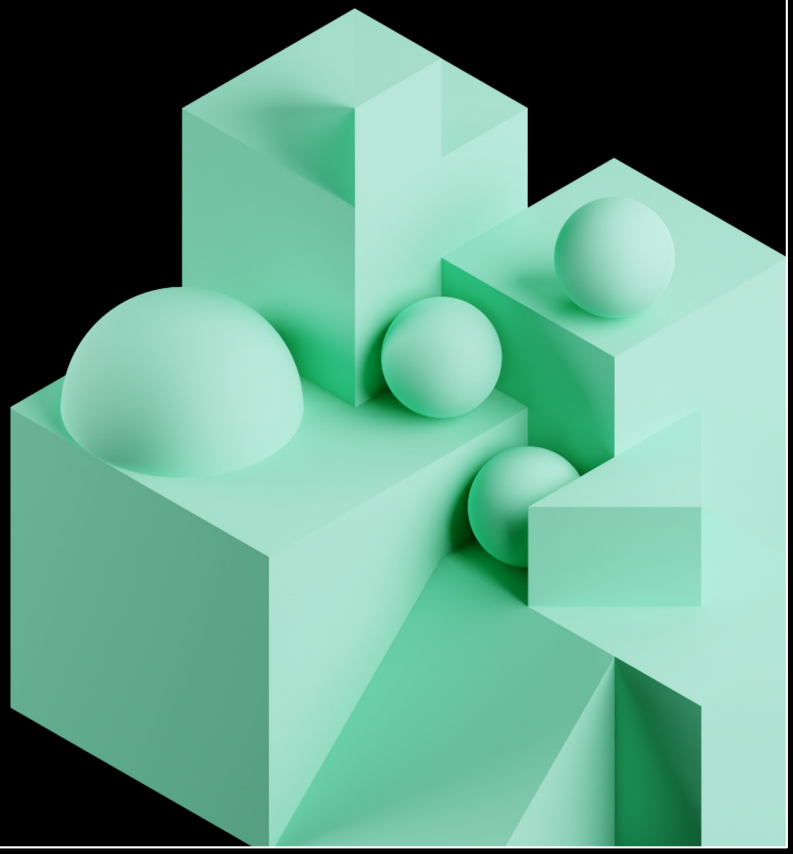
DATA+AI SUMMIT
BY databricks

Delta Sharing: The Key Data Mesh Enabler

Francesco Pizzolon,

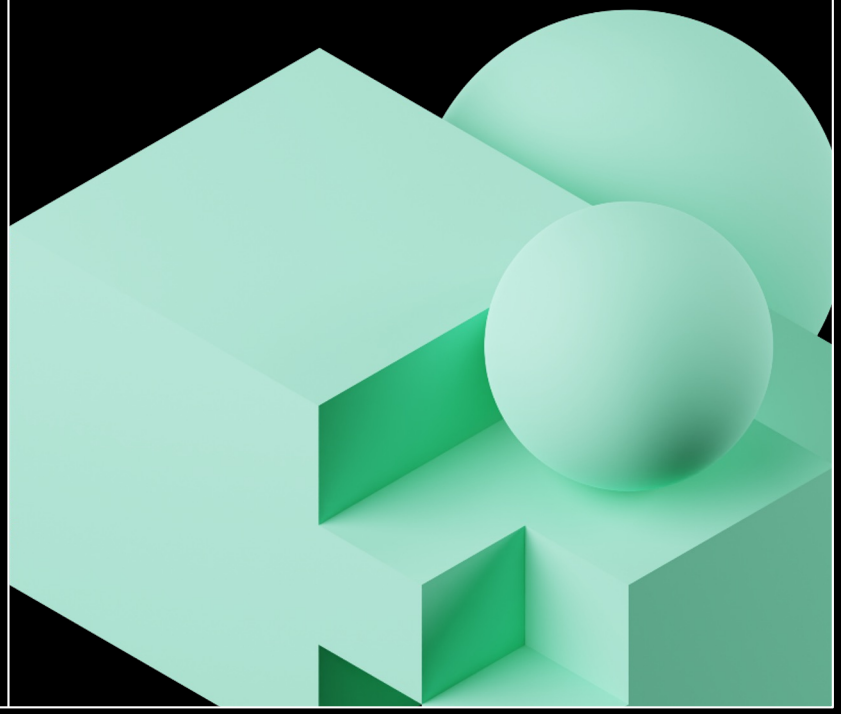
Avanade



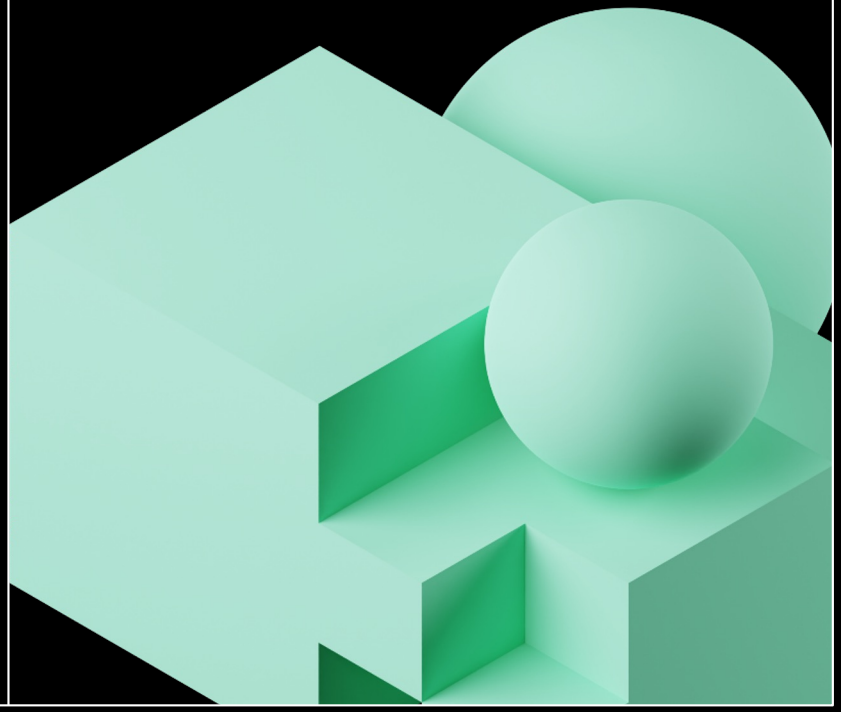
Databricks
2023

Agenda:

1. Introduction
2. The Client and Their Challenge
3. Solution Design
4. The Data Mesh Standards
5. Results
6. What's Next?



