

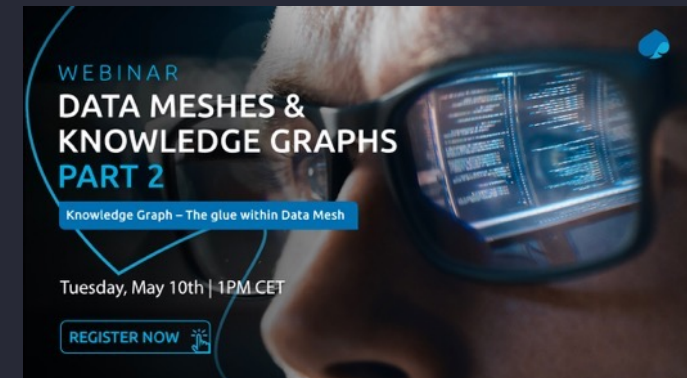
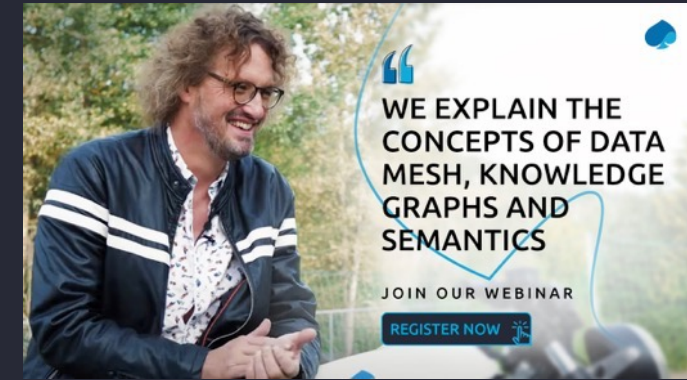
# INSIGHTS & DATA

Webinar Series – Data Meshes & Knowledge Graphs

# Data Meshes & Knowledge Graphs

## Webinar Series

1. **Part 1** – “How you can add the human way of understanding data better”  
April 6th, Robert Engels  
Replay available
2. **Part 2** – “Knowledge Graph – The glue within Data Mesh”  
May 10th, Arne Rossman  
Today's webinar
3. **Part 3** – “Semantic data mesh - The missing links in your data platform”  
May 20th, speaker Aniruddha Khadkikar  
Hosted by PTC with Capgemini as a partner at the NordicTalks webinar





# DATA MESHES AND KNOWLEDGE GRAPHS PART 2

Knowledge Graph – The glue within Data Mesh

Arne Rossmann



# WEBINAR DATA MESHES & KNOWLEDGE GRAPHS

Presented by Capgemini, *Insights & Data*



Today you will learn about:

- Why the main pitfall of Data Mesh is on the Governance Layer
- How KnowledgeGraph can solve that issue
- What building blocks you need to enable users

Speaker:

Arne Rossmann

*Data & AI Foundation Lead Intelligent Industry for Capgemini, Insights & Data*





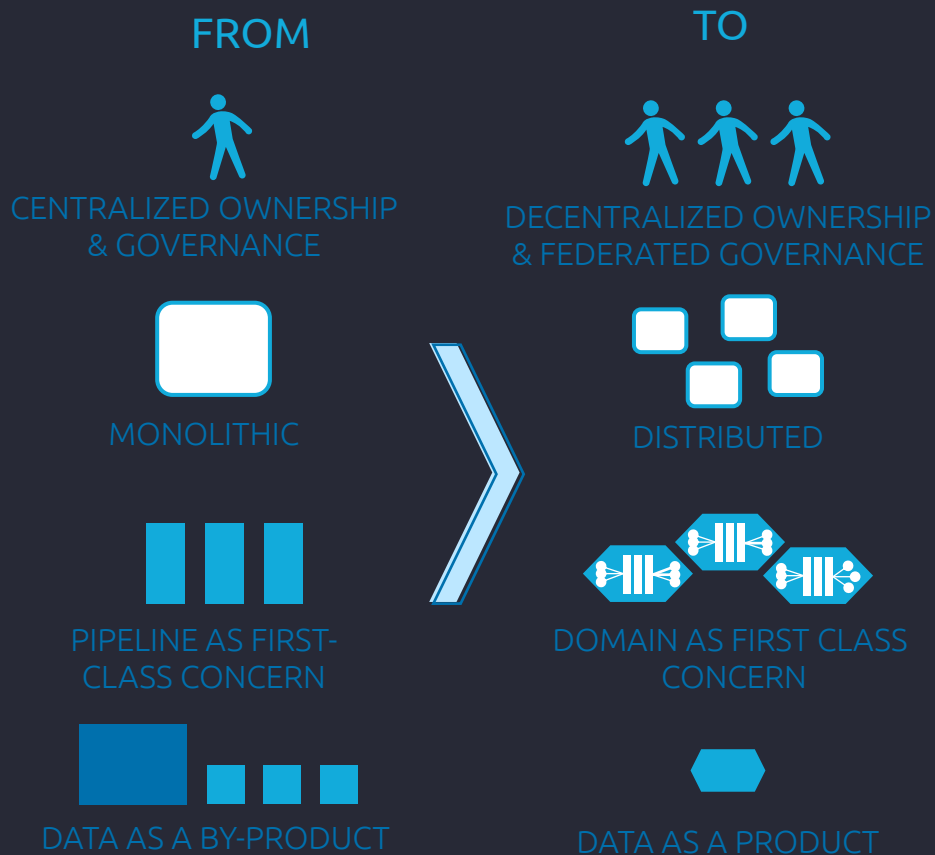
# SHORT INTRODUCTION TO DATA MESH





# DATA MESH ON A NAPKIN

*The sociotechnical approach to share, access and manage analytical data in complex and large-scale environments - within or across organizations.*



## Data Mesh Principles

### DOMAIN ORIENTED DECENTRALIZATION



Data Mesh essentially refers to the concept of breaking down data lakes and siloes into smaller, more decentralized portions. Much like the shift from monolithic applications toward microservices architectures in the world of software development, Data Mesh can be described as a data-centric version of microservices

