Data Mesh at Gruppo Hera

Our Data Strategy journey from 2021

12 december 2023

WHO AM I

Lisa Mazzini
Data Governance Manager

I work in the **Data Analytics & Intelligent Automation unit** (DAIA for friends), under the Innovation Director of Gruppo Hera.

Our goal is to coordinate and harmonize Data Analyitics projects all around the Group and to promote a shared Data Strategy to extract and exploit the value hidden inside corporate big data.

GRUPPO HERA

Gruppo Hera is a big italian multi-utility based in Bologna but working in many other cities around Emilia Romagna, Veneto and Marche.
It manages the supply of **energy**, **water** and **environmental services**, as well as **public lighting** and **telecommunications** to citizens and businesses.

Today Gruppo Hera is composed by more than **20 different companies.**

Market position

1st PLACE
Waste treated

2nd PLACE
Water invoiced

3rd PLACE
Sale of Gas and Electricity

1° OPERATOR
in the environmental sector regarding waste treatment and disposal

2° OPERATOR
in the water cycle industry, in terms of water supply

3° OPERATOR
in the sale of gas and electricity, in terms of number of customers

4° OPERATOR
in the distribution of gas, in terms of amount supplied

5° OPERATOR
in the field of electricity distribution, in terms of quantity.

5° OPERATOR
in the field of public lighting, in terms of number of light sources managed

AGENDA

01. What is Data Mesh

02. Why we chose Data Mesh and how

03. Data Mesh In practice: a real use case

04. Conclusion and lesson learnt

01. What is Data Mesh

OPERATIONAL PLANE VS ANALYTICAL PLANE

Data flows in a controlled and safe way

OPERATIONAL PLANE

All the systems where the data is generated by daily processes. Here is where data is created, modified and deleted following the business operational needs. The business continuity MUST be granted.

ANALYTICAL PLANE

The platform built in order to perform analysis on business data. On this «side», data is historicized, analyzed but should not be modified or deleted. Data can be analyzed without compromise the business continuity.

Improve and optimize operational processes

FOUR PILLARS OF DATA MESH

DATA AS A PRODUCT

Data is not simple file extraction: data should be treated as a product that each domain should provide it to consumers with a specific SLA and quality level.

FEDERATED GOVERNANCE

In order to avoid data silos and difficult integrations, there must be a governance group that defines the guidelines and standards to achieve interoperability.

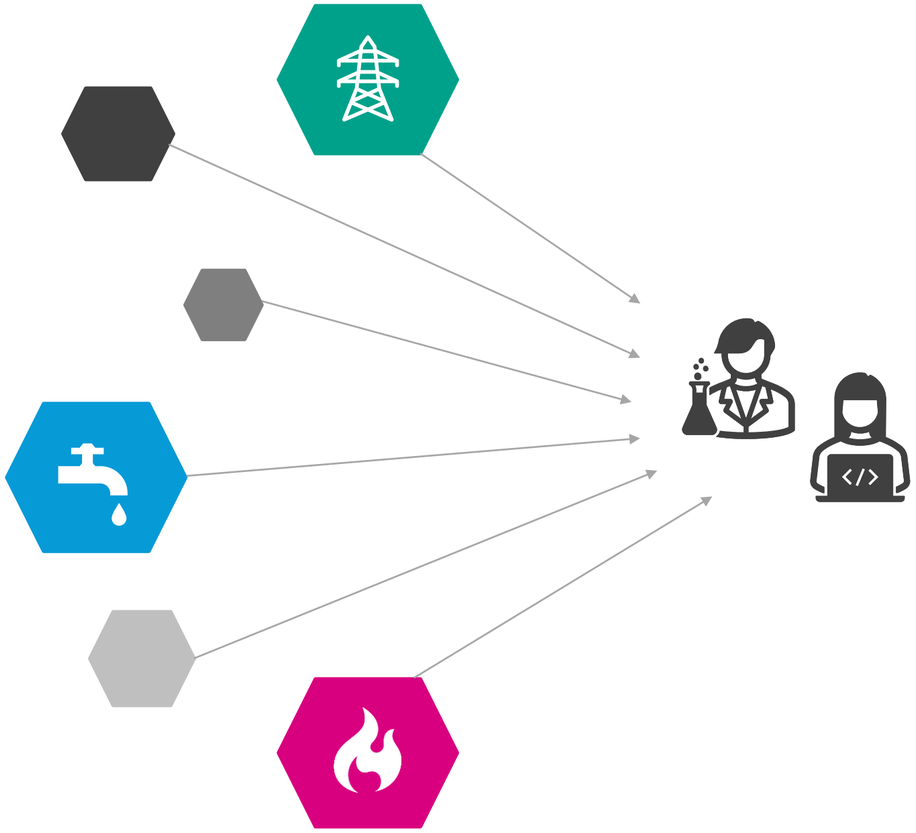
DOMAIN OWNERSHIP

Ownership of data, in its content and quality, is not assigned a central data team, but it's distributed on each domain team, which can evaluate it using its own domain knowledge.

SELF-SERVE DATA PLATFORM

Data products should live on a platform that provides all the tools needed to build and maintain them, granting the respect of interoperability and security policies.

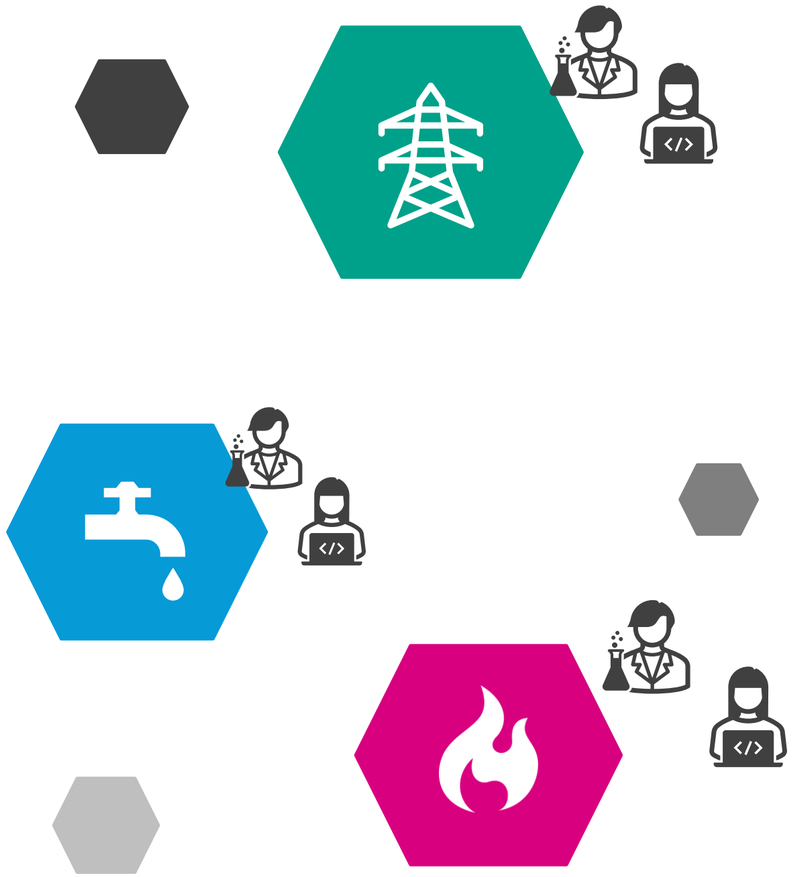
DECENTRALIZATION AND SCALABITY



CENTRAL “DATA TEAM”

- Lack of specific domain knowledge, making interpretation and cleaning of data not efficient
- Bottleneck effect: often in big organization, the team is not big enough to handle all the work from all domains

DECENTRALIZATION AND SCALABITY



DECENTRALIZED MODEL

- Each domain is now **accountable and owner** of its data in terms of quality, using its own domain knowledge.
- Data can be shared between different domains, but under the approval of the right Owner.
- This approach is **far more scalable**: each domain has its “data people”, from analyst to domain-expert data scientist. *(the so-called citizen data scientist)*

“The accountability of data quality shifts upstream as close to the source of the data as possible”

Zhamak Dehghani

DEMOCRATIZATION AND OWNERSHIP

ONE OF THE MAIN GOALS OF A DATA MESH IS TO REACH THE DATA DEMOCRATIZATION.