1. Introduction



Introduction

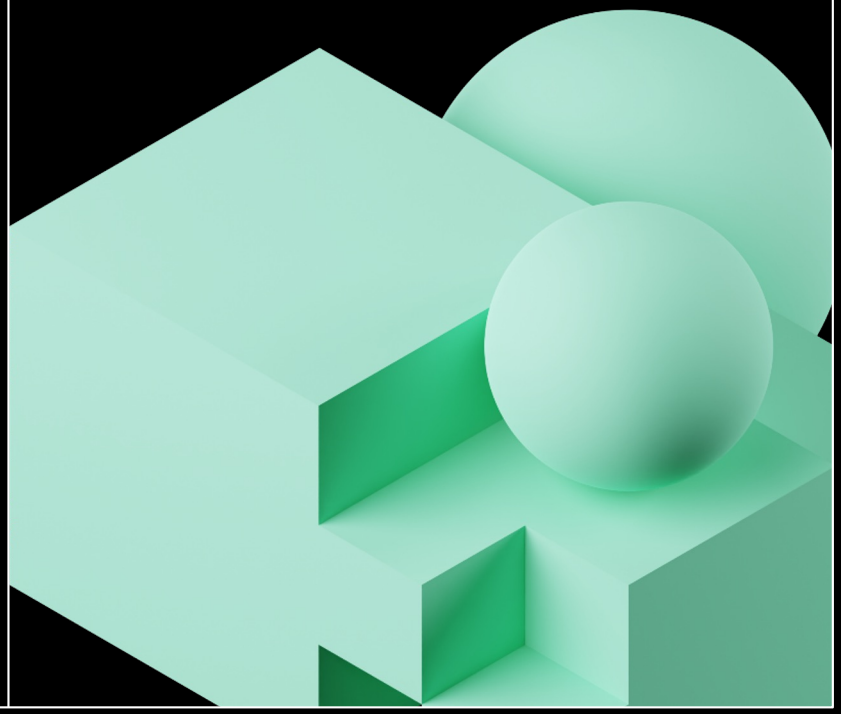
Welcome!

Francesco Pizzolon

francesco.pizzolon@avanade.com

- Director, Data Engineering @ Avanade UK
- 12+ years in Data Engineering
  - Banking, Retail, Transportation, Energy, Public Sector
- BSc & MSc in Computer Science

2. The Client and their Challenge



The Client and their Challenge

Sizewell C (SZC) - A New Nuclear Power Station for Britain

Sizewell C
The power of **good for Britain**

- Sizewell C (SZC) will be a **3.2-gigawatt power station** generating low-carbon electricity for around 6 million homes.
- It will be a **close copy of Hinkley Point C** which is already under construction in Somerset (UK).
- These new power stations will play a **key role in the UK's energy future**, supplying reliable, clean electricity for at least 60 years.

The Client and their Challenge

Typical phases for a Nuclear Power Station Project

Phase	Timeline (7-8 decades)
Project Initiation	
Site Selection and Evaluation	
Licensing and Regulatory Approval	
Detailed Design and Planning	
Procurement & Manufacturing	
Construction	
Installation and Testing	
Commissioning	
Operational Handover	
Ongoing Operations & Maintenance	
Decommissioning	

The Client and their Challenge

Organisational Setup

Regulatory (ONR)

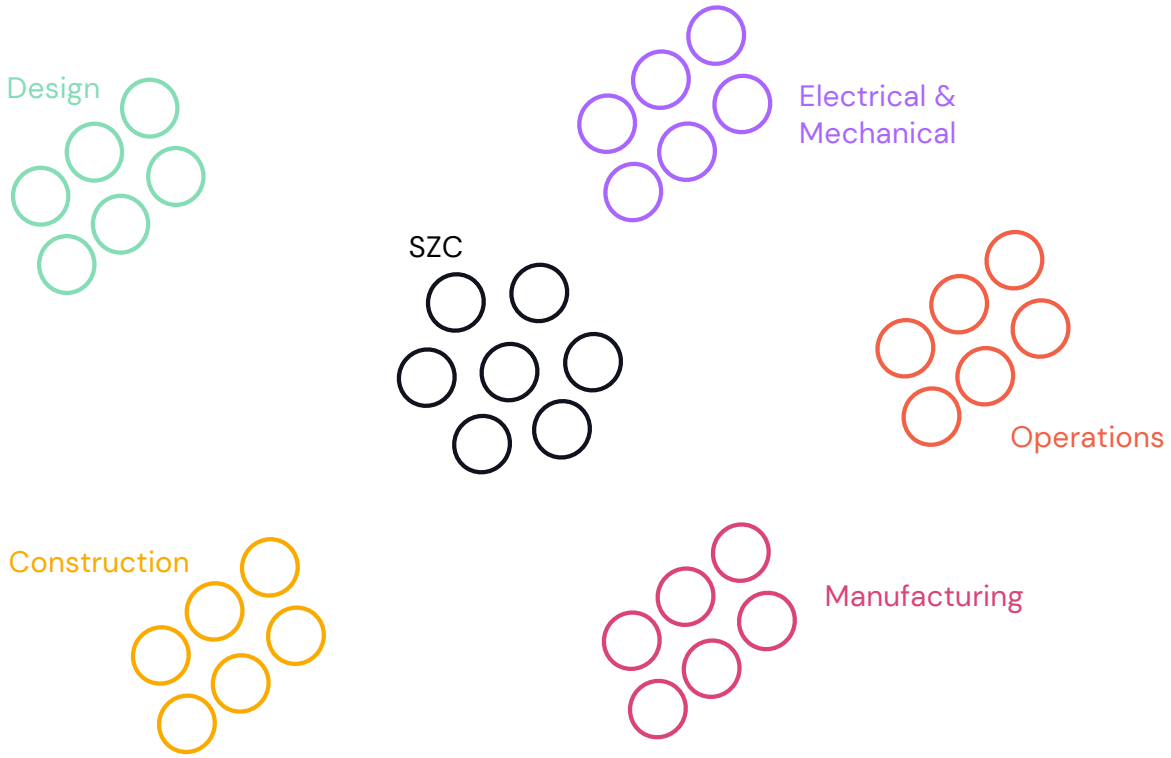
Sizewell C
The power of good for Britain