### DATA AS A PRODUCT



A data product is a product that primarily uses data (e.g. Legacy Data) to contribute value to the achievement of an organization's objectives

### SELF-SERVE DATA INFRA AS A PLATFORM



Data self-service depends on the technology used and can be deployed, for example, as Docker, Kafka Service, or Spark code



# EVERY “ENTERPRISE” HAS IT’S “DOMAINS” I WANT TO NAVIGATE TO





# WITHIN A “DOMAIN” THE “DATA PRODUCT” ARE EXPOSED WITH CLEAR “PROPERTIES”



<https://unsplash.com/photos/TcpYjs6qF9o>



<https://unsplash.com/photos/Gk8LG7dsHWA>



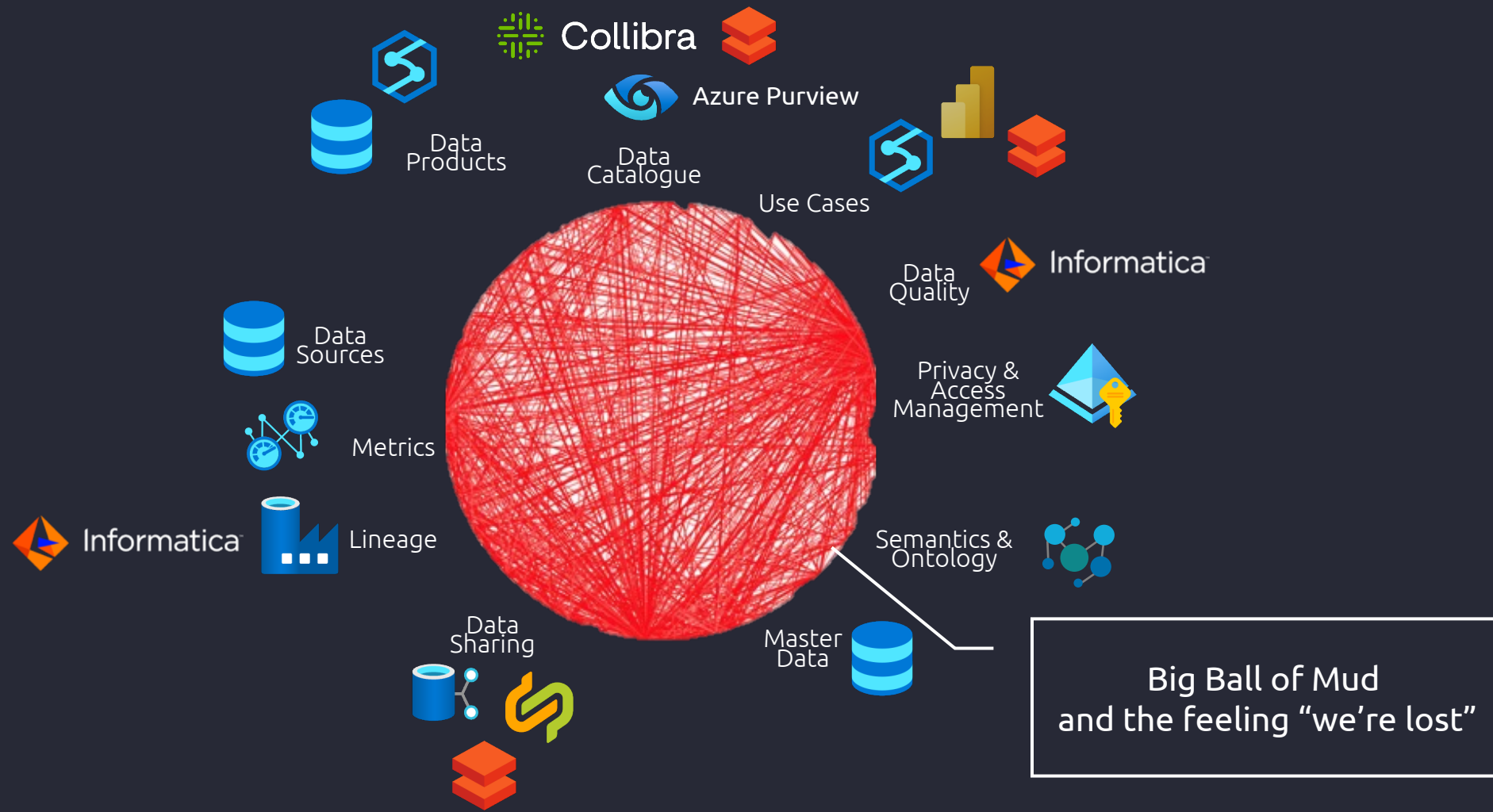


# WHY WE NEED A KNOWLEDGE GRAPH?



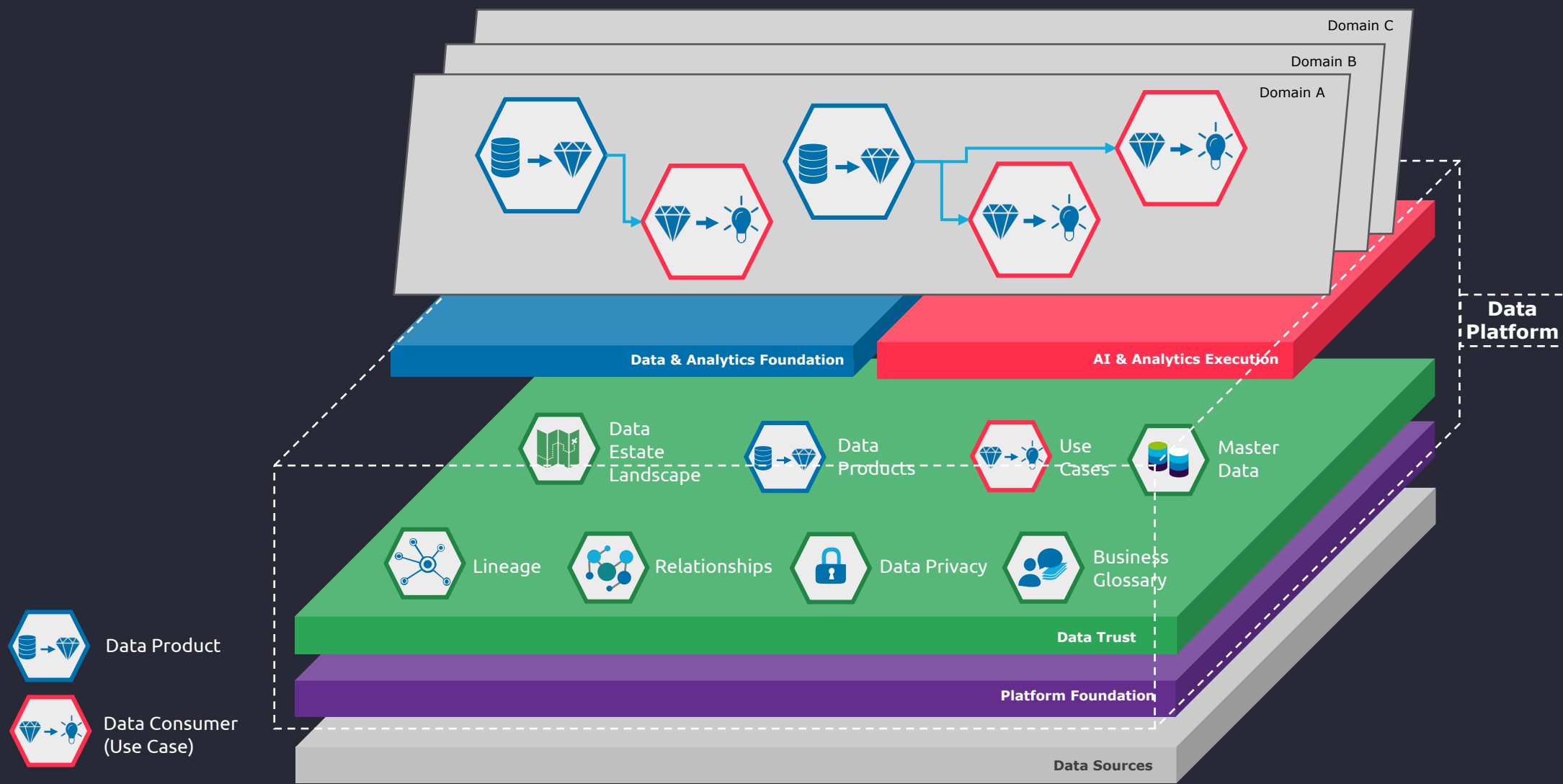


# AT ENTERPRISE LEVEL A MULTITUDE OF TECHNOLOGIES HAS TO BE COMPROMISED FOR DATA GOVERNANCE



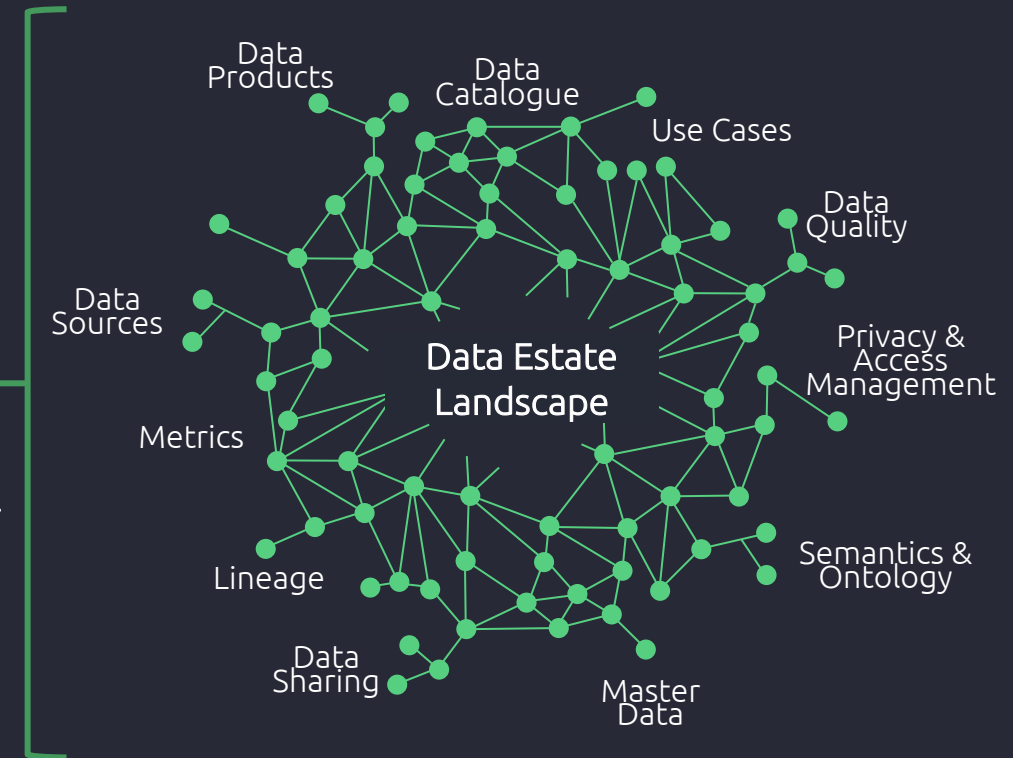
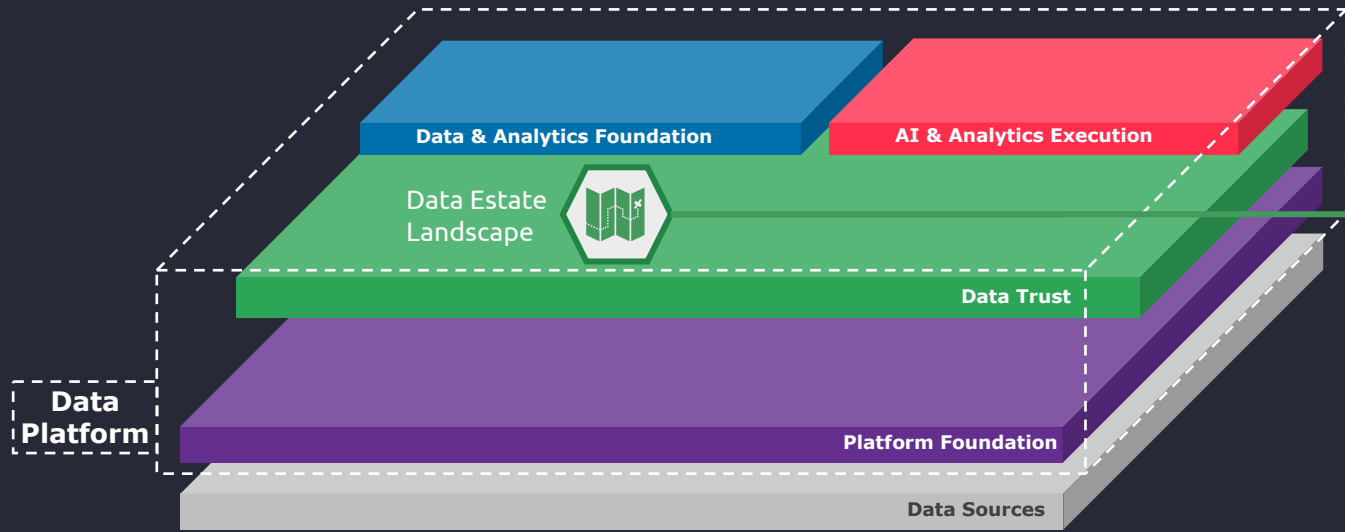


