



# 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  
Todays 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





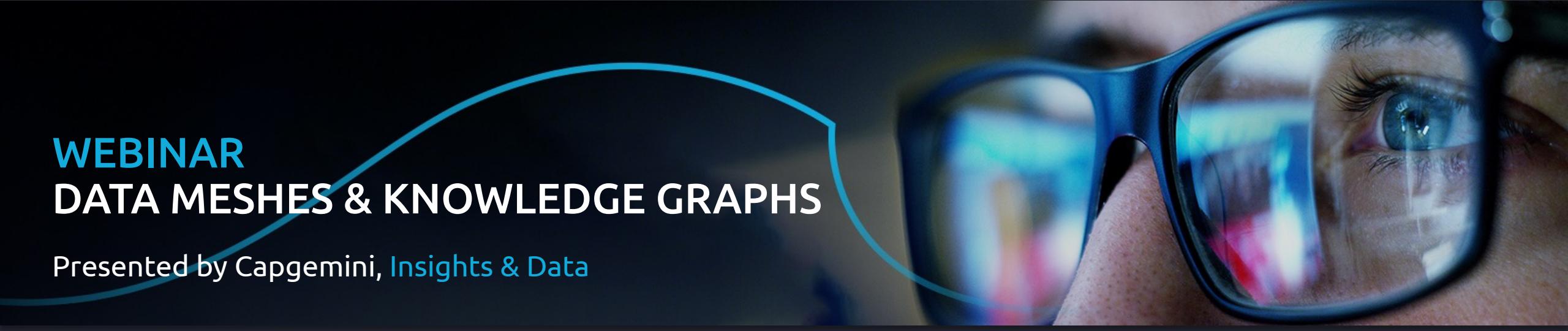
# DATA MESHES AND KNOWLEDGE GRAPHS

## 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*

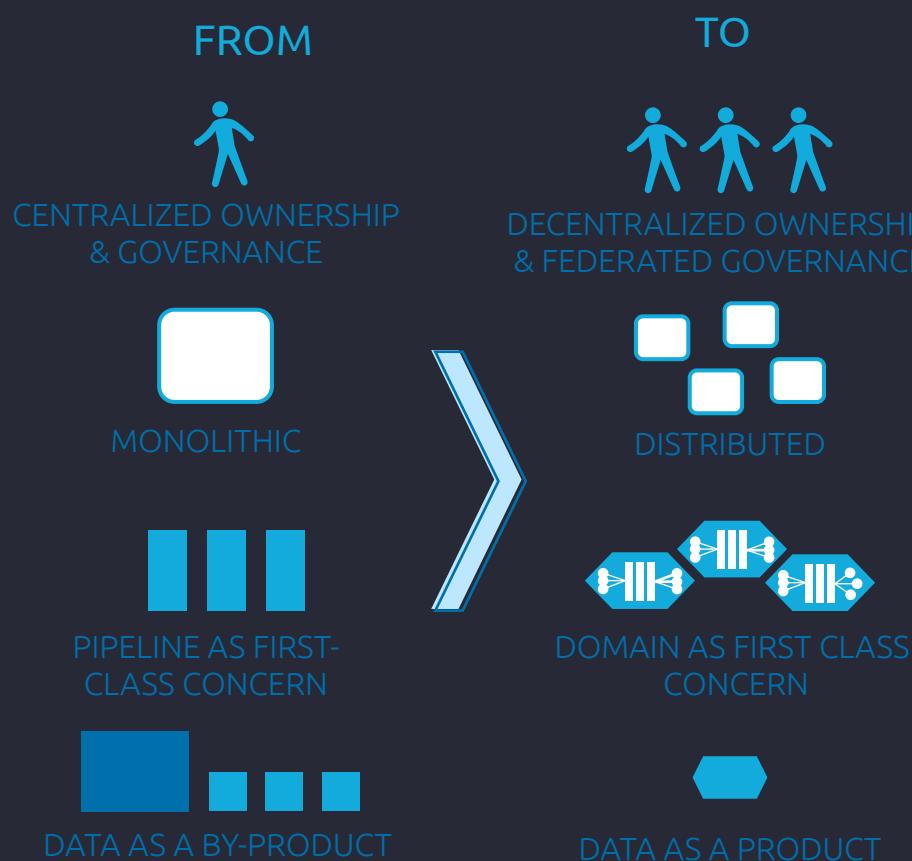


# SHORT INTRODUCTION TO DATA MESH



# 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



<https://unsplash.com/photos/AG0iadQfRDg>



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