Partners

- Design
- Construction
- Electrical & Mechanical
- Manufacturing
- Operations

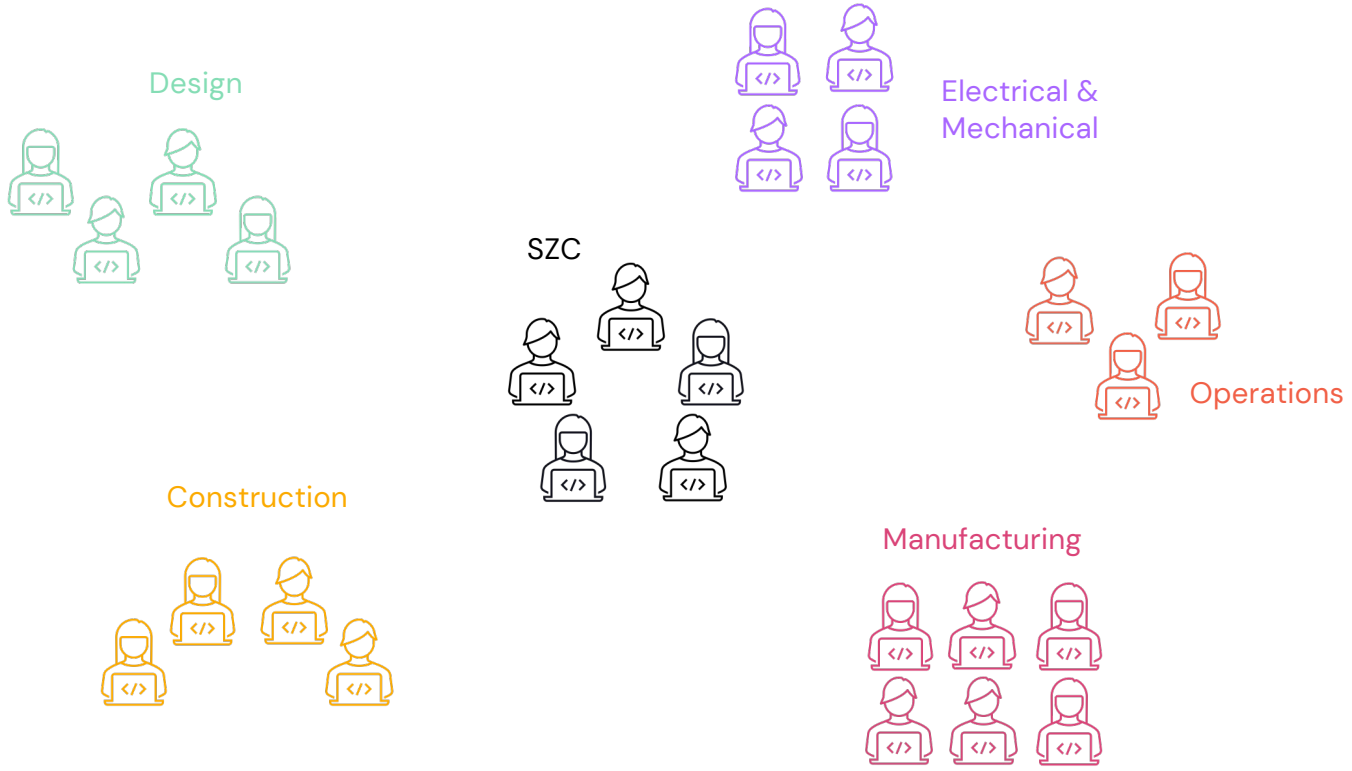
The Client and their Challenge

Operational Systems



The Client and their Challenge

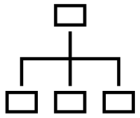
Data Teams



The Client and their Challenge

User Personas

Dashboard Users consume data insights from predefined reports and dashboards that will help them make informed decisions their job and role require.



Data Model Users self-serve on predefined data models that include tables, relationships between tables, and business certified measures. They can create their own reports and derived measures from the predefined data models.

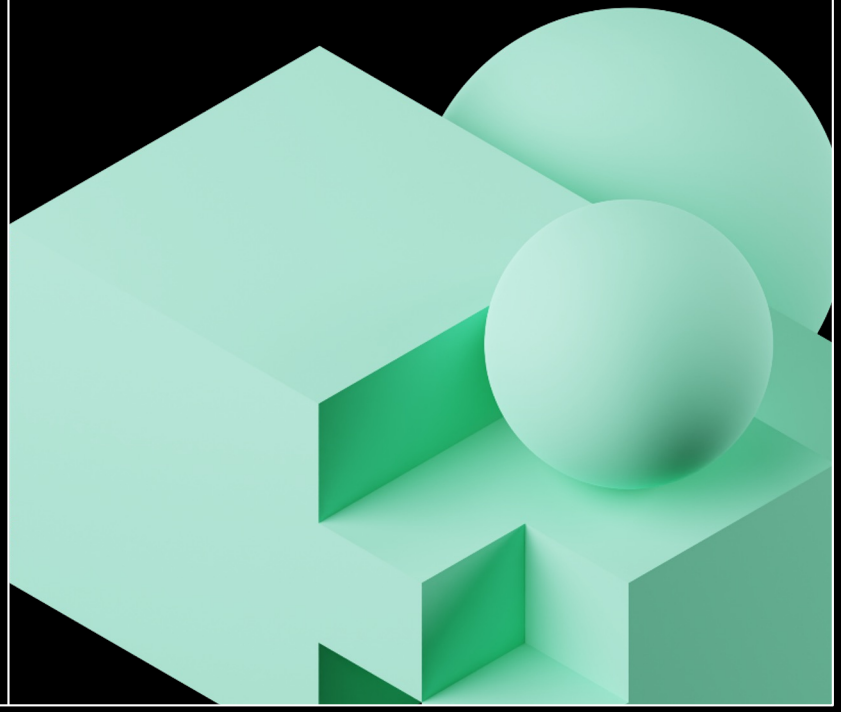
Data Analysts consume data via an SQL interface – they understand how to query the data via SQL and how to join tables and perform aggregations.

The Client and their Challenge

Non-Functional Requirements

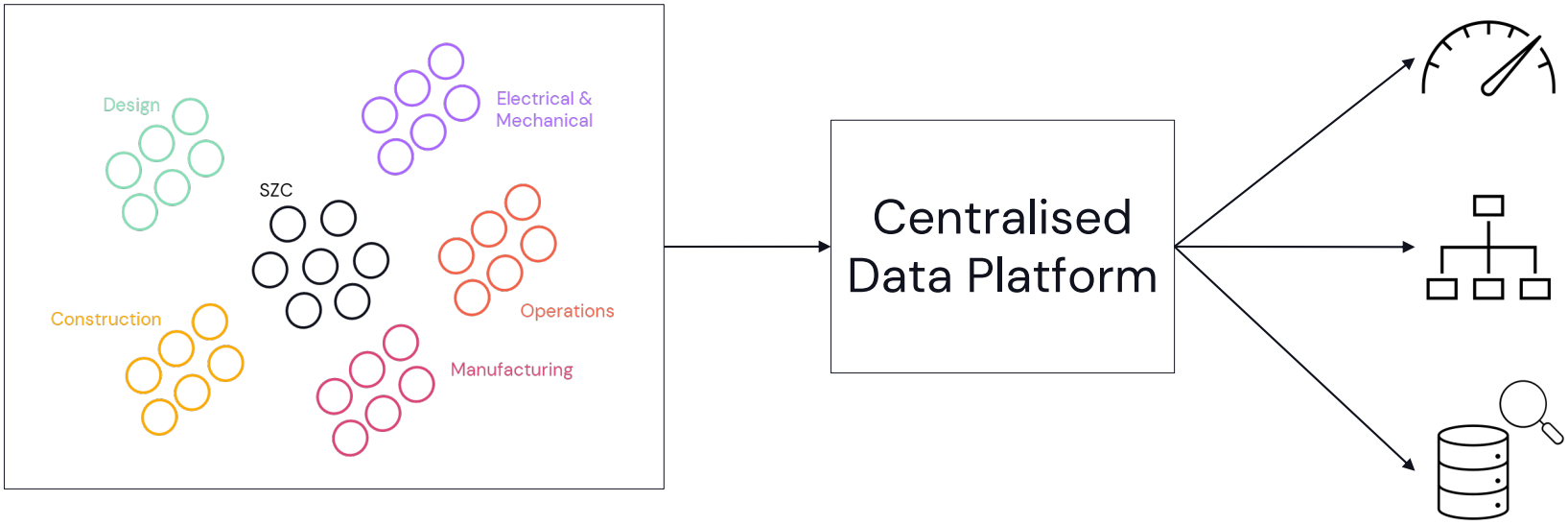
Title	Description
Data Location	**Data** related to this project must be **hosted within the United Kingdom**
Data Sovereignty	All **analytical data** related to SZC **should be hosted in SZC's Cloud Tenant**
Environments	SZC's **partners should use SZC's environment** to deliver data insights and products.
Operating Model	The data environments to be appropriately designed so that: • The **maximum level of isolation in a cloud environment** can be offered to SZC and their partners' data • All **access to data assets is audited** and **intellectual property** for SZC and their partners **is safeguarded** • The system to be flexible enough so that **future partners can easily integrate** with it **should they choose to operate outside SZC's Cloud Tenant**
Data Sharing	SZC and partners to **share data in a reliable, secure, and auditable way**
Time to Value	SZC and partners will deal **with a lot of new operational source systems as they get rolled out** – the data operational model must be **flexible enough to accommodate this**