# THE DATA TRUST LAYERS ENABLES GOVERNANCE & TRANSPARENCY WITHIN THE ORGANIZATION





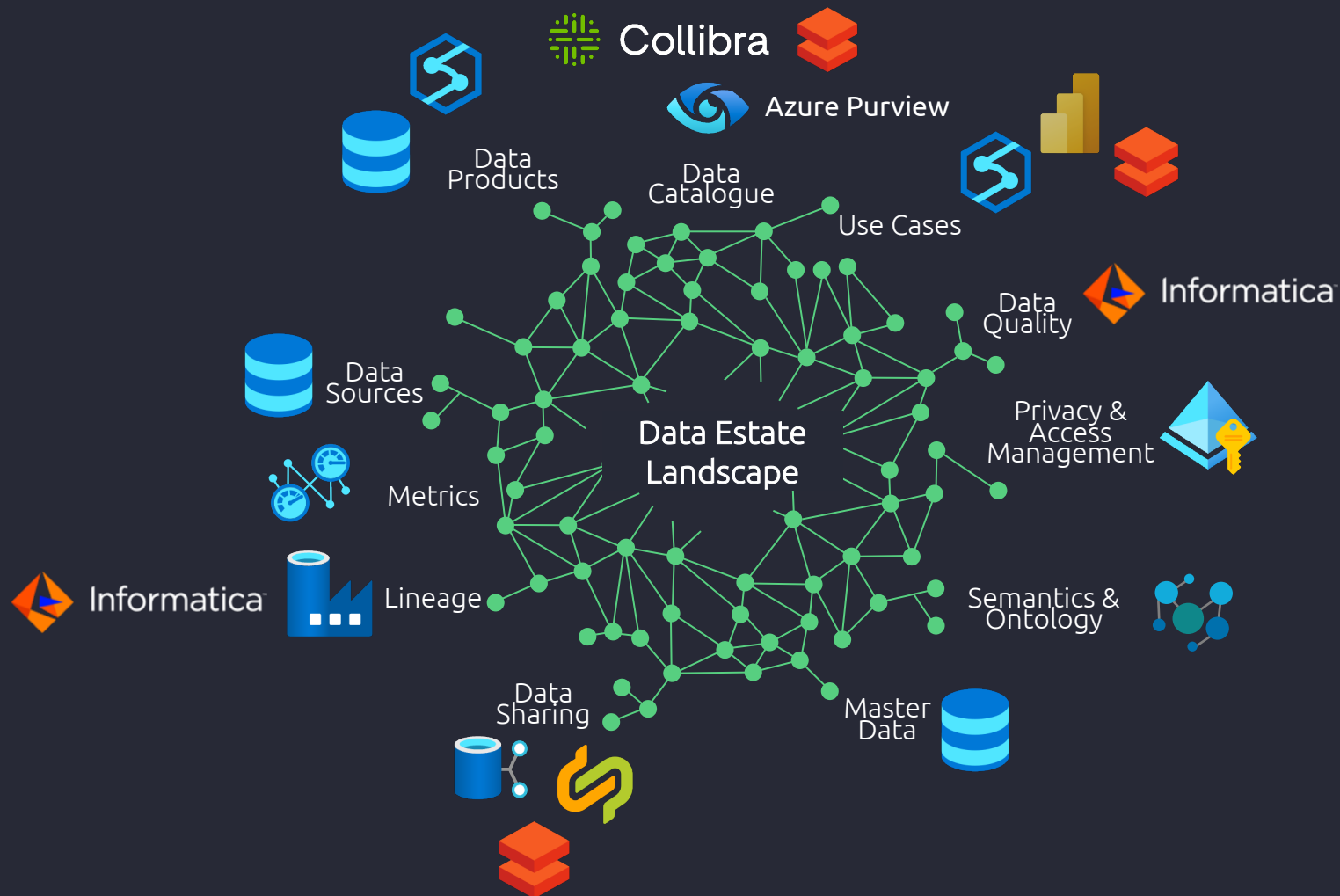
# A GRAPH BASED DATA ESTATE LANDSCAPE ADDRESSES THESE ISSUES AND PROVIDES TRANSPARENCY AND AGILITY



- The Data Estate Landscape brings transparency to the Data Platform Ecosystem by maintaining the relationship between Sources, Data Products and Use Cases
- This enables faster Use Case development and reuse of data and resources
- Establishment of the Data Estate Landscape ensures overall governance over the Data Ecosystem