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



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

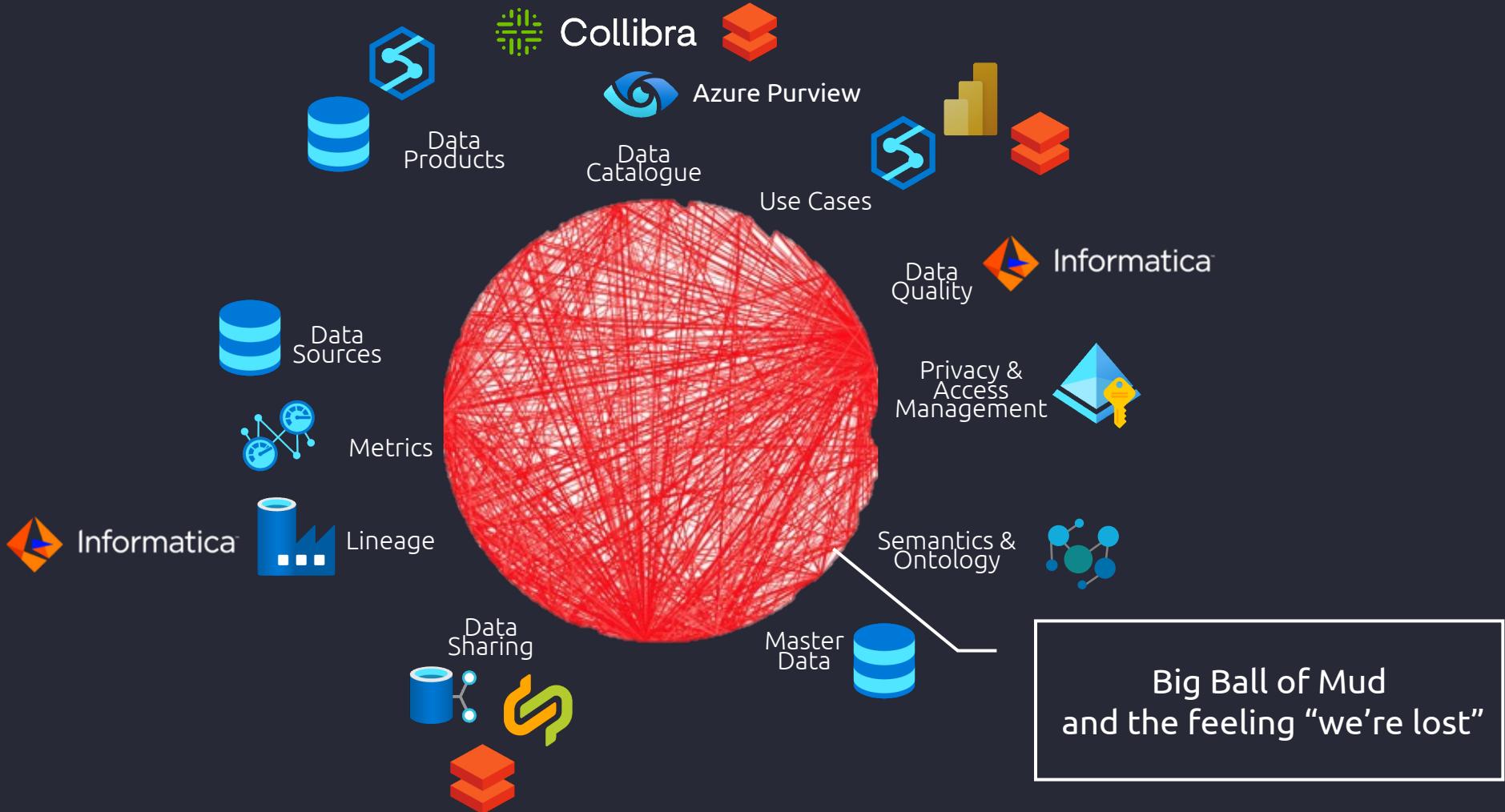


# WHY WE NEED A KNOWLEDGE GRAPH?



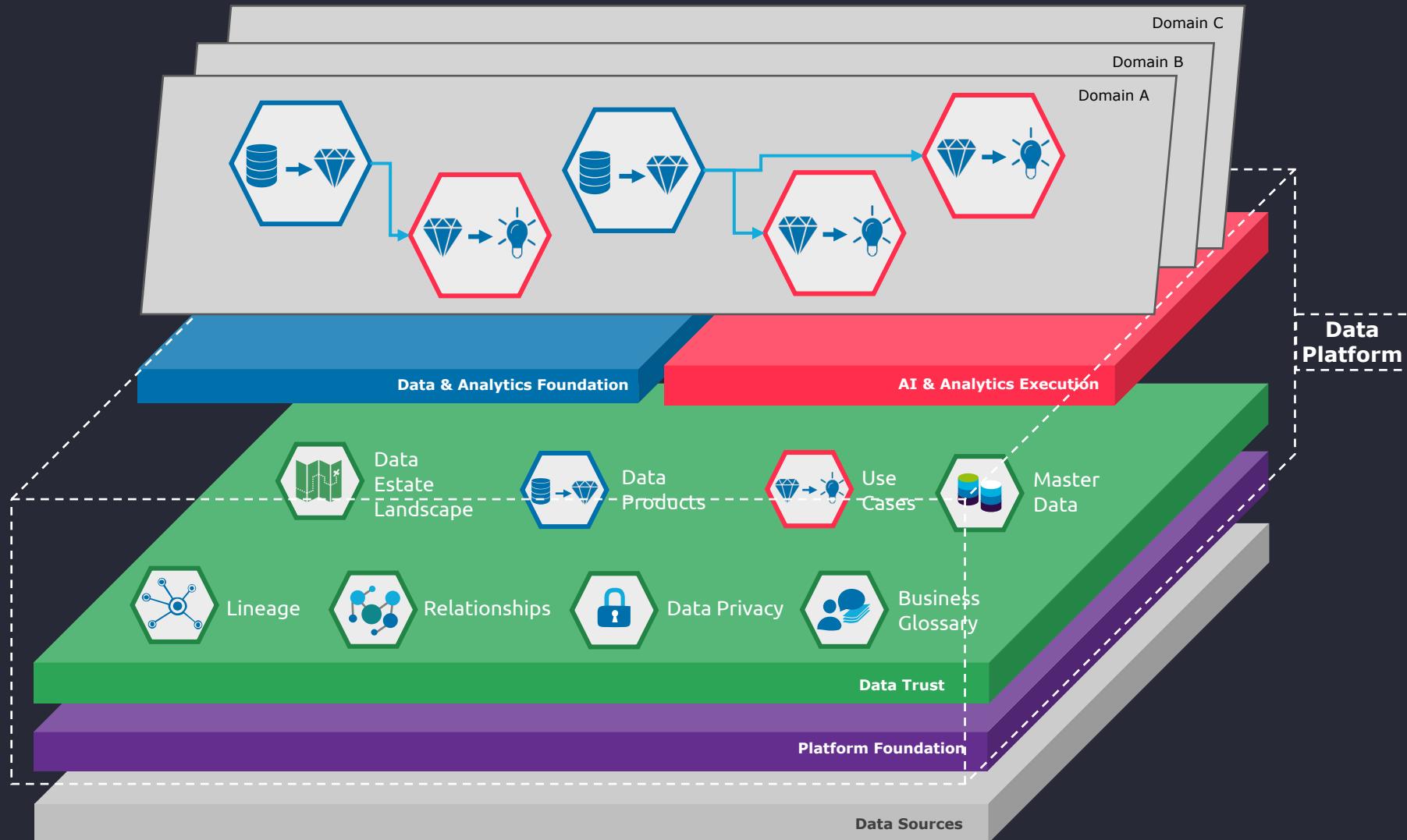


# AT ENTERPRISE LEVEL A MULTITUDE OF TECHNOLOGIES HAS TO BE COMPROMISED FOR DATA GOVERNANCE



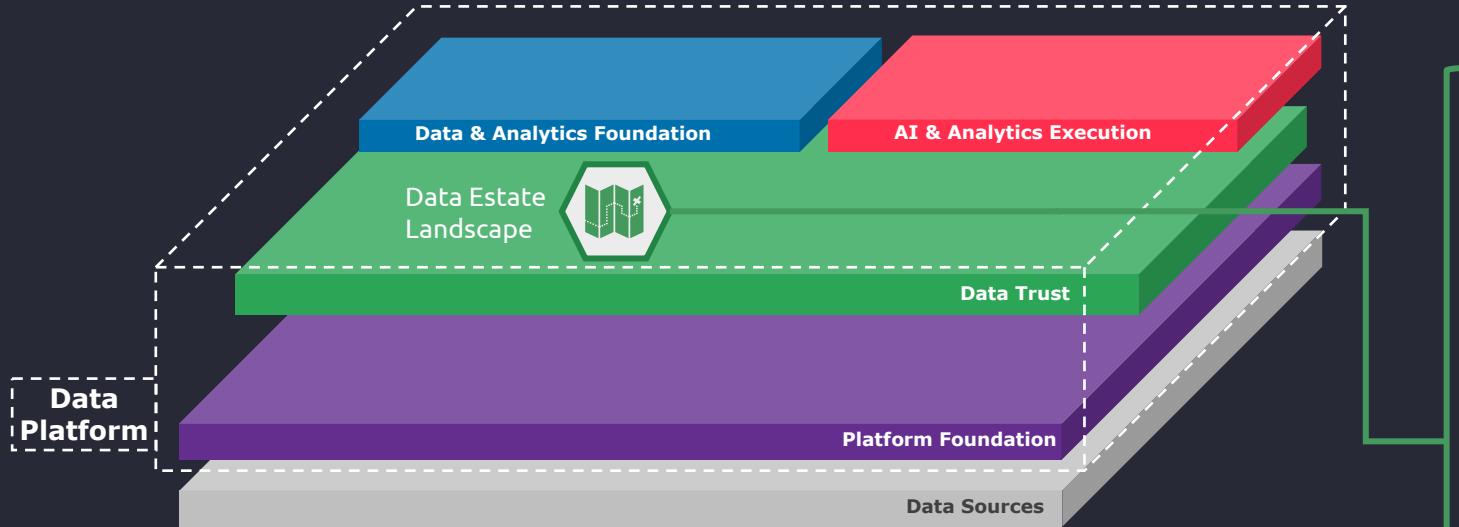


# THE DATA TRUST LAYERS ENABLES GOVERNANCE & TRANSPARENCY WITHIN THE ORGANIZATION

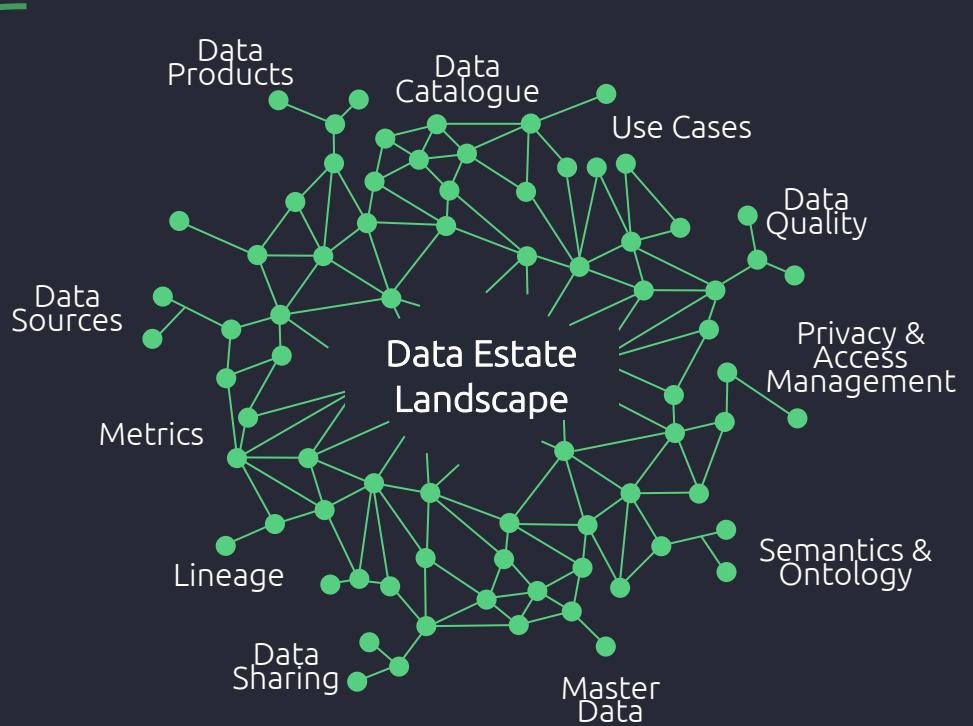