BUT WHAT DOES IT MEAN?

Often, data is **locked** inside the operational systems under the control of the IT department.
Often, data is modelled following the requirements of the operational system, which not always can be understood easily by business people (es. SAP).

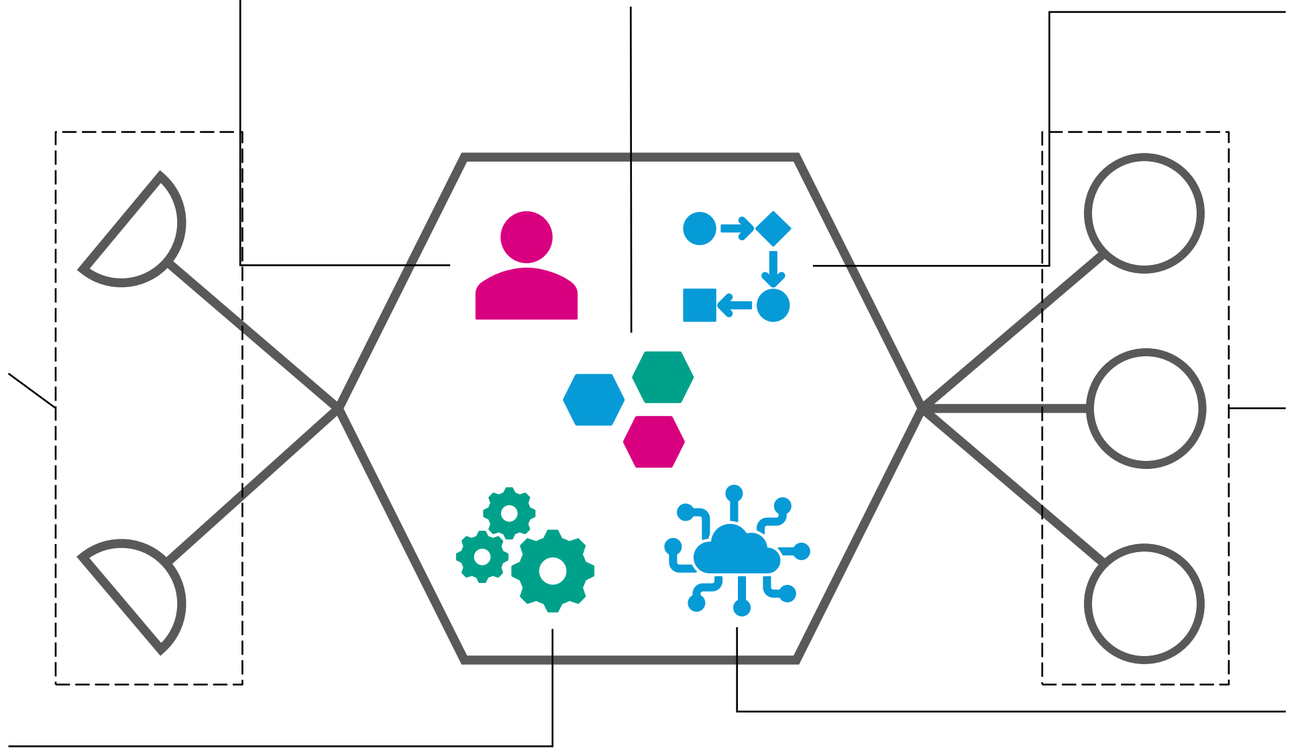
The **democratization** aims to bring data closer to its real owner and to make it understandable
By making it understandable, the owner can also perform better data quality controls on data and improve the overall quality of the processes.

NOT JUST A DATASET..

OWNERSHIP
Every Data Product has a Owner: the owner has a deep understanding of the process that produce the data and they can evaluate and decide how to measure the data quality.

DATA
Data on the analytical plane must be immutable and historicized in order to preserve all the information.

METADATA
In order to consume the Data, the Owner and stewards should provide complete business metadata, along with technical metadata and lineage (that usually can be detected with specific tools, i.e.Data Catalog)



INPUT PORT
Data is acquired from the operational plane or from other Data Products, using their respective output ports

OUTPUT PORT
Data produced should be accessible using its Output Ports. Depending on the use cases, a consumer can access data in different ways (SQL DB, data lake, streaming, ...)

FLOWS & LOGIC
The engine of the Data Product: all the pipelines (code or low-code) that manipulate and prepare data to be consumed. All the best practices of software development should be followed.

INFRASTRUCTURE
Data Products should be built in a common Data Platform that can guarantee technical and security policies.

...BUT A DATA *PRODUCT*

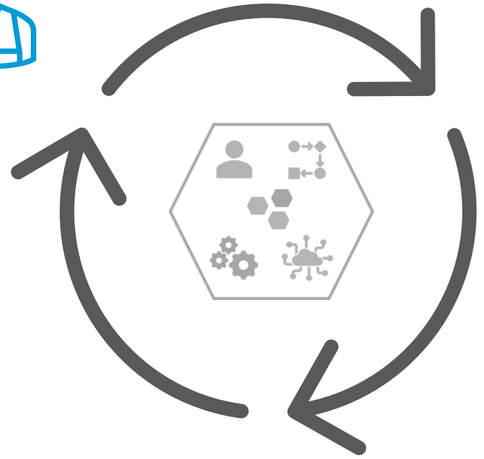
A Data Product is born when there's a business need that requires its data. It will be then designed and developed...

IDEA & PROJECT

DEVELOPING

MAINTENANCE

EVOLUTION



IMPROVEMENT

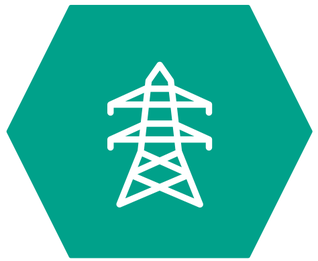
OUT OF BUSINESS

When the product ceases to bring value, it will be put out of business and «turned off». This is the end of its life cycle.

...but these are only the first steps of its life cycle! As every other product, it must be maintained, evolved and improved constantly, in order to keep bringing value to its consumers.