3. Solution Design



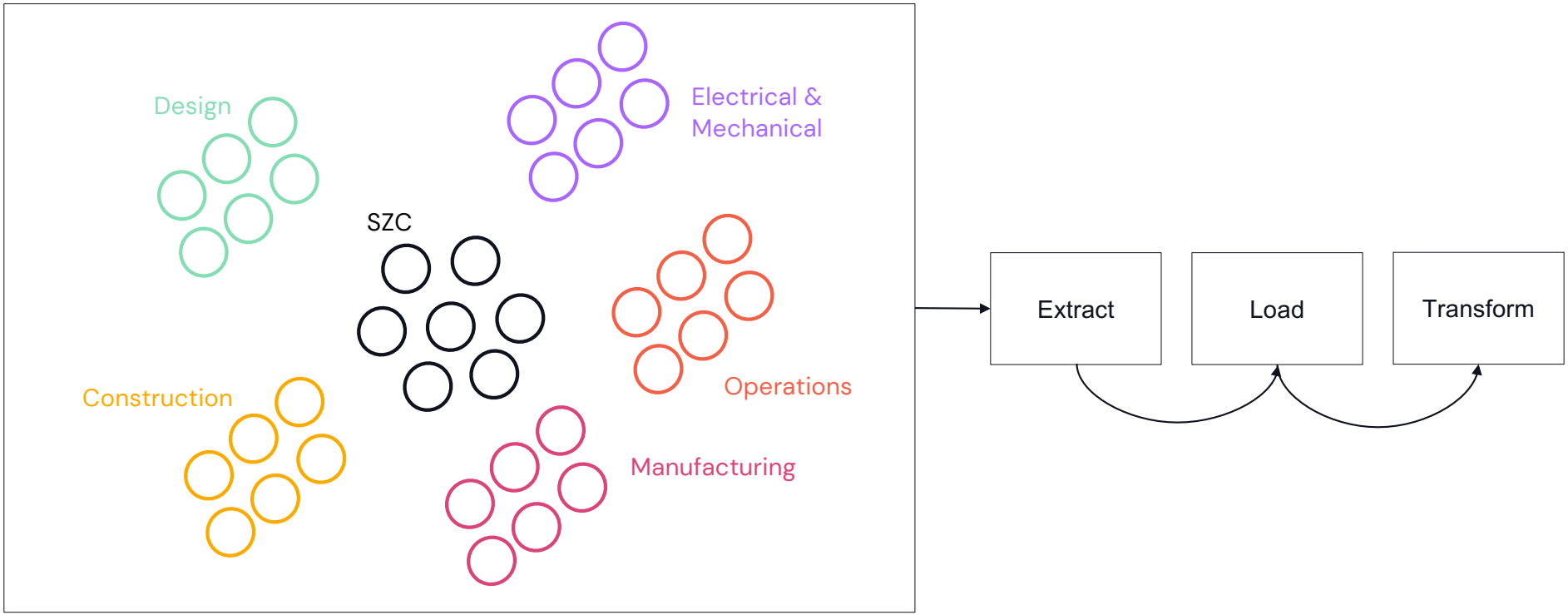
Solution Design

Centralised Data Platform



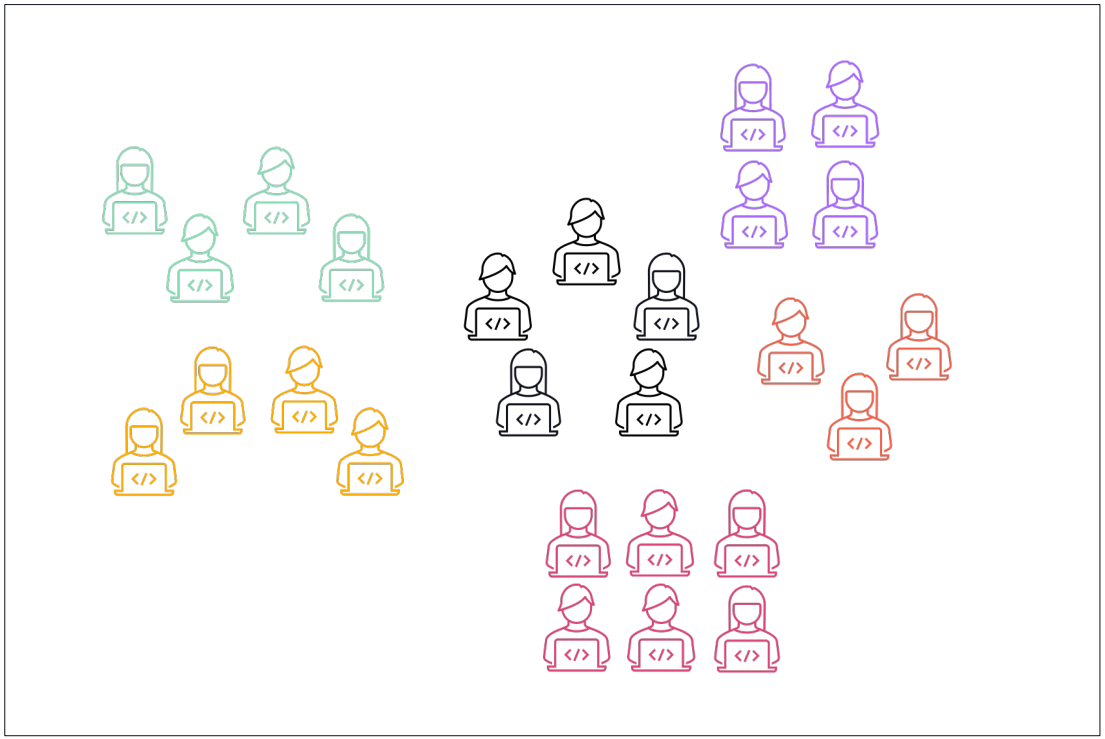
Solution Design

Centralised Data Platform / Operational Systems



Solution Design

Centralised Data Platform / Data Teams



Backlog

SZC

Partner1

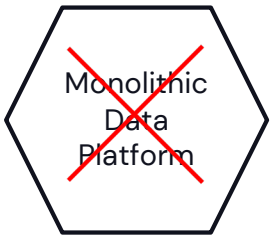
Partner2

…

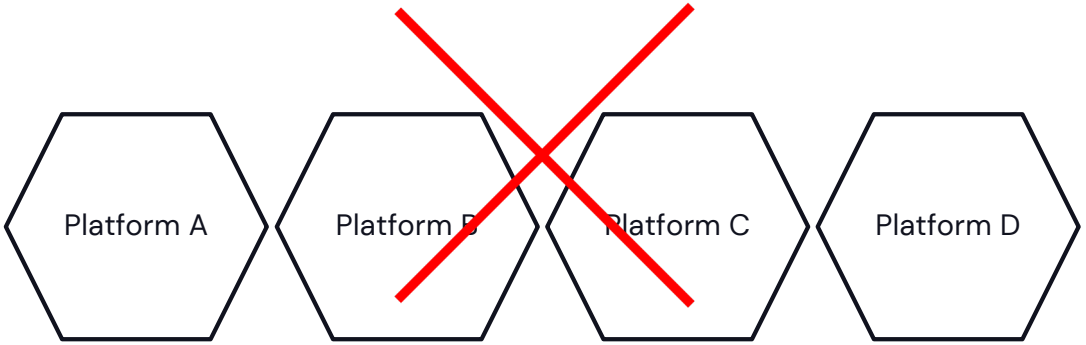
PartnerN

Solution Design

Distributed Data Platform aka Data Mesh



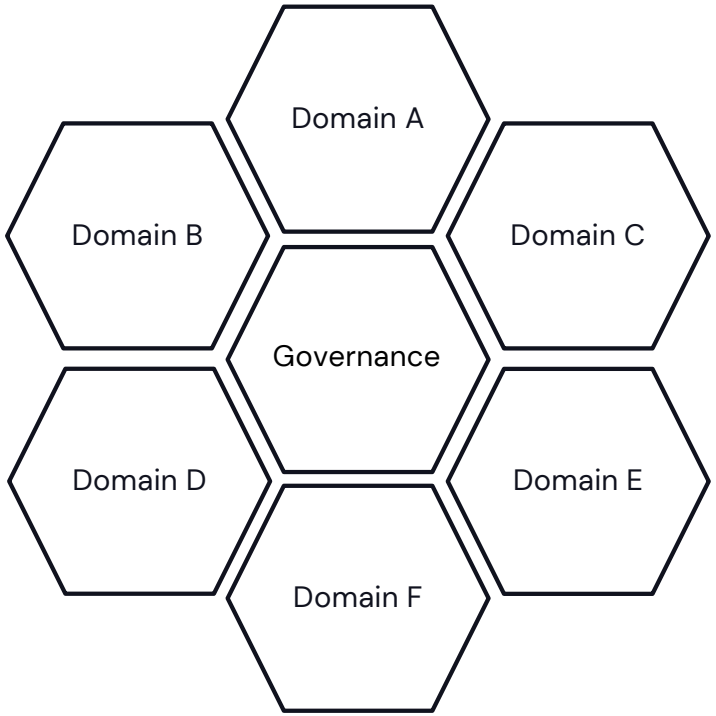
One Monolithic Data Platform
Enforces consistency at the expense of speed and agility



Many Independent Data Platforms
Enables flexibility, but forces data duplication, inconsistency across platforms, and rework to process data