# A GRAPH BASED DATA ESTATE LANDSCAPE ADDRESSES THESE ISSUES AND PROVIDES TRANSPARENCY AND AGILITY

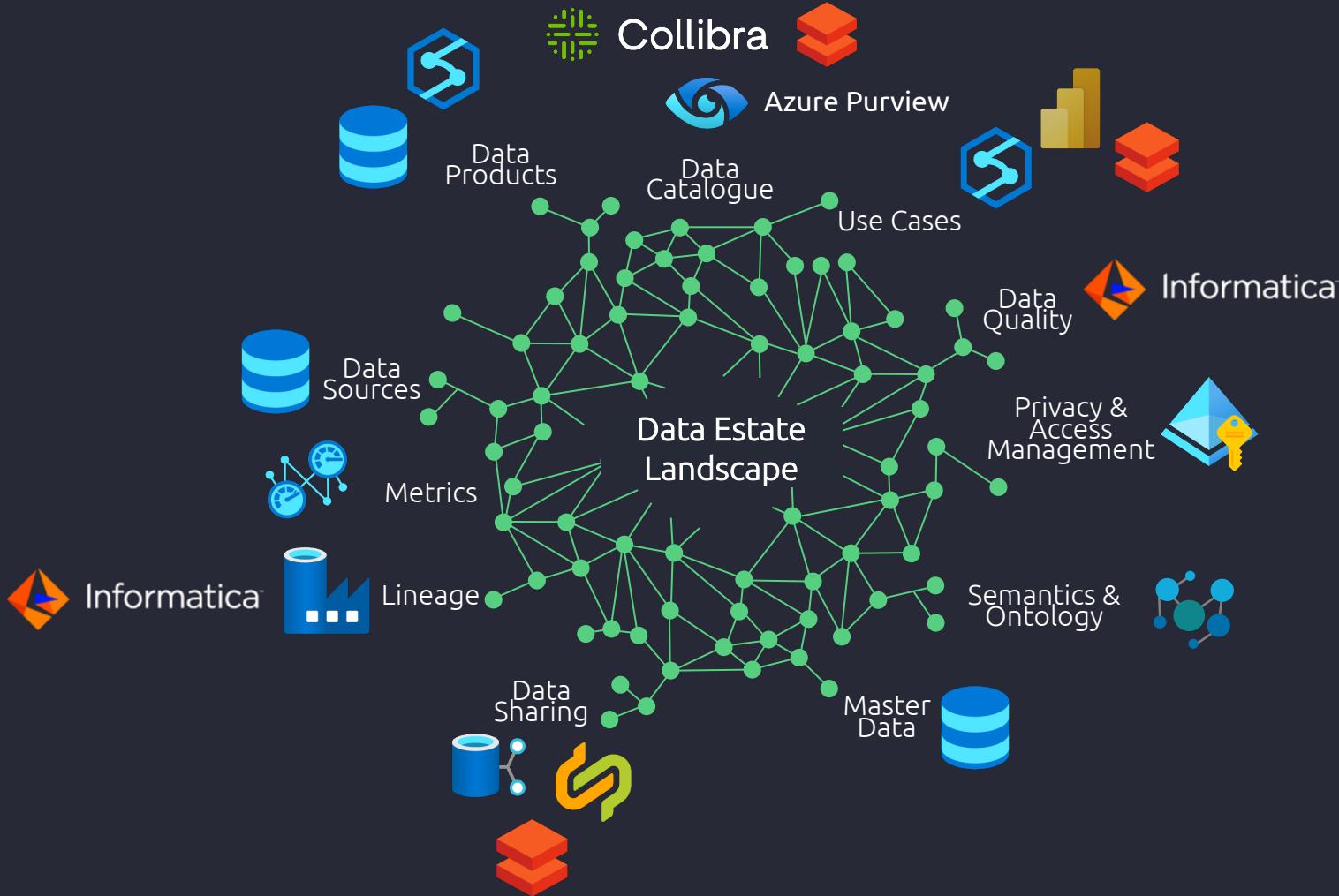


- The Data Estate Landscape brings transparency to the Data Platform Ecosystem by maintaining the relationship between Sources, Data Products and Use Cases
- This enables faster Use Case development and reuse of data and resources
- Establishment of the Data Estate Landscape ensures overall governance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