CHARACTERISTIC OF A DATA PRODUCT

- **DISCOVERABLE**: potential consumers need to be able to explore all the available Data Products.
- **ADDRESSABLE:** Data Products should have a permanent and unique address to access it, assuring the continuity of usage
- **UNDERSTANDABLE:** consumers need to understand the underlying data, both in terms of its syntax and semantic.
- **TRUSTWORTHY:** data exposed by a Data Product should be truthful and it should represent the facts of business correctly



CHARACTERISTIC OF A DATA PRODUCT

- **ACCESSIBLE:** Data Products need to make it possible to various data users to access and read data in their native mode of access (dashboarding, APIs, db queries…)
- **INTEROPERABLE:** Data Products should follow standards and harmonization rules that allow to link data across domains easily.
- **VALUABLE:** data exposed should be valuable and meaningful on its own, without the need of being joined and correlated with other data products.
- **SECURE:** each Data Product should implement security policies in terms of access control, confidentiality levels and regulations.

DATA MARKETPLACE

As every other product, to be discovered and consumed should be available on a **Marketplace**.

The **Data Marketplace** is a tool that can be used by potential consumers to find out:

- What Data Products are **available** to use
- What kind of **information** is contained inside
- What are the **Service Level Agreements** (data freshness, quality level, …)
- Who’s the **owner**
- **Ask** the owner to **access** their Data Products.

DATA CATALOG

One of the main tool in Data Management
is the **Data Catalog**.

Data Catalog is where all the information about the data (but not the data per sè) is available to be consulted by potential consumer.
Inside a Data Catalog can be found information about:

- Technical metadata
- Business metadata
- Data lineage
- Business glossary

DATA CONTRACTS

The consumer that wants to access to a Data Product ask the permission to the Owner, who will then write a **Data Contract**, containing:

- What data will be available to the consumer, through which Output Port.
- Terms and Condition of usage
- Quality attributes (freshness, non-null values..)
- Service Level Objectives (availably of data, support, ..)

```yaml
dataContractSpecification: 0.9.0
id: orders-latest-npii
info:
  title: Orders Latest NPII
  version: 1.0.0
  description: Successful customer orders in the webshop. All orders since 2020-01-01.
  owner: Checkout Team
terms:
  usage: Data can be used for reports, analytics and machine learning use cases.
  limitations: Not suitable for real-time use cases.
  billing: 5000 USD per month
  noticePeriod: P3M
schema:
  type: dbt
  specification:
    models:
      - name: orders
        columns:
          - name: order_id
            type: string
            description: Primary key of the orders table
          - name: order_timestamp
            type: timestamptz
            description: The business timestamp in UTC when the order was successfully registered.
          - name: order_total
            data_type: integer
            description: Total amount of the order in the smallest monetary unit (e.g., cents).
quality:
  type: SodaCL
  specification:
    checks for orders:
      - row_count between 1000000 and 3000000
```

SELF-SERVE DATA PLAFORM: CLOUD

If you want to reach a strong Data Democratization and keep data close to the real owners, you need to provide a **Self-Serve Data Platform**. The goal is to enable domain teams to seamlessly create and consume data, in the simplest way possible.

Cloud resources respond really well to this need for mainly two reasons:

- Most cloud providers have **fully managed** services that «hide» all the systemistic effort to users, so that they can focus of data analysis tasks. Many services also **scale up or down seamlessly** based on the workflow, granting both **flexibility** and **speed.**
- Cloud Architecture can be provided through Automation pipelines written using **Infrastructure as Code** frameworks (such as Terraform), **reducing effort** and **time.**

FEDERATED GOVERNANCE

In order to maintain a decentralized model, there must be a **strong Governance group** that defines:

- Global policies that can guarantee **interoperability**, to allow different domain teams to use data products in a consistent way.
- Choose and implement all the **useful tools** seen before (Data Catalog, Data Marketplace and so on) as well as defining the requirements of the Data Platform.
- Define best practices and policies to guarantee data protection and industry specific legal requirements.

Data Mesh suggests this team to be a **Federated Governance team**, formed by representatives from each domain.

02.

Why we chose Data Mesh and how

WHY DATA MESH IN GRUPPO HERA

DATA STRATEGY JOURNEY

JULY 2021

Data Strategy starts here..!
We defined the main guidelines regarding platform, processes and governance.

End of 2021

- Envisioning sessions with Business Units
- Choosing first pilot projects to test the guidelines

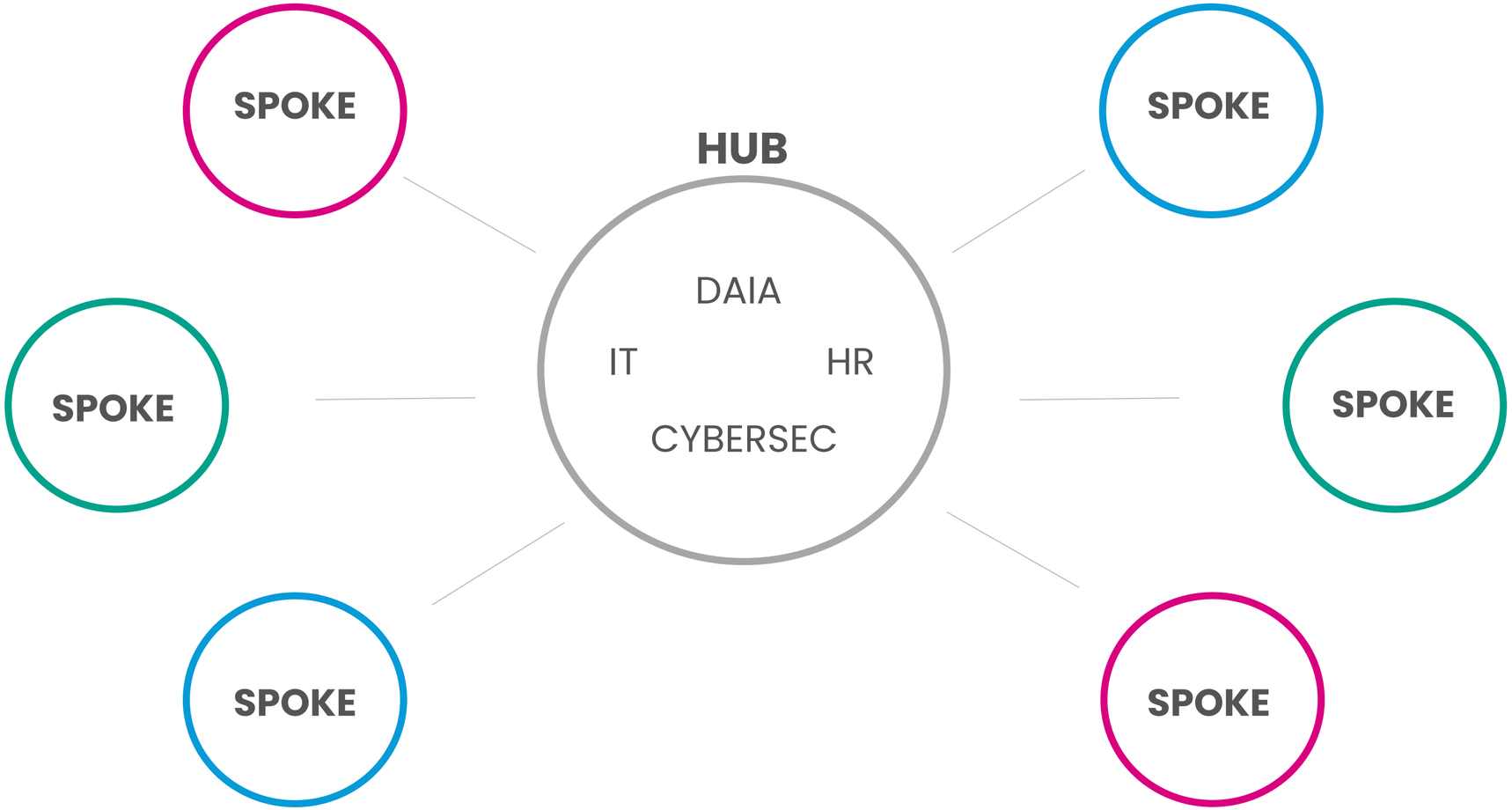
2022

- Building the Data Platform
- Defining automation processes and policies
- Pilot projects started
- Defining change management actions, regarding roles and learning paths.