Solution Design

Distributed Data Platform aka Data Mesh



Core Concepts

Data as a Product: A data product consists of code, data/metadata, and infrastructure, with a Product Owner accountable for data quality and improvement.

Domain-Oriented Data Ownership:
Enables decentralisation and distribution of data responsibility to support scalability and agile change cycles.

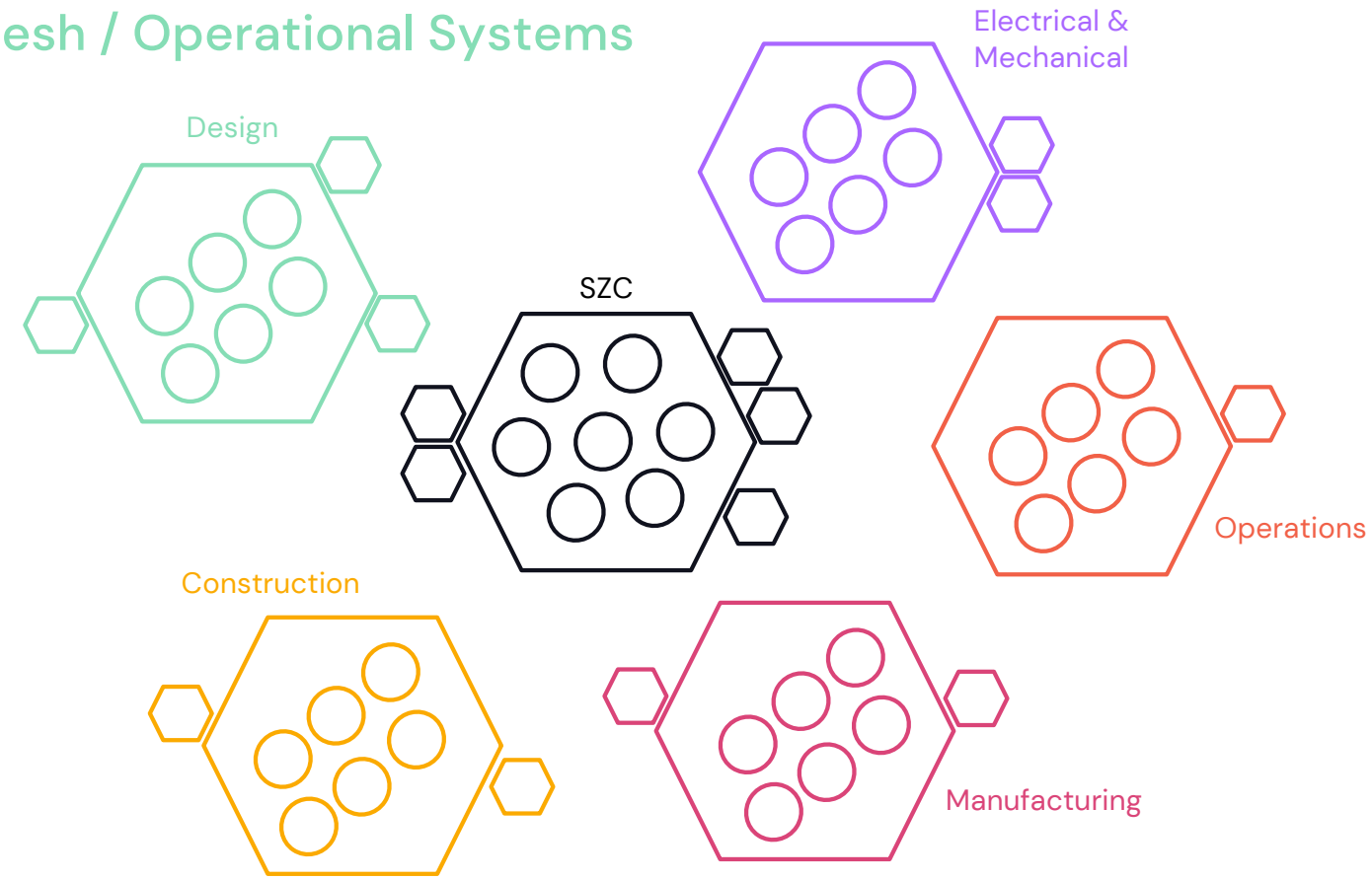
Self-serve Data Infrastructure: Providing tools to capture, store, and serve data products for teams to independently access and manage data.

Federated Governance: Establishing a framework for collaborative decision-making and coordination among data domains.