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



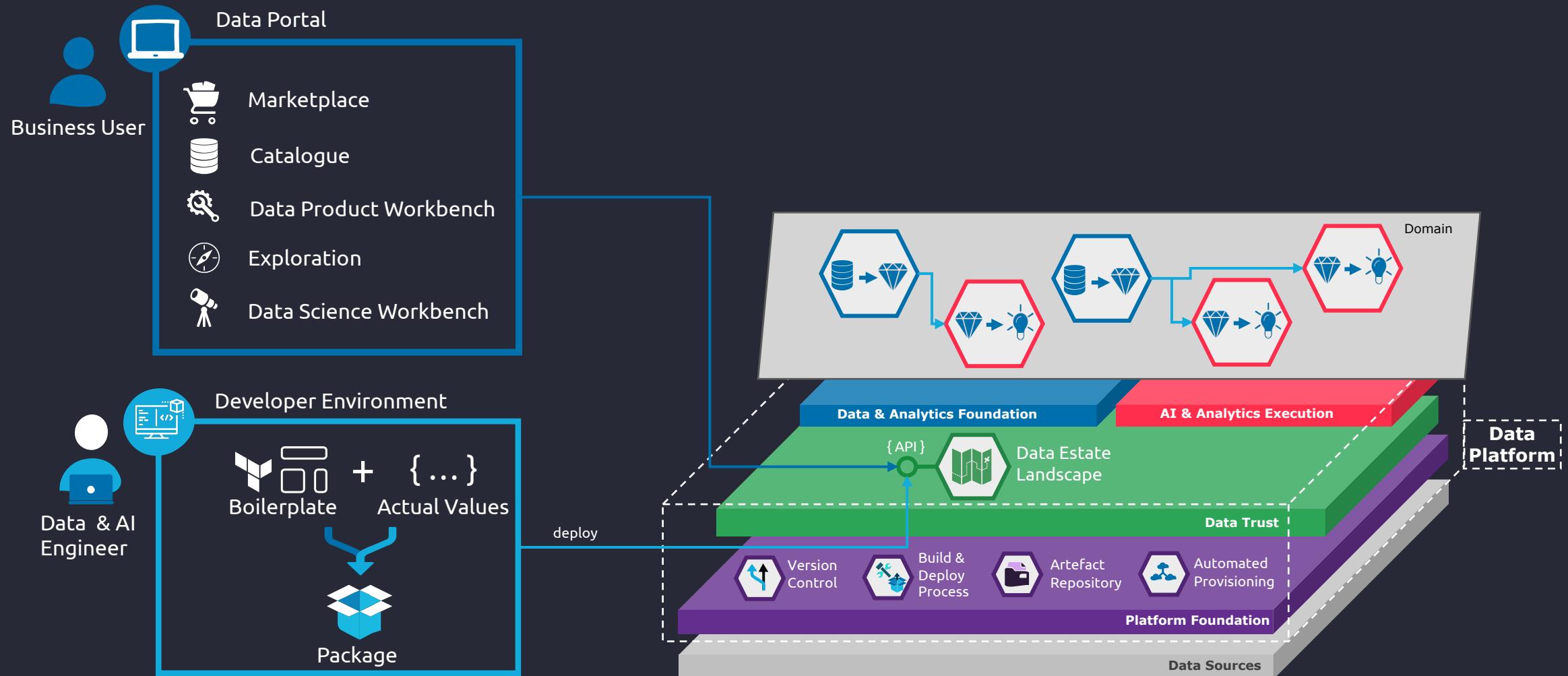


HOW TO ENABLE  
BUSINESS WITH  
THAT  
TECHNOLOGY





# 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

https://data-portal.company.com

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

The diagram illustrates the Data Estate Landscape across three domains: Domain A, Domain B, and Domain C. The landscape is represented by a network of hexagonal nodes connected by arrows. There are two main clusters of nodes. One cluster is located in Domain A, containing four nodes. Another cluster is located in Domain B, containing five nodes. These two clusters are interconnected by a single arrow pointing from the top node of Domain A to the bottom node of Domain B. Additionally, there are two isolated nodes in Domain C, one above the other, which are also interconnected by a single arrow pointing from the top node to the bottom node.



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

The screenshot shows a dark-themed Data Portal interface with a sidebar and a main content area.

**Left Sidebar:**

- Three circular icons at the top.
- Home icon (blue)
- Cart icon (blue, selected)
- Dash icon (grey)
- Telescope icon (grey)
- Gears icon (grey)
- Bell icon (grey) at the bottom.

**Top Bar:**

<https://data-portal.company.com>

**Main Content Area:**

- Production Data** (Blue Hexagon icon):
  - Owner: [redacted]
  - Data Source: [redacted]
  - ENV: [redacted]
  - Description: [redacted]
  - Structure: [redacted]
  - ★★★★★ (Rating)
- Car Master Data** (Blue Hexagon icon):
  - Owner: [redacted]
  - Data Source: [redacted]
  - ENV: [redacted]
  - Description: [redacted]
  - Structure: [redacted]
  - ★★★★★ (Rating)
- Production Overview** (Red Hexagon icon):
  - Owner: [redacted]
  - Domain: [redacted]
  - Source: [redacted]
  - Description: [redacted]
  - Business Value: [redacted]
  - ★★★★★ (Rating)
- Customer Master Data** (Blue Hexagon icon):
  - Owner: [redacted]
  - Data Source: [redacted]
  - ENV: [redacted]
  - Description: [redacted]
  - Structure: [redacted]
  - ★★★★★ (Rating)



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

The screenshot shows a dark-themed Data Portal interface with a sidebar and a main content area.

**Left Sidebar:** Contains icons for Home, Shopping Cart, Databases (highlighted in blue), Location, Telescopium, Magnifying Glass, and Bell.

**Header:** Shows the URL <https://data-portal.company.com>.

**Main Content - Production Data:**

- Owner:** Dagobert Duck
- Data Source:** MES Siemens
- Type:** Relational
- ENV:** DEV

**Description:** Lorem ipsum sit dolor

**Metrics:**

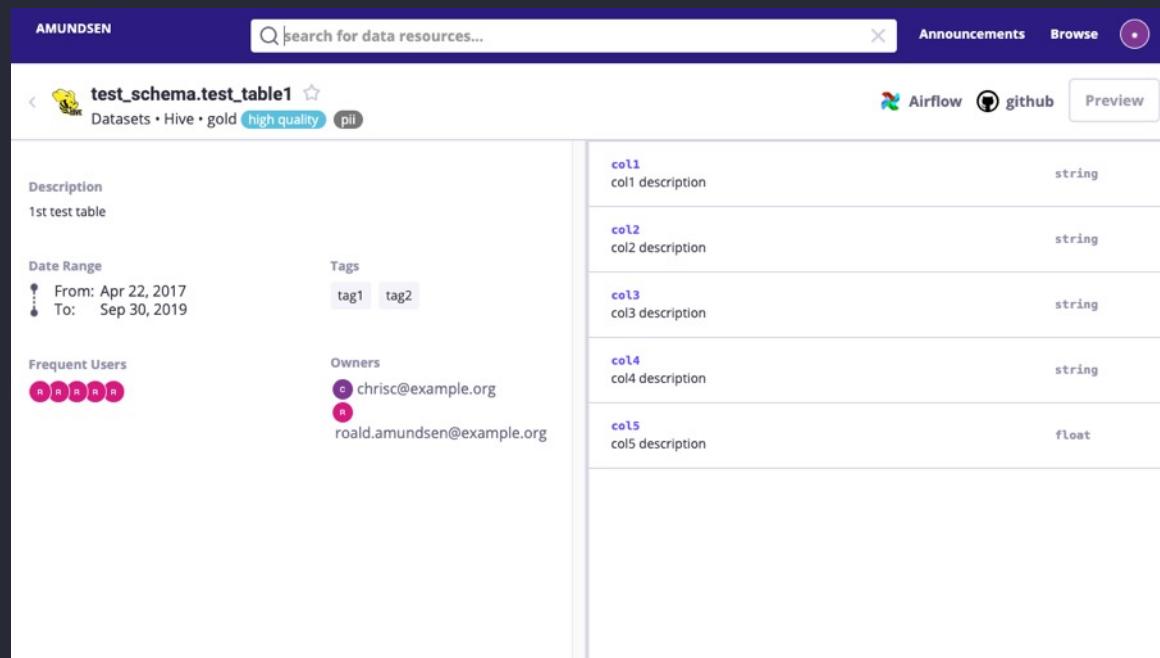
- Last loaded: 2022-04-12-21:30:00
- Load frequency: every hour
- Data quality: 100% matched records