2023

- Bringing the first set of Data Products to life
- Building some use cases that can leverage the Data Products available
- Improving the Data Platform with better infrastructure and tools.

HUB & SPOKE ORGANIZATIONAL MODEL



“AGILE ROOM”

What we call an Agile Room in Hera is a small team composed by the key people that are essential to build a great Data Product.
The main goal is to **reduce the distance** between the owner that has the domain knowledge and every other actor that is needed to build the Data Product.

It is fundamental to build a so-called **ubiquitous language,** which is a common dictionary of terms to describe the domain that is shared with all the team.

Inside the Agile room, there can be up to 8 different profiles:

- Business Sponsor
- Data Product Owner
- Data Product Steward
- Data Architect & Engineer
- Data Security Engineer
- Data Governance Manager
- Data Scientist
- Data Visualization Developer

PROFILES

DATA PRODUCT STEWARD

DATA VISUALIZATION DEVELOPER

BUSINESS SPONSOR

DATA PRODUCT OWNER

DATA SCIENTIST

DATA ARCHITECT & ENGINEER

DATA SECURITY ENGINEER

DATA GOVERNANCE MANAGER

DATA PRODUCT OWNER

The **Data Product Owner** is the one who:

- has a deep knowledge of the business process, because is part of their daily work.
- can define the macro-requirements of the Data Product and can then assess the compliance.
- Knows who can or cannot access to their Data Products and approves the requests defining the right Data Contracts.
- Is accountable for the Data Quality of their Data Products.

DATA PRODUCT ARCHETYPES

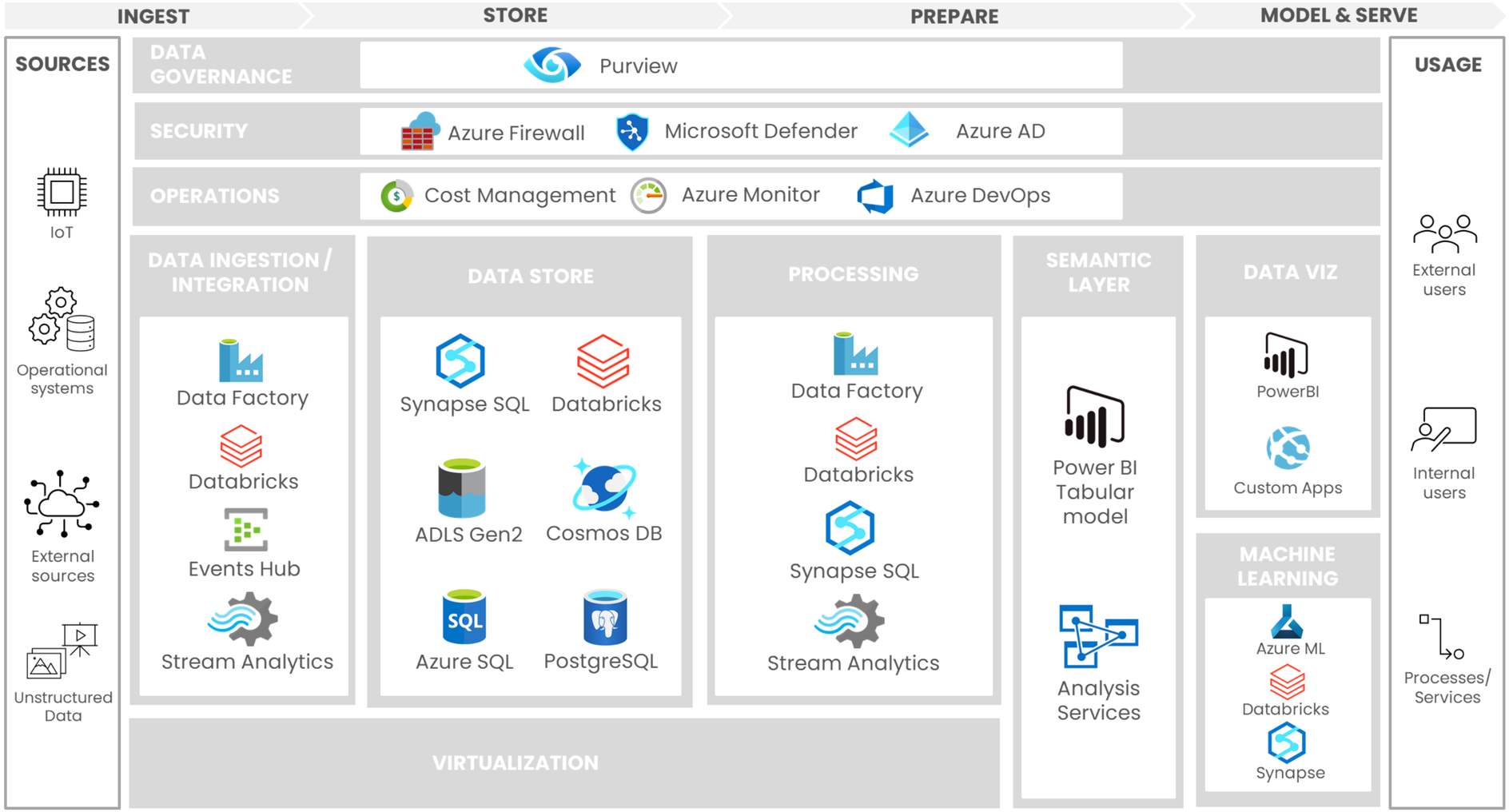
DOMAIN DATA PRODUCT

This Data Product brings the data from the operational systems to the analytics plane. The best practice is to keep the data model as general purpose as possible.

CONSUME DATA PRODUCT

Built from the elaboration of Domain Data Products, this kind of Data Product answers to a specific use-case (can be a dimensional model for BI, but also the application of a ML model, i.e. failure prediction)

INFRASTRUCTURE STACK



03.