# AT ENTERPRISE LEVEL THE DATA ESTATE LANDSCAPE HAS TO COMPROMISE A MULTITUDE OF TECHNOLOGIES

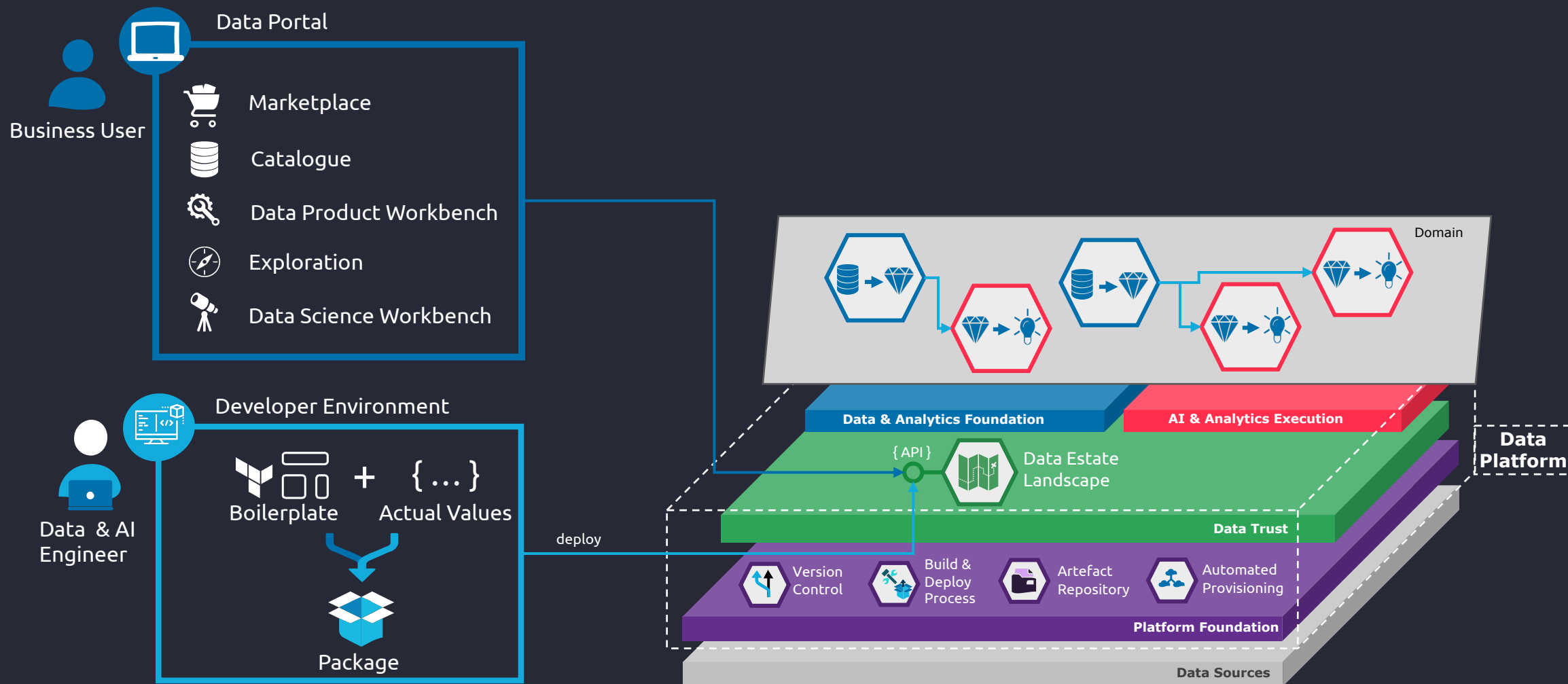




# HOW TO ENABLE BUSINESS WITH THAT TECHNOLOGY



# BY PROVIDING API ENDPOINTS THE DATA ESTATE LANDSCAPE SUPPORT DEVELOPERS AND BUSINESS USERS ON THEIR WORK





# THE DATA PORTAL ACTS AS THE ENTRY POINT FOR THE USERS

The screenshot shows a web browser window with the URL `https://data-portal.company.com`. The interface is divided into a left sidebar and a main content area.

**Left Sidebar:**

- Home icon
- Shopping cart icon
- Database icon
- Compass icon
- Telescope icon
- Gears icon
- Bell icon

**Main Content Area:**

**Introduction**

This is the Data Portal in which you can find:

- Marketplace
- Catalogue
- Data Product Workbench
- Exploration
- Data Science Workbench

**Tutorial**

With this tutorial you'll get a step-by-step guide on how to work within the Data Portal and find the data you're looking for. Click [here](#) to continue.

**Data Estate Landscape**

Domain A	Domain B	Domain C

The diagram illustrates a data flow from Domain A to Domain B and then to Domain C. It shows a sequence of data sources (cylinders) feeding into processing steps (hexagons) across three domains. Some hexagons are highlighted with blue borders, while others have red borders. Arrows indicate the direction of data flow between these components.





# THE DATA PORTAL ACTS AS THE ENTRY POINT FOR THE USERS

The screenshot displays a web browser window with the URL `https://data-portal.company.com`. The interface features a left-hand navigation sidebar with icons for Home, Shopping Cart, Database, Search, Telescope, Gear, and Notification. The main content area is titled "Search" and includes three dropdown filters: "Domain", "Environment", and "Refreshed recently".

Four data cards are displayed in a grid:

- Production Data** (blue hexagon icon):
  - Owner: Data Source: ENV:
  - Description:
  - Structure:
  - Rating: ★★★★★
- Car Master Data** (blue hexagon icon):
  - Owner: Data Source: ENV:
  - Description:
  - Structure:
  - Rating: ★★★★★
- Production Overview** (red hexagon icon):
  - Owner: Domain: Source:
  - Description:
  - Business Value:
  - Rating: ★★★★★
- Customer Master Data** (blue hexagon icon):
  - Owner: Data Source: ENV:
  - Description:
  - Structure:
  - Rating: ★★★★★



# THE DATA PORTAL ACTS AS THE ENTRY POINT FOR THE USERS

https://data-portal.company.com

Search

Domain

Environment

Type

### Production Data

Owner: Dagobert Duck

Data Source: MES Siemens

Type: Relational

ENV: DEV

Description: Lorem ipsum sit dolor

Metrics: Last loaded: 2022-04-12-21:30:00  
Load frequency: every hour  
Data quality: 100% matched records

Structure:

Lineage:

The interface displays a sidebar with navigation icons (Home, Dashboard, Data, Analytics, Reports, Settings, Notifications) and a search filter section. The main content area shows details for 'Production Data', including its owner, data source, type, and environment. It also provides a description, metrics (last loaded, load frequency, data quality), and visualizations for its structure and lineage. The structure diagram shows a hierarchy of tables, and the lineage diagram shows a flow from a database through three processing steps to an API endpoint.