**Structure:** A hierarchical tree diagram showing data tables branching from a central root node.

**Lineage:** A diagram illustrating the flow of data from a central database source through various processing steps (represented by gears) to an API endpoint.



# EXAMPLES OF DATA CATALOGUES



**AMUNDSEN**

test\_schema.test\_table1 ★

Datasets · Hive · gold · high quality · pii

Description  
1st test table

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

Frequent Users

Owners  
chrisc@example.org · roald.amundsen@example.org

Tags  
tag1 · tag2

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

**BigQuery Dataset: bigquery-public-data.covid19\_public\_forecasts.county\_14d**

Schema · Documentation · Properties · Lineage · Queries · Stats · Validation

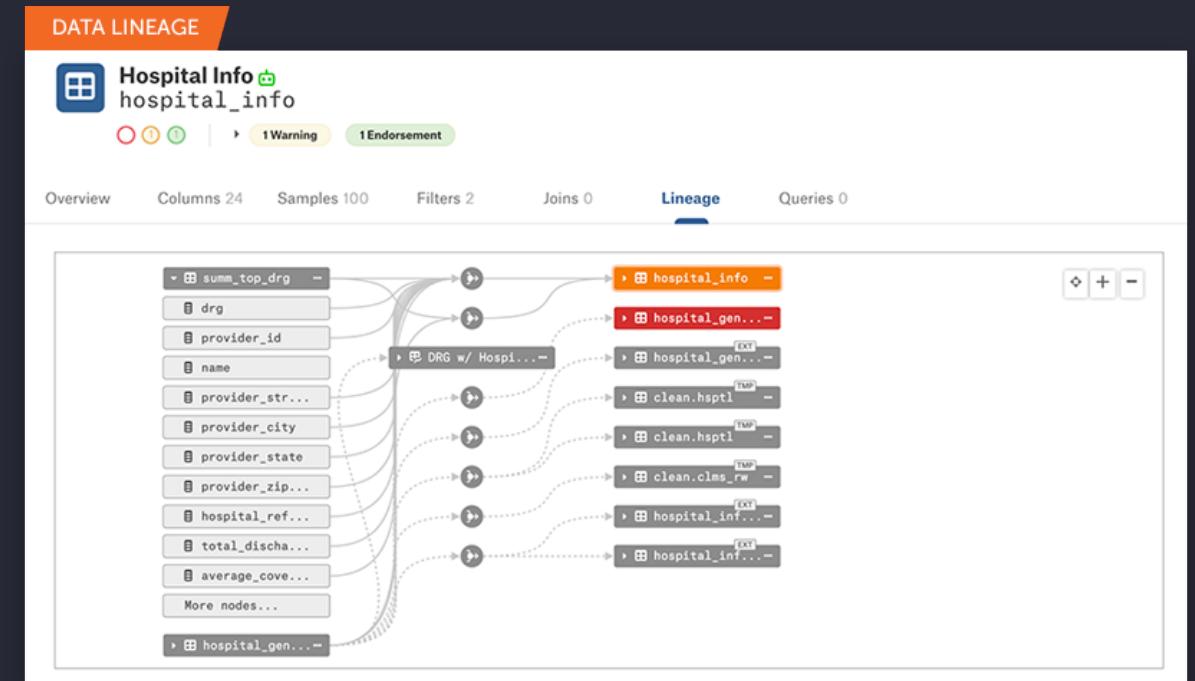
Reported at 24/08/2021, 08:00 CEST

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



# EXAMPLES OF DATA CATALOGUES

The screenshot shows a data catalog interface with a top navigation bar featuring 'Create', 'Search', and '1 task'. Below the navigation is a breadcrumb trail: Business Analysts Community > Schemas > Customer Data. The main area displays a table titled 'customer data.csv' with 11 columns. The columns are: #, Name, Data Classification, Is Primary Key, Data Type, represented by, and Empty Values. The 'Data Classification' column includes percentages such as 97% for Street address and 98% for Country. The 'Is Primary Key' column shows 'Address' as the primary key. The 'Data Type' column shows all entries as Text. The 'represented by' column contains placeholder icons. The 'Empty Values' column shows counts of 0 for all rows. On the left sidebar, there are sections for Summary, Details, Columns (selected), Sample data, Diagram, Pictures, Responsibilities, References, History, and Files. A modal window is open over the table, showing a row for 'Last Name' with a dropdown menu containing 'Ethnicity' and 'Accept' buttons.