Data Mesh In practice: a real use case

REAL USE CASE: EMERGENCY RESPONSE PROCESS

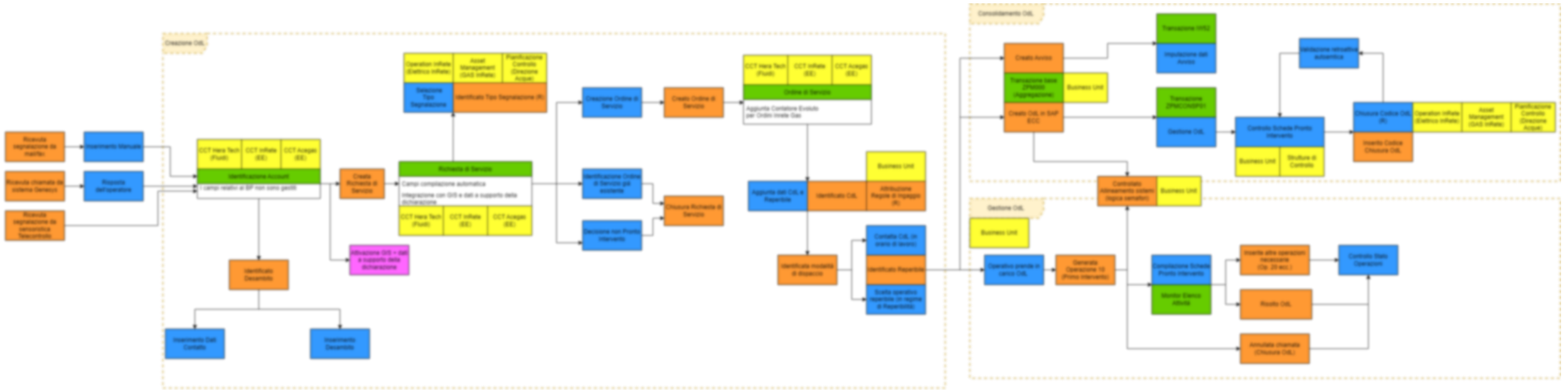
This project was one of the four main pilot projects of our Data Strategy. The goal was to build the Domain Data Products regarding the **Emergency Response process** for water, electricity and gas.

Clearly this is one of the **most important** processes for Gruppo Hera and its clients and it’s also **strictly regulated**. Being able to access and analyze this data can be extremely valuable both in terms of quality of service given to customers, but also regarding our performance.

PROCESS ANALYSIS

First step was to **describe and analyze the detailed process** of the Emergency response, starting from the fault report from a customer, for example, to the intervention and finally to the validation and evaluation of the work.

The output is a clear vision of every actor involved and which operational systems host the data regarding the process.



PROCESS ANALYSIS

This seems easy on a first look...

...but it's not!

3	6	3
DIFFERENT ORGANIZATIONS	BUSINESS UNITS	OPERATIONAL SYSTEMS

Different roles divided in «**line**» (the ones that actually handle the intervention)
and «**staff**» (technical people who work on systems and handle the data).

WHO’S THE OWNER?

«Line»

- People who *do the job*
- They know exactly how the process works because it’s their daily job
- They may not have a deep understanding of the importance of the data and its quality.

«Staff»

- People who handle and analyze data regarding the process.
- They have technical and analytical skills to elaborate the data
- They may not know all the detailed steps of the process.

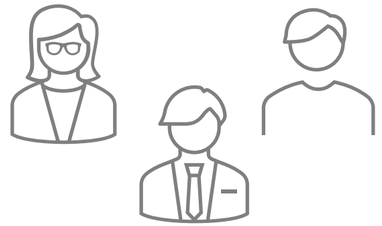
WHO’S THE OWNER?

DATA PRODUCT OWNER

«Line»

- People who *do the job*
- They know exactly how the process works because it’s their daily job
- They may not have a deep understanding of the importance of the data and its quality.

DATA PRODUCT STEWARD



«Staff»

- People who handle and analyze data regarding the process.
- They have technical and analytical skills to elaborate the data
- They may not know all the detailed steps of the process.

PROCESS ANALYSIS

The first output of the process analysis was the definition of **three main domains**.

Now that the domains are clear, the Data Product Owners and Data Product Stewards must be defined, in order to leverage their domain knowledge and experience during the design.
To find and «elect» the right people, the **HR department** was asked to give their opinion and help in this decision.

LET’S BUILD DATA PRODUCTS

Now that the team is completed with Owners and Stewards, **we started building our Data Products**.

First of all, let’s move data from the operational system to the analytical plane.
As said, data is often «locked» inside the system, modeled in a technically-efficient way that is not easy for the business people to understand. So the first step is to explore the operational system and search where all the needed information is stored.

Once you find it, you have to extract the data **without compromising the system’s business continuity**. This aspect is, of course, fundamental to efficiently separate the operational plane from the analytical plane.

LET’S BUILD DATA PRODUCTS

For example: we want to model the concept of a **Service Request**: a Service Request is born when the call center receives a call from a customer that wants to report a issue on the network.

The call center operator registers a set of information from the call, for example:

- Date & time
- Issue type
- Priority
- Geolocation of the caller
- … And many more

All this data can be stored in **different tables**, using different reference data, and so on.
The Agile Room must work together to understand what data is needed and where to find it .

LET'S BUILD DATA PRODUCTS

CRM Operational System

Service Request

Status Description

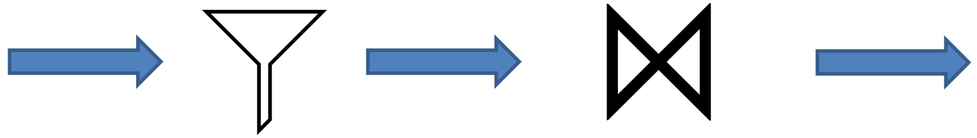
Issue Description

[Reference table]

[Reference table]

Very technical model, not well separated, hard to read for the business.

Filter only the emergency requests
And join all the technical tables.



Filter only on emergency requests

Join on key fields

Emergency Process Service Request

Create a representation that is closer to the business language, can be easily read and it's general purpose

LET'S BUILD DATA PRODUCTS

Service Request

FILTRI		
Tabella	CRMD_ORDERADM_H	
Campo	PROCESS_TYPE	ZZSERVIZIO
Valore	ZSRQ	CCTF

JOIN					
Tabella Master	Tabella Slave	Chiavi Master		Chiavi Slave	
		Campo chiave 1	Campo chiave 2	Campo chiave 1	Campo chiave 2
CRMD_ORDERADM_H	ZCCTI_DLIVURG	ZZURGENZA		ZCDPRIORITA	
CRMD_ORDERADM_H	ZCCTI_INDVINC	ZZVINCOLO		ZCDINDVINC	
CRMD_ORDERADM_H	ZCCTI_PMTB0014	ZZTIPO_SEGNALAZ	ZZDESAMBITO	ZTPSEGN	ZCDDSM
CRMD_ORDERADM_H	ZCCTI_STATI	ZZSTATO_CHIAMATA		CODICE_STATO	

