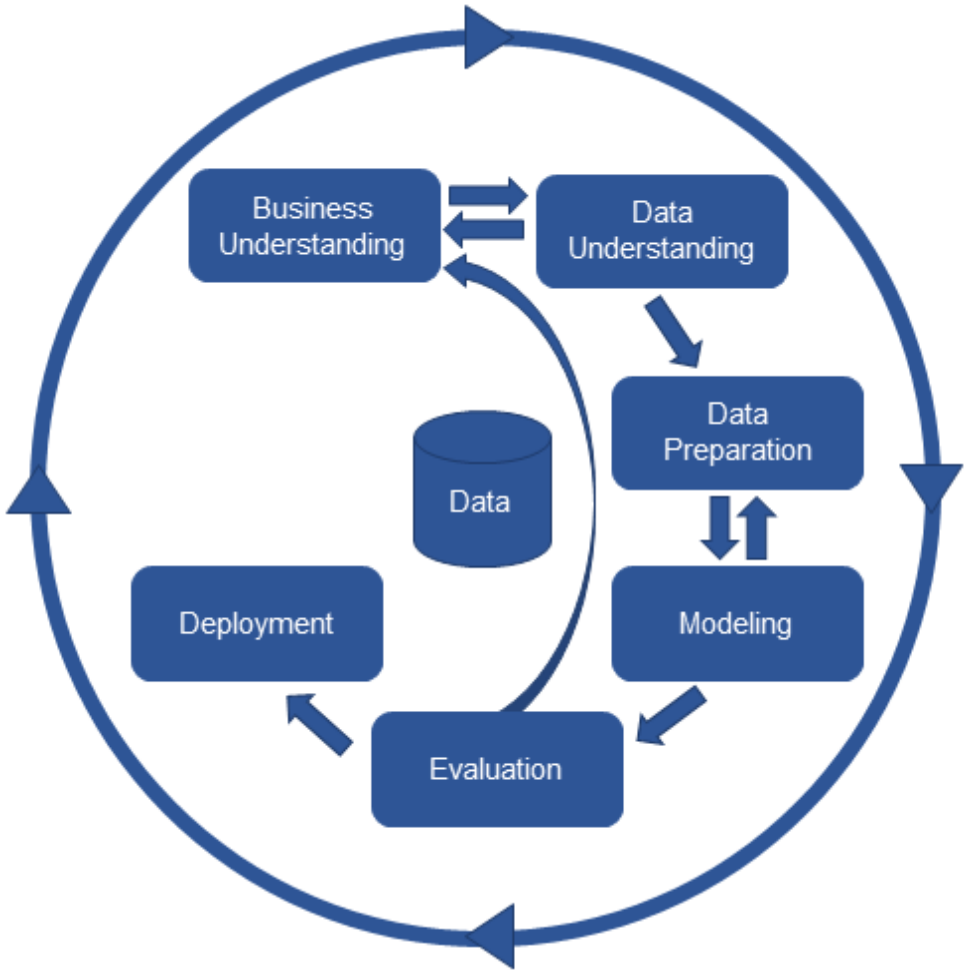
Phase 2 – POC

Phase 3 – Pilote

AI Project management

Agile iterative sprints

- Goals
  - Avoid “infinite R&D” mode on data
  - Reduce the time to production (TTM) of the use case
  - Reinforce the Dev/Ops mindset of the team
  - Integrate business users (create an enterprise-level data culture)
- Cinematic :
  - Framing and definition of the business problem, targeting of useful data to the use case
  - Analysis of data quality and content, understanding of business concepts behind the data
  - Cleaning, formatting, creation of features
  - Choice of technique/architecture, model configuration, training
  - Evaluation of the performance of the trained model



At the end of the cycle, demo to show the results/difficulties and readjust the objectives with the business users if necessary

Quizz

What we've learned

Question				
Truth point of the business data that will be consumed by other services of the company	Metadata	Master data	Reference data	Golden data
A limit of Datalakes is the lake of knowledge sharing between Platform engineers and domain experts	Y	N		
Data Mesh use big data and datalake technologies	Y	N		
Data Product should inherit of the specific/closed formats and protocols of the technology serving them	Y	N		
A Data Product owner is focused on data consumer satisfaction, developping trust and quality	Y	N		
The Standard data layer is composed of Raw data that have been structured and normalized	Y	N		
The Release phase of the ML lifecycle is focused on the industrialization and validation of the model	Y	N		
Feature store are used only in the Discovery phase when doing feature engineering	Y	N		
Model registry can retrain models when needed	Y	N		
The only profile needed for AI agile sprints are data scientists	Y	N		

Quizz

What we've learned

Question				
Truth point of the business data that will be consumed by other services of the company	Metadata	Master data	Reference data	Golden data
A limit of Datalakes is the lake of knowledge sharing between Platform engineers and domain experts	Y	N		
Data Mesh use big data and datalake technologies	Y	N		
Data Product should inherit of the specific/closed formats and protocols of the technology serving them	Y	N		
A Data Product owner is focused on data consumer satisfaction, developping trust and quality	Y	N		
The Standard data layer is composed of Raw data that have been structured and normalized	Y	N		
The Release phase of the ML lifecycle is focused on the industrialization and validation of the model	Y	N		
Feature store are used only in the Discovery phase when doing feature engineering	Y	N		
Model registry can retrain models when needed	Y	N		
The only profile needed for AI agile sprints are data scientists	Y	N		

They should be based on open standards
Feature store is also used in model serving during inference
Model registry is only a passive store
Data Engineers, Business Owner (or proxy BO), Ops are also necessary for AI sprints => create an enterprise-level data culture

In Practice

Lab Content

- Exo1 Data search & discovery with Catalog
  - Find where raw data is located using search
  - Create a transformation view on this data (ELT approach)
  - Visualize lineage into the catalog
  - Add metadata to this newly created data
- Exo2 Use feature store
  - From this transformed data, apply already existing features (available in the Feature store)
  - Create and register a new feature
  - Use both features to train a model
- Exo3 Use Model Registry
  - Push the trained model into the registry
  - Retrain changing some HP, push to registry and compare both models

- Go further :
  - Use the features and the model in a batch inference pipeline in KFP
  - Use the features and the model in a streaming inference pipeline in Kserve

EVIDEN