# EXAMPLES OF DATA CATALOGUES

This screenshot shows the Microsoft Azure Purview Data Catalog interface. At the top, there's a search bar with the query "Revenue" and a user profile for "contoso@contoso.com". Below the header, the "Sources" section displays a grid of data collections. There are five main collections: "NorthAmericaDataCenter", "EuropeDataCenter", "AzureAndBINorthAmerica", "AmazonNorthAmerica", and "AzureEurope". Each collection contains several data sources, such as "OnPremSQLServer-FinData", "SAP-S4HANA-Procurement", "AzureDataLakeStorage", "AmazonS3-HRData", and "AzureDataLakeStorage". Each source item has a "View details" button.

This screenshot shows the Apache Atlas interface for creating business metadata. The title bar says "Create Business Metadata". The form includes fields for "Name" (set to "AcmeDataAssetManagement"), "Description" (set to "Attributes to manage data assets of Acme"), and "Add Business Metadata attribute". A detailed configuration panel is open, showing fields like "assetID" (Type: string, Search Weight: 5), "assetName" (Type: string, Search Weight: 5), and "assetType" (Type: string, Search Weight: 5). There are also sections for "Enable Multivalues", "Max length", and "Applicable Types" (listing "adls\_gen2\_container", "aws\_s3\_bucket", "hbase\_table", "hive\_table", "kafka\_topic", and "rdbms\_table"). At the bottom right are "Cancel" and "Create" buttons.