CONTENUTO DATA PRODUCT			
Campi DDP Tecnici	Descrizione Campi DDP	Campi SAP CRM	Nome Tabella SAP CRM
CRMD_ORDERADM_H-GUID	GUID	GUID	CRMD_ORDERADM_H
CRMD_ORDERADM_H-OBJECT_ID	ID Oggetto	OBJECT_ID	CRMD_ORDERADM_H
CRMD_ORDERADM_H-PROCESS_TYPE	Tipo Operazione	PROCESS_TYPE	CRMD_ORDERADM_H
CRMD_ORDERADM_H-POSTING_DATE	Data Registrazione	POSTING_DATE	CRMD_ORDERADM_H
CRMD_ORDERADM_H-CREATED_AT	Data Creazione	CREATED_AT	CRMD_ORDERADM_H
CRMD_ORDERADM_H-CREATED_BY	Autore	CREATED_BY	CRMD_ORDERADM_H
CRMD_ORDERADM_H-CHANGED_AT	Data Modifica	CHANGED_AT	CRMD_ORDERADM_H
CRMD_ORDERADM_H-CHANGED_BY	Autore Modifica	CHANGED_BY	CRMD_ORDERADM_H
CRMD_ORDERADM_H-HEAD_CHANGED_AT	Data Modifica Format	HEAD_CHANGED_AT	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZID_CHIAMATA	ID Chiamata	ZZID_CHIAMATA	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZDATA_CHIAMATA	Data Chiamata	ZZDATA_CHIAMATA	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZNUMERO_VERDE	Numero Verde	ZZNUMERO_VERDE	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZCODACENTRALINC	Coda Centralino	ZZCODACENTRALINC	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZSERVIZIO	Tipo Servizio	ZZSERVIZIO	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZCHIAMANTE	Codice Segnalante	ZZCHIAMANTE	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZBP	Business Partner	ZZBP	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZCONTATTO	Contatto	ZZCONTATTO	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZCODICE_SOT	Codice SOT	ZZCODICE_SOT	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZCIVICO	Civico	ZZCIVICO	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZCIVICO_NOSIST	Civico Non A Sistema	ZZCIVICO_NOSIST	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZINCROCIO	Incrocio	ZZINCROCIO	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZONA	Zona	ZZONA	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZNOTE_UBIC_CALL	Note Ubicazione	ZZNOTE_UBIC_CALL	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZSETT_MERC_ODL	Settore Merceologico	ZZSETT_MERC_ODL	CRMD_ORDERADM_H
CRMD_ORDERADM_H-ZZDESAMBITO	Desambito	ZZDESAMBITO	CRMD_ORDERADM_H

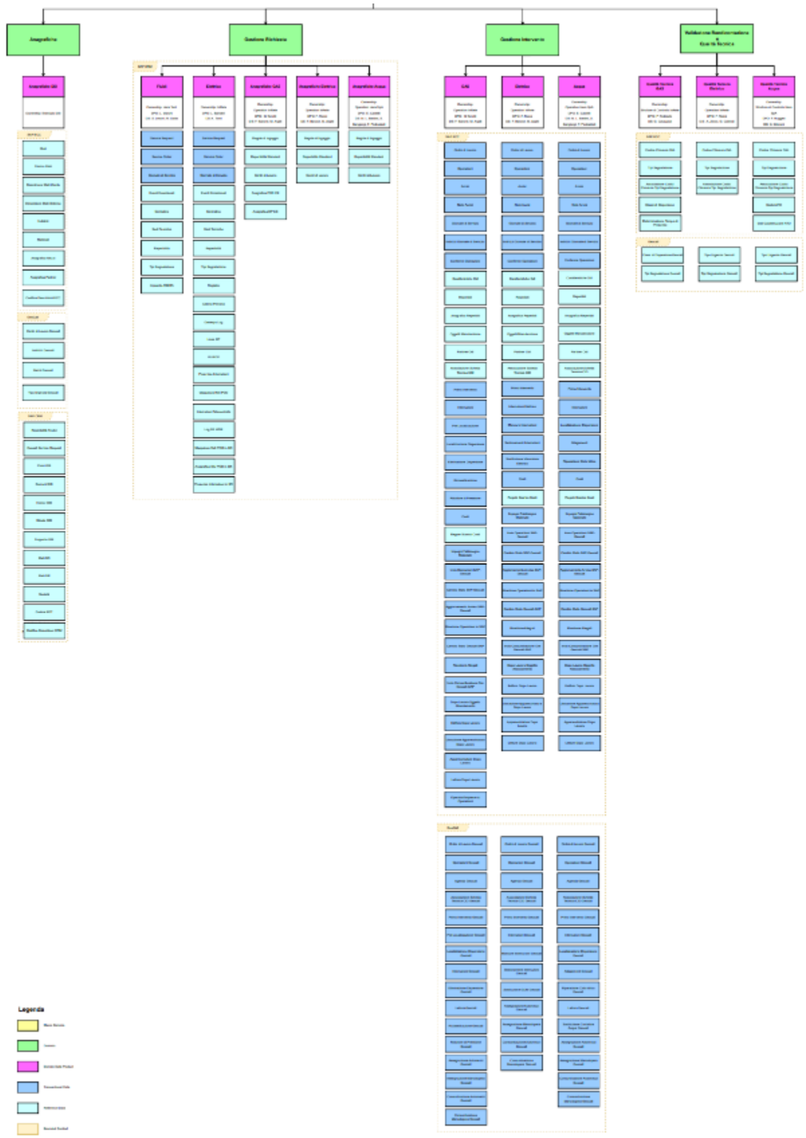
+ other 60 fields

LET’S BUILD DATA PRODUCTS

Now, imagine doing this operation, for many concepts, distributed among hundreds of tables, from three different operational system.. **Not that easy!**

In the end, we identify **12 different Domain Data Products**, that are composed by many datasets that represent different concepts from the same «portion» of the domain.

For example, the entity Service Request is just one of the many identities that cover the «Request of intervention» portion of the Emergency response process.



LET'S BUILD DATA PRODUCTS

This is a high-level representation of the Cloud Architecture we chose:

OPERATIONAL PLANE

ANALYTICAL PLANE

AZURE DATA FACTORY

Orchestration tool to build and orchestrate data pipelines to move and manipulate data.

AZURE STORAGE ACCOUNT

Storage service that is durable and highly available, secure, and **massively scalable**