# EXAMPLES OF DATA CATALOGUES

AMUNDSEN  Announcements Browse

test\_schema.test\_table1   
 Datasets • Hive • gold high quality pii

Airflow Preview

Description  
1st test table

Date Range  
From: Apr 22, 2017  
To: Sep 30, 2019

Frequent Users

Tags  
tag1 tag2

Owners  
chris@example.org  
roald.amundsen@example.org

col1	col1 description	string
col2	col2 description	string
col3	col3 description	string
col4	col4 description	string
col5	col5 description	float

Search Datasets, People, & more

Analytics Domains Users & Groups Policies

bigquery-public-data.covid19\_public\_forecasts.county\_14d

Reported at 24/06/2021 08:00:00 EST Normal None 0.02 - 11 months ago

Field	Description	Tags	Terms
county_fips_code (string)	5-digit unique identifier of the county		
county_name (string)	Full text name of the county		
state_name (string)	Full text name of the state in which a given county lies		
forecast_date (date)	Date of the forecast		
prediction_date (date)	Predicted date of the given metrics		
new_confirmed (number)	Predicted number of new confirmed cases on the prediction_date. This is not cumu... <a href="#">Read More</a>		
cumulative_confirmed (number)	Predicted number of cumulative deaths on the prediction_date. This is cumulative. <a href="#">Read More</a>		
new_confirmed_7day_rolling (number)	The seven day rolling average of new confirmed cases.		
new_deaths (number)	Predicted number of new deaths on the prediction_date. This is cumulative over 1... <a href="#">Read More</a>		
cumulative_deaths (number)	Predicted number of cumulative confirmed cases on the prediction_date. This is... <a href="#">Read More</a>		
new_deaths_7day_rolling (number)	The seven day rolling average of new confirmed cases.		
hospitalized_patients (number)	Predicted number of people hospitalized on the prediction_date. This is not cumu... <a href="#">Read More</a>		
recovered (number)	Predicted number of people documented as recovered on the prediction_date. This... <a href="#">Read More</a>		
new_confirmed_ground_truth (number)	Actual number of new confirmed cases according to the ground truth data. This is... <a href="#">Read More</a>		
cumulative_confirmed_ground_truth (number)	Actual number of cumulative confirmed cases according to the ground truth data... <a href="#">Read More</a>		
new_deaths_ground_truth (number)	Actual number of new deaths according to the ground truth data. This is not cumu... <a href="#">Read More</a>		
cumulative_deaths_ground_truth (number)	Actual number of cumulative deaths according to the ground truth data. This is c... <a href="#">Read More</a>		
hospitalized_patients_ground_truth (number)	Actual number of people hospitalized according to the ground truth data. This is... <a href="#">Read More</a>		
recovered_documented_ground_truth (number)	Actual number of people hospitalized according to the ground truth data		
county_population (number)	Total population of the county		

About  
This predicts the value for key metrics for COVID-19 impacts over...

Tags  
+ Add Tag

Glossary Terms  
+ Add Term

Owners  
+ Add

Domain  
Set Domain



# EXAMPLES OF DATA CATALOGUES

customer data.csv

#	Name	Data Classification	Is Primary Key	Data Type	represented by	Empty Values
2	Address	Street address 97%		Text		0
3	City			Text		0
5	Country	Country 98%		Text		0
6	Email	Email 96%		Text		0
10	First Name	First name 90%		Text		0
7	Gender	Gender 75%		Text		0
8	Job Title			Text		0
9	Language	Ethnicity		Text		0
11	Last Name	Last name		Text		0
1	Name	Full name 65%		Text		0
4	State	US state 70%		Text		0

DATA LINEAGE

Hospital Info  
hospital\_info

1 Warning 1 Endorsement

Overview Columns 24 Samples 100 Filters 2 Joins 0 Lineage Queries 0

Diagram showing data lineage from source tables (e.g., summ\_top\_drg, drg, provider\_id, name, provider\_str..., provider\_city, provider\_state, provider\_zip..., hospital\_ref..., total\_discha..., average\_cove...) to target tables (e.g., hospital\_info, hospital\_gen..., clean.hsptl, clean.clms\_rw, hospital\_inf...).



# EXAMPLES OF DATA CATALOGUES

The screenshot shows the Microsoft Purview Data Catalog interface. At the top, there's a navigation bar with "Microsoft Azure", "Purview", and "Adatum Corp". A search bar contains the word "Revenue". Below the navigation, there's a "Sources" section with options to "Register", "New collection", and "Refresh". It indicates "Showing 5 collections, 1134 sources". The main area is a grid of data source cards, each with a "View details" link. The cards are organized into five regional collections: North America, Europe, Azure and North America, Amazon North America, and Azure Europe. Each collection contains various data sources like SQL Servers, SAP systems, Azure Data Lake Storage, Amazon S3, and Teradata.