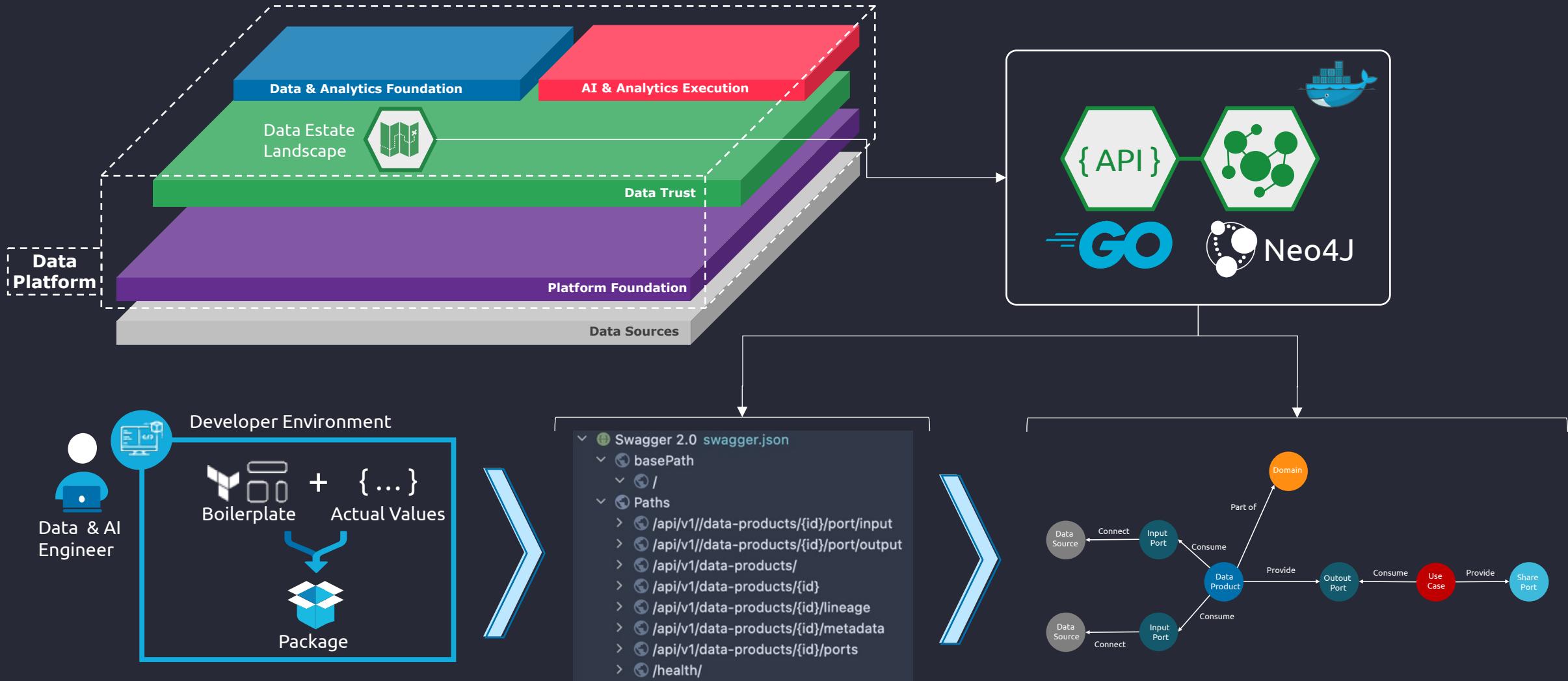


# EXEMPLARY IMPLEMENTATION



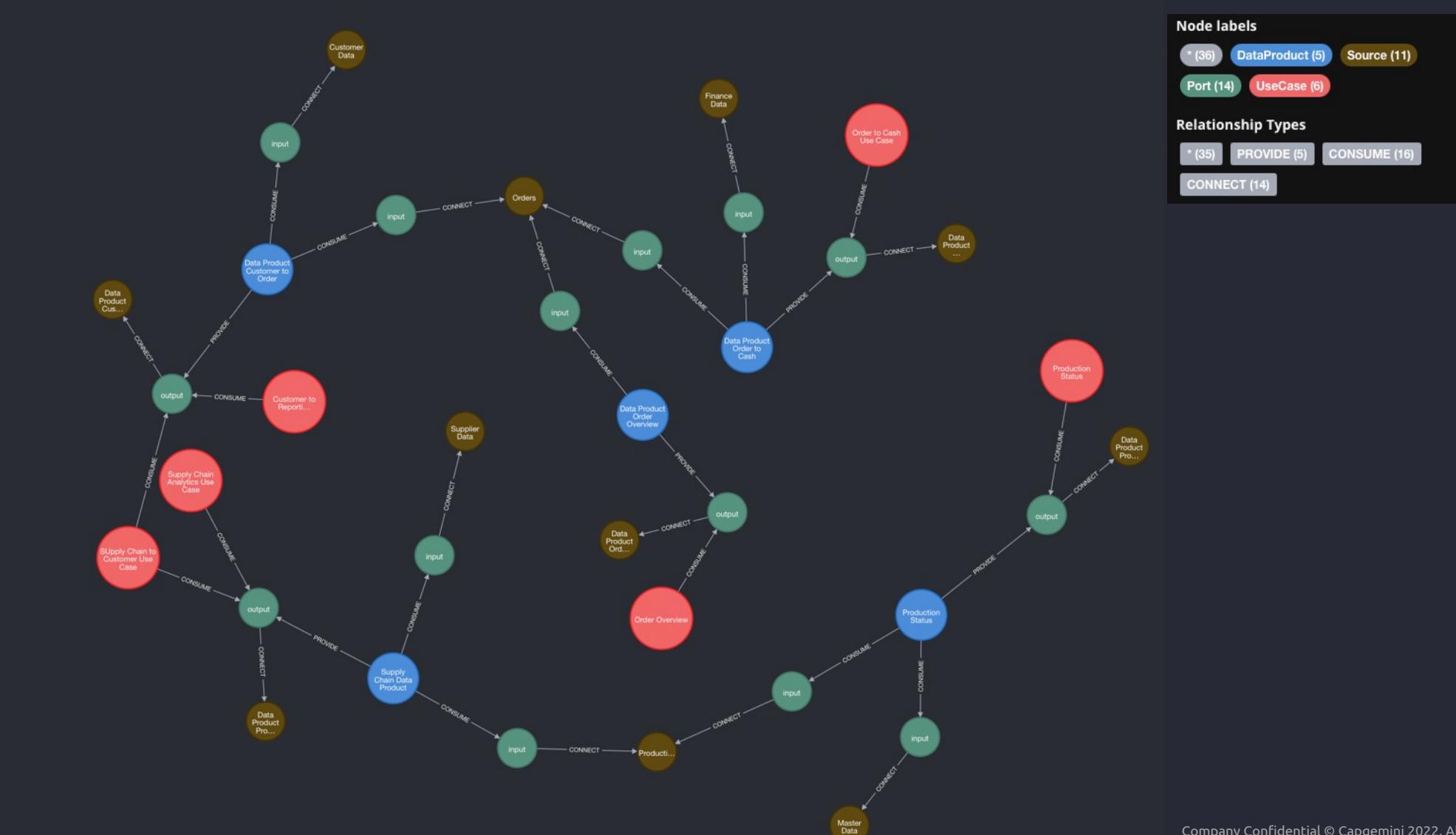


# A GRAPH BASED DATA ESTATE LANDSCAPE ADDRESSES THESE ISSUES AND PROVIDES TRANSPARENCY AND AGILITY





# KNOWLEDGE GRAPH OF EXEMPLARY DATA MESH





# GETTING ALL INVOLVED DATA SOURCES FOR SELECTED USE CASE

The screenshot shows the Neo4j browser interface with the following details:

- Query:** neo4j\$ `match(u:UseCase)--(p:Port)--(d:DataProduct)--(ip:Port)--(s:Source) WHERE u.name='Supply Chain to Customer Use Case' return s`
- Graph View:** Shows four nodes: "Supplier Data", "Customer Data", "Orders", and "Production Data".
- Table View:** Displays the results of the query as a table.
- Text View:** Shows the raw JSON results of the query.
- Code View:** Shows the executed Cypher query.
- Overview Panel:** Shows the following information:
  - Node labels:** \* (4) Source (4)
  - Displaying 4 nodes, 0 relationships.

| Source          | Description                    | Type | URL                                |
|-----------------|--------------------------------|------|------------------------------------|
| Supplier Data   | All Supplier information       | jdbc | jdbc://server/supplier_information |
| Customer Data   | Customer Data                  | jdbc | jdbc://server/database/customer    |
| Orders          | All Orders from Customers      | mqtt | mqtt://server/orders               |
| Production Data | Production Data from the plant | mqtt | mqtt://server/production           |

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