Solution Design

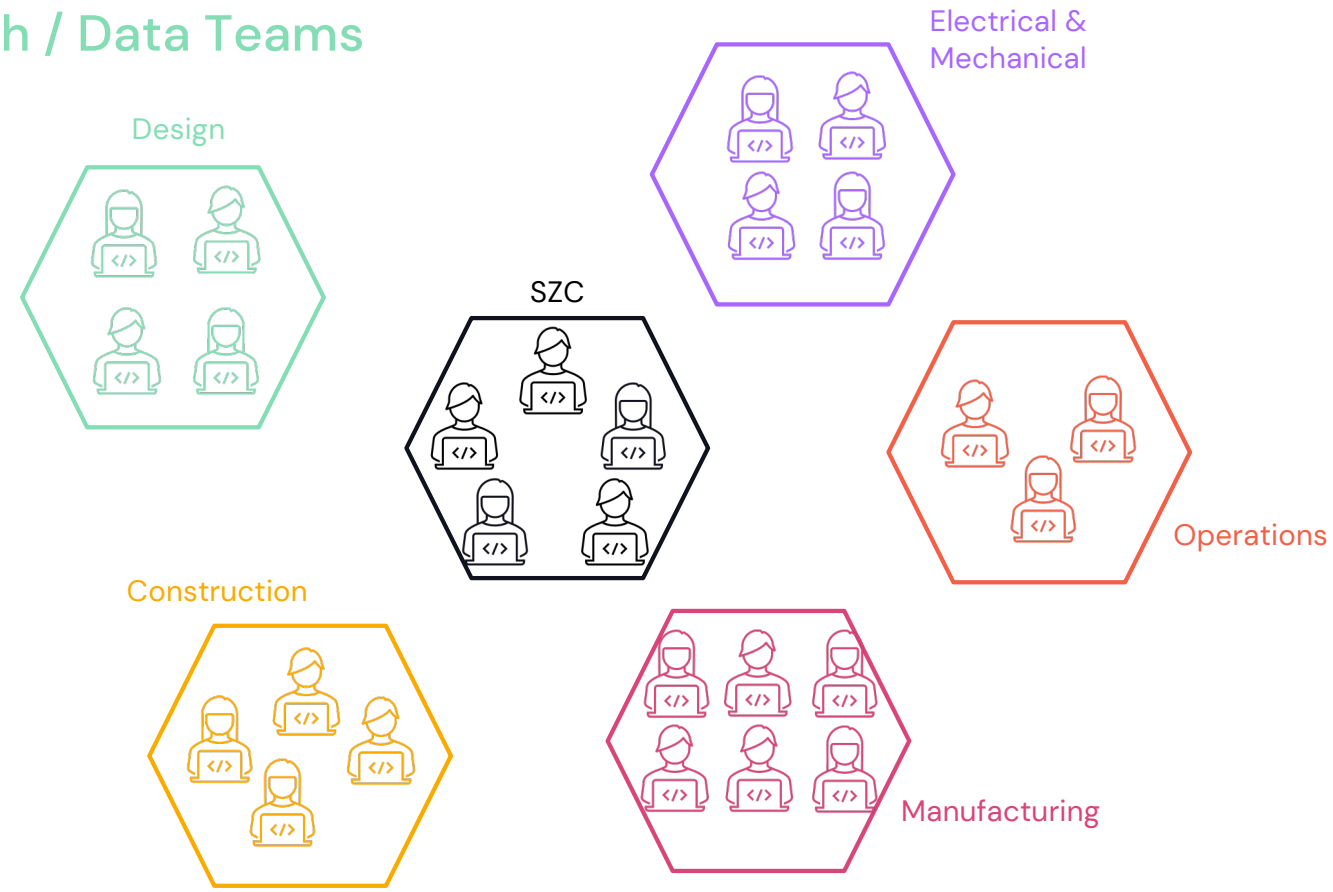
Data Product

Data Mesh / Operational Systems



Solution Design

Data Mesh / Data Teams



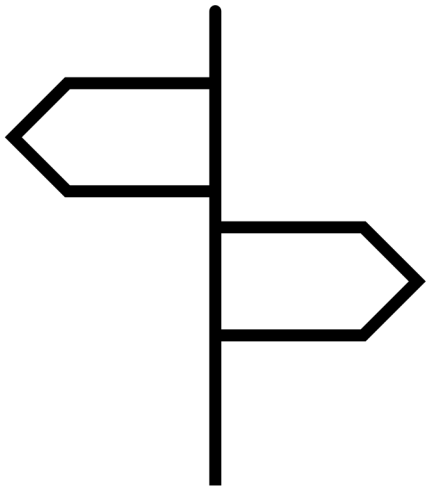
Solution Design

Pros & Cons

	Pros	Cons
Centralised Data Platform	• Great level of control • One centralised infrastructure to code & maintain • Highest level of consistency • One source of the truth	• Potential bottleneck because of the central data engineering team • Slower response to change • Partners' segregation in a single data lake will be difficult and admin-intensive • No network segregation between partners
Data Mesh	• Partners' segregation achieved by design • Reacts well on changes in the data landscape, proliferation of sources of data, diversity of data use cases • Promotes reusability	• Initial network setup can be complicated • Requires a shift from the centralised team mindset to a distributed operating model

Solution Design

Decision Time!



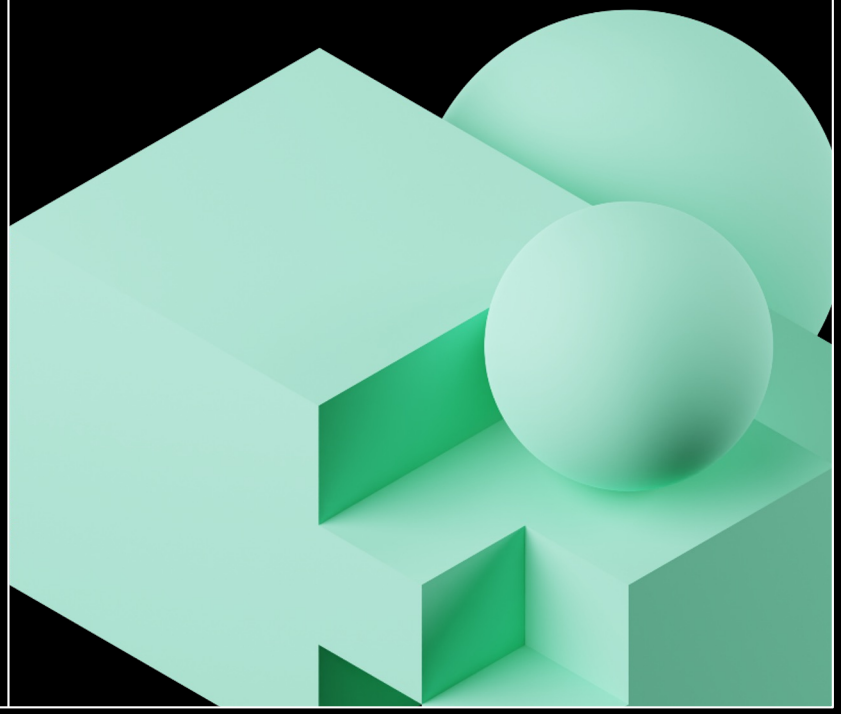
Given

- The uniqueness of the SZC joint venture business setup
- The NFRs, particularly network segregation between partners and data sovereignty

We choose:

Data Mesh!

4. The Data Mesh Standards



The Data Mesh Standards

Lots of standards!

- The Data Format standard
- The Data Sharing standard
- The Operational Plane standard
- The Infrastructure Plane standard
- The Data Catalog standard
- The Data Quality standard
- The Data Lineage standard

The Data Mesh Standards

The Data Format standard: Delta Lake!

Schema enforcement
Automatically handles schema variations to prevent insertion of bad records during ingestion.

ACID Transaction on Spark
Serializable isolation levels ensure that readers never see inconsistent data.

Time travel
Data versioning enables rollbacks, full historical audit trails, and reproducible machine learning experiments.

Cost Savings
Based on the Apache Parquet file format and the snappy compression algorithm, optimised for data engineering and data analytics workloads.

Upserts and deletes
Supports merge, update and delete operations to enable complex use cases like change-data-capture, slowly-changing-dimension (SCD) operations.

Open
Can be implemented regardless of the cloud provider.

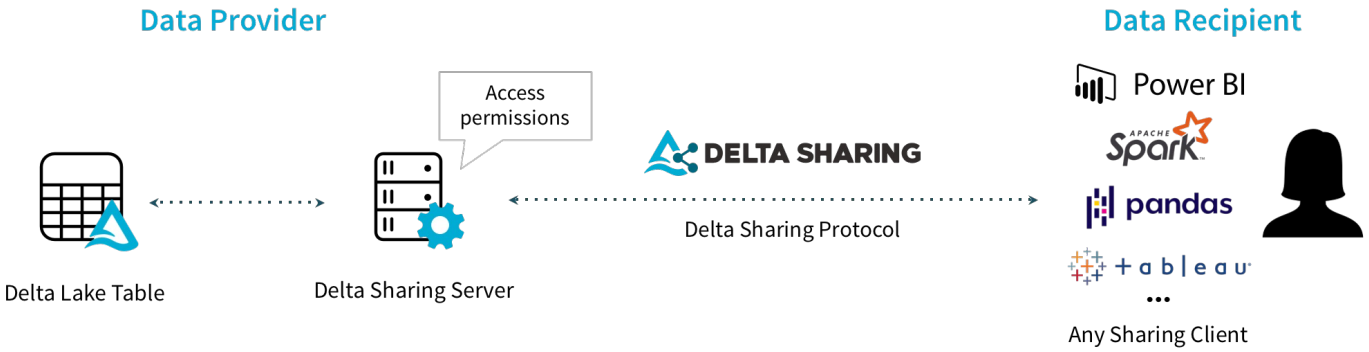
The Data Mesh Standards

The Data Sharing standard: Delta Sharing!