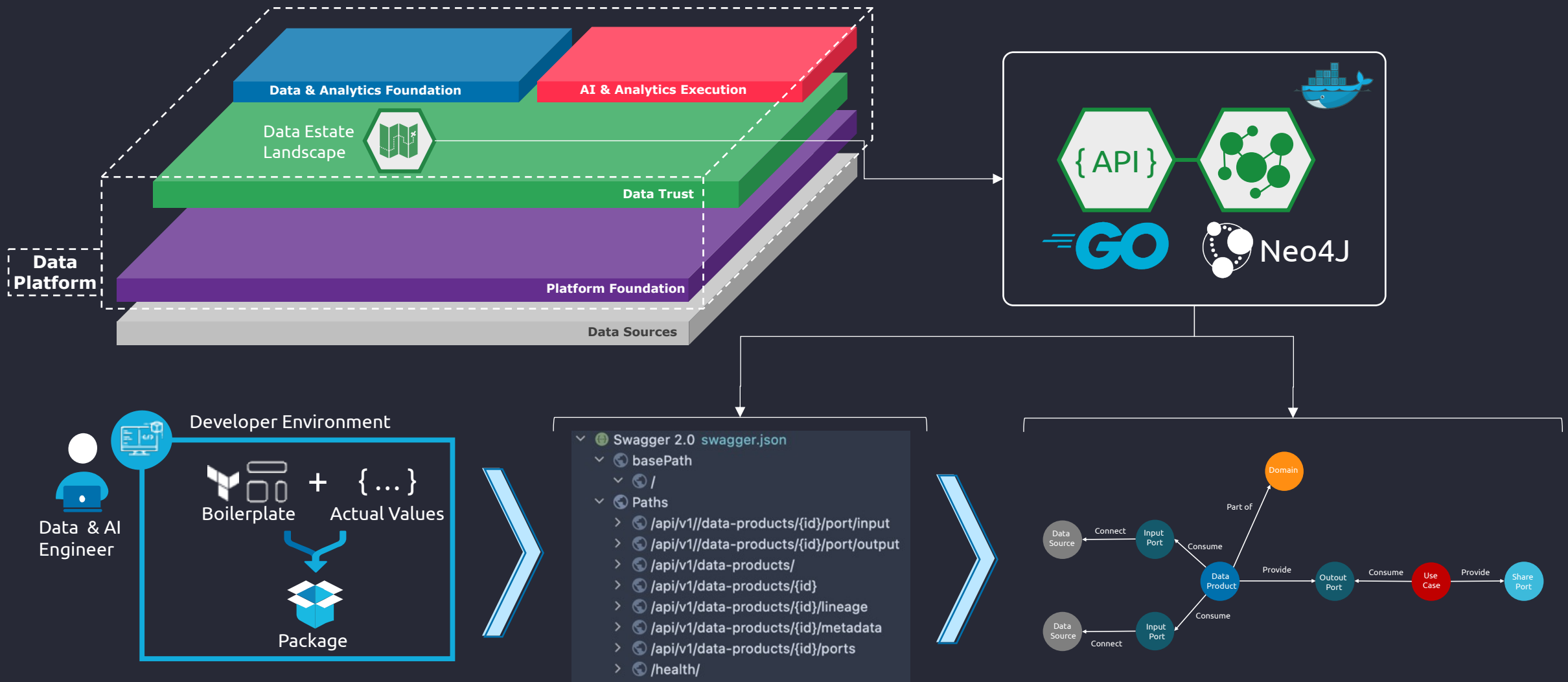
The screenshot shows the Apache Atlas interface. At the top, there's a search bar and navigation tabs for "SEARCH", "CLASSIFICATION", and "GLOSSARY". Below the navigation, there's a "Create Business Metadata" form. The form has a "Name" field with the value "AcmeDataAssetManagement" and a "Description" field with the value "Attributes to manage data assets of Acme". There's a "+ Add Business Metadata attribute" button. Below this, there are several fields for defining the metadata attribute: "Name" (assetID), "Type" (string), "Search Weight" (5), "Enable Multivalues" (checkbox), and "Max length" (50). The "Applicable Types" field contains several tags: \*adis\_gen2\_container, \*aws\_s3\_bucket, \*hbase\_table, \*hive\_table, \*kafka\_topic, and \*rdbms\_table. At the bottom, there are "Cancel" and "Create" buttons.