The diagram illustrates the relationships between various data products and use cases. Nodes include:

- Customer to Report...
- Data Product Order to Cash
- Data Product Customer to Order
- Order to Cash Use Case
- Supply Chain to Customer Use Case
- Order Overview

Overview

Node labels

\* (7) UseCase (4) DataProduct (3)

Displaying 7 nodes, 0 relationships.

| "u"  | "d"   |
|--|---|
| {"owner":"Dagobert Duck","createdDate":"2022-03-07 10:00:00","deputy":null,"name":"Customer to Order Reporting","description":"Having the overview on orders","lastModified":"2022-03-07 10:00:00","id":"uc00004","version":"1.0.0"}                   | {"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"}                   |
| {"owner":"Dagobert Duck","createdDate":"2022-03-07 10:00:00","deputy":null,"name":"Order to Cash Use Case","description":"Making the CFO happy","lastModified":"2022-03-07 10:00:00","id":"uc00005","version":"1.0.0"}                                 | {"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"} |
| {"owner":"Dagobert Duck","createdDate":"2022-03-07 10:00:00","deputy":null,"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"} | {"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"}    |

# CONCLUSIONS



# DATA MESH WITHOUT DATA GOVERNANCE

Will end up in muddy games





A large, thin blue line forms a circular arc at the top right, and another line curves from the bottom left towards the center, creating a dynamic, swoosh-like graphic element.

**GET THE  
FUTURE  
YOU WANT**

[capgemini.com](http://capgemini.com)



## 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.