In-place Data Sharing
Allows data access and collaboration directly from its original location, eliminating need for data copying or movement.

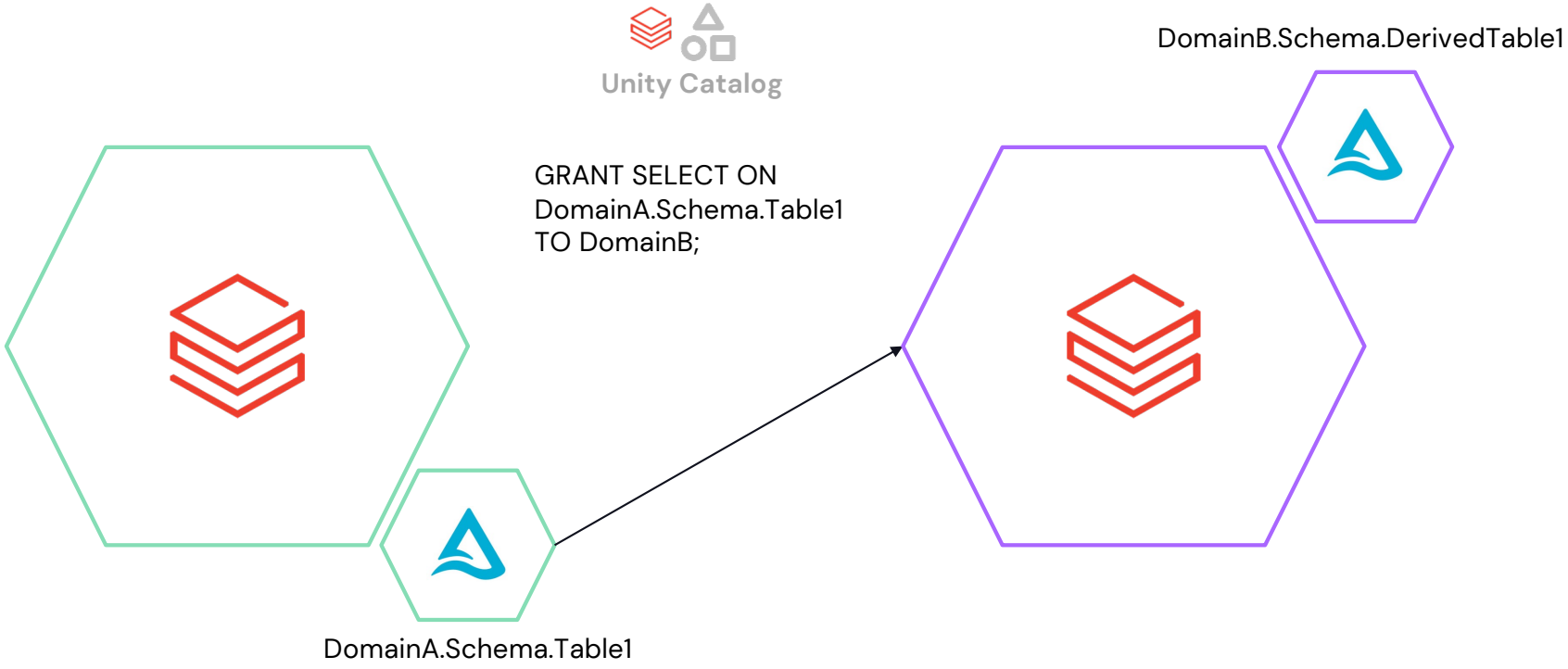
Secure and Auditable
Encryption and access control mechanisms. Auditability maintained with comprehensive logs of data transactions for compliance and review.

Open
Recipients don't need to be on the same platform as the provider, or even in the same cloud provide -- sharing works across clouds and even from cloud to on-premise users.