AZURE DATABRICKS

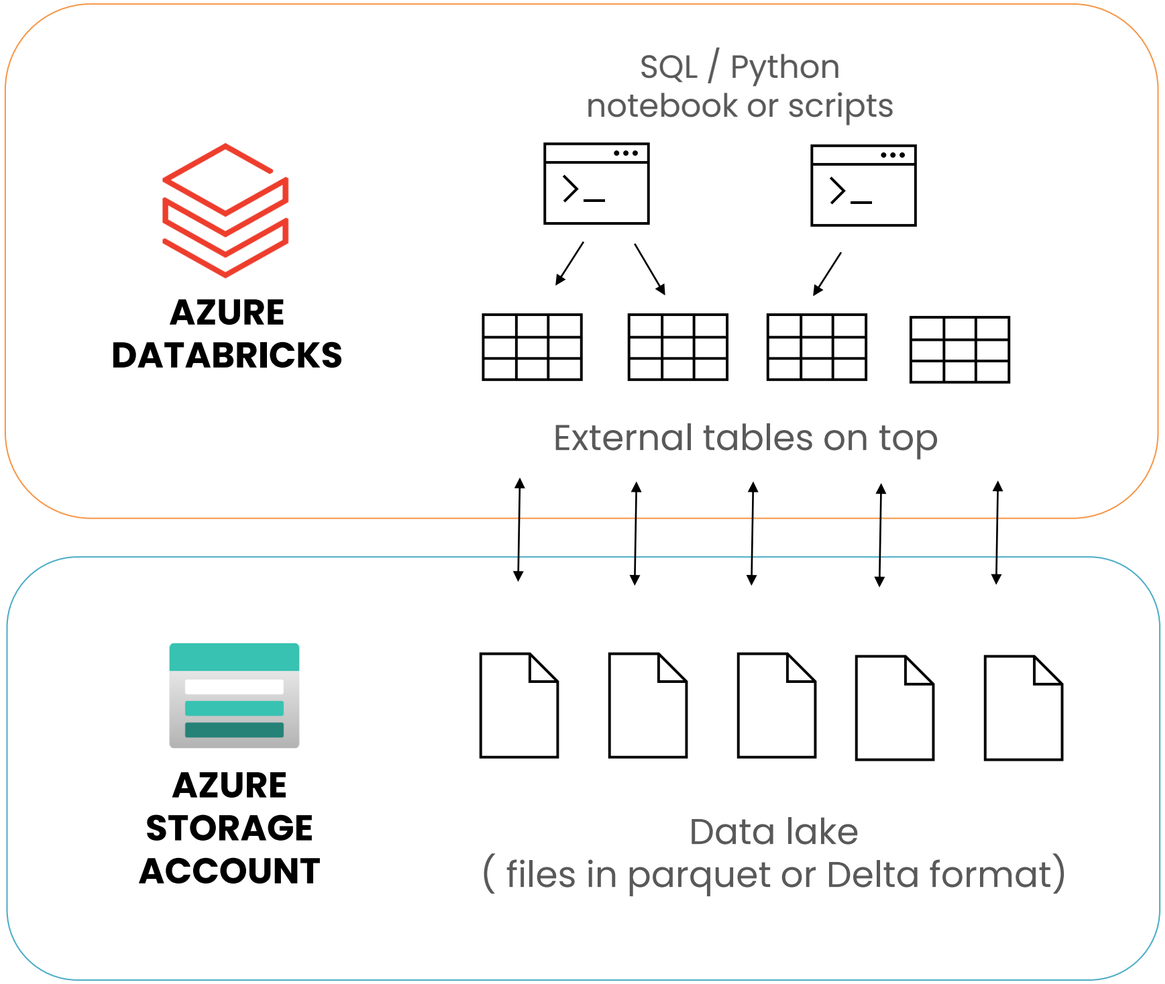
Data Engineering tool to process and transform large quantities of data, using **Spark engine**

LET’S BUILD DATA PRODUCTS

The technical paradigm we are choosing is the **Data Lakehouse**.

A Data Lakehouse is a pretty new architecture that combines *the best of both worlds*: the **flexibility** and **scalability** of a data lake, with the **functionalities** and **structures** of a Data Warehouse.

This paradigm satisfies the accessibility of a Data Product, since it can be queried using SQL from Stewards, but also explored by code from Data Scientist to build a statistical model.



TESTING AND GO LIVE

Once a Data Product is ready, then Owners and Stewards can perform UAT (User Acceptance Tests), to verify that:

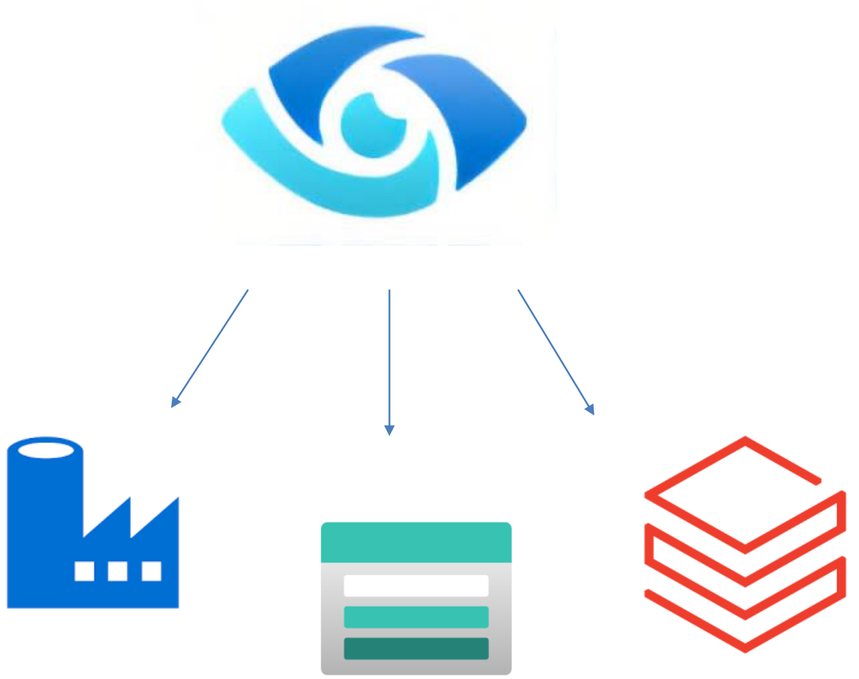
- All requirements are fullfilled
- All the information is correctly included
- Data formats are correct
- Eventual pipeline errors
- ... And so on.

This moment is **crucial**, especially when we are building the first Data Products: *finally, Owners and Stewards can see and appreciate the potential of finally have access to their data.*

METADATA AND CATALOG

Now that Data Products are ready, there are two last steps that must be fullfilled.

First of all, Data Products must be **registered on the Data Catalog** as new assets available for the all Group to explore and ask access to. In our case, right now we are following a «pull» approach, meaning that our Catalog will scan the Cloud resources selected and will retrieve all the data assets available (databases, tables, files, etc). In this phase, the **technical metadata** is also retrieved **automatically** from assets.



METADATA AND CATALOG

- Once the assets are registered in the Catalog, business people can add business metadata that can be useful to potential consumers, for example:

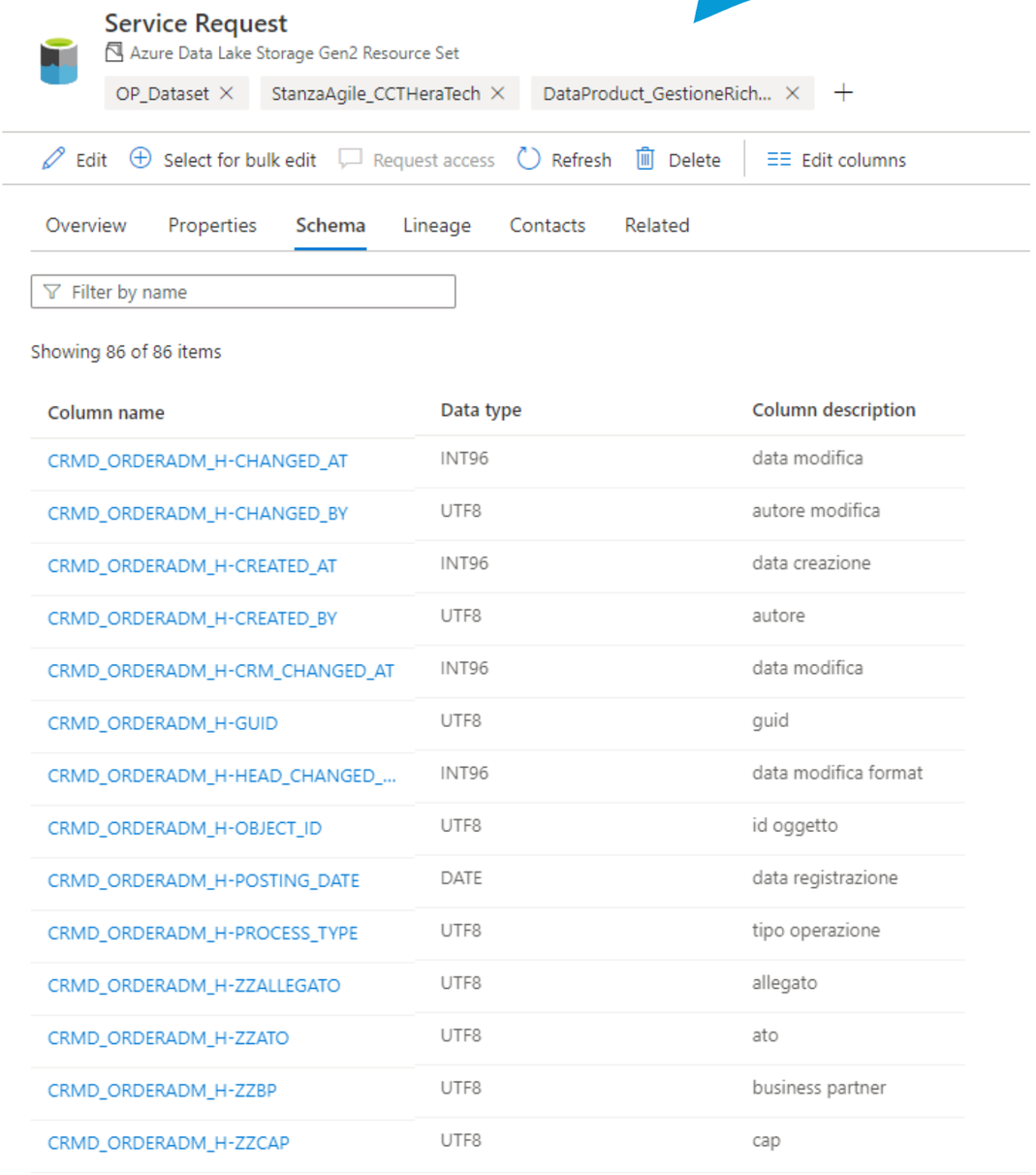
Data product name: Request of Intervention – Electric Domain

Description: All the information registered on the CRM when the call center operator receives an emergency call reporting a failure in the electric network.

Freshness: Data is refreshed daily at 6am.

Output ports available:

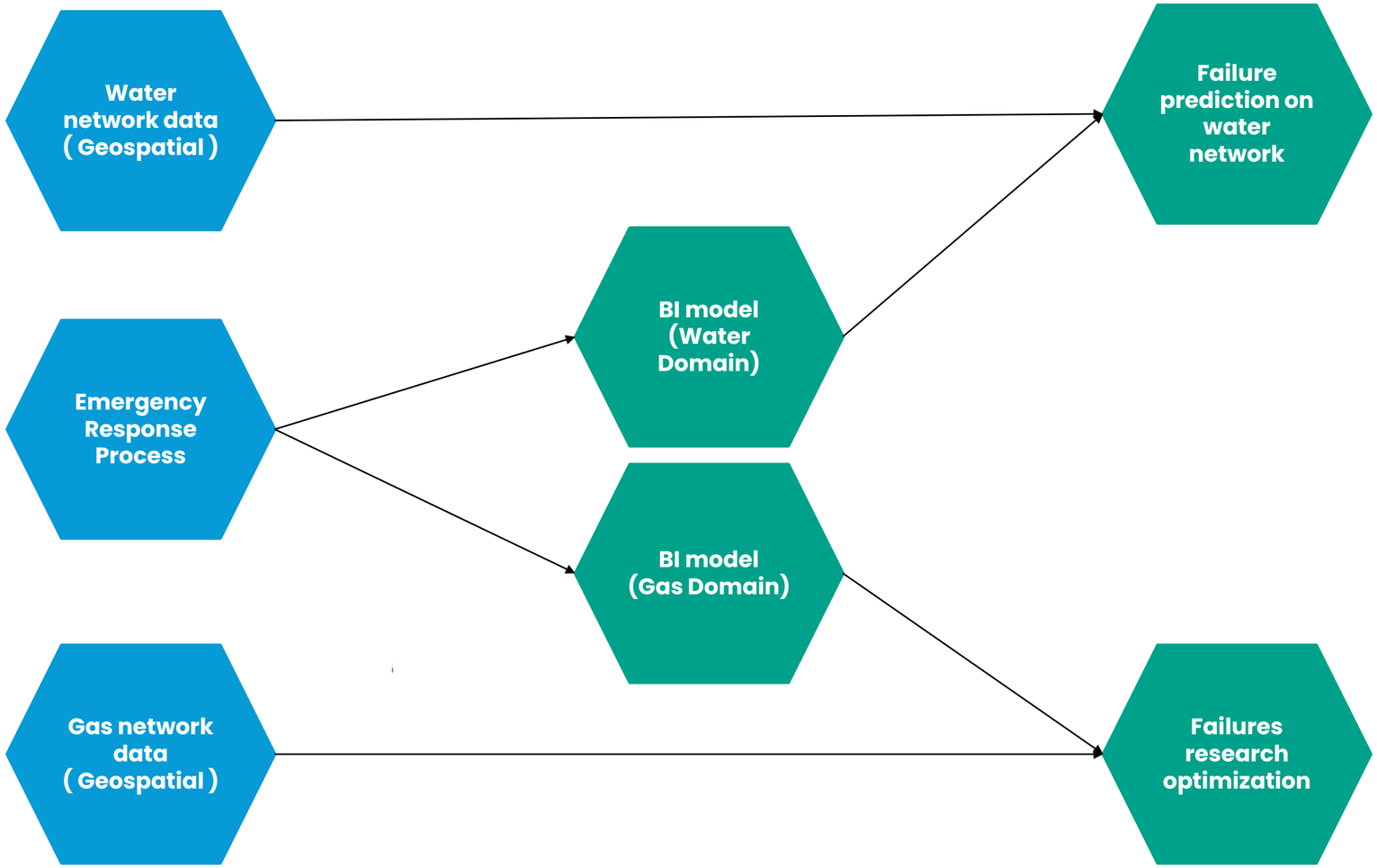
- Data lake (parquet files)
- Data lakehouse (external tables)



04.

Conclusions & Lesson learnt

ONE DATA PRODUCT LEADS TO ANOTHER...



DOMAIN DATA PRODUCT

CONSUMPTION DATA PRODUCT

LESSON LEARNT

- The real challenge is mostly **cultural and organizational**.
- Since this paradigm is socio-technical, there is a strong «**human factor**» that cannot be ignored if you want to be successful: at first, people will be pretty confused and skeptical. It's normal and it must be handled.
- There must be a strong buy-in both from C-levels and from business people.
- Specially for pilot projects, from a Governance perspective **don't take guidelines as «written in stone».** Always question if guidelines are completed or may be improved.
- Definitely better to **start small than big**.

THANKS!

Any questions? :)