# EXEMPLARY IMPLEMENTATION

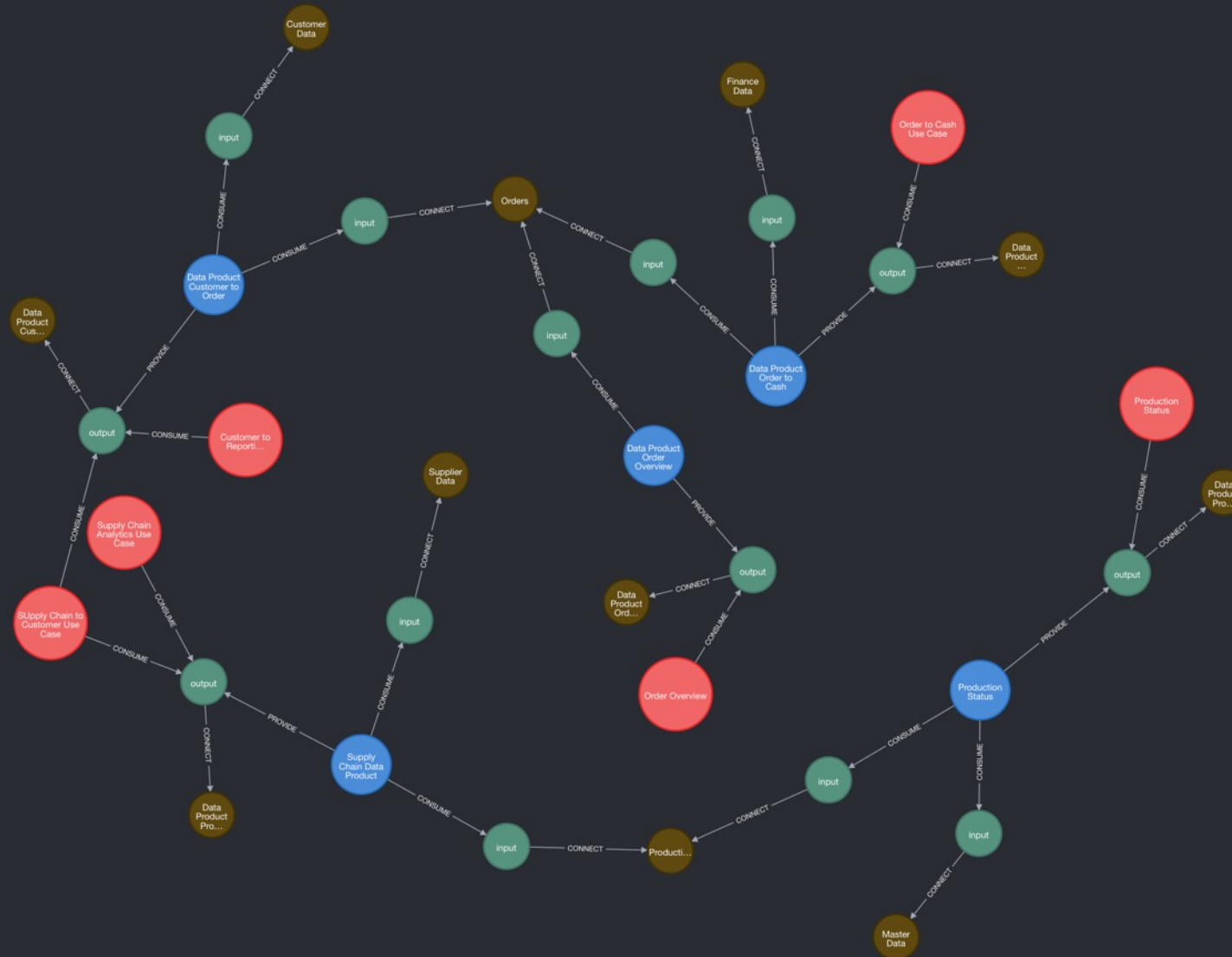


# A GRAPH BASED DATA ESTATE LANDSCAPE ADDRESSES THESE ISSUES AND PROVIDES TRANSPARENCY AND AGILITY





# KNOWLEDGE GRAPH OF EXEMPLARY DATA MESH



**Node labels**

- \* (36)
- DataProduct (5)
- Source (11)
- Port (14)
- UseCase (6)

**Relationship Types**

- \* (35)
- PROVIDE (5)
- CONSUME (16)
- CONNECT (14)





# GETTING ALL INVOLVED DATA SOURCES FOR SELECTED USE CASE

```
neo4j$ match(u:UseCase)--(p:Port)--(d:DataProduct)--(ip:Port)--(s:Source) WHERE u.name='Supply Chain to Customer Use Case' return s
```

Graph visualization showing four nodes: Supplier Data, Orders, Customer Data, and Product Data.

```
"s"  
{"name":"Orders","description":"All Orders from Customers","id":"ds0006","type":"mqtt","url":"mqtt://server/orders"}  
{"name":"Customer Data","description":"Customer Data","id":"ds0003","type":"jdbc","url":"jdbc://server/database/customer"}  
{"name":"Supplier Data","description":"All Supplier information","id":"ds0005","type":"jdbc","url":"jdbc://server/supplier_information"}  
{"name":"Production Data","description":"Production Data from the plants","id":"ds0002","type":"mqtt","url":"mqtt://server/production"}
```



# GETTING ALL DATA PRODUCTS AND USE CASES OF DATA SOURCE

neo4j\$ `match(u:UseCase)--(p:Port)--(d:DataProduct)--(ip:Port)--(s:Source) WHERE s.name='Orders' return u,d`

Graph

Table

Text

Code

Overview

Node labels

(7) UseCase (4) DataProduct (3)

Displaying 7 nodes, 0 relationships.

"u"	"d"
<code>{"owner":"Dagobert Duck","createdDate":"2022-03-07 10:00:00","deputy":"Donald Duck","name":"Order Overview","description":"Having the overview on orders","lastModified":"2022-03-07 10:00:00","id":"uc00006","version":"1.0.0"}</code>	<code>{"owner":"Darth Sidious","createdDate":"2022-03-14 10:00:00","collectedSince":"2022-03-14 10:00:00","name":"Data Product Order Overview","description":"Data Product for Order Overview","lastModified":"2022-03-07 10:00:00","id":"dp00005","version":"1.0.0"}</code>
<code>{"owner":"Dagobert Duck","createdDate":"2022-03-07 10:00:00","deputy":"Donald Duck","name":"Order to Cash Use Case","description":"Making the CFO happy","lastModified":"2022-03-07 10:00:00","id":"uc00005","version":"1.0.0"}</code>	<code>{"owner":"Darth Sidious","createdDate":"2022-03-14 10:00:00","collectedSince":"2022-03-14 10:00:00","name":"Data Product Order to Cash","description":"Data Product for making orders to cash transparent","lastModified":"2022-03-07 10:00:00","id":"dp00004","version":"1.0.0"}</code>
<code>{"owner":"Dagobert Duck","createdDate":"2022-03-07 10:00:00","deputy":"Donald Duck","name":"Customer to Order Reporting","description":"Customer to Order","lastModified":"2022-03-07 10:00:00","id":"uc00004","version":"1.0.0"}</code>	<code>{"owner":"Sith Loard","createdDate":"2022-03-14 10:00:00","collectedSince":"2022-03-14 10:00:00","name":"Data Product Customer to Order","description":"Data Product for connecting Customer to Orders","lastModified":"2022-03-07 10:00:00","id":"dp00003","version":"1.0.0"}</code>
<code>{"owner":"Dagobert Duck","createdDate":"2022-03-07 10:00:00","deputy":"Donald Duck","name":"Supply Chain to Customer Use Case","description":"Connecting Supply Chain and Customer Data","lastModified":"2022-03-07 10:00:00","id":"uc00003","version":"1.0.0"}</code>	<code>{"owner":"Sith Loard","createdDate":"2022-03-14 10:00:00","collectedSince":"2022-03-14 10:00:00","name":"Data Product Customer to Order","description":"Data Product for connecting Customer to Orders","lastModified":"2022-03-07 10:00:00","id":"dp00003","version":"1.0.0"}</code>

# CONCLUSIONS



# DATA MESH WITHOUT DATA GOVERNANCE

Will end up in muddy games



<https://www.dodgycoder.net/2013/01/big-ball-of-mud-design-pattern.html>



**GET THE  
FUTURE  
YOU WANT**



## About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of over 325,000 team members more than 50 countries. With its strong 55-year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2021 global revenues of €18 billion.

Get The Future You Want | [www.capgemini.com](http://www.capgemini.com)



This presentation contains information that may be privileged or confidential and is the property of the Capgemini Group.

Copyright © 2022 Capgemini. All rights reserved.