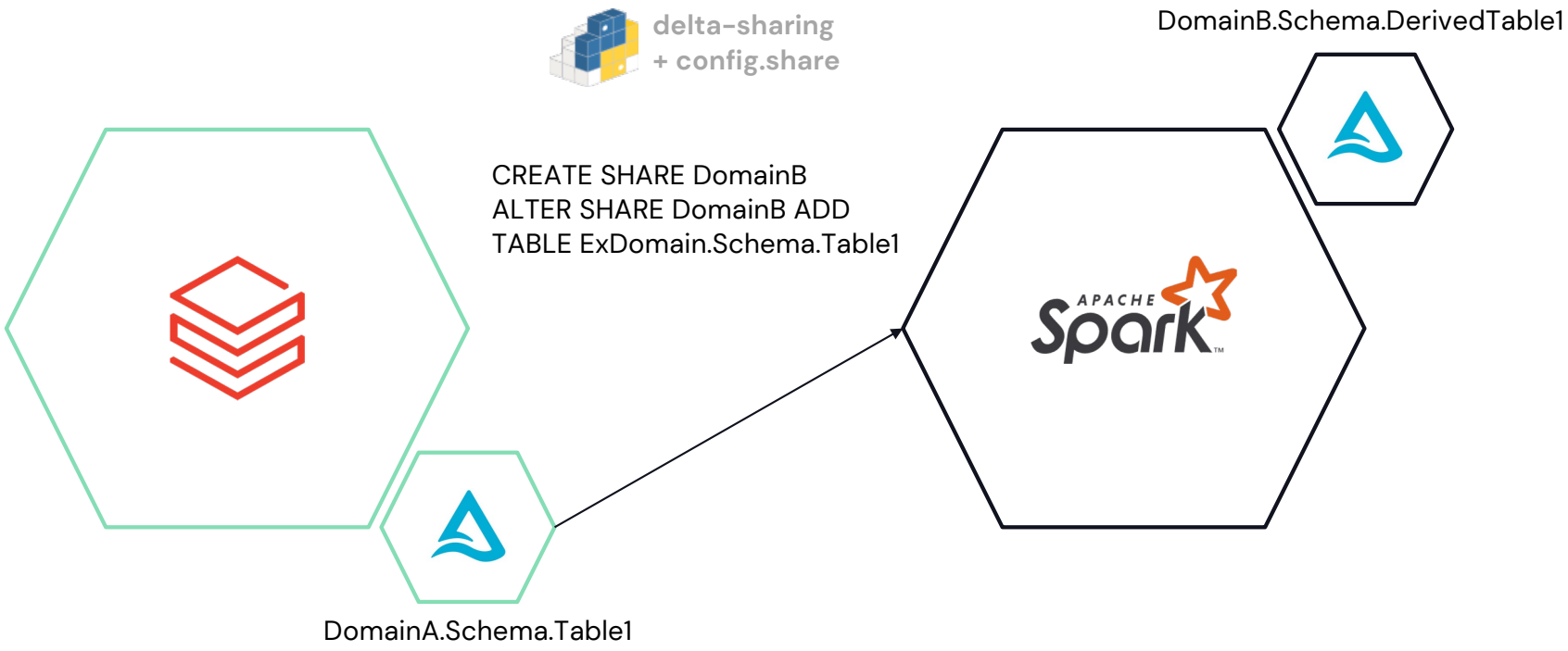
The Data Mesh Standards

The Data Sharing standard / SZC Azure Tenant

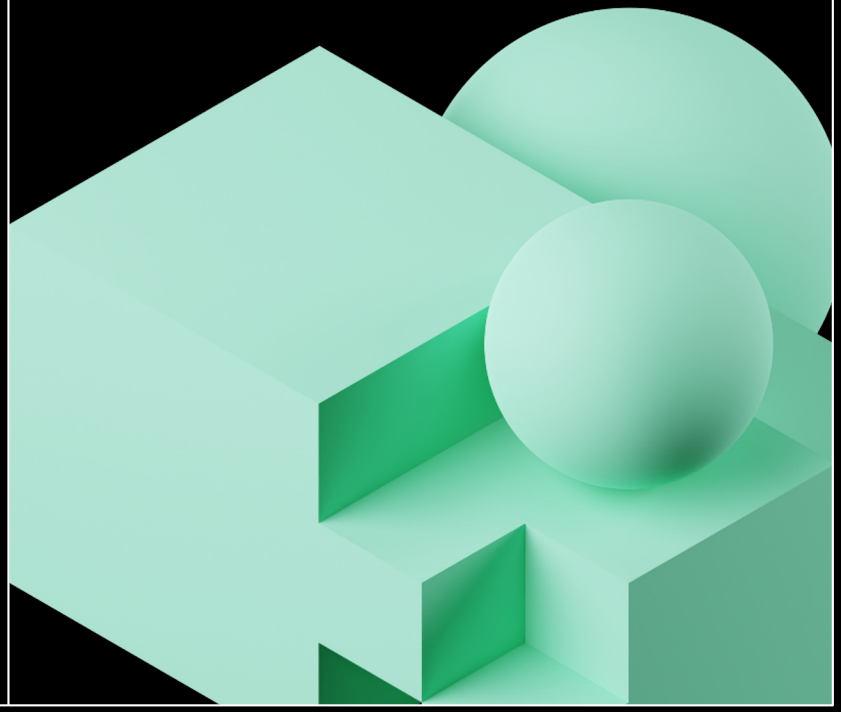


The Data Mesh Standards

The Data Sharing standard / External Tenant



5. Results



Results

Data Mesh Success

Data Products

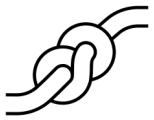
- 20+ shared data products
- 150+ developed data products
- New operational system's data to source-aligned data products in days

Data Governance

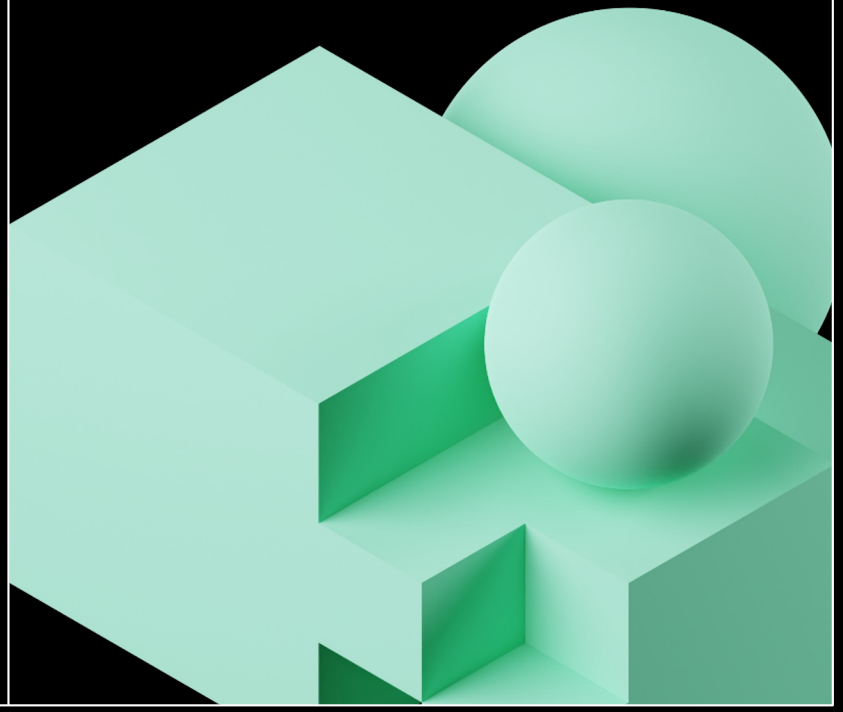
- 100% Data Product Ownership – products don't get released in production if missing an owner
- Data Quality issues resolved in days, not weeks/months

SZC & Partners Empowerment

- 100% partners so far have decided to use the SZC tenant for data analytics, future and prospect partners are confident they can use the platform to deliver their data workloads in isolation



6. What’s Next?



What's Next

Future of SZC's Data Mesh

Databricks Marketplace

- Ability to monetise assets:
  - Data products
  - Notebooks
  - Solution accelerators

ML Data Products

mlflow

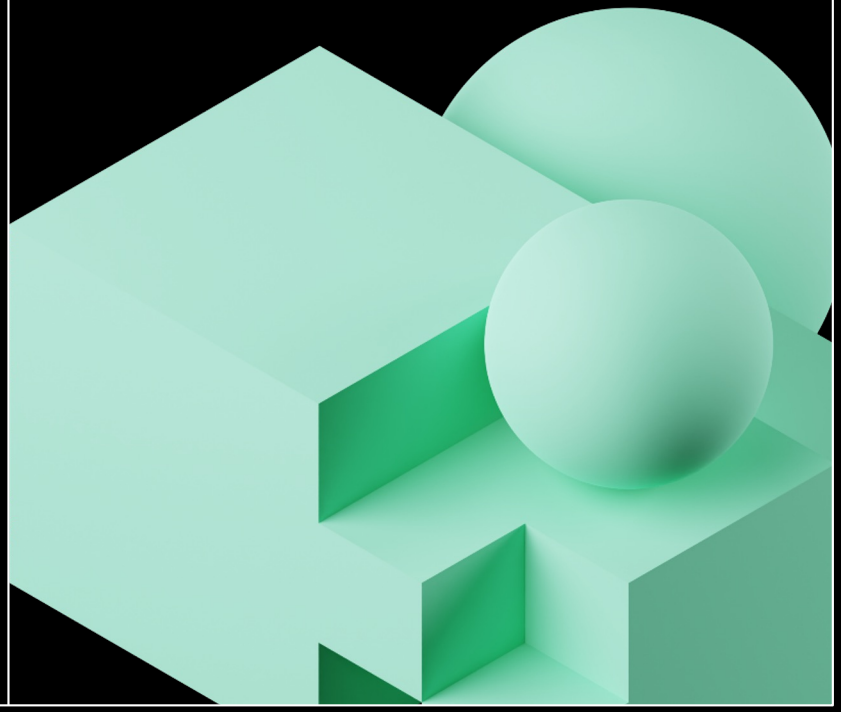
- Focus on machine learning opportunities, e.g.:
  - Summarisation of technical documents
  - Predictive maintenance
  - Public Relations & Communication (complex technical details in layman's terms)

Deeper integration with Microsoft

Microsoft Azure

- SZC is actively participating in several private previews features involving the integration between Unity Catalog and Microsoft Purview for better data governance

If you Q I will A!




databricks